



Universiteit
Leiden
The Netherlands

Design and Implementation of FAIR principles in digital health data in Uganda, Africa

Basajja, M.

Citation

Basajja, M. (2026, September 8). *Design and Implementation of FAIR principles in digital health data in Uganda, Africa*. Retrieved from <https://hdl.handle.net/1887/4309278>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4309278>

Note: To cite this publication please use the final published version (if applicable).

Design and Implementation of FAIR principles in digital health data in Uganda, Africa

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 8 september 2026
klokke 11:30 uur

door

Mariam Basajja

geboren te Kampala, Uganda

in 1993

Promotor: Prof. Dr. Ir. F.J. Verbeek

Co-promotor: Dr. K.J. Wolstencroft

Promotiecomissie

Prof. Dr. M.M. Bonsangue

Prof. Dr. H.C.M. Kleijn

Prof. Dr. B. Mons

Prof. Dr. M.R. Spruijt

Dr. R. Baguma (Makerere University, Kampala, Uganda)

Dr. L. Stork (University of Amsterdam)

Copyright © 2026 Mariam Basajja. All rights reserved. No part of this thesis may be reproduced in any form or by any means without permission of the author.

The research presented in this dissertation was financially supported by KIU.

Backcover: Cover designed by Mariam Basajja.

Contents

1	Introduction	9
1.1	General Introduction	10
1.1.1	The Sustainable Development Goals (SDGs) that Foster Uganda's Digital Health Agenda	11
1.1.2	Digital Structural Gaps in Primary Healthcare	11
1.1.3	Uganda's Digital Health Ecosystem Key Players	12
1.2	FAIR Data and Digital Health	13
1.3	VODAN Africa	14
1.4	Research Questions and Contributions	15
1.4.1	Main Research Question	15
1.4.2	Specific Research Questions	15
1.5	Thesis Outline	15
2	VODAN Africa and Asia	19
	Abstract	20
2.1	Introduction	21
2.2	Implementing FAIR Principles in Africa and Asia	21
2.2.1	Challenges in Digital Health Implementation	22
2.2.2	Understanding the Adoption of FAIR Principles	23
2.2.3	Conceptual Framework	25
2.2.4	The Acceptance of FAIR Data principles and their Implementation	26
2.2.5	Conclusion	28
3	Information Streams in VODAN-Africa Health Facilities: The Case of Uganda	31
	Abstract	32
3.1	Introduction	33
3.2	Research Question and Objectives	36
3.3	Methods and Materials	37
3.3.1	Study Design	37
3.3.2	Sampling Procedure	37
3.4	Results	39
3.4.1	Ideal Health Information Management in Uganda	40

3.4.2	Actual State of Health Information Management	40
3.4.3	General Observations	42
3.4.4	Stakeholder Insights	43
3.5	Conclusion	44
3.6	Discussion	45
4	FAIR Equivalency with Regulatory Framework for Digital Health in Uganda	49
	Abstract	50
4.1	Introduction	51
4.1.1	Background	51
4.1.2	Main Objective	51
4.1.3	Specific Research Questions and Objectives	52
4.2	Materials and Methods	52
4.2.1	Theoretical Considerations	53
4.2.2	Study Design	53
4.2.3	Historical Perspective on the Evolution of ICT and Health-Related Policies in Uganda	53
4.2.4	Uganda's National e-health Vision	55
4.2.5	Identification of Relevant Documents	56
4.2.6	Mention of the FAIR principles in Policy Documents	57
4.2.7	Level of FAIR Equivalency in Policy Documents	58
4.3	Results	61
4.3.1	Mention of FAIR principles in Policy Documents	61
4.3.2	FAIR Equivalency in Policy Documents	61
4.4	Conclusion	64
4.5	Discussion	64
4.5.1	FAIR Equivalency	65
4.5.2	Facet A1: Standardized Communications Protocol for Accessibility	65
4.5.3	Facet I1: Formal, Accessible, Standardized Controlled Vocabularies	66
4.5.4	Facet F1: Globally Unique and Persistent Identifiers for Findability	67
4.5.5	FAIR as a Solution to Challenges in Uganda's e-health Sector	68
4.5.6	Challenges of the Study	69
5	Possibility of Enhancing Digital Health Interoperability in Uganda through FAIR Data	71
	Abstract	72
5.1	Introduction	73
5.2	Research Questions and Objectives	74
5.3	Two Digital Health Hubs	75
5.3.1	Digital Health Atlas Uganda	75
5.3.2	Uganda Digital Health Dashboard	76
5.4	Methods	76

5.5	Results: Level of FAIRness of Digital Health Hubs in Uganda	77
5.5.1	Uganda Digital Health Dashboard (UDHD) Results	78
5.5.2	Digital Health Dashboard Uganda (UDHD) Results	79
5.5.3	Digital Health Atlas Uganda (DHA-U) Results	80
5.5.4	Digital Health Interventions (DHIs) FAIRness	81
5.5.5	After FAIRification	81
5.6	Conclusion	84
5.7	Discussion	84
6	Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic	87
	Abstract	88
6.1	Introduction	89
6.2	Relevance of the Study	89
6.3	FAIR Guidelines	90
6.4	Research Questions and Objective	91
6.5	Development of VODAN In a Box	93
6.5.1	Components	93
6.5.2	Technical Specifications.	95
6.5.3	Metadata Component Choices	96
6.6	Proof of Concept	98
6.6.1	Deployment in A Real-life Situation	98
6.6.2	Procedure for Proof of Concept	99
6.7	SWOT Analysis	100
6.7.1	Strengths	100
6.7.2	Weaknesses	101
6.7.3	Opportunities	102
6.7.4	Threats	102
6.8	Conclusion	103
6.9	Discussion	104
7	Privacy-Preserving Health Data Analytics in Uganda: A Comparative Study of DHIS2 and FAIR VODAN-Africa Architecture	105
7.1	Introduction	107
7.1.1	Motivation	109
7.1.2	Research Significance	110
7.1.3	Research Objectives	110
7.1.4	Research Questions	110
7.2	Related Work	111
7.2.1	Overview of DHIS2	111
7.2.2	FAIR Data Principles	112
7.3	Methodology	113

7.3.1	Data Privacy and Security Considerations and Identification of the Key Statistics of the Selected Datasets	113
7.3.2	Data Collection	114
7.3.3	Application of VODAN-Africa and Asia Health Management Information System Architecture	117
7.3.4	Standardization of the VODAN-Africa Controlled Vocabulary Terms and Semantic Properties	118
7.4	Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard	124
7.4.1	Key statistics Data Sources for the VODAN-Africa Local and Community Dashboards	125
7.4.2	Data Analytics on the AllegroGraph Repositories	126
7.5	Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda	130
7.5.1	DHIS2's Role in Transforming Uganda's Health Information Systems	131
7.5.2	FAIR's Role in Transforming Uganda's Health Information Systems	133
7.6	Conclusion	139
8	Machine Learning Models in FAIR Health Data from Uganda	141
	Abstract	142
8.1	Introduction	143
8.1.1	Research Questions and Objectives	147
8.1.2	Our contributions	147
8.2	Related Work	148
8.2.1	Machine Learning in Healthcare	148
8.2.2	Federated Machine Learning	151
8.2.3	Federated Architecture in Healthcare Application	151
8.2.4	Data Science and Big Data in Health	153
8.3	Use cases, Data and Methods	153
8.3.1	Data Sources and Code	153
8.3.2	Federated Learning Architecture for Local Hospital Systems	154
8.3.3	Machine Learning Modeling Workflow	157
8.3.4	Use Case 1: FAIR Outpatient Department Uganda Diagnoses with K-Means Clustering Analyses	159
8.3.5	Use Case 2: FAIR Integrated Antenatal Care (ANC) Uganda Data Anomaly Detection and Patient Data Clustering Modelling	168
8.3.6	Use Case 3: FAIR ANC Uganda Data Pregnancy Risk Factors and Fetal States Determinants' Clustering Analysis	171
8.3.7	Use Case 4: FAIR ANC Uganda Data Fetal State Prediction Tool Modelling	177
8.4	Discussion	190
8.5	Conclusion	191

9 Discussion, Outlook and Conclusion	195
9.1 Observations Answering the Research Questions	196
9.2 Implications for Practice	197
9.3 Limitations of the Study	198
9.4 Call to Action	199
9.5 Future Research Directions	199
9.6 Conclusion	201
Bibliography	202
A Abbreviations	240
B Summary	243
C Samenvatting	246
D Curriculum Vitae	249
E Acknowledgements	250

Chapter 1

Introduction

1.1 General Introduction

In this research, we present the use of FAIR principles in digital health in Africa. To achieve the third Sustainable Development Goal, which is "to ensure the quality of life and prosperity for all at all ages", it is important to focus on health care in Africa [1]. In the broader context of this goal, improving digital health across the African continent has emerged as a key priority. The integration of digital health systems with seamless health care delivery is the key to data reuse [2, 145]. The delivery of integrated care that facilitates the exchange of patient information requires integrated solutions that are tailored to the unique healthcare system in Uganda and, more broadly, Africa. Moreover, a strong governance structure must be established. There should be ethical and legal considerations for digital health systems in order to protect patient privacy, support the integrity of data, and promote responsible use of data [4]. The digital health landscape is undergoing a transformation that will have a significant impact on healthcare systems. The integration of digital technologies into healthcare has become a global trend that has transformed the methods that are used to deliver, access, and manage medical services. Digital healthcare technology offers innovative solutions to the deficiencies in the structure of delivering healthcare through improvements such as enhancing stock-management systems and streamlining blood donations through GPS-equipped drones.

The adoption of digital health technologies is driven by a shared vision to improve access, quality, and efficiency [5] of healthcare. Governments, organizations, and innovators around the world are using digital tools to solve age-old challenges in healthcare. But this journey is not without its challenges, and the reuse of digital health information is a recurring concern [6, 7]. In response to the challenges posed by the rapid growth of digital health, the Ugandan government took prompt steps by launching the Centralized Center for Digital Health in 2016 [8]. The center aims to provide digital health solutions to facilitate and address the challenges within the delivery of national healthcare. The Ugandan experience reflects a global trend in which countries grapple with the dual task of harnessing the power of digital health and ensuring its long-term viability.

The global digital health landscape encompasses a range of sectors and services, each of which contributes to the goal of improving health outcomes through technology. In Uganda, 49 identified providers were involved in the Centralized Center for Digital Health [8]. Similar challenges are witnessed worldwide, necessitating an understanding of the dynamics at play in digital health ecosystems. Many initiatives around the world have focused on the applicability of digital health solutions, emphasizing the importance of understanding the unique challenges faced by these initiatives. The emergence of mobile health (mHealth) applications and digital health interventions underscores the transformative impact of even simple implementations [9]. The use of mobile phones to connect healthcare providers and people through voice, video, or messaging services has proven to be particularly impactful, transcending geographical barriers and improving healthcare access in both urban and rural settings. This is evident in 2013 statistics showing that nearly 54% of medical calls received by the center were resolved effectively by telephone consultation [10].

Countries such as Rwanda and Kenya have shown significant progress in prioritizing digital health in their regional strategies. However, Uganda does not have an equally focused ICT program with ICT-based infrastructure [145] provides a special case for examining the underlying components that support absence-data reuse. Globally, digital health experiences ongoing challenges despite the remarkable growth within this sector. The rapid growth in technology has led to innovative advancements that offer unprecedented opportunities, which have led to a transformation in healthcare delivery. Despite this, there are concerns regarding the existing problems of connectivity and data reuse. Uganda's experience is similar to the current global situation. Therefore, this highlights the need for comprehensive strategies to ensure that the success of digital health initiatives is lasting. To address that, this thesis explores possible solutions to overcome and control these challenges.

1.1.1 The Sustainable Development Goals (SDGs) that Foster Uganda's Digital Health Agenda

Uganda's commitment to using technology to improve healthcare is aligned with the Sustainable Development Goals (SDGs). It specifically contributes to SDG 3, which focuses on Good health and well-being. Uganda is actively working towards ensuring healthy lives and well-being for all people of all ages by focusing on improving their access to healthcare services, promoting health education, and using technology to strengthen its healthcare systems [11]. Moreover, employing digital tools such as telemedicine addresses the shortage of healthcare workers, which is a significant barrier to achieving SDG 3. This action is also in harmony with SDG 9, which focuses on Industry, innovation, and infrastructure through aiming to build resilient infrastructure, promote sustainable industrialization, and foster innovation [12]. Uganda's Digital Health Program is expanding its impact on marginalized communities. This effort contributes to SDG 10, which focuses on reducing inequalities. Reducing such disparities in healthcare is achieved through promoting the development of health services for underserved populations by using digital tools such as telemedicine [12, 13]. Since it acknowledges the need for action together internally and externally, Uganda's digital health strategy is in line with SDG 17, which focuses on sharing goals through creating and maintaining global partnerships on sustainable development [12]. In summary, Uganda's digital health agenda is linked with SDGs 3, 9, 10, and 17 because it promotes fostering innovation and global partnerships as well as adopting technological advancements to reduce inequalities within the healthcare systems, and promoting partnerships. Uganda is in a position to make significant contributions to these global goals. This is in line with the Ministry of Health's objective of "Building a Strong and Interoperable Digital Health Information System for Uganda" [12, 14].

1.1.2 Digital Structural Gaps in Primary Healthcare

Digital health emerges as a transformative force addressing crucial structural gaps within primary healthcare. This leads to improved health service delivery across various di-

mensions, as outlined in Table 1.1. A concrete solution to a structural gap has been implemented through the utilization of technologies such as drones for delivering HIV drugs to remote islands in Kalangala District, Uganda. This case serves as a testament to the efficacy of digital health in rectifying structural shortcomings by ensuring that essential healthcare requirements can be met through utilizing digital health in challenging areas [15].

1.1.3 Uganda’s Digital Health Ecosystem Key Players

In guiding Uganda’s dynamic digital health system, it is important to align with key stakeholders who have contributed significantly to improving healthcare. Understanding the specific roles of these organizations not only shapes digital health policy but also ensures alignment with global policies. The key players in Uganda’s digital health ecosystem include government agencies such as the Ministry of Health (MoH), the National Information Technology Authority (NITA), and the Uganda Communications Commission (UCC). These agencies play a critical role in shaping policies, providing regulatory frameworks, and advocating for digital health solutions [22]. Using digital tools such as electronic health records (EHRs), healthcare providers contribute significantly to improving patient care and communication among professionals [85]. Private sector entities like Rocket Health actively contribute by offering services such as telemedicine, e-pharmacy, and health information systems development [24]. Non-governmental organizations like USAID, IntraHealth International, and Elizabeth Glaser Paediatric AIDS Foundation provide funding, expertise, and resources to support digital health solutions. Academic and Research Institutions like Kampala International University and Makerere University School of Public Health engage in research and training, thereby contributing to capacity building in the sector and for Patients. Patients, communities, and Community Health Workers (CHWs) also play a pivotal role in ensuring healthcare accessibility in remote areas, making them integral to Uganda’s digital health ecosystem. While digital health research projects show great promise, they struggle with significant challenges, especially compliance with global policies. These challenges include navigating a complex regulatory framework, such as the Health Insurance Portability and Accountability Act (HIPAA) [25] and the European General Data Protection Regulation (GDPR) [26]. Although the GDPR is a European law, its impact extends beyond the EU, affecting digital health initiatives globally. Complex processes are required to comply with HIPAA and GDPR. This poses some challenges to digital healthcare design and implementation that are clear when trying to collaborate and navigate potential digital divides in the healthcare sector. The rapid deployment of digital tools to address the COVID-19 pandemic demonstrated the adaptability and resilience of Uganda’s health systems. The country’s journey in adopting and harmonizing digital health solutions is a reflection of the collaborative efforts of various stakeholders. While working towards overcoming challenges and aligning with global initiatives, Uganda’s digital health transformation is at the crossroads of innovation, process, and collaboration. It is crucial to ensure that Uganda’s digital health initiatives are locally inclusive and aligned with international standards and expecta-

tions. Such alignment ensures that these projects remain sustainable and successful in a global context. Uganda Industrial Research Institute (UIRI) serves as a center of excellence for the East African Community (EAC) in industrial research. The institution has evolved into a competence and capability centre which champions innovations, translates applied research results into practical applications, which include products and industrial processes, imparts technical skills, and deploys technology across Uganda. However, the existing research policies and guidelines are silent on research data management and data reuse [26]. This has resulted in limited accessibility, ownership, and data interoperability, leading to fragmented digital health data, disconnected research data, and over-reliance on manual analytics within Uganda’s health facilities.

1.2 FAIR Data and Digital Health

In 2014, a collective of academics, industry leaders, funding agencies, and scholarly publishers collaborated to introduce the FAIR principles [27]. These guidelines, designed to enhance data management practices, aim to make both data and metadata easily discoverable by both humans and machines. The four foundational principles of FAIR are ‘Findable’, ‘Accessible’, ‘Interoperable’, and ‘Reusable’ [28]. The absence of FAIR attributes makes deep and integrative data analysis challenging, often requiring significant time for data aggregation. Given the constant evolution of data types and formats, ensuring data is machine-actionable becomes increasingly critical. Harmonizing and integrating data across general-purpose and special-purpose repositories, thereby making data FAIR, holds the key to extracting maximal benefits, saving time, and reducing research costs [29].

Originating in Europe, the FAIR principles gained momentum in 2014 with support from stakeholders in life sciences. Major entities, including the European Union, G7, G20, and the US-based Big Data to Knowledge (BD2K), endorsed these guidelines. In 2017, the European Union established the European Open Science Cloud (EOSC), mandating FAIR-based data stewardship for all publicly funded scientific research from 2020 onward [119, 30]. Although adoption of FAIR has progressed in some regions around the world since 2014, its implementation in Africa, particularly in the realm of digital healthcare, has been gradual. Scholars, such as Van Reisen and others, have emphasized the potential of FAIR principles to address systemic issues in African healthcare, highlighting the need for contextual awareness in implementation [31]. Despite its clear benefits, FAIR uptake in Africa has been limited and slow [31].

Semantic web technologies can be leveraged to make data interoperable by machines. Several platforms aligned with the FAIR principles have been developed. For example, FAIRDOME combines SEEK [32], a web-based resource for sharing heterogeneous scientific research datasets, models, or simulations, processes, and research results, and openBIS, a system for managing biological data. Both systems implement a semantic representation of data using the Resource Description Framework (RDF) data model. Resources such as UniProt have converted the entire database into this semantic representation, allowing both humans and machines to mine these resources efficiently [33].

Furthermore, various efforts have been performed on the FAIRification of existing datasets, making them more fair [34]. FAIRification refers to the process of enhancing existing datasets to meet FAIR standards, ensuring that both resource data and metadata are machine-readable. This involves making metadata clearly describing how data can be accessed and reproduced while ensuring metadata is Findable by machines. The data FAIRification process includes original data retrieval, dataset identification and analysis, the definition of the semantic model, data transformation, license assignment, metadata definition, and FAIR data resource development (data, metadata, license), which involves datasets being generated directly according to FAIR principles, in which data and metadata are in a linked format. This is achieved through the development and use of computational applications and workflows in which data can be consistently appended with metadata.

A noteworthy milestone in FAIR implementation in Africa was the establishment of the East Africa Open Science Cloud for Health (EAOSCH) in 2019 by the East African Community (EAC) [31]. This regional initiative aims to facilitate the sharing of health data among EAC member states, showcasing the advantages of data reuse and interoperability across borders. The East African Health Research Commission (EAHRC) endeavours to set common standards for digital health, emphasizing regional principles for data sharing, system interoperability, and digital tool design, aligned with FAIR principles [35, 31].

1.3 VODAN Africa

The VODAN-Africa architecture is designed to support the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) research data. This research became part of the VODAN-Africa and Asia implementation network in March 2020, which aims to address current and immediate challenges in the use and connectivity of digital health data [290]. The VODAN-Africa and Asia initiative is central to an expanding network of equal partnerships between African and European universities. These partnerships produce or utilize high-quality, high-volume, machine-actionable data, bringing African health insights into the global medical knowledge base to improve health and healthcare in Africa. It is integrated with repositories where datasets are stored and made available for discovery, access, and use under clearly defined terms. This study is driven by the need to bridge gaps in digital health in Africa by exploring the transformative potential of FAIR data principles. It aims to leverage FAIR data support design to overcome the prevalent concerns about the interoperability and reusability of digital health data for quality treatment at the point of care in Uganda. This is achieved by implementing remote querying facilities across Uganda and other VODAN-Africa member countries, ensuring that data remains within health facilities while enhancing the quality of care. Understanding the need for FAIR data requires recognizing its potential to transform the current data landscape. Adherence to FAIR principles ensures that data becomes more findable, accessible, interoperable, and reusable. The research questions outlined below guide the exploration of how FAIR data impacts overcoming interoperability challenges,

enhancing data accessibility, and contributing to a more cohesive and usable digital health ecosystem in Uganda.

1.4 Research Questions and Contributions

The specific focus of our research was to examine the adoption of FAIR data in the digital health landscape in Uganda. Understanding the complexities of digital health challenges, particularly in the context of Africa and Uganda, is essential to formulate meaningful research questions.

1.4.1 Main Research Question

How can data integration, supported by the FAIR data principles, address the challenges of data interoperability and reuse within the African digital health sector?

1.4.2 Specific Research Questions

How does the lack of policies on data management, data sharing, and data analysis contribute to the under-utilization of digital health data in Uganda?

How can the systematic application of FAIR principles address the challenges related to interoperability and reusability in the development and maintenance of digital health data systems in Uganda?

How can data analytics of health data be effectively propagated, integrated, and maintained within health facilities in Uganda through the application of FAIR data principles?

Can the adoption of FAIR principles support and strengthen privacy-preserving access to diverse health data types, particularly for applications in machine learning?

1.5 Thesis Outline

This thesis explores the application of the FAIR principles to digital health data in Africa, with a specific focus on Uganda. It is structured into nine chapters, beginning with an introduction in Chapter 1, which outlines the research problem, the objectives, and the significance of the principles of FAIR data. Chapter 2 reviews the VODAN Africa and Asia initiative and the global implementation of FAIR principles. Chapter 3 investigates data management practices and information flow in Ugandan health facilities through focus group discussions and interviews. Chapter 4 assesses Uganda's regulatory framework for digital health and examines the feasibility of adopting FAIR principles. Chapter 5 explores how FAIR principles can enhance interoperability in Uganda's digital health initiatives. Chapter 6 demonstrates a proof of concept using the VODAN-in-a-Box (ViaB) system for COVID-19 data transformation and exchange. Chapter 7 compares

Thesis Outline

FAIR data with DHIS2 for privacy-preserving health analytics. Chapter 8 examines the integration of machine learning models with FAIR health data for enhanced analytics and decision-making. Finally, Chapter 9 presents conclusions and recommendations for the further application of FAIR principles in Uganda's digital health ecosystem.

Thesis Outline

Table 1.1: Roles, Benefits, and Challenges of Digital Health in Addressing Structural Gaps in Primary Healthcare

Service	Role of Digital Health	Benefits	Challenges
Human Resources for Health (HRH) / Integrated Human Resources Information System (iHRIS) [16, 17]	iHRIS stands at the forefront of addressing HRH challenges. It ensures efficient workforce deployment and management, optimizing the availability of experts and consultants in the healthcare sector.	Improved staffing decisions, streamlined HR processes, and enhanced workforce efficiency contribute to robust healthcare provision.	Resistance to digitalization, data privacy concerns, and the need for extensive training pose challenges to the seamless implementation of iHRIS.
Community Health Workers (CHWs) / Village Health Teams (VHTs) [18]	Digital Health empowers CHWs/VHTs in administering crucial services like family planning, antenatal care, vaccinations, and Integrated Community Case Management (iCCM).	Digital tools facilitate the coordination and effectiveness of front-line healthcare workers, bridging gaps in service delivery in remote and underserved areas.	Limited digital literacy among CHWs, connectivity issues, and the need for ongoing technical support can hinder the effective use of digital tools.
Health Service Delivery in Remote Areas [19, 20]	Digital health interventions in telemedicine break geographical barriers, extending healthcare services to remote or underserved areas.	Telemedicine ensures that individuals in distant locations have access to healthcare expertise and services, reducing disparities in health service delivery.	Limited infrastructure, poor connectivity, and the need for investment in digital infrastructure can pose challenges to the widespread implementation of telemedicine.
Supply Chain Management and Stock Availability[21]	Digital health interventions in supply chain management ensure the availability of essential health supplies such as blood and drugs.	Innovations like drone delivery (e.g., Zipline) provide efficient and timely transportation of crucial medical resources, addressing logistical challenges in remote or isolated regions.	Regulatory hurdles, and the initial cost of implementing health technology, pose challenges to the scalability of such solutions.

Chapter 2

VODAN Africa and Asia

This chapter is partially based on the following publications:

Mirjam van Reisen, Mia Stokmans, Mariam Basajja, Antony Otieno Ong'ayo, Christine Kirkpatrick, Barend Mons; Towards the Tipping Point for FAIR Implementation. *Data Intelligence* 2020; 2 (1-2): 264–275. doi: https://doi.org/10.1162/dint_a_00049

Mirjam van Reisen, Mia Stokmans, Munyaradzi Mawere, Mariam Basajja, Antony Otieno Ong'ayo, Primrose Nakazibwe, Christine Kirkpatrick, Kudakwashe Chindoza; FAIR Practices in Africa. *Data Intelligence* 2020; 2 (1-2): 246–256. doi: https://doi.org/10.1162/dint_a_00047

Abstract

This chapter provides the global perspective on the transformative role of FAIR principles in scientific data management and stewardship with a specific focus on Africa. This highlights the persistent challenges within African healthcare systems, as illustrated in Table 2.2, underscoring the urgent need for inclusive and equitable FAIR practices in diverse scientific domains. A detailed analysis of FAIR Data implementation across continents and scientific fields was presented, revealing significant disparities. The concentration of initiatives in biological and natural sciences, primarily in Europe and America, raises critical concerns about the global inclusivity of FAIR principles in Africa and Asia. Building on this context, the chapter examines how the Virus Oubreak Data Network(VODAN) initiative addresses key challenges, including barriers to collaboration, limitations in data reuse, sustainable development, and the co-creation of knowledge. Kingdon’s theory of agenda setting is applied to the African context to explore the adoption of FAIR Data, emphasizing the pivotal role of policy actors in facilitating this transition. Within this framework, the chapter positions VODAN-Africa as a leading advocate for FAIR principles in Africa, highlighting its strategic efforts to address these challenges. Reference is made to the GO BUILD programme, which enhances stakeholder awareness and promotes the development and use of FAIR Data. These efforts underscore VODAN-Africa’s unwavering commitment to fostering the adoption of FAIR principles, ultimately contributing to a more inclusive and sustainable data ecosystem in the region.

Keywords: FAIR Data, health, digital health, digital health in Africa, mHealth, data-driven science, FAIR Implementation networks, GO-FAIR

2.1 Introduction

The global landscape of scientific data management and stewardship has been transformed by the implementation of the FAIR principles [30]. Designed to enhance the findability, accessibility, interoperability, and reusability of data, these principles represent a significant milestone in promoting effective data management and collaboration across diverse scientific disciplines [31]. The adoption of FAIR practices is a key factor that has emerged to provide the necessary innovative remedies rooted in digital health frameworks for the issues faced in African healthcare systems [37]. The FAIR principles are vital as they offer a structured framework for managing data, which is essential to address challenges in disease surveillance and healthcare delivery.

While the global impact of FAIR principles has grown considerably [37], Africa's healthcare systems present unique challenges that demand special attention. This is evident in the dominance of FAIR data management within current health systems and the bio-ecology field in Africa, despite limited application of FAIR principles in other scientific disciplines and communities across the continent [38]. To address this, African institutions must increase the implementation of FAIR principles to ensure that best practices outlined in the FAIR guidelines are consistently applied across diverse disciplines.

To resolve the challenges facing existing healthcare systems, FAIR principles must be integrated into digital healthcare. Van Reisen and others highlight a geographical imbalance in the adoption of FAIR initiatives, with widespread implementation in the biosciences of Europe and the United States. In contrast, the systems in Africa remain in the early stages of adopting FAIR principles [39]. This chapter, therefore, focuses on introducing VODAN (Virus Outbreak Data Network), which is a program that is designed to address the challenges that are associated with data on viral outbreaks in Africa and Asia. Furthermore, the chapter highlights the significance of VODAN Africa and Asia in addressing data management challenges across these continents. By presenting this information, the chapter contributes to a broader understanding of the global application of FAIR Data.

2.2 Implementing FAIR Principles in Africa and Asia

The Virus Outbreak Data Network (VODAN) is a pioneering initiative driving transformations in data management, aligned with advancing the adoption of FAIR principles across Africa and Asia. To understand the context of FAIR Data use, it is essential to examine the distribution of efforts across scientific disciplines and continents, as illustrated in Table 2.2. This table¹ provides valuable insight into the current state of FAIR adoption worldwide, emphasizing the importance of broad coverage across scientific disciplines.

¹**Source:** van Reisen et al. [30]

Implementing FAIR Principles in Africa and Asia

Table 2.1: Global distribution of FAIR Data implementation efforts across continents and scientific disciplines (n=100)

	Africa	America	Europe	Europe+	Other	Unknown	Total
Bio-science	1	11	53	5	2	12	84
Natural science	0	3	5	1	0	2	11
Social science	0	0	3	0	1	1	5
Total	1	14	61	6	3	15	100

The distribution of FAIR Data implementation efforts reveals significant disparities across continents and scientific disciplines. Notably, most biological science programmes are concentrated in Europe. In contrast, only one programme is situated in Africa, raising critical concerns about the inclusiveness of FAIR principles in the African context. In this context, VODAN is pivotal in advancing FAIR Data processing and management. VODAN primarily targets viral outbreak data in Africa and Asia, addressing critical gaps in data management for infectious diseases. The programme aligns with FAIR principles, contributing significantly to efforts aimed at ensuring data is findable, accessible, interoperable, and reusable. The subsequent sections delve into the challenges VODAN faces in implementing FAIR principles and examine the broader implications of these efforts for digital health solutions in Africa and Asia.

2.2.1 Challenges in Digital Health Implementation

Several inherent challenges shape the effectiveness of digital health initiatives in Africa. This section outlines these challenges and examines their impact on data management and utilisation within the framework of FAIR principles. As shown in table 2.2, these challenges include collaboration challenges, data reuse issues, limited scalability issues, lack of knowledge sharing, and insufficient short-term effort.

Table 2.2: Challenges in Digital Health Implementation in Africa

Challenges	Description
Interoperability challenges [4-8]	Diverse digital health systems face obstacles in achieving seamless data exchange and integration, hindering collaborative efforts.
Data reuse issues [2, 7, 9]	Challenges that are associated with data reuse remain, including disparate systems, inconsistent documentation, and divergent data standards, preventing a unified digital health system.
Limited sustainability [2, 10-13]	Digital healthcare often has difficulty maintaining long-term impact and relevance. Many initiatives focus on short-term goals, preventing the development of robust, scalable solutions.
Lack of co-knowledge building [2, 14, 15]	Collaborative knowledge-building efforts among stakeholders are insufficient, limiting collective impact.

Collaboration challenges present a major obstacle, as digital health systems often fail to achieve seamless data exchange and integration. These challenges hinder collaborative efforts and restrict the ability to achieve comprehensive and integrated healthcare solutions. Data reuse challenges exacerbate the situation, as varying formats, incompatible documentation, and disparate data standards impede the development of a unified digital health ecosystem. Limited sustainability remains a pervasive issue, with digital healthcare initiatives frequently needing support to maintain long-term impact, relevance, and scalability. Such limitations hinder the development of robust and scalable solutions. Additionally, insufficient knowledge sharing among stakeholders constrains collaboration and diminishes the overall impact of digital health interventions.

2.2.2 Understanding the Adoption of FAIR Principles

Kingdon’s theory, also known as the multiflow model [48], is a framework that is used to explain the administrative process. It can also be applied to the process of technology adoption in organizations. Technology adoption occurs when all three strands come together at the same time. This requires the presence of a problem to address, a viable technical solution, and the political will to adopt the solution. The theory also recognizes the importance of structure involving professionals who consist of individuals or groups working to focus attention on a problem or a solution and mandates adoption, the planning staff in technology adoption, such as IT professionals, consultants, or other stakeholders who have basic technical skills, and other stakeholders who advocate for the adoption of a particular technology. Overall, Kingdon’s theory offers a valuable framework for comprehending the complex process of technology adoption. It also highlights

Implementing FAIR Principles in Africa and Asia

the need to address problems, devise viable solutions, and navigate the political landscape to successfully implement new technologies within organizations. This discussion examines how VODAN aligns with Kingdon's theory of abundant water, emphasizing its role in lobbying for FAIR principles and highlighting the program's function as a policy agent advocating for their adoption in the region.²

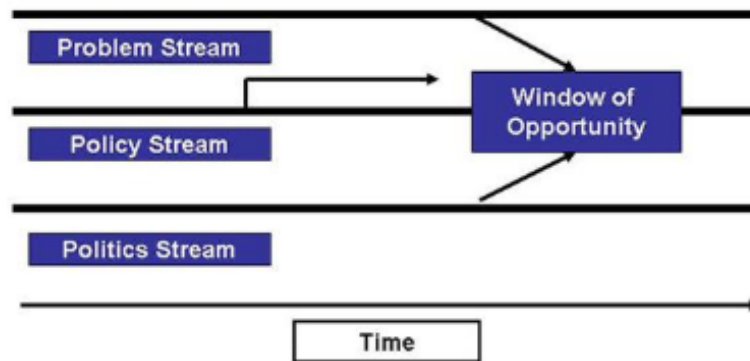


Figure 2.1: Conceptual representation of the Kingdon's multiple streams framework, illustrating the convergence of the problem, policy, and politics streams to create a window of opportunity for policy change

The theoretical framework underpinning this thesis is partially derived from Kingdon's three-stream model. Kingdon posited that the process of policy agenda-setting comprises the problem stream, the policy stream, and the political stream. Kingdon identified that new ideas could enter the policy agenda when the problem, potential policy solutions, and political relevance of the problem are aligned. At such a point of alignment, a policy window opens, which is a specific moment in time, when a new idea can enter the policy agenda [49]. The dynamic process that Kingdon describes allows for the investigation of how policy changes and the role of policymakers and other policy entrepreneurs, who are 'members' of policy networks in a certain domain [39, 49].

Van Reisen and others explain that the discrepancy between the FAIR Principles and the opinions and beliefs of researchers and practitioners who are expected to use these Guidelines can be considered cultural entropy [42]. The authors go on to argue that if cultural entropy is large, it will hinder the acceptance of the FAIR Principles, as it is likely to result in a discrepancy in terms of what is perceived as the problem and what the solution could be. Kingdon's theory has been used to analyze the FAIR Principles [30]. It has also been applied to the implementation of the FAIR Principles when managing scientific data [48, 50]. The framework that is presented in Kingdon's theory has been used widely since the publication of Kingdon's "Agendas, Alternatives and Public Policies" in 1984. This theory has also been used to examine the cause of the gaps that are in modern

²Source: <https://www.policynl.ca/>

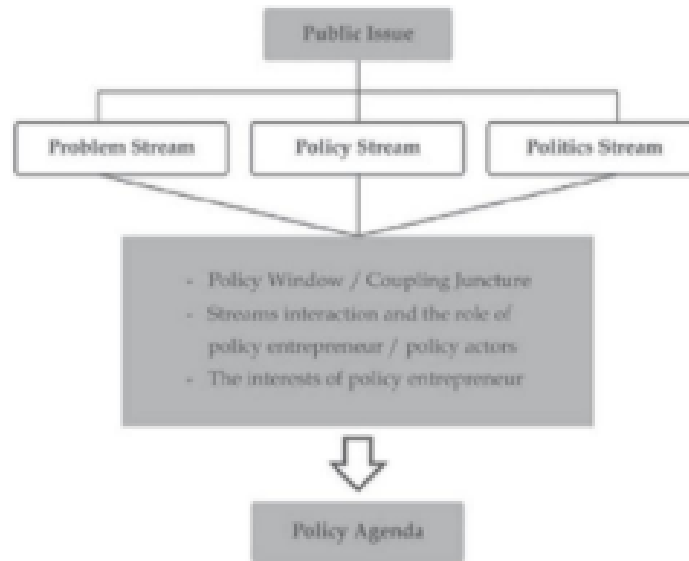


Figure 2.2: Kingdon's multiple streams theory of agenda setting.

slavery policies [49]. It is often used to understand the reasons for policy changes and the processes that are followed in making those adjustments. Kingdon's theory starts by acknowledging that the process of change is not linear or logical and that actors causing this change are limited by bounded rationality. Kingdon's framework is relevant to a wide range of governance systems and policy applications, including implementation in both national and sub-national layers of governance.

Initiatives like VODAN are pivotal in advancing FAIR principles through actively advocating for the adoption of these principles and facilitating the implementation of FAIR-compliant data management practices. Lastly, the "political" stream encompasses the broader socio-political landscape that influences policy decisions. Political support and commitment from key stakeholders are important to driving the systemic adoption of FAIR principles.

2.2.3 Conceptual Framework

This study uses Kingdon's multi-stream theory of agenda setting as a conceptual framework. Kingdon's model, which is illustrated in 2.2, provides a comprehensive understanding of policy and policy establishment. These three elements are interrelated in the theory: problem streams, policy process, and politics.

Source: [271]

In the problem stream, issues are identified and recognized as problems that require solutions. For example, this study recognizes the inefficient use of data and the need to effectively manage data in emergencies. Planning strategies are then developed to

address the problems that were identified. The policy stream involves developing policy alternatives to address the identified problems. Here, the FAIR Principles are a solution to the under-utilization of data. The political stream also identifies the political will to shape how a specific issue is addressed. In the context of this research, political support from entities like the European Commission and African governments helps to support FAIR Principles.

The intersection of these three streams creates a "policy window," which is a critical moment where new ideas, such as the global adoption of FAIR principles, can enter the policy agenda. This framework provides guidelines for analyzing the global adoption and implementation of FAIR principles. It considers the interplay between identified problems, proposed solutions, and the political will to adopt these solutions on a global scale. This conceptual framework provides a lens through which the dynamics that influence the adoption of FAIR Principles can be examined. In turn, this emphasizes the role of policy actors, the urgency of the problems that have been identified, and the alignment of political will for the FAIR principles to be implemented worldwide in the Internet-based FAIR Data and Services (IFDS) [53].

2.2.4 The Acceptance of FAIR Data principles and their Implementation

The factors that influence the acceptance and successful implementation of FAIR principles include perceived usefulness, ease of use, and the alignment of FAIR design and implementation with the changes that are constantly occurring within the Information and Communication Technology (ICT) sector.

The GO FAIR initiative [117], originally launched on their official website during its actively funded phase (2017-2020), emerged as a global network of offices and implementation networks aimed at kick-starting FAIR-aligned data infrastructures, initially within the context of the European Open Science Cloud (EOSC). Over time, GO FAIR evolved beyond its original European mandate, with offices and implementation activities spreading globally, reflecting the broader and continuing need for FAIR guidance across domains and regions.

In its current phase, often referred to as GO FAIR, GO FAIR functions as a loosely federated international network of offices, supported and coordinated through the Dutch GO FAIR Foundation (GFF), which now acts as the primary legal entity sustaining the initiative. The GFF founded the GO FAIR brand and logo and plays a central role in maintaining continuity, coordination, and strategic direction in the absence of large-scale central funding. Rather than evolving into a bureaucratic organization, GO FAIR has intentionally remained lightweight, emphasizing facilitation, interpretation, and practical implementation of FAIR principles, particularly with a strong focus on machine-actionable aspect of FAIR.

Strategically, GO FAIR continues to operate through three interrelated pillars: GO CHANGE, GO TRAIN, and GO BUILD. GO CHANGE focuses on socio-cultural transformation by promoting policies, incentives, and stakeholder engagement that support

Implementing FAIR Principles in Africa and Asia

the adoption of FAIR principles. This pillar addresses the diverse beliefs, motivations, and concerns of stakeholders involved in data collection, sharing, and reuse. GO TRAIN coordinates awareness-building and skills development, with an emphasis on cultivating competent data stewards capable of designing and implementing FAIR-compliant data management plans. GO BUILD concentrates on aiding to choose, align and coordinate FAIR-aligned technical standards, best practices, and interoperable infrastructure components required for effective FAIR data implementation. Together, these three pillars form a coherent strategy that positions GO FAIR as a facilitative framework guiding communities toward robust FAIR adoption.

The ICT architecture exemplified by GO BUILD helps shape perceived usefulness and ease of use, which influences the implementation of the FAIR principles for digital health data in Africa [30, 31, 117]. In the process of building applications, decisions are made about including specific research characteristics and how they should be implemented within a framework that is aligned with the FAIR principles. These decisions, driven by human choices, contribute to the utility of FAIR for specific research purposes within defined frameworks. However, it is important to acknowledge that the current design of the FAIR principles may not fully account for the diversity of research practices and contexts in the real-world. This implies that it is necessary for FAIR Data implementation to be adaptable and applicable to a broader range of research settings beyond the controlled environments in which it may have been initially conceptualized or tested, that is, bio-science, natural science, and social science. These oversights pose challenges to the broader acceptance of FAIR in diverse research practices, especially within FAIR Data Implementation [30, 31].

In the context of African healthcare care, the acceptance of FAIR faces obstacles related to contextual understanding, limitations in digital connectivity, and skepticism about the intentions of this European idea. Historical interactions with European initiatives, marked by extraction and exploitation, have created a perception that Europeans, in this case, are extracting and storing data out of reach of African communities. Moreover, existing data gathering practices are often Eurocentric as they fail to consider African perspectives, objectives, and epistemology. To counteract these challenges, the African scientific community must become a major player in co-designing, co-implementing, and co-monitoring the FAIR architecture for Africa. This has been done by VODAN Africa in collaboration with Universities, Hospitals, and Ministries in policy formulation [56].

To advance FAIR globally, it is crucial to explore the barriers and opportunities associated with its implementation. Two key perspectives for analysis are the public policy agenda-setting that shaped the FAIR Data architecture and the acceptance of FAIR Principles by researchers and practitioners in their daily practices.

From a public policy perspective, the acceptance of FAIR Data depends on opening a policy window that aligns with societal needs. This also requires a synergy between the problem stream, policy stream, and political stream, as described by Kingdon [30, 31, 49]. Social entrepreneurs play a vital role in identifying challenges, promoting ideas, and activating political will to embrace FAIR. Events that change the dominant issues on the

agenda can trigger synergy among the streams, emphasizing the importance of focusing on such events for the worldwide implementation of FAIR. Motivations, beliefs, and opinions of different stakeholders influence the acceptance of FAIR Principles. Social entrepreneurs must be aware of these varying motivations and objectives among policy-makers, data managers, researchers, and practitioners. The GO CHANGE initiative is critical in motivating potential users and fostering acceptance. It does this by addressing diverse beliefs and opinions about data collection, use, and reuse. In order for FAIR to gain broad global acceptance, it is essential that cultural entropy is addressed, diverse research domains are considered, and alignment with the beliefs and practices of the research community is ensured. When initiatives like GO CHANGE and GO BUILD are implemented together, they can significantly contribute to overcoming barriers and facilitate the global adoption of FAIR Principles.

2.2.5 Conclusion

The FAIR principles have achieved significant acceptance and implementation, representing a major milestone in the advancement of data management and stewardship. Nevertheless, the current implementation landscape demonstrates that FAIR adoption remains disproportionately concentrated in the Western hemisphere, particularly within the bio- and natural sciences. Expanding FAIR adoption globally, therefore, requires broader disciplinary coverage and contextual adaptation, ensuring that FAIR applications are relevant across diverse research domains and epistemologies. Such inclusivity is essential not only for improving the generalizability of research outcomes but also for enabling the development of robust, context-sensitive research and implementation models. For FAIR to achieve sustained global uptake, it is imperative that researchers, practitioners, and policymakers collectively recognize its value while acknowledging variations in digital infrastructure, research cultures, and data governance norms.

Within this evolving landscape, the GO FAIR movement continues to play a pivotal facilitative role. Following its initial centrally funded phase, GO FAIR has transitioned into a lightweight, internationally federated network of offices, coordinated and sustained through the GO FAIR Foundation. Rather than operating as a prescriptive or bureaucratic authority, GO FAIR increasingly functions as a FAIR co-pilot, supporting communities in interpreting, implementing, and aligning FAIR principles with local needs and capacities. The GO CHANGE initiative remains particularly significant in this regard, as it addresses socio-cultural diversity, stakeholder motivations, and resistance to change through awareness-building and engagement. The training and coordination of Implementation Networks (INs) under this pillar continue to be central to fostering acceptance and enabling FAIR-aligned practices across domains. As architects of interoperable FAIR data spaces, INs contribute to the development of foundational frameworks that support FAIR integration across heterogeneous academic and applied research environments.

VODAN-Africa and Asia represent critical regional contributions to the global FAIR implementation journey. Established in 2019 as African-led initiatives, these networks bring together partners across multiple countries to address region-specific challenges

Implementing FAIR Principles in Africa and Asia

such as contextual misalignment, infrastructural constraints, and concerns regarding data ownership and sovereignty. Their work demonstrates how FAIR principles can be operationalised in ways that respect local priorities through approaches certified as Findable, Accessible, Interoperable, and Reusable with Ownership, Localization, and Regulatory Compliance (OLR). By engaging in initiatives aligned with GO BUILD and participating actively in policy formulation alongside universities, hospitals, and ministries, VODAN-Africa and Asia exemplify how FAIR can be co-designed, co-implemented, and co-governed. Such regionally grounded engagement underscores the importance of combining technical guidance with socio-political sensitivity, reinforcing the role of GO FAIR and its partners in enabling equitable, sustainable, and globally relevant FAIR data ecosystems.

Chapter 3

Information Streams in VODAN-Africa Health Facilities: The Case of Uganda

This chapter is based on the following publication:

Basajja, M., Nambobi, M.: Information streams in health facilities: The case of Uganda. Data Intelligence 4(4) (2022). doi: https://doi.org/10.1162/dint_a_00177

Abstract

In this chapter, we discuss the current challenges of health data management in Uganda, the limitations of paper-based systems, and the potential role of FAIR architecture in improving data accessibility and reuse. Amid the prevailing COVID-19 pandemic, the lack of digitally recorded and interconnected health data presents significant challenges for effective analysis. Virus outbreaks, such as the current pandemic, highlight the potential for optimizing and reusing data to enhance the management of future outbreaks. However, there is a general lack of knowledge about the actual flow of information in health facilities, which is also the case in Uganda. In Uganda, where this case study was conducted, comprehensive knowledge is lacking about the types of data collected and the methods used to collect it throughout a patient's journey in a health facility. This study investigates information flows of clinical patient data in health facilities in Uganda. The findings revealed that nearly all health facilities in Uganda store patient information in paper files, typically organized on shelves. Hospitals in Uganda rely on paper-based tools, including reporting forms, registers, and manuals, to collect district data in aggregate form, which are then submitted as digital reports to the Ministry of Health Resource Centre. These reporting forms are not digitised and are therefore not machine-actionable. Hence, it is not easy for health facilities, researchers, and others to find and access patient and research data. Additionally, reusing and integrating these data with global digital health datasets is difficult, often leading to the erroneous perception that health data in Uganda are scarce. Implementing a FAIR architecture offers the potential to address these challenges and support the transition from paper-based to digital records within Uganda's healthcare system.

Keywords: Data management, Information flow, FAIR principles, Hospital information system(HIS), Electronic medical record (EMR), Electronic health record (EHR), Patient health record (PHR)

3.1 Introduction

The crucial role of rapid data sharing in accurately diagnosing Coronavirus 2 (SARS-CoV-2), the causative agent of the ongoing new coronavirus disease (COVID-19) pandemic, has been underscored [57]. In response, the World Health Organization (WHO) called for global scientific cooperation and equitable sharing of knowledge, data, and intellectual property [58]. WHO's initiative to ensure that all COVID-19-related data is available immediately and is accessible with proper attribution is noteworthy [57]. Virus outbreaks such as the COVID-19 pandemic have made it necessary to optimize and reuse data so that future outbreaks can be effectively managed. However, there are persistent challenges in accessing and making data interoperable and reusable. This is demonstrated by the difficulties that have been encountered with data from past epidemics, such as Ebola [59].

The healthcare system in Uganda embodies a dynamic partnership between public healthcare and private healthcare providers, with the Ministry of Health serving as the authority [60]. The National Policy on Public-Private Partnerships in Health has formalized this collaboration. Various types of healthcare providers exist within the private sector. They are categorized into private-not-for-profit health providers (PNFPs), private health practitioners (PHPs), and traditional and complementary medicine practitioners (TCMPs). This categorization includes non-governmental and faith-based organizations, such as the Uganda Catholic Medical Bureau, Uganda Protestant Medical Bureau, Uganda Orthodox Medical Bureau, and Uganda Muslim Medical Bureau [60]. Beyond these recognized categories, there is an informal health sector, comprising individuals with limited or no formal health training engaged in treating patients and selling drugs in small pharmacy stores. Recognition within the private health sector requires registration with relevant regulatory bodies and obtaining the necessary licenses. Oversight is provided by entities such as the Uganda Medical and Dental Practitioners Council, Uganda Nurses and Midwives Council, Allied Health Professionals Council, and Pharmacy Council [17].

The Ugandan healthcare system functions across five distinct levels, each assigned a specific function (see figure 3.1). At the community or village level, the facilities of the Health Center I (HC I) and the Village Health Teams (VHTs) play an important role. Moving up, the parish and sub-county levels are served by HC II facilities (often referred to as dispensaries) and HC III facilities, respectively. HC IV facilities and general hospitals are located at the county or district level, while regional referral hospitals (RRHs) and national referral hospitals operate at the regional and national levels, respectively. In particular, HC I to HC IV facilities fall under the jurisdiction of the local government, while regional referral and national referral hospitals maintain a semi-autonomous status [145].

In this chapter, the intricate details of information flow within health facilities in Uganda were delved into. The practices and tools used to manage the data within these facilities were explored through interviews conducted in prominent health facilities, including Lira Regional Referral Hospital, Case Clinic Hospital, and Kampala International University Teaching Hospital. Furthermore, data flow from health facility levels to the national Health Management Information System (HMIS) was analyzed. As these as-

Introduction

pects were studied, the insights of the current literature were integrated. This provided a comprehensive understanding of the challenges and opportunities in healthcare data management within Uganda’s healthcare system.

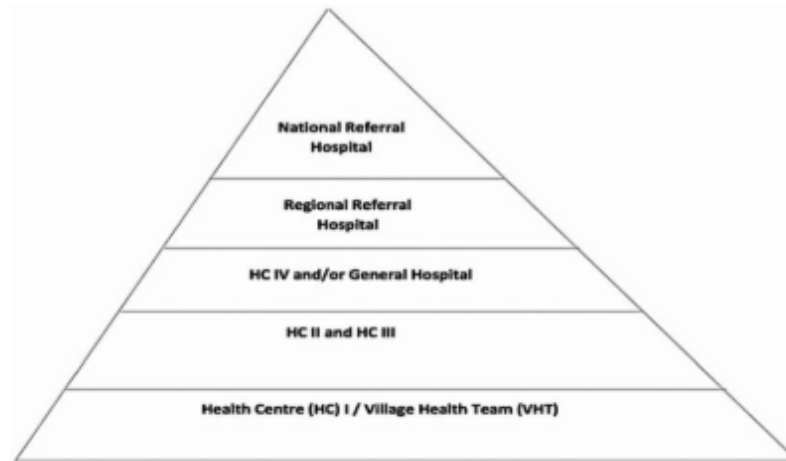


Figure 3.1: Pyramid diagram depicting the hierarchical levels of the Uganda’s Healthcare system including community health services, health centres, general hospitals, regional hospitals and national referral hospitals

Uganda’s healthcare system operates on the District Health System, a referral system that uses community-based services as the first point of delivery for primary healthcare. In this context, primary healthcare includes preventive, curative, and health promotion programmes. The District Health System [61]. In Uganda, VHTs, also known as community health workers, are the first point of contact between community members and the healthcare system. VHTs then refer patients to parish and sub-county-level health units (HC I and II), respectively. These lower-level health workers can then refer patients for more specialised treatment to the HC IVs and general hospitals at the county and district level, respectively [62]. However, the rapid proliferation of political units (parishes, sub-counties, counties, and districts) has resulted in some areas lacking corresponding health facilities. For example, if the parish does not have an HC II, patients are referred to the next level of healthcare, e.g., an HC III.

Adding another layer to the context, Uganda boasts a total of 6,937 health facilities, encompassing public, private, and not-for-profit entities. Government-owned facilities constitute 45.16% (3,133), Private and Not for Profit (PNFP) establishments make up 14.44% (1,002), and Private for Profit (PFP) facilities contribute 40.29% (2,795), with a minimal 0.10% (7) representing community-owned facilities. Notably, Government and PNFPs predominantly operate higher levels of health facilities, while Private For-Profit establishments are primarily involved in lower levels, such as HC IIs and clinics.

Even though the adoption of the District Health Information Software (DHIS) 2 as part of Uganda’s core Health Management Information System (HMIS) has facilitated the transition from paper-based reporting and storage to an electronic web-based system,



Figure 3.2: Piechart illustrating the distribution of Health Facilities in Uganda by ownership category including government-owned, private not-for-profit, private for-profit and community owned facilities

most of the reporting tools at the lower-level health facilities are still paper-based [64, 65]. Some of the data captured on the paper-based HMIS forms are not incorporated in the online DHIS2. This exacerbates the general lack of knowledge of the actual flow of information in health facilities in Uganda. Additionally, records personnel are often tasked with collecting and capturing information that exists outside the paper-based system. This labor-intensive process, requiring double entry of HMIS data into both the online DHIS2 platform and paper-based versions, is not unique to Uganda and significantly demotivates health personnel [309]. Furthermore, such data offers little value for data analytics at the point of care. The burden of data entry is, therefore, of no benefit to the health facility that produces the data [290].

A noteworthy challenge lies in the tedious and time-consuming data entry process, where records personnel are tasked with collecting and capturing information that exists solely in paper-based form. This cumbersome double-entry process, both on the online DHIS 2 platform and in paper-based versions, proves demotivating to health personnel [273, 64]. Moreover, the observed data entry practices often fall short in delivering tangible benefits to the originating health facility, particularly in terms of facilitating data analytics at the point of care.

To understand the actual flow of health information in Uganda's Health Management Information System (what data is collected, how it is stored, transferred and accessed in hospitals and/or clinics in Uganda), interviews were conducted at three major health facilities: Lira Regional Referral Hospital (LRRH) (Lira District, Northern Uganda), which is a public health facility; Case Clinic Hospital (Kampala District, Central Uganda), which is a private health facility; and Kampala International University Teaching Hospital

Research Question and Objectives

(KIU-TH) (Bushenyi District, Western Uganda), which is also a private health facility. LRRH is the only regional referral facility in the sub-region offering both specialised clinical care, including care for HIV/AIDS, and general clinical services. The hospital is a free hospital serving numerous districts (all eight districts in the Lango sub-region in Northern Uganda).

The overall goal of this study was to understand the data management practices and information systems used at selected healthcare facilities (hospitals/clinics) in Uganda. Emphasis was put on documenting the tools used and data collected at every step along the patient's journey/care timeline (e.g., registration → history and physical examination → laboratory testing → diagnosis → prescription), discovering ways to transform the data into FAIR data and establish a FAIR Data Point. This study specifically looked at whether the data is digitised and what tools are used to capture patient data at every point in the data flow, from the health facility level to the national HMIS level.

3.2 Research Question and Objectives

To evaluate the general flow of information within Ugandan health facilities and examine the potential for implementing FAIR Guidelines in the country's e-health sector, the following research questions and sub-research questions are proposed.

Research Questions

Main Research Question: What are the current data management practices and information systems employed in Uganda?

Sub-Research Questions:

Sub-Research Question 1:

How does the ideal state health information management in Uganda compare to its actual implementation in various healthcare facilities?

Sub-Research Question 2:

What perspectives and experiences do different stakeholders hold regarding Uganda's healthcare information management system?

Sub-Research Question 3:

How do health information management practices differ among various types of healthcare facilities in Uganda?

The objectives of this study are as follows:

To document the data management processes in selected Ugandan health facilities and evaluate the findability, accessibility, interoperability, and reusability (FAIR) of the collected data.

To assess the readiness of Uganda's current health data management system for implementing FAIR principles, specifically examining whether data are digitized

Methods and Materials

and the tools used to capture patient information at each point in the data flow from health facilities to the national HMIS level.

To evaluate the feasibility of mapping the HMIS reporting forms used in Ugandan health facilities to machine-readable formats (RDF), enabling the creation of machine-actionable (FAIR) digital objects. Additionally, to explore the potential for developing user-friendly, web-based data input forms that support FAIR data and interoperability.

3.3 Methods and Materials

In this section, we present the study design and sampling procedure of the research.

3.3.1 Study Design

This study employs qualitative data collection methods, including observations and interviews, to understand and document the flow of information and data management practices. The study was conducted at three health facilities located in different regions of Uganda. Observation provided context and an understanding of how patient information flows play out in the health facilities. Face-to-face and follow-up interviews were conducted to investigate the underlying meaning behind the observed events. These interviews provided insights into the motivations, preferences, and events that shaped current behaviors and practices. The interviews were held in English and followed a semi-structured format using a questionnaire (see Supplementary Material). Follow-up interviews were conducted via phone and email to clarify issues raised during the initial interviews. The findings informed the development of graphical representations illustrating the flow of patient data within the hospitals. The data collection aimed to evaluate the types of data gathered from patients and the methods employed for data capture.

3.3.2 Sampling Procedure

The selection of health facilities was based on the ease of obtaining authorisation from the respective district local government authorities and district health officers. Nevertheless, all health facilities in Uganda adhere to standardized guidelines set by the Ministry of Health [64]. The maximum variation sampling approach [67, 68] was employed to account for the diverse range of data management practices. Two private-sector facilities and one public-sector facility, located in different regions of Uganda, were selected for the study. The participants who took part in the study consisted of nine data management personnel at the respective health facilities (i.e., data clerks, records officers, and statisticians), which is a relatively modest sample. Participants were included based on their knowledge and experience with Uganda’s Health Management Information System (HMIS). Participants were informed that their responses would contribute to understanding data management practices in health facilities as part of the Virus Outbreak Data Network (VODAN) research project. Participants were also informed that their participation was

Methods and Materials

voluntary, and they were under no obligation to answer any questions. Participation was entirely voluntary, and participants could withdraw at any time. Verbal consent was obtained for the initial interviews, while written consent was provided for the follow-up interviews. Respondents were assured that their identities would remain anonymous and that their data would not be published or disclosed.

Interviews

Face-to-face interviews were conducted with data management personnel at four health facilities in different geographic regions in Uganda: KIU Teaching Hospital (Western region), Lira Regional Referral Hospital (LRRH) (Northern Region), the Antiretroviral Therapy (ART) Clinic Lira (Northern region) which is part of Lira Regional Referral Hospital, and Case Clinic Kampala (Central region). With the aim of understanding what data is collected and how it is collected at the level of hospitals and clinics, a questionnaire was designed (see Supplementary Material) to collect information from the data management personnel at the respective health facilities. The first step was to document the actual data practices at the health unit, without interfering with decisions on how the data is managed. The overall goal was to enable the semantic interoperability of clinical data at the hospital or clinic with other datasets available elsewhere. The questionnaire also aimed at obtaining information regarding the regulations and policies governing data at the participating hospital/clinic.

The purpose of the interviews was to discover whether the patient data is recorded in digital or paper form; what templates are used for data collection (if any); who has access to, owns, and makes use or reuse of the data; whether or not an electronic health record system is in place; whether or not an information system for computer-based data storage, retrieval, and analysis is used; whether or not the different data systems in place are interoperable; what security and authorisation is needed to access the data; how digital data is stored; and, finally, the challenges faced in all the above data handling procedures.

The interviews were conducted with 9 participants in the period January - February 2020. The participants were associated with five health facilities. The participants were working as records management personnel, medical records officers, data clerks, data officers, monitoring and evaluation officers, data statisticians, IT specialists, and bioinformatics officers.

Table 3.1: Interview Details

# of Participants	Position
5	Records management personnel
1	Monitoring & evaluation officer
2	Statistician
1	IT specialist

Mariam Basajja conducted the interviews in person at the selected hospitals and clin-

Results

ics. A questionnaire (see Supplementary Material) was used to facilitate the interviews and collect participant data. Three primary questions were posed to the study participants:

How does patient data flow from the moment a patient arrives at the hospital to their discharge?

What types of patient data are collected?

How is patient data collected, and where is it stored?

The questionnaire used in this study was carefully designed to ensure its relevance and effectiveness in capturing comprehensive insights into the complexities of information flow within Ugandan health facilities. Informed by a synthesis of current literature, the research team identified key themes and critical areas for investigation. Subsequently, a draft questionnaire was crafted, incorporating open-ended and closed-ended questions to solicit detailed responses from participants. The draft underwent a rigorous review process involving experts in public health, Ministry of Health officials, data management professionals, and healthcare administrators to improve its clarity, comprehensiveness, and cultural sensitivity..

After refinement, the questionnaire was evaluated for appropriateness and effectiveness through a pilot study. Data from pilot study participants, along with observations and insights from the research team, played a central role in further refining the questionnaire. Ethical approval for the study was subsequently secured from the Kampala International University Ethics Review Committee, ensuring compliance with established research ethics guidelines.

The interviews were conducted in English, the official language of Uganda. This was done to facilitate clear communication with participants. Each interview lasted approximately 120 minutes, allowing participants to elaborate on their experiences and perspectives regarding challenges facing healthcare facilities. The interviewer listened attentively, took notes during the interview sessions, and sought clarification as needed. To deepen understanding, follow-up discussions were conducted via phone and email, enabling participants to elaborate on specific points and ensuring a thorough exploration of the topics covered in the questionnaire. This systematic approach to designing and approving the questionnaire strengthens the robustness and validity of the study's findings.

3.4 Results

This section outlines the findings on health information systems and data management practices within healthcare facilities in Uganda. It begins with an overview of the ideal scenario for health information management and an analysis of the actual state of affairs in each facility study, and concludes with insights from individual stakeholders.

3.4.1 Ideal Health Information Management in Uganda

In an ideal scenario, Ugandan healthcare facilities would implement a unified electronic Health Information System (HIS) to streamline patient data management. This system would incorporate standardized data collection and reporting procedures across all facilities and include robust security measures to safeguard patient data. Seamless integration between local HIS and national systems such as DHIS2 would be essential, supported by reliable infrastructure, including consistent power supply and internet connectivity. Well-trained personnel would oversee data management, adhering to clear guidelines on data ownership, storage, and sharing. The system would be interoperable with national identification systems and support diverse data formats, including multimedia files such as images. Regular audits and data quality assessments would uphold the integrity and reliability of the information.

3.4.2 Actual State of Health Information Management

Kampala International University Teaching Hospital (KIU-TH)

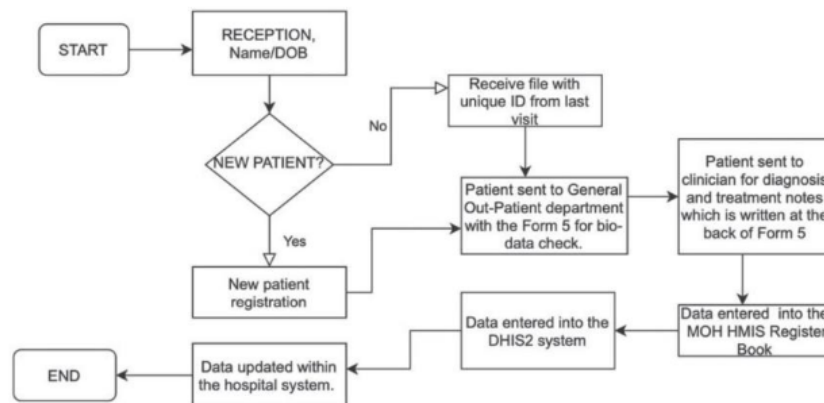


Figure 3.3: Workflow illustrating the flow of patient data through clinical, administrative and reporting processes atKampala International University Teaching Hospital.

KIU-TH operates a hybrid system that combines paper-based and electronic data management. Patient biodata is first recorded on paper forms (MF5 and HMIS Form 031) and later entered into the KIU-TH Health Information System (HIS). Each patient is assigned a unique ID for tracking purposes across departments. Departments maintain separate registers, all of which are in hard copy format. These registers are used to input data into the KIU-TH HIS and to aggregate data for the DHIS2 and RHITES systems on a weekly, monthly, and quarterly basis.

The KIU-TH HIS operates on a local area network, with data backed up on a server in the executive director’s office. While there’s no uninterruptible power supply (UPS) for computers, a generator provides backup power during outages. The system uses

Results

individual usernames and passwords for database access, enhancing security. However, the risk of significant data loss in case of a fire outbreak remains a concern.

Data reporting and integration at KIU-TH face several challenges. The KIU-TH HIS lacks interoperability with DHIS2 or the national identification system, significantly limiting its data-sharing capabilities.. Patient files contain more details than the KIU-TH HIS, including x-rays, clinical letters, and photographs, none of which are captured in the electronic system. Internet speed issues, particularly near reporting deadlines, further complicate the data management process.

Lira Regional Referral Hospital (LRRH)

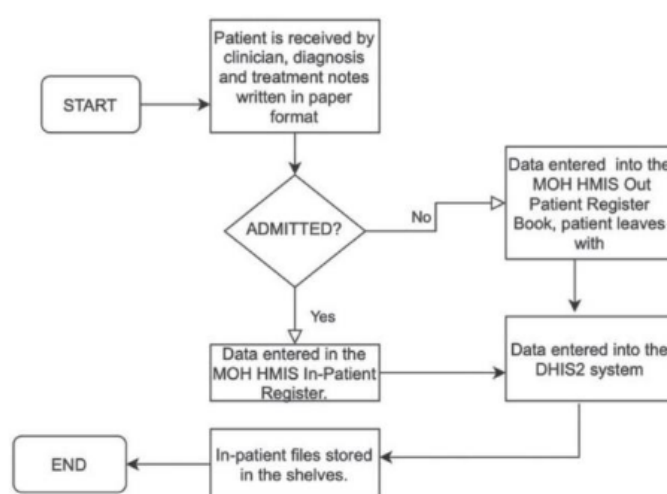


Figure 3.4: Workflow illustrating the flow of patient data through clinical, administrative and reporting processes at Lira Regional Referral Hospital

LRRH relies predominantly on a paper-based system for data collection and management. The hospital uses an outpatient registry for general patient data, with a separate Ministry of Health system to record information from the birth certificate of children. The lack of a comprehensive digital system results in limited storage space for outpatient records, with only inpatient files stored on shelves.

Data reporting at LRRH involves manual entry into DHIS2 from paper registers. Ministry of Health officials periodically verify the DHIS2 data against HMIS registers to ensure accuracy. The hospital faces significant infrastructure challenges, particularly in terms of internet connectivity. Health workers often resort to using personal funds for bundles of internet data to submit patient data to the Ministry of Health on time.

ART Clinic at Lira Regional Referral Hospital

The LRRH ART Clinic uses an electronic medical record (EMR) system run by the Ministry of Health. The system captures comprehensive patient information and retains

Results

it indefinitely, even posthumously.. However, the EMR system faces technical limitations, running on interconnected computers without a dedicated server, which slows down its functionality.. Despite the electronic system, paperwork is still involved when entering data into the EMR, adding to the workload of clinic staff.

Case Clinic Hospital, Kampala

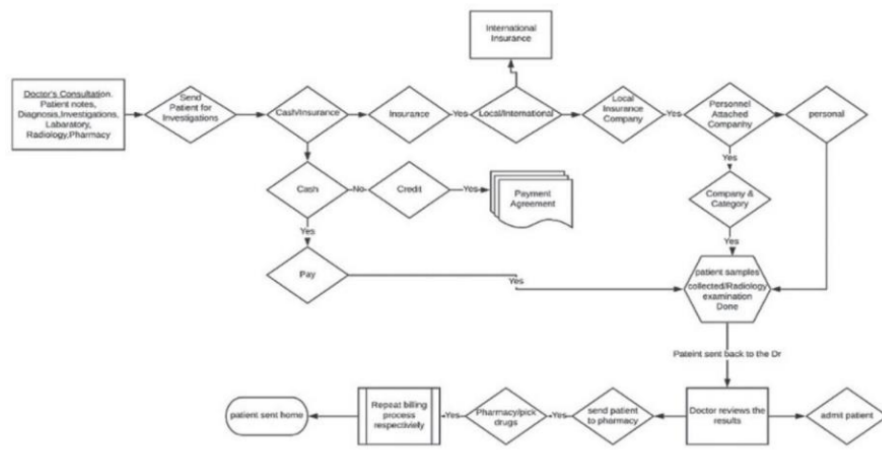


Figure 3.5: Workflow illustrating the flow of patient data through clinical, administrative and reporting processes at Case Clinic Hospital

Case Clinic Hospital in Kampala employs a more advanced electronic system for data management. The hospital uses a local system (Med360) alongside DHIS2 for direct entry of patient data. Certain types of data, including insurance, antenatal, and delivery information, are still collected on paper. The Med360 system primarily serves as a data storage mechanism without analysis capabilities.

Case Clinic boasts a more robust infrastructure and security measures compared to the other facilities included in this study. Patient data is stored off-site using remote servers, with monthly backups conducted. The hospital implements varying access rights for personnel to enhance data security. Notably, the hospital claims ownership of patient data, with patients able to access their information through reports provided by health attendants.

3.4.3 General Observations

Across all facilities, the Health Management Information System (HMIS) established by the Ministry of Health is the common thread. Most facilities use a combination of paper-based and electronic systems, with DHIS2 software mandated for use by all health facilities in Uganda. Standardized HMIS forms facilitate routine reporting, with health information assistants overseeing operations at the sub-district level and biostatisticians

Results

managing processes at the district level. Each facility maintains a box file stored in the District Health Services Directorate, serving as a physical backup of critical information.

3.4.4 Stakeholder Insights

KIU-TH Statistician

“We currently do not have written guidelines or regulations regarding patient data handling, but I am working on developing such guidelines.” (Participant 1, face-to-face interview, 20-01-2020)

KIU-TH Data Management Team Member

“We send our reports directly to the district because we have access to the system (DHIS2 or RHITES system). For health facilities that do not have access to the system, they send their reports to the health sub-district for entry into the system.” (Participant 2, face-to-face interview, 20-01-2020)

“The computers do not have UPS for power back-up, but there is a generator that comes on in case of power outage, although there may be a short period before the generator starts operation.” (Participant 1, face-to-face interview, 20-01-2020)

“The DHIS2 and RHITES databases are online, and they capture the data entered automatically. As for the KIU-TH HIS, it also saves the data entered and will take you back to where you stopped when restarted after a power blackout. However, if there were a fire outbreak, we would lose 90% of our records.” (Participant 2, face-to-face interview, 20-01-2020)

“Sending reports to the district through the DHIS2 system is slow, especially towards the deadline, because that is when most people are sending reports.” (Participant 2, face-to-face interview, 20-01-2020)

LRRH Medical Records Officer

Upon arrival at the hospital, patients are immediately directed to see the clinician. A label (‘male’, ‘female’, or ‘child’) is placed on the Outpatient Register, depending on the patient’s gender or age. After manually entering patient data into the Outpatient Register and Outpatient weekly, monthly, and quarterly reports, the data is then entered into the DHIS2 system.” (Participant 3, face-to-face interview, 24-01-2020)

“However, there is a separate system run and owned by the Ministry of Health to capture the personalized information of children so that they can obtain a birth certificate. Apart from the Ministry of Health systems, the hospital

Conclusion

has no digital system. Apart from the data entered into the Ministry of Health registers, the hospital does not maintain records of patients who are not admitted (outpatients) in the hospital. Outpatients receive a form, while inpatients have files stored in the hospital's shelves. This is due to a lack of space to store all patient files. The Ministry of Health officials then come to verify whether or not the information entered into the DHIS2 system aligns with the data from the HMIS register.” (Participant 3, face-to-face interview, 24-01-2020)

KIU-TH Data Management Team Member on Data Quality

“A team from the Ministry of Health comes twice a year to perform a data quality assessment by checking whether the data entered in the registers is the same as the one in DHIS2. There is also a team from the district that comes quarterly to do supervision and review.” (Participant 8, face-to-face interview, 04-02-2020)

These findings highlight the disparities between the ideal scenario and the current state of health information management in Ugandan healthcare facilities, revealing challenges in infrastructure, integration, and standardization at different levels of healthcare provision. The reliance on paper and electronic systems, coupled with issues of connectivity and data security, presents significant obstacles to achieving a unified and efficient health information system.

3.5 Conclusion

This study explored the dynamics of health information flow across the patient's care trajectory within the selected health facilities in Uganda. The study focused on documenting the tools used and data collected at each stage of the patient journey, including registration, history taking, physical examination, laboratory testing, diagnosis, and prescription. Insights derived from face-to-face and follow-up interviews with data management personnel underscored the dual reliance on the Health Management Information System (HMIS), predominantly the DHIS2 managed by the Ministry of Health, and diverse in-house Health Information Systems (HISs) tailored by health facilities.

These information systems are hybrid as they integrate both paper-based and electronic record systems. The DHIS2 is utilized for routine reporting under the Ministry's purview. Conversely, individual health facilities make use of their HISs to cater to specific operational needs such as patient record management, inventory oversight, and billing. With the use of the different systems, there is a significant hurdle in the lack of interoperability between these individualized systems and the Ministry of Health's DHIS2. This lack of interoperability is hindering seamless communication and data exchange.

Digitizing paper-based systems has the potential to transform the healthcare system, not only by improving resource efficiency but also by significantly enhancing patient care quality across Ugandan hospitals and clinics. Additionally, this change towards the

digitization of managing data within the healthcare system is pivotal for advancing the FAIRification process. This process fosters data that is Findable, Accessible under well-defined conditions, Interoperable, and Reusable (FAIR). Such a transformation addresses limitations on data sharing and reuse attributed to the structural non-uniformity and semantic heterogeneity prevalent in existing information systems. Consequently, this lays the groundwork for the development and establishment of a FAIR Data Point (FDP) in Uganda, enabling machine-assisted querying across diverse datasets, thereby unlocking insights from disparate sources.

3.6 Discussion

To document the data management processes in select health facilities in Uganda and assess the FAIRness and feasibility to transform the data into FAIR data, interviews and follow-up discussions with data management personnel at the selected hospitals shed light on the prevailing information systems, primarily comprising the Health Management Information System (HMIS) established by the Ministry of Health and diverse in-house Health Information Systems (HISs) employed by health facilities. These systems exhibit a dual nature, incorporating both paper-based and electronic record systems, with the Ministry-managed electronic record system predominantly relying on the DHIS2 software. Within the framework of the Ministry-managed HMIS, each health facility is mandated to undertake mandatory routine reporting utilizing standardized HMIS forms. The governance of patient data management, whether collected by public or private health facilities under Uganda's Ministry-controlled HMIS, is encapsulated in the HMIS Manual. This manual undergoes revision approximately every five years, as disclosed during the interview with participant 3. The regulations encompass critical components, including the stipulation of mandatory HMIS forms for completion by health facilities, the designated higher administrative level for form submission (e.g., health sub-district or Ministry of Health Resource Center), obligatory reporting deadlines, required reporting frequency, personnel responsibilities for each HMIS form, prescribed procedures for form completion, summarization protocols into weekly, monthly, quarterly, and annual tables, and the designated format for information entry in each column of the HMIS forms. Notably, the diagnosis must be articulated clearly in the diagnosis column, aligning with standard case definitions and Uganda Clinical Guidelines provided by the Ministry of Health. The structured nature of data management practices under the Ministry-controlled HMIS is underscored by the reliance on standardized forms, reporting protocols, and adherence to comprehensive guidelines. This approach ensures consistency and facilitates the systematic collection of data at various levels of administration. The periodic revision of the HMIS Manual shows a commitment to adaptability in order to stay aligned with healthcare needs as they evolve and the advancements in technology. Overall, the active involvement of the Ministry in overseeing the HMIS has established a cohesive and standardized approach to the management of health data in the health care systems in Uganda.

Table 3.2: HMIS Data Collection and Reporting Forms

HMIS Form	Reporting Frequency	Information on Form
Medical Form Five (MF5)	Daily	Captures patient's details and biodata including: name, outpatient number, age, sex, tribe, village, parish, sub-county, district, telephone contact, next of kin, blood pressure, weight, height, BMI, blood sugar, Z-score, MUAC, diagnosis, and treatment notes.
HMIS Form 105	Monthly (Outpatient Department)	Reports depersonalised outpatient attendance, referrals, diagnoses, maternal and child health, HIV/AIDS counselling and testing, safe male circumcision, essential medicines and health supplies, and outpatient laboratory tests.
HMIS Form 108	Monthly (Inpatient Department)	Reports depersonalised inpatient data including: census information, referrals, major and minor surgical procedures, special service utilisation, admissions and deaths by diagnosis, and maternal conditions.
HMIS Form 107	Annual	Captures depersonalised data on disease surveillance, notifiable diseases, and epidemics at the outpatient level, maternal health, and drug balances.
HMIS Form 033B	Weekly (Surveillance)	Captures depersonalised data on disease surveillance, notifiable diseases, and epidemics at the outpatient level, maternal health, and drug balances.
HMIS Form 031	Daily (Outpatient Register)	Records patient name, residence, age, sex, next of kin, MUAC, BMI, blood pressure, malaria test, tuberculosis, new diagnoses, drugs, and treatment.
HMIS Form 071	Antenatal	N/A
HMIS Form 106a	Quarterly	Captures depersonalised data on HIV/ART services, nutrition (cross-sectional), tuberculosis/leprosy services, and includes comments by the health facility in charge.
HMIS Form 009	Monthly (Nutrition Addendum)	Captures detailed nutrition-related data to supplement routine reporting.

This study was limited in scope, focusing on one regional referral hospital (LRRH), one

Discussion

district general hospital (KIU-TH), and two clinics (Case Clinic and the ART Clinic), with a modest sample size of nine participants. Despite this limitation, it is important to note that all health facilities in Uganda adhere to the same standardized guidelines for health data collection and processing, as mandated by the Ministry of Health. Consequently, the findings of this study are believed to be representative of data management practices across all health facilities in Uganda.

Chapter 4

FAIR Equivalency with Regulatory Framework for Digital Health in Uganda

This chapter is based on the following publication:

Basajja, M., Van Reisen, M., Oladipo, F.: FAIR Equivalency with regulatory frameworks for digital health in Uganda. Data Intelligence 4(4) (2022). doi: https://doi.org/10.1162/dint_a_00170

Abstract

This chapter explores the intersection between the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles and Uganda's existing regulatory framework for digital health. It highlights the critical challenges paper-based data systems pose in Ugandan health facilities, especially amid health crises such as the COVID-19 pandemic. The study examines the flow of clinical patient data and reveals a significant reliance on non-digitized reporting tools, which hinders timely data access, sharing, and reuse. As a result, national and international health stakeholders face difficulties in integrating Ugandan health data into broader global datasets, often misinterpreting the country's data landscape as sparse. The chapter argues that aligning Uganda's digital health strategy with FAIR principles could significantly improve data governance, interoperability, and reusability. By examining the current gaps and drawing parallels between FAIR standards and Uganda's health data regulations, the chapter proposes a roadmap to transition from fragmented paper records to a cohesive, machine-actionable digital health infrastructure.

Keywords: eHealth, Digital health, GAIR Guidelines, FAIR Equivalency, Findable, Accessible, Interoperable, Reusable, HMIS

4.1 Introduction

Within the introduction, we present the background, main objectives, specific research questions, and objectives.

4.1.1 Background

In the era of digitization, data serves as the cornerstone of modern healthcare systems worldwide. The exponential growth of data, coupled with advancements in technology, has revolutionized the methods used in managing, accessing, and utilizing healthcare information [70, 71]. However, amidst this data deluge, ensuring that information is Findable, Accessible, Interoperable, and Reusable (FAIR) remains a persistent challenge, particularly in the context of Uganda’s digital health landscape.

Uganda, like many nations, faces significant challenges in managing vast amounts of health-related data. Although FAIR principles have gained global recognition, their implementation within Uganda’s digital health regulatory framework presents unique challenges and considerations. It is important that the health data policies, regulations, and existing frameworks within Uganda’s digital health ecosystem are examined. Integrating FAIR principles into Uganda’s digital health governance holds potential to unlock new avenues for data-driven decision-making, research, and innovation in healthcare delivery.

As global perspectives on FAIR principles evolve, Uganda is uniquely positioned to leverage these principles to enhance the interoperability of its digital health infrastructure. Through a systematic evaluation of Uganda’s regulatory framework and the existing digital health policies, this study aims to explore the feasibility of integrating FAIR principles and the implications of this integration on the country’s healthcare ecosystem. This research aims to foster equitable digital health frameworks in Uganda by bridging the gap between theoretical concepts and practical implementation. Collaborative efforts within the Virus Outbreak Data Network (VODAN) [56] offer Uganda an opportunity to align with global best practices while tailoring FAIR principles to its socio-economic and healthcare contexts.

This chapter examines the alignment between FAIR principles and Uganda’s regulatory framework for digital health. Through analyzing Uganda’s digital health policies, research objectives, and theoretical considerations, this study aims to develop a roadmap for seamlessly integrating FAIR principles into Uganda’s digital health governance.

4.1.2 Main Objective

This study was conducted within the Virus Outbreak Data Network (VODAN) [56] as part of a project that focuses on developing a FAIR Data Point for Uganda [73]. The primary aim was to assess the conditions of Uganda’s digital health ecosystem, focusing on health data policies and documents, to evaluate the feasibility of initiating a policy window for adopting and implementing FAIR principles. The study specifically explores how these guidelines could address challenges confronted by Uganda’s e-health sector.

4.1.3 Specific Research Questions and Objectives

The primary research question aimed to evaluate the governance framework for digital health policies in Uganda and explore the feasibility of creating a policy window to adopt FAIR principles in the e-health sector.

The following are the objectives of the study: **To examine the governance framework for digital health in Uganda and the standards employed in managing digital health data:** This objective entailed analyzing Uganda's governance framework for digital health policies, focusing on health policies and managing digital health data within the context of information and communication technology (ICT).

To investigate the conditions within the digital health ecosystem in Uganda and assess the feasibility of opening a policy window for adopting and implementing FAIR principles: This required a comprehensive review of key policy documents related to digital health governance in Uganda. The review aimed to identify factors influencing the introduction of FAIR principles and to evaluate strategies for maximizing their adoption.

Research Questions

Main Research Question: To what extent does the absence of policies on data management, sharing, and analytics contribute to interoperability and information flow challenges within Uganda's digital health solutions? **Sub-Research Questions:**

Sub-Research Question 1:

To what extent are FAIR (Findable, Accessible, Interoperable, Reusable) principles referenced and implemented in Uganda's digital health policy documents?

Sub-Research Question 2:

How has the evolution of ICT and health-related policies shaped the current state of digital health solutions in Uganda??

Sub-Research Question 3:

What is the level of FAIR equivalency in Uganda's existing policy documents related to digital health?

Sub-Research Question 4:

How can FAIR principles address the challenges in Uganda's e-health sector, particularly regarding standardized communication protocols and unique identifiers?

4.2 Materials and Methods

We address the theoretical considerations, study design, and historical perspective on the Evolution of ICT and Health-Related Policies in Uganda, Identification of Relevant

Documents, Mention of the FAIR principles in Policy Document, and Level of FAIR Equivalency in Policy Documents.

4.2.1 Theoretical Considerations

Kingdon's model finds the connection between a problem, potential policy solutions, and political relevance, and creates a policy window that enables the introduction of new ideas into the policy agenda [49]. This dynamic process, involving the interaction between policy changes and the roles of policymakers and policy entrepreneurs within specific domains, was explored to understand how policy evolves [49, 74]. Van Reisen and others introduced the concept of cultural entropy to explain the reasons for the acceptance of FAIR principles potentially being hindered. If there is a significant cultural entropy — that is, the discrepancies between the guidelines in a framework and the beliefs of practitioners implementing the framework - the adoption of the FAIR principles may be impeded [30, 31].

In the context of this study, the awareness of cultural entropy is crucial as significant misalignment between the FAIR principles and the practices of stakeholders may lead to substantial obstacles to the successful adoption of FAIR principles. Therefore, exploring cultural entropy is an integral theoretical consideration that offers insights into the factors influencing policy acceptance and the implementation of FAIR principles within the digital health sector in Uganda.

4.2.2 Study Design

This cross-sectional study analyzed policy documents related to ICT and health that were published by the Government of Uganda between 2009 and 2019.

4.2.3 Historical Perspective on the Evolution of ICT and Health-Related Policies in Uganda

The historical evolution of ICT and health-related policies was explored to identify the foundational policy documents and legislation shaping digital health/e-health policy in Uganda. In this section, the study focused on understanding the steps leading to the establishment of e-health reforms in Uganda by looking into the key reforms that were initiated in 2000 in both sectors. In the early 2000s, the Ugandan Government recognized emerging issues due to the Internet and technological advancements. Responding to these trends, the National ICT Policy Framework was approved in 2003 to guide the developments experienced in the ICT sector. The creation of the Ministry of ICT in 2006 aimed to consolidate ICT responsibilities under one entity [75]. Several policies were implemented to guide the evolving ICT sector, leading to the formulation of e-health reforms in Uganda.

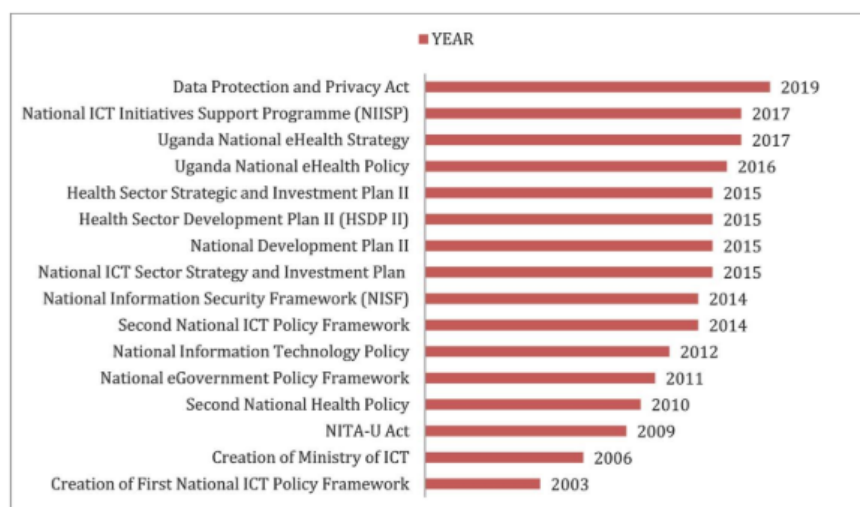


Figure 4.1: Bar chart illustrating the timeline of major ICT, eHealth, and health-sector policy reforms in Uganda between 2000 and 2019

These policies include the following:

- 1. National Information Technology Authority Act (2009):**
Establishment of the National Information Technology Authority-Uganda (NITA-U).
- 2. National e-Government Policy Framework (2011):**
Framework for advancing e-government initiatives across sectors.
- 3. National Information Technology Policy (2012):**
Policy addressing key aspects of information technology.
- 4. National ICT Policy Framework (2014):**
Policy acknowledging emerging areas, including e-health [75]..
- 5. National Information Security Framework (2014):**
Framework focusing on information security.
- 6. National ICT Sector Strategy and Investment Plan (2015):**
Strategic plan for the ICT sector's development and investments.
- 7. National ICT Initiatives Support Program (2017):**
Support program fostering ICT initiatives.
- 8. Data Protection and Privacy Act (2019):**
Legislation addressing data protection and privacy.

Materials and Methods

The National ICT Policy, launched in 2014, recognized the dynamic nature of the ICT sector and highlighted emerging areas that required policy attention, including e-health [75]. Consequently, the Ministry of Health formulated policies specific to e-health. This included the drafting of the National e-health Policy (2013) and the National e-health Strategy (2013), leading to the final versions: National e-health Policy (2016) [76] and National e-health Strategy (2017) [77]. These policies aimed to guide on leveraging ICT for healthcare transformation, emphasizing information access, operational support, and decision-making within the healthcare sector [76].

4.2.4 Uganda's National e-health Vision

An investigation was conducted to evaluate the conditions within Uganda's digital health/e-health ecosystem, aiming to determine the feasibility of initiating a policy window for the adoption and implementation of the FAIR principles. The governance framework for the digital health/e-health sector in Uganda was shaped by the publication of the National e-health Policy (2016) [76] and the subsequent National e-health Strategy (2017) [77]. This strategic initiative was spearheaded by the Ministry of Health, laying the foundation for the National e-health Vision.

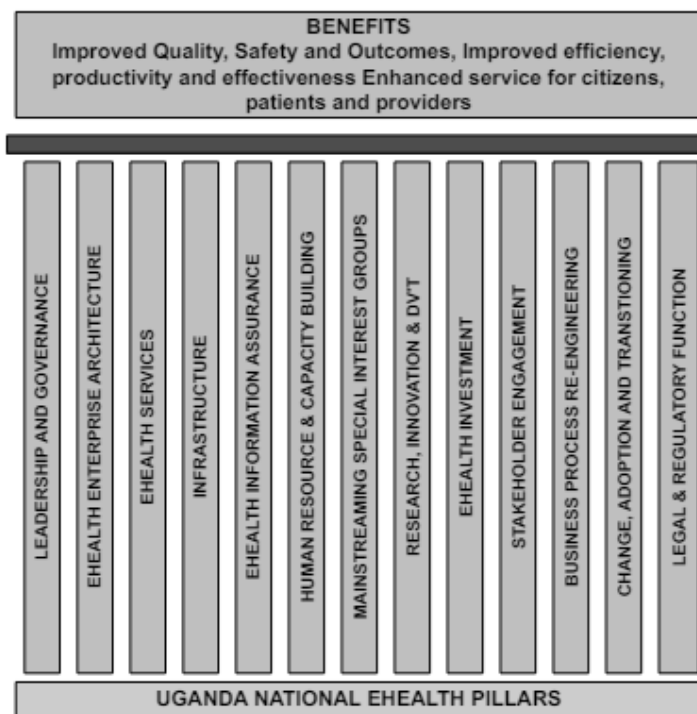


Figure 4.2: Conceptual framework illustrating Uganda's National e-health Vision that constitutes the thirteen strategic e-health Pillars that support digital health governance, implementation and service delivery

Materials and Methods

The National e-health Vision identified 13 priority areas, referred to as pillars, that are crucial for realizing advancements in digital health. Each of these pillars is supported by a specific leadership and governance structure. They are also associated with an action plan of five years that is stated in the National e-health Policy and the National e-health Strategy [77].

Implementation Phases (2016/17-2020/21)

For effective execution, each pillar underwent detailed planning, including budget allocation and action plans spanning three phases: phase I (2016/17-2017/18), phase II (2018/19-2019/20), and phase III (2020/21). This phased approach aimed to systematically implement the strategic objectives and associated initiatives, ensuring the gradual realization of the National e-health Vision over the stipulated five-year period.

4.2.5 Identification of Relevant Documents

A set of 14 policy documents, encompassing plans, policy frameworks, acts, strategies, and programs, was identified for analysis in alignment with the study's research objectives (refer to Table 4.1). These documents, published by the Government of Uganda and pertinent to both ICT and health, constitute the focal point of the investigation. The selection process involved obtaining documents from ICT personnel in the Ministry of Health and extracting others from the websites of relevant government institutions.

SN	Title	Type of Document	Year	Publisher	Source of Document
1	National Information Technology Authority-Uganda (NITA-U) Act	Act	2009	National Information Technology Authority	National Information Technology Authority-Uganda (NITA-U) Act NITA-U Act (Act No. 4 of 2009)
2	Second National Health Policy	Policy	2010	Ministry of Health	Second National Health Policy 2010
3	National eGovernment Policy Framework	Framework	2011	Ministry of ICT	National e-Government Policy Framework 2011
4	National Information Technology Policy	Policy	2012	Ministry of ICT	National IT Policy 2012
5	National ICT Policy Framework	Framework	2014	Ministry of ICT	ICT Policy 2014

Materials and Methods

6	National Information Security Framework	Framework	2014	National Information Technology Authority	National Information Security Framework
7	Second National Development Plan	Plan	2015	National Planning Authority	NDPII Final
8	Second Health Sector Development Plan	Plan	2015	Ministry of Health	Health Sector Development Plan 2015-16_2019-20
9	Uganda National e-health Policy	Policy	2016	Ministry of Health	Uganda National e-health Policy
10	Uganda National e-health Strategy	Strategy	2017	Ministry of Health	National e.Health Strategy
11	National ICT Initiatives Support Program	Program	2017	Ministry of ICT	NIISP Program
12	Data Protection and Privacy Act	Act	2019	Ministry of ICT	Data Protection and Privacy Act 2019

Table 4.1: Policy Documents Analyzed

Source: Compiled from various sources by M. Basajja, 2021.

4.2.6 Mention of the FAIR principles in Policy Documents

An analysis of the policy documents (Table 4.1) related to the e-health sector in Uganda was conducted to determine their explicit or implicit references to the FAIR principles—Findable, Accessible, Interoperable, and Reusable. None of the examined documents explicitly mentioned the FAIR principles [29]. However, semantic equivalents, denoted as 'FAIR Equivalent,' were identified within the documents. The determination of whether a statement found in the policy documents was equivalent to FAIR principles hinged on assessing its intent, purpose, or meaning concerning the original FAIR article [29]. For instance, if a document advocated for the use of standards to facilitate data exchange, it was labelled as equivalent to 'Reusability', a FAIR Guideline emphasizing the necessity for data and metadata to conform to domain-relevant community standards. Keywords and phrases such as common language/vocabulary, harmonise, incorporate, common systems, alignment, collaborate, common infrastructure, coordinate, integrate, open-source, information exchange, one-stop centre, and shared services were associated with the concept of 'Interoperability.' To signify this similarity in policy direction, the term 'FAIR Equivalent' was coined.

A binary code system was employed for tabulating annotations, with '1' signifying the presence of either a 'FAIR Mention' or 'FAIR Equivalent,' and '0' indicating the

Materials and Methods

absence of both (refer to Table 4.2 in Results). This structured approach parallels the framework method, widely employed in health research for qualitative data management and analysis.

Policy Document	FAIR Mention	FAIR- Equivalent
NITA-U Act (2009)	0	0
Second National Health Policy (2010)	0	1
National eGovernment Policy Framework (2011)	0	1
National Information Technology Policy (2012)	0	0
National ICT Policy Framework (2014)	0	1
National Information Security Framework (2014)	0	1
Second National Development Plan (2015)	0	1
Second Health Sector Development Plan (2015)	0	1
Uganda National e-health Policy (2016)	0	1
Uganda National e-health Strategy (2017)	0	1
National ICT Initiatives Support Program (2017)	0	1
Data Protection and Privacy Act (2019)	0	1
Average	0	0.8333333333
Percent	0%	83%
Sum of FAIR-Equivalent (n)	0	10

Table 4.2: FAIR Mentions and FAIR-Equivalents in Policy Documents

4.2.7 Level of FAIR Equivalency in Policy Documents

To gauge the level of FAIR Equivalency in the examined policy documents, an analysis was extended to identify the presence of the equivalent of the 15 sub-criteria, termed ‘FAIR facets,’ that constitute the FAIR principles. These facets encompass four main categories: ‘Findability’ (F1, F2, F3, F4); ‘Accessibility’ (A1, A1.1, A1.2, A2); ‘Interoperability’ (I1, I2, I3), and ‘Reusability’ (R1, R1.1, R1.2, R1.3) [29].

Materials and Methods

Policy Document	F1	F2	F3	F4	A1	A1.1	A1.2	A2	I1	I2	I3	R1	FE-Score
NITA-U Act (2009)	0	0	0	0	0	0	0	0	0	0	0	0	0
Second National Health Policy (2010)	0	0	0	0	0	0	0	0	1	0	0	0	1
National eGovernment Policy Framework (2011)	1	0	0	1	1	1	1	0	1	0	0	0	7
National Information Technology Policy (2012)	0	0	0	0	0	0	0	0	0	0	0	0	0
National ICT Policy Framework (2014)	1	1	0	0	1	1	1	0	1	1	0	1	11
National Information Security Framework (2014)	0	0	0	0	1	0	1	0	0	0	0	0	2
Second National Development Plan (2015)	1	0	0	0	1	1	1	0	1	0	1	0	6
Second Health Sector Development Plan (2015)	0	0	0	0	1	0	0	0	1	0	1	1	5
Uganda National e-health Policy (2016)	1	1	0	1	1	1	1	0	1	0	1	0	11
Uganda National e-health Strategy (2017)	1	1	0	1	1	1	1	0	1	0	1	1	12

Materials and Methods

National ICT Initiatives Support Program (2017)	0	0	0	0	1	1	1	0	1	0	0	0	6
Data Protection and Privacy Act (2019)	0	0	0	1	1	0	0	0	0	0	0	1	4
Average	0.41	0.25	0	0.33	0.75	0.5	0.58	0	0.66	0.08	0.33	0.33	4.25
Percentage	41.7	25	0	33.3	75	50	58.3	0	66.7	8	33.3	33.3	4.25
Sum of FAIR-Equivalent (n)	5	3	0	4	9	6	7	0	8	1	4	4	51

Table 4.3: Summary of FAIR Equivalency assessment results for the analysed Policy Documents, indicating the presence and frequency of FAIR-related facets

A systematic review was conducted on the policy documents with the mention of the equivalent of each FAIR facet labeled by appending the respective FAIR facet code (e.g., F1, A1.2) to the corresponding statements. Using a binary code system ('1' for presence and '0' for absence), a Microsoft Excel spreadsheet (refer to Table 4.3 in Results) was utilized to compute the FAIR Equivalency Score (FE-Score) for each policy document. The FE-Score, representing the sum across all 15 facets, provides insight into the alignment of each document with the FAIR principles. The highest attainable FE-Score per document is 15 (refer to Supplementary Material 3, Basajja, 2021).

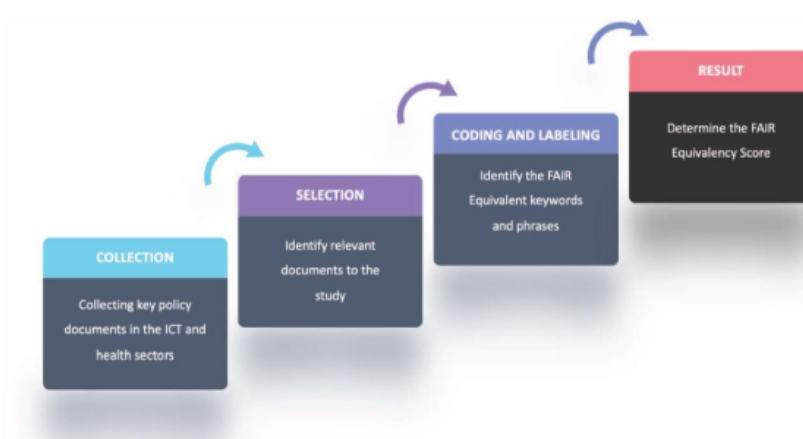


Figure 4.3: Process flow diagram showing the steps undertaken to assess FAIR equivalency within Ugandan policy documents.

4.3 Results

The results of this study are addressed in this section.

4.3.1 Mention of FAIR principles in Policy Documents

The investigation into whether the policy documents explicitly mention the FAIR principles revealed that the principles are not specifically mentioned in any of the documents. However, 10 of the 12 documents mention principles that are equivalent to the FAIR principles, which is a significant majority, see Table 4.2. This indicates that 80% of the documents are connected to the FAIR principles to some extent. Notably, the other two documents, which are relatively early policy documents from 2009 and 2012, do not contain any elements related to the FAIR principles.

4.3.2 FAIR Equivalency in Policy Documents

The FAIR principles consist of 15 FAIR facets. An examination of the presence of each facet in the documents is presented in descending order based on frequency:

1. Most Mentioned:

A1 (75%, n=9):

Accessibility facets are frequently mentioned, indicating the importance placed on data retrievability using a standardized protocol.

I1 (66.7%, n=8):

Interoperability shows emphasis on using a formal, accessible, and commonly shared language to provide knowledge representation is well-represented.

A1.2 (58.3%, n=7):

The protocol allowing for authentication and authorization procedures is also present.

2. Moderately Mentioned:

A1.1, R1.3 (50%, n=6):

Both the Authentication Protocol openness and Reusability aspects are moderately present.

F1 (41.7%, n=5):

Findability, including assigning a globally unique identifier, is relatively less mentioned but still present.

3. Less Frequently Mentioned:

F4, I3, R1, R1.1, R1.2 (33.3%, n=4):

These facets, including metadata registration and using qualified references, are less commonly found.

Results

4. Rarely Mentioned:

F2 (25%, n=3):

Rich metadata description is less emphasized in the documents.

I2 (8%, n=1):

The use of vocabulary is the least mentioned facet.

5. Not Mentioned:

A2, F3 (0%, n=0):

Accessibility of metadata when data is unavailable, and metadata explicitly including data identifiers are not mentioned in any document.

These results are visually represented in a heatmap (Figure 4.5). To assess the potential improvement in FAIR Equivalent representation over time, the FAIR Equivalent Score (FE-Score) was calculated, providing an aggregate score across all documents and facets (Table 4.3 and Figure 4.4).

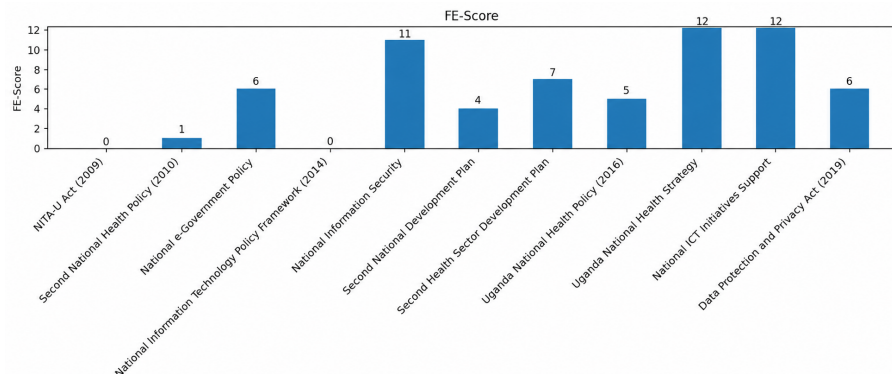


Figure 4.4: Bar chart illustrating the FAIR Equivalency (FE) scores across the analysed Ugandan Digital Health policy documents published between 2009 and 2019.

Results

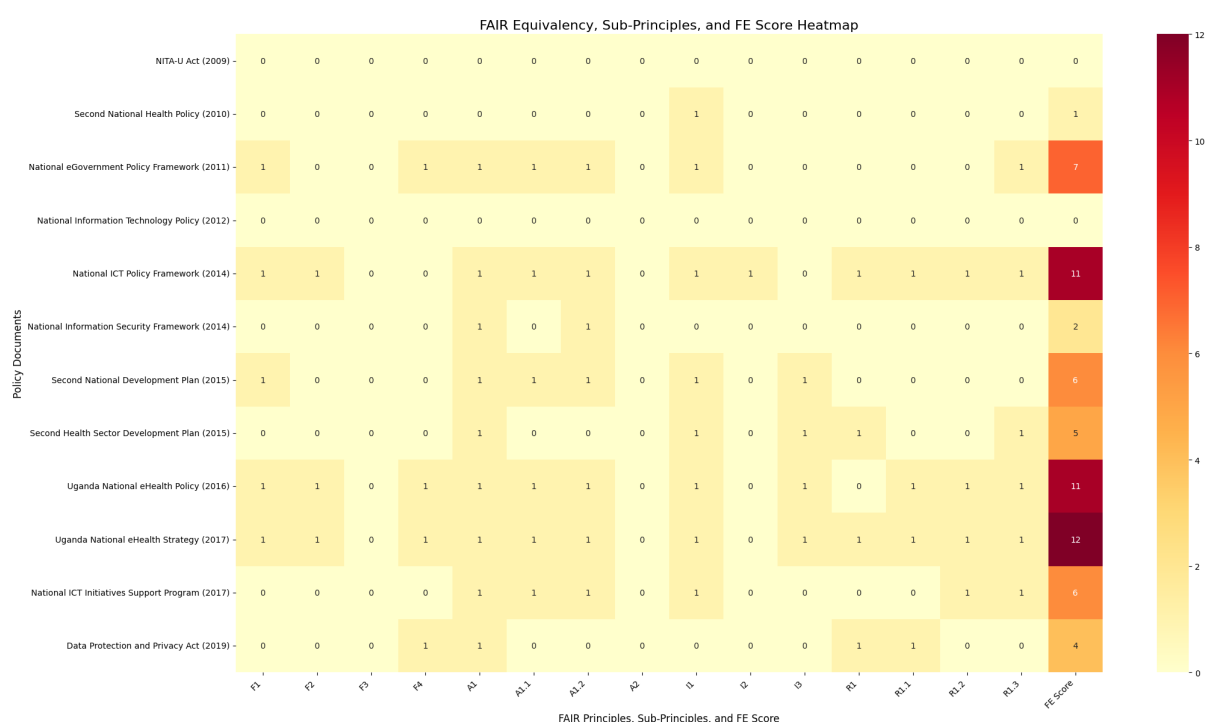


Figure 4.5: Heatmap illustrating the frequency of the FAIR equivalent concepts identified across Ugandan digital health policy documents mapped against the FAIR principles and their corresponding sub-principles ‘Findability’ (F1, F2, F3, F4), ‘Accessibility’ (A1, A1.1, A1.2, A2), ‘Interoperability’ (I1, I2, I3) and ‘Reusability’ (R1, R1.1, R1.2, R1.3) per policy document

Conclusion

Figure 4.4 shows the FE-Score for each document by the year of its publication. The figure shows that the National e-health Policy (2016) and the National e-health Strategy (2017) have the highest FE-Scores, followed by the National ICT Policy Framework (2014). The charts show that FAIR Equivalency does not consistently increase over time. Instead, the highest FAIR Equivalency is observed for the National e-health Policy (2016), and the National e-health Strategy (2017), both have an FE-Score of 12, while the National ICT Policy Framework (2014) has an FE-Score of 11. The three documents that exhibit the highest FAIR Equivalency form the core of the framework of governance for Uganda's e-health sector and the foundation of Uganda's digital health policies.

4.4 Conclusion

Applying the FAIR principles to Africa's digital health services presents a transformational opportunity to address critical issues within the health sector on the continent. Such issues include: sustainability challenges and fragmentation of health service solutions. This study aimed to evaluate Uganda's framework of governance for digital health policies. It also investigated the feasibility of opening a policy window for adopting and implementing the FAIR principles in the country's e-health sector.

Key policy documents that are central to the emerging e-health sector, encompassing information, communication, and technology (ICT) and health, were identified and analyzed. A binary coding system using keywords and phrases was employed to assess whether these documents explicitly mention the FAIR principles or their equivalent, spanning 15 FAIR sub-criteria or facets. The resulting FAIR Equivalency Score (FE-Score), representing a total of 15 facets, was calculated for each document. Despite none of the documents directly referencing the FAIR principles, 83% of them acknowledged efforts equivalent to FAIR principles.

The highest FE scores were associated with Uganda's National e-health Policy, National e-health Strategy, and National ICT Policy Framework, indicating a higher alignment with FAIR principles. Notably, these three documents were among the most recently drafted of the selected policy documents. While the explicit mention of the FAIR principles was absent, the prevalent mention of FAIR Equivalency efforts, such as the adoption of the National Identification Number and planned integration of information systems, suggests a favourable environment for FAIR adoption in Uganda's Health sector within e-health.

In conclusion, the study indicates that the necessary conditions are favourable for advocating the adoption and implementation of the FAIR principles in Uganda's e-health sector. The FAIR Equivalency efforts reflected in policy documents demonstrate a foundational alignment with FAIR principles, paving the way for the potential acceptance and integration of these guidelines into Uganda's digital health policies.

4.5 Discussion

Although the analyzed documents do not explicitly reference the FAIR principles, the use of FAIR Equivalent terminology indicates a relatively small cultural entropy. This suggests that stakeholders in the e-health sector may not strongly resist the acceptance of FAIR principles, as the foundational principles align closely with existing policy language and objectives.

4.5.1 FAIR Equivalency

The FAIR Equivalency analysis was conducted using a systematic approach to identify terms and concepts in policy documents that align with the FAIR principles. The analysis revealed that in all examined policy documents, the facets of 'Accessibility' and 'Reusability' were most frequently referenced, with frequencies of 22 and 18, respectively. In contrast, the facets of 'Findability' and 'Interoperability' appeared less frequently, with frequencies of 12 and 13, respectively. The most frequently referenced FAIR Equivalent facets include:

A1: "(meta)data are retrievable by their identifier using a standardised communications protocol."

I1: "(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation."

Conversely, the least-referenced FAIR Equivalent facets are F3, A2, and I2. These frequencies may be attributed to several factors:

Policy focus: The higher frequency of Accessibility and Reusability facets likely reflects a policy emphasis on making data easily accessible and reusable for multiple purposes.

Technical complexity: The lower frequency of Findability and Interoperability facets might reflect the technical challenges associated with implementing these principles, or a lack of awareness about their importance.

Historical context: The emphasis on certain facets may be influenced by historical data management practices and an evolving understanding of data-sharing needs.

Stakeholder influence: The frequencies could reflect the priorities of key stakeholders involved in policy development.

Regarding the least-mentioned FAIR Equivalent facets, Wilkinson and colleagues [29] described these facets as follows: F3—"metadata clearly and explicitly include the identifier of the data it describes"; A2—"metadata are accessible, even when the data are no longer available; I2—"meta(data) use vocabularies that follow FAIR principles". The reason why these FAIR Equivalent facets (F3, A2, and I2) are the least referenced in the policy documents is that they are data-centric, meaning that they specifically deal with details about data and metadata, which is beyond the scope of the documents analyzed. As they are policy-level documents, they do not explicitly mention such things. Therefore, it is only after the policies or guidelines in the documents have been implemented in a particular context that one can ascertain whether or not the particular implementation conforms to these facets. serves

4.5.2 Facet A1: Standardized Communications Protocol for Accessibility

Facet A1 of the FAIR principles emphasizes the need for a standardized communications protocol, ensuring both human and automated users can easily retrieve data and metadata through identifiers. While acknowledging exceptions for highly sensitive data, maintaining accessibility through contact details is considered FAIR.

Examples of Mention in Policy Documents:

1.Uganda National e-health Policy (2016):

Highlights the Ministry of Health's Knowledge Management Portal serving as a single access point for health-related information, supporting evidence-based decision-making.

2.National e-Government Policy Framework and National ICT Policy Framework:

Emphasizes implementing an open data system for the government to streamline how citizens access information on the government, aligning with the priority of e-government applications.

3.National Development Plan (NDP) II Public Investment Plan:

Outlines the development of a Monitoring and Evaluation Web Portal with a database, acting as a centralized information hub for government program monitoring.

4.Health Sector Development Plan (HSDP) II:

Focuses on establishing a functional community information system, emphasizing the development of a National Health Information System Framework (NHISF) for reliable and accessible health data management.

These examples demonstrate a commitment to standardized communication protocols, reflecting a broader strategy to enhance accessibility and data retrieval. The incorporation of technology, portals, and frameworks aligns with the FAIR principle of A1, ensuring data and metadata accessibility through standardized protocols.

4.5.3 Facet I1: Formal, Accessible, Standardized Controlled Vocabularies

In terms of the FAIR principles, this is reflected in the first aspect I1: the use of formal, understandable, and broadly accepted language for knowledge representation; this is achieved by adopting standardised vocabularies, ontologies, and clear data models. This requirement can be met by taking up the Enterprise Architecture and Interoperability Framework for e-government and e-health implementations, which has been mentioned in policy documents such as the National e-health Policy, National e-health Strategy, and the National e-Government Policy Framework.

The National e-health Policy (2016) endorses the use of the Enterprise Architecture and Interoperability Framework for the development of e-health systems. They highlight the need to integrate systems for interoperability and avoid the lack of standards in managing health information.

In the National e-health Strategy published in 2017, it was also mentioned that an e-health Enterprise Architecture(EA) and Interoperability Framework should be created that would govern all investments in e-health. This emphasizes the role of EA in avoiding fragmentation and duplication and promoting integration in digital health solution architectures. The National e-Government Policy Framework also discusses the adoption of the Enterprise Architecture and Interoperability Framework for e-government and e-health, underlining the ability to use the simplified structures to enhance interoperability.

The importance of integration work in this context depicts the value of integrating information systems with various health information systems, such as HRHIS, CIS, and LSMIS. The intended integration of such systems complies with the 'Interoperability' principle of FAIR.

These examples show an adherence to the outlined frameworks and languages for knowledge representation to structure the FAIR interoperability component of I1 with reference to a common language. That having been said, the emphasis on integrated systems can be seen as corresponding with efforts that have been made in the development of the e-health system toward improving the availability and usefulness of health information.

4.5.4 Facet F1: Globally Unique and Persistent Identifiers for Findability

Facet F1 of the FAIR principles emphasizes the importance of assigning globally unique and persistent identifiers to data and metadata, ensuring their findability. This facet is foundational for the subsequent principles of Accessibility, Interoperability, and Reusability within the FAIR framework. Here are examples from the policy documents highlighting initiatives aligned with FAIR's 'Findability'. Both the National e-health Policy and National e-health Strategy recognize the crucial role of a unique health identifier in e-health systems. Acknowledging the absence of a unique identifier for health records and the issue of record duplication, the policies propose the use of the National Identification Number (NIN) as a unique identifier [76, 77].

The National Identification and Population Databank Project, represented by the NIN, provides a unique identification number for every individual in Uganda. The policies advocate for reusing the NIN as the unique health identifier for electronic medical records, electronic health records, and personal health records [76, 77]. The National e-health Strategy outlines the strategic objective of adopting the NIN as the unique identifier for patients/clients. It also emphasizes the enforcement of its use in electronic medical records systems. Additionally, there are existing plans that include the adoption and enforcement of unique identifiers for health facilities and providers.

To establish an electronic medical records system, the Ministry of Health has adopted the Open Medical Records Systems (OpenMRS), an open-source electronic health record system. OpenMRS facilitates the sharing of an individual's medical record electronically across different medical departments, aiding in patient tracking and addressing specific health concerns [81, 82]. The Ministry of Health's plan includes rolling out OpenMRS in a significant number of health facilities to contribute to the findability and accessibility of medical records. This initiative aligns with FAIR's 'Findability' principle by establishing a system with globally unique and persistent identifiers [83]. These examples demonstrate the commitment to creating globally unique and persistent identifiers for health records. This improves their findability and lays the foundation for achieving the FAIR principles comprehensively. Additionally, the emphasis on using the already established identification systems, such as the NIN, reflects a strategic approach to aligning e-health initiatives with FAIR principles.

4.5.5 FAIR as a Solution to Challenges in Uganda's e-health Sector

The FAIR principles are rooted in the belief that publicly funded data is for the public good. Therefore, the principles emerge as a solution to the challenges facing Uganda's e-health sector. An essential step towards adopting FAIR principles involves aligning them with Uganda's National e-Health Vision. This alignment aims to address existing barriers while strategically leveraging opportunities within key e-health pillars.

In Leadership and Governance, the FAIR principles must be represented as a solution to challenges in e-health implementation. The GO CHANGE Pillar plays an essential role in this by engaging stakeholders to align the 'eHealth Information Assurance' pillar with FAIR principles. A key initiative involves the formulation and enforcement of the information security guidelines for e-health and the privacy guidelines [77].

The e-health Enterprise Architecture pillar provides an avenue for adopting FAIR by applying its principles in areas such as data, information, knowledge sharing agreements, as well as data standards. Collaboration with digital health providers through the GO BUILD Pillar is necessary to develop mobile software that adheres to FAIR principles when establishing an e-health data, information, and knowledge management system [77].

FAIR integration in developing mHealth infrastructure must be explored so that the services can be aligned with the emerging technologies within e-health Services. Collaborative efforts by the GO BUILD and GO TRAIN Pillars are needed to ensure the development of mobile software that complies with FAIR principles. A strategic initiative involves developing guidelines for data ownership, data quality, and data analysis [77].

The infrastructure pillar emphasizes the continuous assessment of mature and emerging technologies for supporting e-health. The GO CHANGE Pillar should advocate for the adoption of FAIR principles by aligning them with both mature and emerging technologies, considering those that alter the business and social environment substantially [77].

Addressing e-health Information Assurance involves including FAIR principles in the review of information security frameworks and data protection laws. The GO CHANGE and GO TRAIN Pillars should work collectively with stakeholders to embed FAIR within data handling procedures and protocols, especially in scenarios where data accessibility is restricted due to privacy concerns [77].

The Human Resources and Capacity Building pillar presents a unique opportunity for FAIR adoption by proposing its inclusion in the e-health curriculum for workforce development. The GO TRAIN Pillar should engage stakeholders to advocate for the inclusion of FAIR principles in training programs. Strategic goals include developing accredited courses, providing ICT training, and establishing e-health centres of excellence [77].

In the context of change, adoption, business process re-engineering, and transitioning, aligning the digital health transitioning strategy with FAIR principles for standardization and interoperability is essential. Collaboration with the Ministry of Health Resource Centre is necessary for motivating FAIR adoption, especially considering the objective of managing rapid technological changes through adherence to e-health 'Enterprise Architecture' [77].

In order to effectively engage stakeholders, it is required that FAIR principles are incorporated into mass campaigns alongside e-health services. Additionally, it is critical to establish a framework for measuring the level of effectiveness in engagement and awareness of FAIR to ensure the adoption of the FAIR principles [77]. Positioning FAIR as an international best

practice is a vital step to further open the policy window. With entities like the World Health Organization backing it, FAIR can become an integral part of e-health governance, aligning with global standards.

Health authorities globally benchmarking their e-health systems against international best practices will lead to the rapid adoption of FAIR principles in ways that reflect a commitment to data management excellence. This strategy leverages existing benchmarks used in adopting e-health Enterprise Architecture to promote harmonization with international standards and practices. Establishing a regional FAIR Center of Excellence that could reinforce FAIR principles' visibility and offer a collaborative space for stakeholders in East Africa will lead to enhanced understanding and application of FAIR principles.

4.5.6 Challenges of the Study

The study faced challenges in obtaining two potentially relevant documents. Despite extensive searches on relevant NGO and government institution websites and formal requests submitted to key personnel, these documents remained inaccessible, as they were neither available online nor provided upon request. This limitation may have affected the comprehensiveness of the governance overview. The research focused exclusively on national-level documents published by Uganda's Government. This selection criterion comprised policies, frameworks, strategies, plans, legislation, and programs, and was applied to obtain a governance overview that aligns with the research objectives. However, the exclusion of annual health sector performance reports, evaluation reports, survey reports, and program reports may limit the depth of the analysis. These limitations should be considered when interpreting the study's findings, as they may affect the generalisability and depth of the governance framework overview in Uganda's digital health sector.

Discussion

Chapter 5

Possibility of Enhancing Digital Health Interoperability in Uganda through FAIR Data

This chapter is based on the following publications:

Basajja, M., Nambobi, M., Wolstencroft, K.: Possibility of enhancing digital health interoperability in Uganda through FAIR data. *Data Intelligence* 4(4) (2022). doi: https://doi.org/10.1162/dint_a_00178

Suchánek, M., Basajja, M.: VODAN in a Box: Proof of concept [Online]. Zenodo (27 November 2020). doi: <https://doi.org/http://doi.org/10.5281/zenodo.4321626>

Abstract

This chapter examines the potential for improving digital health interoperability in Uganda through the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. Despite the increasing need for efficient data exchange in healthcare—especially highlighted during the COVID-19 pandemic—Uganda’s health information systems remain largely paper-based and fragmented. These limitations hinder the ability of health facilities, researchers, and policy makers to access, share, and reuse clinical data effectively. The chapter explores existing data collection and reporting practices in Ugandan health facilities, identifies key barriers to digital interoperability, and evaluates how FAIR-aligned strategies could overcome these challenges. By mapping FAIR principles onto the current health data ecosystem, the chapter proposes practical pathways for achieving improved data interoperability, ultimately enabling more coordinated, data-driven decision-making across the country’s healthcare system. This chapter also explores the interoperability assessment of two crucial digital health dashboards: the Digital Health Dashboard. Atlas Uganda (DHA-U) and the Uganda Digital Health Dashboard (UDHD), along with their associated vertical digital health tools within the Ugandan context. Efforts were made to enhance the FAIRness of the digital health hubs. To achieve this, a FAIR Data Point was used to facilitate integration and alignment with FAIR principles. The level of FAIRness was reassessed after the process of FAIRification, which was done using the FAIR Evaluation Services.

Keywords: FAIR, FAIRness level, Interoperability, Integration, Common standards, Health information exchange

5.1 Introduction

Interoperability has emerged as an indispensable element in the implementation of large-scale data management systems, particularly those intended for adoption by national health systems. This chapter analyses the potential augmentation of digital health interoperability through the FAIR (Findable, Accessible, Interoperable, Reusable) principles. We conduct an analysis that is centred on two digital health dashboards that were piloted in Uganda to scrutinize how sustainable they are. The value and potential scalability of an integrated system are significantly increased by the capability of a digital health product to easily and effectively exchange information with external systems.

Establishing interoperability with the National Health Management Information System (HMIS) enables seamless access to and utilization of the data that is collected by a digital health product by the country's Ministry of Health. This integration enhances the value of the health product as a tool that is being used to improve the administration of healthcare services. It also proves how critical it is for these health products that are employed to align with health system structures [84].

There has been significant progress in the evolution of digital health systems that were primarily constructed from Electronic Health Records (EHRs). In the early 2000s, medical records were confined to Microsoft (MS) Access spreadsheets within individual hospitals. This became insufficient shortly afterwards as the volume of data being managed continued to grow. Subsequently, a more efficient and robust system based on Java web applications and MySQL databases was created [85]. The Open Medical Record System (OpenMRS), which was introduced around 2005, was designed as a free and open-source system that could support health-care delivery systems through a generalizable EHR, particularly in developing countries. Having been deployed in over 25 predominantly low-income countries, including Uganda, OpenMRS has proven to be effective, especially in recording the treatment of millions afflicted by HIV/AIDS and tuberculosis.

There has been an increase in initiatives that explore how digital health applications can offer solutions, which has been influenced by the potential of digital health to address structural challenges in African healthcare. This surge is evident in initiatives such as the Ghana Health Service telemedicine program, which is backed by the Novartis Foundation. This program leverages mobile phones to connect remote primary healthcare facilities and community health workers to a teleconsultation centre [86]. Despite these advancements, there are persistent concerns over the sustainability of digital health solutions, as discussed by the Broadband Commission Working Group [87]. The Commission emphasizes that there is a scarcity of policies and infrastructure that are necessary to embed these pilot solutions within a sustainable environment.

This apprehension aligns with findings from a systematic literature review [88], which covers mHealth for community health in Africa from 2005 to 2015. The review shows weak evidence regarding the sustainability of mHealth interventions, with technology-related issues and circumstantial complications being significant barriers to the expected impact. This classification of the challenges being faced explains the complexities that are associated with mHealth. Consequently, it validates the claims about its weak evidence base.

Recognizing the digital health agenda in the East African Community, particularly emphasized in the 2017 adoption of an action plan [290], some countries like Rwanda and Kenya have made substantial progress with active government support for ICT-led development. However,

Research Questions and Objectives

Uganda, lacking a comparable focused ICT-policy orientation, provides an intriguing case study. The absence of central government focus often correlates with challenges to sustainability resulting from ICT-based infrastructure, making Uganda an apt subject for investigating the reasons behind this phenomenon and exploring potential remedies [89].

In this chapter, we scrutinize two digital health dashboards piloted in Uganda, evaluating their sustainability. Furthermore, we assess the extent to which the incorporation of FAIR data protocols could serve as a solution to structural challenges impeding the development and maintenance of sustainable digital health solutions in Uganda.

5.2 Research Questions and Objectives

To assess the level of interoperability and potential of enhancing digital health interoperability through FAIR in Uganda's eHealth sector, the following research questions were posed:

Research Questions

Main Research Questions:

Q1. What factors are currently hindering the interoperability of digital health solutions in Uganda?

Q2. Can the introduction of the systematic use of a FAIR data protocol through a FAIR Data Point help overcome structural challenges to the development and maintenance of sustainable digital health solutions in Uganda?

Sub-Research Questions:

Sub-Research Question 1:

What is the current level of FAIRness in Uganda's major digital health hubs, specifically the Digital Health Atlas Uganda and Uganda Digital Health Dashboard?

Sub-Research Question 2:

How do the FAIRness levels differ among various digital health interventions (DHIs) in Uganda?

Sub-Research Question 3:

What improvements in FAIRness can be achieved through the process of FAIRification for Uganda's digital health hubs?

Sub-Research Question 4:

How can enhancing the FAIRness of digital health hubs contribute to improved interoperability of digital health solutions in Uganda?

The main objectives of the study were:

To analyse the level of FAIRness of the two digital health dashboards piloted in Uganda before and after FAIRification.

To determine the level of the digital interventions in the two digital health dashboards

To examine the extent to which the introduction of FAIR data protocols would help overcome structural challenges to the development and maintenance of sustainable digital health solutions in Uganda

5.3 Two Digital Health Hubs

This section explores the interoperability assessment of two crucial digital health dashboards: the Digital Health Atlas Uganda (DHA-U) and the Uganda Digital Health Dashboard (UDHD), along with their associated vertical digital health tools within the Ugandan context. The primary objective of this assessment is to examine methods that can be used to enhance the sustainability of digital health interventions by improving the coordination between these digital hubs. In order to achieve this, this study investigated the existing situation in Uganda's healthcare system and suggested and then tested solutions that focus on improving how data is used and managed in the country's digital health centres. The step that followed was checking the effectiveness of the methods that were used to manage data within the specified digital health systems. This was done to ensure that the FAIR principles were implemented in those healthcare systems. Before discussing the results that were obtained from analyzing the DHA-U and the UDHD, this study will present an overview of these two digital health hubs.

5.3.1 Digital Health Atlas Uganda

The Digital Health Atlas (DHA), as mentioned earlier, is described as an open source that can be accessed from the web. It has a repository of digital health interventions (DHIs) and has records of similar interventions implemented globally. The primary purpose of the DHA is to support an efficient organization of the interventions that have been made within the fields associated with digital health systems, which are in line with the FAIR principles for data. It is worth noting the other supportive role of the DHA, which is to support the activities of advanced planning, timing, monitoring, and evaluation of digital health services. Some of the tools include mHealth Assessment and Planning for Scale (MAPS) that the DHA makes available to governments, service providers, and donors to determine the track of advancement of the digital health systems development [91]. DHA is considered a global public good as it has support from influential organizations, including the WHO's Division of Reproductive Health and Research, USAID, United Nations Foundation, UNFPA, Johns Hopkins University Global mHealth Initiative, and Digital Health and Interoperability Working Group.

Its mission includes enabling service providers to list their digital healthcare business globally. In this way, DHA acts as an inventory of digital health products. It helps to create a comprehensive understanding of the global digital health landscape, working towards minimizing duplication, enhancing cooperation, and promoting integration among software projects. This aligns with the FAIR Guidelines' emphasis on interoperability. The DHA facilitates strategic alignment and reduces redundancy in the efforts being made to improve healthcare systems by allowing investors access to crucial information on DHIs, such as geographic scope, health focus areas, project duration, implementation phase, readiness for scale, and financial support details [91].

In Uganda, the DHA-U consists of 38 DHIs, 25 of which receive government investment, while 13 of them are funded by other stakeholders. These digital health solutions address various health focuses such as HIV/AIDS, malaria, maternal health, tuberculosis, surveillance,

Methods

and adolescent and sexual reproduction as they operate across various health facilities in Uganda [91].

5.3.2 Uganda Digital Health Dashboard

The Uganda Digital Health Dashboard (UDHD), also known as the Health Enabled Dashboard, was established in 2016. It is regarded as a pivotal hub for digital health initiatives in Uganda. The dashboard was created by the Government of Uganda to address various challenges within the country's healthcare provision, with a focus on coordinating digital health solutions. The dashboard has a catalogue of 57 providers of Digital Health Interventions (DHIs). In this way, it exhibits the diverse actors who are engaged in digital health in Uganda and the existing fragmentation in service delivery [92].

Comprising four national digital health systems and 53 organizations implementing digital health projects, the UDHD plays a crucial role in consolidating and streamlining digital health efforts in the country. The four national digital health systems featured on the dashboard include:

District Health Information System 2 (DHIS2): Adopted in 2010 as a national tool for collecting and reporting health data.

Human Resource for Health Information Software 3 (HRHIS3): Utilized for data collection, storage, analysis, reporting, and the dissemination of human resource information for evidence-based decision-making.

mTRAC: An mHealth solution designed to track essential medicines and enhance health service delivery. The mTRAC Stop Malaria Program, a component of mTRAC, enables health workers to input data on essential malaria drug stock levels.

Stop Malaria Program: Collaborating with districts to procure supplies.

However, a notable limitation of the 57 DHIs on the UDHD is the absence of a data export function. The health focus areas of these digital health solutions span diverse domains, including HIV/AIDS, malaria, maternal health, tuberculosis, surveillance, adolescent and sexual reproduction, clinical decision support, virtual health education, m-health, eHealth, Health Management Information System (HMIS), and preventable diseases. The UDHD, like the DHA-U, has identified instances of DHI duplication, indicating a need for streamlined coordination [92].

5.4 Methods

This case study employed a comprehensive methodology to evaluate the FAIRness of two digital health hubs and the associated vertical digital health solutions within these hubs. The study, conducted as a desk study based on documents related to digital health solutions in Uganda, encompassed the following key steps:

- Step 1: FAIRness Assessment Using the FAIR Evaluation Services Tool [93]: The initial phase involved assessing the level of FAIRness of the Digital Health Atlas Uganda

Results: Level of FAIRness of Digital Health Hubs in Uganda

(DHA-U), the Uganda Digital Health Dashboard (UDHD), and the vertical digital health solutions within these hubs using the FAIR Evaluation Services tool, which is a web-based automated service. This tool offers a structured evaluation of digital resources as it checks and validates metadata based on the FAIRMetrics and FAIRSharing groups. It also uses a collection of maturity indicator tests in order to ensure compliance with the FAIR Guidelines of Findability, Accessibility, Interoperability, and Reusability.

- Step 2: FAIRification and Integration of the Hubs: Based on the findings from Step 1, some efforts were made to enhance the FAIRness of the digital health hubs. The FAIRification process involved making the hubs more Findable, Accessible, Interoperable, and Reusable. To achieve this, a FAIR Data Point was used. This facilitated integration and alignment with FAIR principles. Subsequently, the level of FAIRness was reassessed after the process of FAIRification, which used the FAIR Evaluation Services Tool.

This methodological approach aimed to evaluate the FAIRness of digital health data in Uganda comprehensively while focusing on improving interoperability and accessibility through the process of FAIRification. The subsequent sections provide the details of the outcomes of these assessments and the implications of those results for sustainable digital health solutions in Uganda.

5.5 Results: Level of FAIRness of Digital Health Hubs in Uganda

The FAIR Evaluation Service was used to assess the level of FAIRness of the Uganda Digital Health Dashboard (UDHD) and Digital Health Atlas Uganda (DHA-U). A globally unique identifier (GUID) was used for the evaluation, and the results are summarized in Table 5.1.

Results: Level of FAIRness of Digital Health Hubs in Uganda

Table 5.1: Results of Maturity Indicator Tests for UDHD and DHA-U [13, 14]

Principle	Metric Test Description	UDHD	DHA-U	FAIR Data Point
F1	Unique identifier presence	1	1	1
F1	Identifier persistence	0	0	0
F1	Data identifier persistence	0	0	0
F2	Machine-findable metadata	1	1	1
F2	'Grounded' metadata	1	1	1
F3	Unique identifier in data	0	0	1
F3	Unique identifier in metadata	0	0	1
F4	Machine-discoverable resource	1	0	0
A1.1	Open and free data protocol	0	0	1
A1.1	Open and free metadata protocol	1	1	1
A1.2	Authentication capability	0	0	1
A1.2	Metadata authentication capability	0	1	1
A2	Persistence policy presence	0	0	0
I1	Broadly applicable language	1	1	1
I1	Semantically-grounded language	1	1	1
I2	Resolvable linked data terms	0	0	1
I2	Linked data terms resolve to FAIR data	0	0	1
I3	Links outward to third-party resources	1	0	1
R1.1	Explicit pointer to license	0	0	1
R1.1	Metadata pointer to license	0	0	1

5.5.1 Uganda Digital Health Dashboard (UDHD) Results

The UDHD was evaluated using the FAIR Evaluation Services based on its unique identifier ¹

. The results that are illustrated in Figure 5.1 demonstrate a varying level of FAIRness across different principles.

¹Source: <http://healthenabled.org/wordpress/uganda-digital-health-dashboard/>

Results: Level of FAIRness of Digital Health Hubs in Uganda

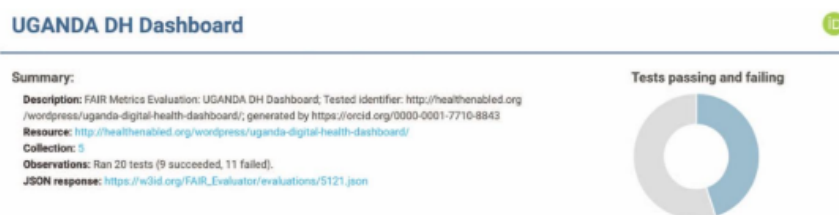


Figure 5.1: Pie chart illustrating the FAIR Evaluation results for DUganda Digital Health Dashbord showing the distribution of successful and unsuccessful maturity indicator tests

Source: [95]

Upon detailed analysis, the UDHD showed the lowest FAIR metrics test pass percentage in the 'Reusable' category at a score of 0%. This is because an explicit pointer to the license was absent in the linked data metadata. Conversely, the 'Interoperable' category achieved the highest FAIR metrics test pass percentage of 60%. The reason for this successful result is that interoperability is positively impacted by the use of a formal language that is broadly applicable for knowledge representation and the presence of links to third-party resources.

5.5.2 Digital Health Dashboard Uganda (UDHD) Results

After evaluating the DHA-U, conducting the FAIR metrics evaluation using the FAIR Evaluation Services and maturity indicator tests, the results show that out of 20 tests, 9 succeeded, and 11 failed. The detailed results are shown in Figure 5.2.

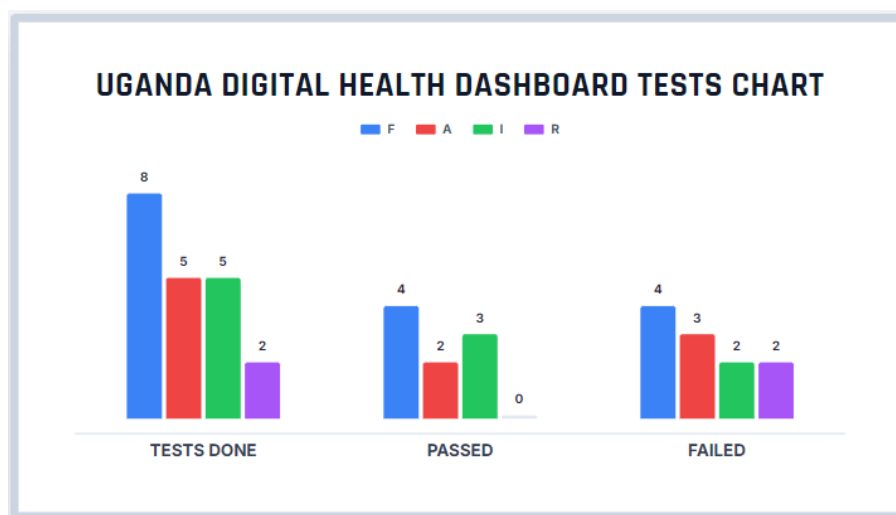


Figure 5.2: Bar chart illustrating the results of FAIR maturity indicator assessments conducted for the Uganda Digital Health Dashboard(UDHD) across the four FAIR principles

Results: Level of FAIRness of Digital Health Hubs in Uganda

The study showed that the performance of DHA-U in different formulations varied. The results indicated that the 'Interoperable' and 'Findable' categories were more successful compared to the other two categories. The reason for the success of the 'Interoperable' category is that the language that was used is formal and common, leading to knowledge representation. However, the 'Findable' category faced challenges because of the complexity of the machine-discovery resources that are used in search engines.

These results illustrate the current state of FAIRness for UDHD and DHA-U. The insights from this study form the basis for further discussion on how to improve the integration and sustainability of digital health solutions in Uganda.

5.5.3 Digital Health Atlas Uganda (DHA-U) Results

The Digital Health Atlas Uganda (DHA-U) was evaluated using its unique identifier ² through the FAIR Evaluation Services.



Figure 5.3: Pie chart illustrating the FAIR Evaluation results for Digital Health Atlas Uganda showing the distribution of successful and unsuccessful maturity indicator tests

Source: [96]

The maturity indicator test results for the DHA-U show that out of the 20 tests that were conducted during the assessment, 7 succeeded, and 13 failed. More detailed results are exhibited below.

²Source: <https://www.digitalhealthatlas.org/en/ug/>

Results: Level of FAIRness of Digital Health Hubs in Uganda

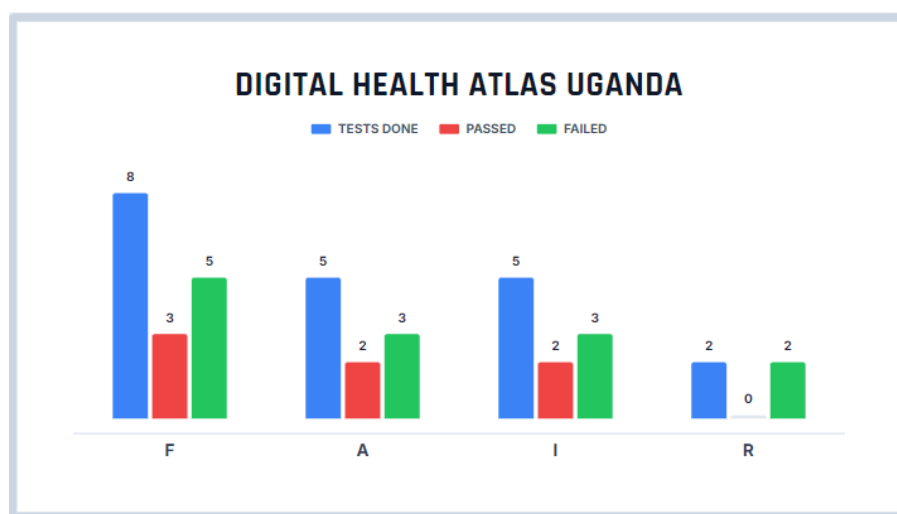


Figure 5.4: Bar chart illustrating the results of FAIR maturity indicator assessments conducted for the Digital Health Atlas(DHA-U) across the four FAIR principles

The analysis showed that DHA-U scored the lowest in the ‘recyclable’ category, receiving a 0% FAIR metrics test pass. This deficiency was attributed to the absence of an explicit pointer to the license in both linked data metadata and metadata. On the other hand, DHA-U excelled in the ‘Findable’ category by achieving a 50% FAIR metrics test pass. This success is because the unique indicators for metadata processing are available. Additionally, this success was due to the ability of devices to search through structured geographic metadata.

5.5.4 Digital Health Interventions (DHIs) FAIRness

To assess the quality of each digital health intervention (DHI) in the two digital dashboards, UDHD and DHA-U, the unique identifier for each DHI was examined. The data that was assessed was obtained and evaluated using the FAIR Evaluation Services comprised 98 digital health interventions from both dashboards. The results are presented in the table below.

The results in this table show that the majority of DHIs had a score of under 10, which is below the average success score. Only one DHI attained the score of 12, which is the highest success score out of all the other digital health interventions. Considering these low scores in the digital health interventions that were assessed, it is clear that there is a need to enhance the FAIRness of DHIs in Uganda. This can be done in various ways that include improving the Findability of each DHI through assigning persistent unique identifiers (PIDs) to metadata and data resources.

5.5.5 After FAIRification

A FAIR Data Point, which serves as a distributed data repository that hosts machine-actionable data and metadata that is aligned with the FAIR Guidelines, was implemented at Kampala International University, Uganda. This was done to improve the FAIRness of the Uganda Digital Health Dashboard (UDHD) and the Digital Health Atlas Uganda (DHA-U). Metadata

Results: Level of FAIRness of Digital Health Hubs in Uganda

Table 5.2: Level of FAIRness of Digital Health Interventions

Group	Success Score	Failure Score	Number of DHIs
1	3	17	19
2	7	13	15
3	8	12	11
4	9	11	7
5	10	10	6
6	11	9	3
7	5	15	1
8	12	8	1
9	0	20	1
10	4	16	1

configuration for the FAIR Data Point included a catalogue, layer, dataset layer, and distribution layer, with a specific catalogue created for the digital health hubs in Uganda.

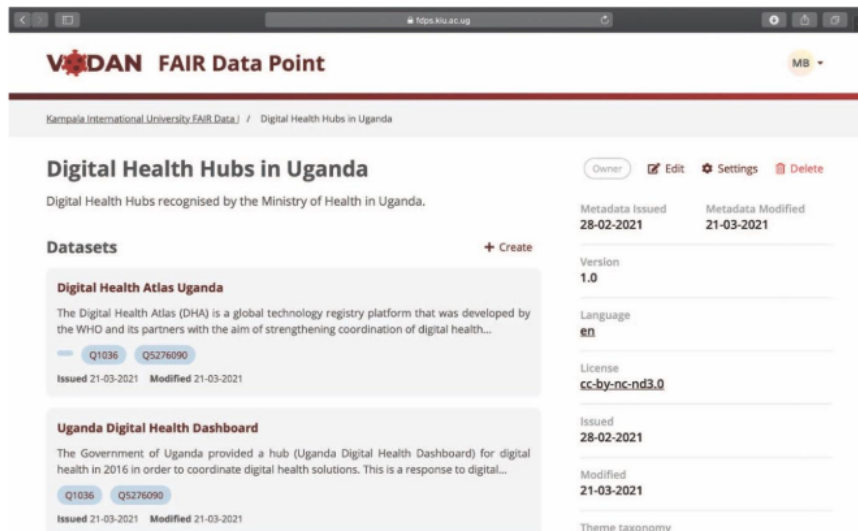


Figure 5.5: Screenshot of the FAIR Data Point catalogue for Ugandan digital health hubs, illustrating the organisation and publication of machine-actionable metadata in accordance to the FAIR Principles

Source: Screenshot taken by M. Basajja, 2021

A subsequent FAIR metrics evaluation was performed on the FAIR Data Point Catalog to assess the new level of FAIRness. The results of this evaluation are presented in Figure 6.

Results: Level of FAIRness of Digital Health Hubs in Uganda



Figure 5.6: Shows FAIR Data Point (UDHD and DHA-U catalogue) test results using FAIR Evaluation Services

Source: [97]

The maturity indicator tests, out of 20, demonstrated a substantial improvement in FAIRness after FAIRification, with 16 tests succeeding and 4 failing. A detailed breakdown of the tests per FAIR Guideline is illustrated in Figure 5.7.

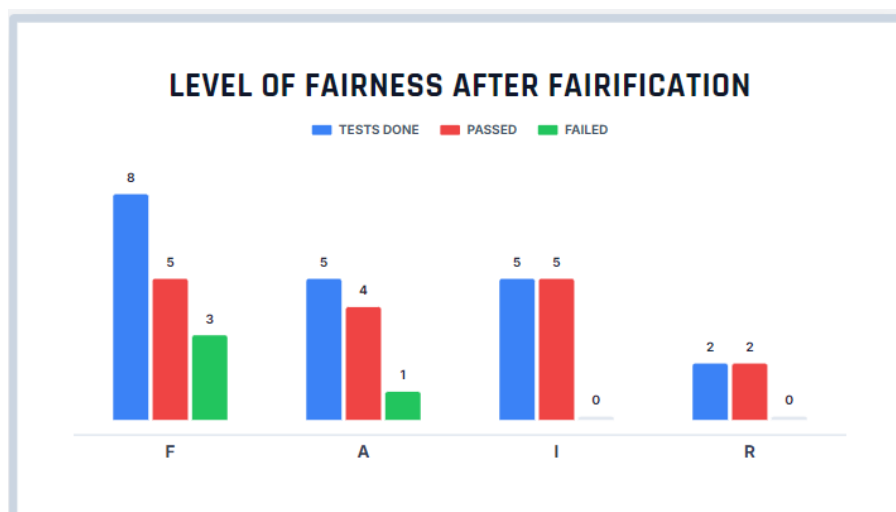


Figure 5.7: Bar chart illustrating the results of FAIR maturity indicator assessments conducted, passed and failed by the digital health hubs in Uganda after FAIRification.

Following FAIRification, the digital health hubs in Uganda exhibited a noteworthy improvement in FAIRness. The 'Reusability' and 'Interoperability' aspects reached 100%, a considerable enhancement from the initial 0%. While 'Findability' and 'Accessibility' experienced slight challenges, with 3 and 1 failed tests, respectively, due to the absence of a persistent identifier and a defined persistence policy, improvements in the number of passed tests were observed. In summary, the introduction of FAIR data protocols, as demonstrated by the FAIR Data Point implementation, significantly addressed structural challenges in the development and maintenance of sustainable digital health solutions in Uganda. This improvement showcases the effectiveness of FAIR principles in enabling Findability, Accessibility, Interoperability, and Reusability.

5.6 Conclusion

The FAIR Guidelines have been emphasized as a general framework for achieving FAIRness in data rather than a rigid standard since they were first developed and implemented. For this reason, the guidelines should be implemented responsibly to make technical decisions that are aligned with the principles of making data Findable, Accessible, Interoperable, and Reusable (FAIR). A detailed FAIR metrics evaluation that used the FAIR Evaluation Services revealed distinct FAIRness characteristics between the Uganda Digital Health Dashboard (UDHD) and the Digital Health Atlas Uganda (DHA-U).

Surprisingly, the UDHD demonstrated higher FAIRness compared to the DHA-U, particularly excelling in 'Findability' and 'Interoperability.' Both hubs achieved equal scores for 'Accessibility' and 'Reusability,' with 'Reusability' scoring the lowest due to inadequate metadata and data description, limiting indexation in searchable repositories. The absence of well-described licenses further restricted data readability to human users.

In particular, none of the digital health interventions assessed exceeded half of the FAIR maturity tests conducted. These interventions operated independently, lacking inherent support for FAIR principles. Data collected by the District Health Information System (DHIS) faced challenges in terms of Findability, Accessibility, Interoperability, and Reusability, hindering easy retrieval, access, and reuse.

To address these challenges, the Ugandan government is being asked to instill a degree of Fairness in all digital health tools. While acknowledging the diverse software and technology used by different Digital Health Interventions (DHIs) due to varied health focus areas and implementation challenges, standardizing data formats and adopting Application Programming Interfaces (APIs) could enhance interoperability. A systematic incorporation of FAIR data protocols through the implementation of a FAIR Data Point is proposed as a solution to overcome structural barriers and promote sustainable digital health solutions in Uganda. This strategic shift towards FAIRness has the potential to foster collaboration, reduce duplication of efforts, and improve the overall effectiveness of digital health initiatives in the country.

5.7 Discussion

The Uganda Digital Health Dashboard (UDHD) and Digital Health Atlas Uganda (DHA-U) both aim to serve as comprehensive repositories of digital health solutions. UDHD and DHA-U can promote collaboration and address the challenges facing sustainability in Uganda's healthcare system. The DHA-U is recognized for its contributions to providing information on interoperability and best practices within the healthcare field. Surprisingly, the UDHD showed higher levels of FAIRness compared to the DHA-U in terms of 'Findability' and 'Interoperability.' However, the results on both hubs showed that they had equal levels of accessibility and reusability, with 'Reusability' scoring the lowest out of all the other categories.

Despite the DHA's aim of encouraging reusability and interoperability, the FAIR metrics evaluation underscored the need for improvements to be made. Only seven of 20 tests passed. This indicated specific areas that require attention. It should be noted that DHA offers valuable information on interoperability and standards for each registered digital health project, including compatibility with various systems. However, there are persistent challenges, with many projects indicating 'N/A' for interoperability. This raises concerns about the transparency of

the details of the implementation of this repository.

To be registered, a digital health project in the DHA is required to complete global and country-specific fields that cover project overview, implementation details, technology, and interoperability. Although DHA encourages comprehensive information sharing, the completeness of the data that is collected is dependent on the diligence of the project team. The lack of details of interoperability may stem from a lack of awareness or understanding of the technology architecture and software used in other systems.

The metaphorical description of the current wave of mHealth interventions as 'black boxes,' indicating a lack of documentation on implementation details, resonates with the findings in Uganda. Despite efforts by the DHA-U, many projects lacked interoperability information, pointing to a broader issue in the documentation of implementation specifics. To address this, governments could play a pivotal role by mandating policies requiring digital solutions to provide comprehensive information, thereby incentivizing adherence to standards, formats, and ontologies in common use.

It is essential to note that FAIR principles emphasize machine-actionable data and metadata without explicitly prescribing specific frameworks. While RDF and linked data represent popular solutions, achieving FAIRness is viable using diverse frameworks. The FAIR Guidelines, however, underscore the importance of machine-readability for comprehensive application.

In line with the principles advocated by Tomlinson et al.[98], there is a critical need for concerted efforts by governments, funders, and private enterprises to establish standards for interoperable platforms, ensuring the scalability and sustainability of mHealth initiatives. Open mHealth architectures that are based on robust platforms with application development standards are recommended, as this aligns with the core of the FAIR Guidelines that promote interoperability and scalability.

Guidance tools such as the mHealth Assessment and Planning for Scale (MAPS) Toolkit focus on using best practices to scale up mHealth innovation and advocate for standardization and interoperability. The lack of specified standards and communication protocols in mHealth systems can lead to competing and duplicated datasets that hinder the integration and improvement of these systems.

In a 2011 paper, Lemaire reiterates the importance of data standardization and communication infrastructure to successfully scale mHealth systems [99]. Suggestions in this regard include developing a framework type, reference systems, and the use of specific data and technical standards will be encouraged. Governments are encouraged to be more efficient in implementing these policies and standards to align with the objectives of the FAIR Directive.

Thus, promoting data standardization and collaboration is important as this has an impact on the implementation of projects that are aimed at scaling up mHealth systems in developing countries. These principles are consistent with the WHO's recommended integrated approaches based on standards expected within the sector.

Discussion

Chapter 6

Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic

This chapter is based on the following publication:

Basajja, M., Suchanek, M., Taye, G.T., Amare, S.Y., Nambobi, M., Folorunso, S., Plug, R., Oladipo, F.O., Van Reisen, M.: Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic. *Data Intelligence* 4(4) (2022). doi: https://doi.org/9.1162/dint_a_00179

Abstract

Rapid and effective data sharing is necessary to control disease outbreaks, such as the current coronavirus pandemic. Despite the existence of data sharing agreements, data silos, a lack of interoperable data infrastructures, and different institutional jurisdictions hinder data sharing and accessibility are hindered. To overcome these challenges, the Virus Outbreak Data Network (VODAN)-Africa initiative is championing an approach in which data never leaves the institution where it was generated, but instead, algorithms can visit the data and query multiple datasets in an automated way. To make this possible, FAIR Data Points—distributed data repositories that host machine-actionable data and metadata that adhere to the FAIR Guidelines (that data should be Findable, Accessible, Interoperable, and Reusable)—have been deployed in participating institutions using a dockerised bundle of tools called VODAN in a Box (ViB). ViB is a set of multiple FAIR-enabling and open-source services with a single goal: to support the gathering of World Health Organization (WHO) electronic case report forms (eCRFs) as FAIR data in a machine-actionable way, but without exposing or transferring the data outside the facility. Following the execution of a proof of concept, ViB was deployed in Uganda and at Leiden University. The proof of concept generated a first query, which was implemented across two continents. A SWOT (strengths, weaknesses, opportunities, and threats) analysis of the architecture was carried out, and established the changes needed for specifications and requirements for the future development of the solution.

Keywords: Digital health, Data in residence, FAIR Guidelines, Machine-actionable, VODAN-Africa

6.1 Introduction

The emergence of the novel coronavirus (COVID-19) pandemic has highlighted the critical need for rapid data sharing to enable an effective pandemic response [100]. Acknowledging this urgency, the World Health Organization (WHO) has emphasized the vital role of public access platforms in disseminating complete viral genome sequences and assay protocols. These efforts have significantly contributed to early and accurate infection diagnosis during the ongoing global emergency [101, 102]. However, the sensitive nature of real-time data collected, referred to as Real World Observation Data, poses challenges in centralized data warehousing, such as at the WHO [100, 103]. In response, a distributed data-sharing approach has emerged, enabling data querying within the country or institution where it is collected. The approach aligns with the principle of making data “as open as possible, as closed as necessary” [103].

The Ebola virus outbreak in Africa (2013–2016) highlighted the critical role of efficient information sharing during public health crises (2013–2016) [104, 105]. Advances in data-sharing mechanisms during this crisis fostered a global consensus on the necessity of transparent information sharing, particularly during emergencies [102, 105] but [2015]. An insight at a WHO advisory meeting in September that merely agreeing on information-sharing principles is not enough highlighted the need for practical implementation of technical challenges, including improving the accuracy of data capture, ensuring data quality, and aligning data structures [104, 105]. Drawing from the lessons of the West African Ebola outbreak, it became evident that individual deployment of various digital solutions for data collection lacked interoperability. This lack of coordination was seen in communication and efficiency within the international response to the epidemic [101, 105].

Limited public access to medical information is a major challenge in Africa during the current COVID-19 pandemic. Restricted access to this information hinders understanding and analysis of epidemic progress. For example, although the first reported outbreak in East Africa occurred in Kenya on March 13, 2020, the lack of publicly available SARS-CoV-2 genome data from East African Member States, including Tanzania, Burundi, Uganda, Rwanda, and South Sudan, raises concerns. Similarly, the first case was reported in Nigeria on February 2, 2020, but SARS-CoV-2 genome data remain elusive. The delay in sequencing and sharing Ugandan SARS-CoV-2 genomes until 2 July 2020 underscores the difficulties in obtaining critical genomic, epidemiological, and clinical data essential for an effective pandemic response [101, 106]. The absence of comprehensive, digitally recorded, and interconnected health data poses a substantial obstacle to understanding, analyzing, and effectively responding to the complexities of the COVID-19 pandemic. The following sections delve into a proof of concept and potential horizons for deploying FAIR Data Points as a solution to address these challenges.

6.2 Relevance of the Study

The urgent need for reliable and accurate data to combat the global spread of COVID-19 has prompted national and international cooperation [107]. To support data-driven decision-making, the dissemination of information, whether to the public, research institutions, hospitals, or government offices, must adhere to usefulness, usability, and desirability principles. Traditional data collection methods have introduced health risks for domain experts, data collectors, and stewards, particularly in isolation measures affecting groups and individuals. As the

COVID-19 crisis unfolds, technological solutions have become essential for mitigating risks associated with the collection, access, and distribution of high-quality data. Yet, navigating these challenges is further complicated by data generated across multiple jurisdictions, each governed by varying policies on data sharing, accessibility, and compliance with the EU General Data Protection Regulation (GDPR).

Contrary to expectations given Africa’s communal lifestyles, reported COVID-19 cases on the continent have remained relatively low [108]. Potential factors contributing to this unexpected trend include minimal imported cases from COVID-19 hotspots and the advantageous experience gained from combating previous outbreaks like Ebola [109, 110]. However, the looming threat of COVID-19 on Africa’s already fragile health systems highlights the critical need for meticulous data collection, storage, and analysis within sovereign local repositories. The pivotal question emerges: How can data be effectively utilized to drive transformative change in Africa’s fragmented health systems? Collaborative efforts among all stakeholders are essential to leverage valuable data for understanding COVID-19 trends and geographical distribution. A significant hurdle in addressing these challenges lies in the poor data handling and reporting mechanisms prevalent in Africa [111]. The lack of accurate, reliable, and timely data hinders the effective monitoring and evaluation of COVID-19 interventions. Electronic health information systems, such as Ethiopia and Uganda, remain highly fragmented and lack interoperability, hindering the seamless exchange of information. Despite the plethora of data collected by non-governmental organizations and foreign researchers in Africa, a critical limitation arises as the data is often relocated to Western research centres, rendering it underutilized for local improvements [112].

The existing data collection methods, whether through paper-based health management information systems (HMISs) or automated tools like electronic medical records (EMRs), often fall short of being machine-readable. In hierarchical systems where data is not federated, health professionals contribute data but seldom receive feedback or have control over their data. Establishing a system that empowers health professionals to own and utilize their data, presented in a machine-readable format, offers the potential for local data engineering and analytics. This approach not only minimizes duplicate efforts but also addresses the imperative need for FAIRification [170]. The subsequent sections present a proof of concept and explore the potential for deploying FAIR Data Points during the ongoing COVID-19 pandemic, aiming to transform the landscape of data management and utilization in Africa.

6.3 FAIR Guidelines

At the heart of making Real World Observation Data ‘Findable,’ ‘Accessible,’ ‘Interoperable,’ and ‘Reusable’ (FAIR) lie the FAIR Guidelines [102, 29], an essential framework championed by a diverse group of data stewards and stakeholders. These guidelines underscore the importance of achieving data localisation, ensuring data is reachable over the Internet, adopting machine-readable formats using linked-data approaches for human and machine comprehension, and fostering data convergence through controlled vocabularies within collaborative data communities. The resulting machine-readable information, known as ‘metadata,’ encapsulates a set of data that describes and informs about other data. However, the realization of these objectives entails navigating diverse political, institutional, and regulatory landscapes governing privacy, security, confidentiality, and the rights associated with health and personal data. Such complexities often act as impediments to seamless data sharing for research and other purposes.

Research Questions and Objective

The existence of data silos and non-interoperable data infrastructures further exacerbates the challenge by hindering the machine-assisted querying of multiple datasets.

Addressing these challenges, VODAN-Africa, in collaboration with the GO FAIR Foundation, has introduced a groundbreaking paradigm for data curation, aiming to render data FAIR. In a FAIR ecosystem, data ownership remains within the local context, granting sovereignty over the data [114]. This paradigm shift empowers algorithms to visit data for analysis. The VODAN-Africa initiative, spearheaded by researchers from six African countries—Ethiopia, Uganda, Kenya, Zimbabwe, Tunisia, and Nigeria—in partnership with Leiden University, specifically focuses on combating COVID-19. The primary objective is to establish distributed access to COVID-19 data through the Internet of FAIR Data and Services (IFDS) [114]. IFDS envisions a scalable and transparent framework for routing data, tools, and services [115], aligning with the principles of FAIR to enhance the accessibility and utility of critical COVID-19 information.

6.4 Research Questions and Objective

Research Questions

Main Research Questions: How can the deployment of FAIR Data Points in the context of the COVID-19 pandemic enhance data interoperability and facilitate distributed analytics across different geographical areas and jurisdictions? **Sub-Research Questions:**

Sub-Research Question 1:

What are the essential components and technical specifications necessary for developing and deploying a VODAN-In-a-Box solution for FAIR data management?

Sub-Research Question 2:

How effective is the deployment of FAIR Data Points in real-life situations, particularly in African countries such as Ethiopia and Uganda, in generating and accessing machine-readable semantic clinical patient data?

Sub-Research Question 4:

What are the strengths, weaknesses, opportunities, and threats (SWOT) of the FAIR Data Point architecture in managing COVID-19 data across diverse jurisdictions?

The primary objective of the VODAN-Implementation Network (VODAN-IN) is to establish a robust, international, and interoperable distributed data network infrastructure. This infrastructure addresses the immediate need for evidence-based responses to the ongoing viral outbreak while also facilitating the long-term reuse of data and service infrastructures in future epidemic scenarios [103]. As part of their collaborative efforts within VODAN-IN [102, 116], four major international data organizations—Committee on Data for Science and Technology (CODATA) [117], Research Data Alliance (RDA) [118], World Data System (WDS) [119], and GO FAIR [120]—have united under the consortium named 'Data Together.' Their collective commitment aims to optimize the global research data ecosystem, identifying opportunities and needs that will drive federated infrastructures to meet the evolving demands of data-driven sci-

Research Questions and Objective

ence [115, 121]. Recognizing data and science as a global public good, Data Together, through its 'Data Together COVID-19 Appeal and Actions,' emphasizes the imperative of aligning data and science platforms with the FAIR Guidelines to provide the best available science for pivotal decision-making in the current and future crises [121].

Amidst the urgency of the COVID-19 pandemic, the accelerated implementation of a FAIR ecosystem becomes imperative [31]. This ecosystem is envisioned to enhance the ability to combine, visualize, and utilize data from diverse sources, facilitate nuanced data access and protection, and enable decentralized and machine-assisted analysis [30]. The formation of VODAN-IN stems from the recognition that the valuable data amassed from past and ongoing epidemics is not consistently accessible across different populations and countries [100, 122].

The FAIR-Implementation Network Africa and the VODAN-Africa Research Network were established with a shared mission—combating and containing the COVID-19 pandemic through distributed access to critical data from Africa. The primary focus of VODAN-Africa is to build resources facilitating digital access to COVID-19 pandemic-associated data, adhering to existing governance regulations [290]. The overarching objective of VODAN-IN Africa is to ensure the FAIRness of COVID-19 (meta)data. This FAIR data can then be analyzed in conjunction with globally available data to effectively combat the ongoing COVID-19 pandemic and future infectious disease outbreaks.

Employing participatory ethnography as the research methodology, the study aimed to understand how data analytics of digital health data in residence could be propagated, integrated, and sustained through a FAIR Data Point (FDP). As a participatory ethnography design, the researchers actively engaged in the development of the solution, providing access to key documents and involving critical stakeholders in the process [290]. The investigation delved into the deployment of a FAIR Data Point, accompanied by the installation of Data Stewardship Wizard (DSW) software prepared for this effort, to produce FAIR data and test the hypothesis that visiting data held in repositories in health facilities would be possible. The proof-of-concept criteria included:

Production of machine-readable semantic clinical patient data in residence.

Visitation of metadata through the Internet by deploying FAIR Data Points in six African countries and Leiden University Medical Center.

Querying FAIR Data Points installed in different geographies and jurisdictions using distributed analytics.

Integration of COVID-19 data between FAIR Data Points through data visiting, validating the feasibility of an Internet of FAIR Data and Services (IFDS).

Testing the architecture and tools in a real-life situation.

To assess the practical aspects of the deployment, this article provides a detailed exploration of the deployment experience in two of the six countries, namely Ethiopia and Uganda. Additionally, a SWOT analysis was conducted on the architecture to discern its strengths, weaknesses, opportunities, and threats (SWOT) and pave the way for potential future development [290].

6.5 Development of VODAN In a Box

To fortify data ownership in Africa, VODAN-Africa advocates an innovative approach where data remains within the local institution's underlying database. In this paradigm, algorithms, encapsulated in virtual machines, visit the data locally to conduct analyses addressing specific inquiries. The outcome is pseudonymized aggregate data, allowing combination with results from analogous queries at other locations, contingent upon the authorization of the respective institution. Addressing this challenge, the technical team of the Data Stewardship Wizard (DSW) developed tools with the following specifications:

Emphasis on data ownership and residency in health facilities.

Secure data-visiting over the Internet.

Focus on COVID-19 clinical patient data.

Integration with COVID-19 research data on prevalence.

Deployability in diverse real-life African settings.

This comprehensive toolset, coined "VODAN in a Box" (ViB), encompasses vital components aimed at capturing COVID-19 clinical data and publishing metadata for the datasets [123].

6.5.1 Components

The ViB consists of three essential components: the FAIR Data Point (FDP), the Case Report Form (CRF) Wizard, and the triple repository/query interface, illustrated in Figure 6.1. [123] The ViB tools, DSW, World Health Organization-sponsored (WHO) Electronic Case Report Forms (eCRFs). It allows you to read what machines can read. Features include:

FAIR Data Point (FDP), a tool for publishing definitions for FDP, and FAIR metadata for catalogues, data types, and classifications. It acts as a metadata repository, enabling data owners to disclose the metadata of their digital products fairly, and making it easier to find for consumers,[124].

Case Report Form (CRF) Wizard: This is a tool for managing semantic data. The data entry form is enclosed with the template K. WHO COVID-19 eCRF. It enables the implementation of efficient data management systems based on FAIR data management and provides highly searchable, accessible, usable, and reusable data [125].

The AllegroGraph Triple Store: it works as a knowledge graph repository that provides query interfaces that allow the retrieval of data [126].

The DSW, originally designed for data management planning, now supports FAIR data stewardship. The ViB's components, including the FDPs deployed in each local institution, allow virtual machines to visit them akin to trains docking at stations. This innovative approach enables virtual machines to visit multiple FDPs to obtain answers to their queries [122, 127]. ViB, therefore, functions as a collection of applications supporting the acquisition and exposure

Development of VODAN In a Box

of [290]. In essence, ViB encapsulates the technical infrastructure needed for the FAIRification of data, emphasizing interoperability and secure data access within a distributed framework.

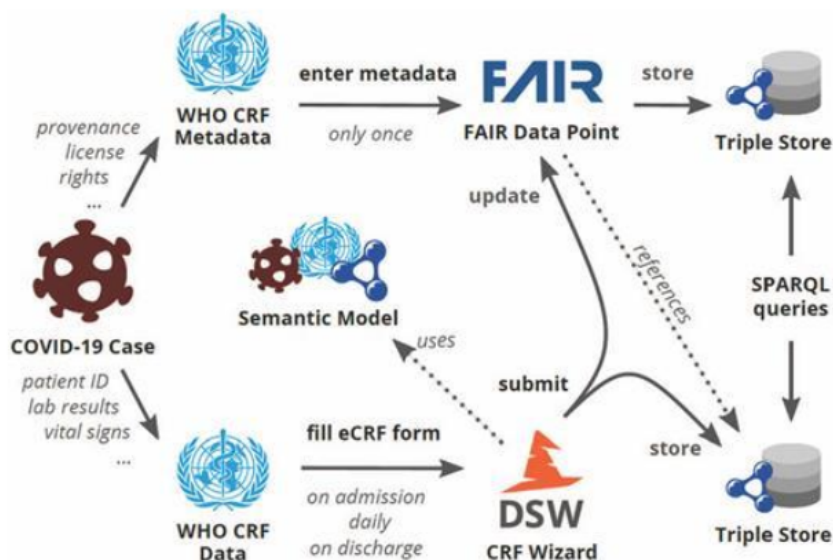


Figure 6.1: Architectural framework for VODAN in a Box platform showing the flow of clinical data, metadata, and semantic annotations within the FAIR data ecosystem

Source: [123]

The COVID-19 electronic Case Report Form (eCRF) is an integral part of the VODAN in a Box (ViB) framework, designed to facilitate machine interpretation of patient data. The eCRF comprises three-page forms, each housing three modules: admission, follow-up, and discharge. These modules group logically related data elements, encompassing vital signs, demographic characteristics, and laboratory test results. The VODAN-Africa initiative sought to develop an eCRF template capable of generating machine-readable data in compliance with FAIR Guidelines.

The ViB deployment architecture is deliberately designed to be user-friendly, even while integrating multiple interconnected services. Hosted as an open-source GitHub repository [128], ViB includes essential configuration files and a Docker-compose.yml. To run ViB, Docker and Docker-compose are prerequisites, applicable for both local testing and production mode. The latter enhances security and availability, offering options such as an HTTPS proxy or a triple store with authentication. ViB seamlessly connects the CRF Wizard, a FAIR Data Point (FDP), and an AllegroGraph triple store. Both the FDP and DSW are actively developed, highly configurable open-source projects. Transitioning ViB to production mode requires preparatory steps such as creating DNS records and generating SSL certificates to ensure secure access. Some configurations involving security measures and local adjustments require manual handling. Collaborative efforts within the network, supported by comprehensive documentation [129], enabled a smooth deployment process, with only minor issues related to misconfigurations or certificate acquisition. On July 22, 2020, Kampala International University successfully deployed the first ViB instance in production.

The AllegroGraph triple-store database, responsible for storing linked data subject predicate object triples, empowers sophisticated queries using the SPARQL query language, including federated queries [126, 130]. By default, the FDP uses an in-memory store, with alternative options including the Native store, AllegroGraph Repository, Graph DB Repository, and Blaze Graph Repository [131]. In Uganda, the ViB implementation specifically utilizes AllegroGraph triple-store storage. User account information for the FDP resides in the MongoDB database. Packaged as a Docker image under FAIR data/FDP, the FDP serves as the core component for managing business logic and operations related to semantic data. The FDP provides an application programming interface (API) for data manipulation in various formats and offers a user interface for human interaction. In a production deployment setup, a reverse proxy ensures secure connections to the FAIR Data Point by handling HTTPS certificates.

6.5.2 Technical Specifications.

The development of the VODAN in a Box (ViB) initiative commenced during the early stages of the COVID-19 pandemic, driven by a strategic focus on COVID-19-related data. In response, the project primarily concentrated on producing machine-readable metadata, with a specific emphasis on the COVID-19 World Health Organization (WHO) electronic Case Report Form (eCRF). Consequently, the scope of ViB was limited to the COVID-19 WHO eCRF, aligning with the project's overarching goal.

The ViB's Case Report Form (CRF) Wizard, designed as a web-based form derived from WHO's Rapid COVID-19 eCRF, facilitates the export of data in the Resource Description Framework (RDF) using the semantic data model of the WHO COVID-19 eCRF. This RDF data can be downloaded as a file or directly submitted to the pre-configured triple store attached to the system. The triple store is specifically designed to facilitate distributed querying across multiple datasets. Although the CRF Wizard is inherently connected to the integrated triple store, its versatility allows for seamless configuration with any triple store that supports a standard SPARQL endpoint for CRF data insertion.

The ViB's CRF Wizard serves as a user-friendly interface for data entry, streamlining the process of converting patient data into the machine-actionable format specified by the CRF. Patient data, derived from sources such as Case Report Forms (CRFs), remains localized, residing within the jurisdiction of the originating hospital or clinic. In the routine workflow of ViB, data entry users or data stewards can:

Enter Data into the CRF Wizard: The interface facilitates the entry of patient data into the CRF Wizard.

Generate a CRF Report: The system enables the generation of a comprehensive CRF report based on the entered data.

Submit the CRF Report to the Triple Store: Once created, the CRF report can be exported to the triple store, which is pre-configured and associated for later storage and retrieval.

The produced data, which is now in a machine-readable format, becomes usable for queries through data-visiting algorithms. It is important to note that successful data visiting relies on the access and control arrangements that were established, although the complexity of these arrangements is beyond the scope of this article.

6.5.3 Metadata Component Choices

The selection of metadata components for the VODAN in a Box (ViB) program was achieved through a collaborative effort by the VODAN-Africa team, conducted over six months of weekly Metadata for Machines (M4M) workshops. The process resulted in nine metadata products with ViB FAIR Data, so we developed standard Points (FDPs), which have been implemented in six African countries. The selected metadata components described in Table 6.1 were intended to meet a variety of requirements to ensure that the data is findable, accessible, interoperable, and reusable in compliance with FAIR principles.

Table 6.1: VODAN metadata component choices for Kampala International University, Uganda FDP 1.1 [29].

#	Metadata Component	FAIR Guidelines	Question for the domain community participating in the Metadata for Machines (M4M) workshop	Kampala International University Uganda FDP 1.1
1	F1	What is the persistence policy for the identifier systems used for digital assets?	Policy statement from the domain community regarding the requirements for identifier systems.	PURL/hospital persistent URL.
2	F2	What is the minimum acceptable profile needed to ensure findability?	Examples: people, institutions, geographical locations, date/time, funding source, publishers, repositories.	Geographical locations, institutions, date/time, repositories, publisher.
3	F4	What are the minimum acceptable indexing requirements and limitations?	Metadata that complies with indexers and search algorithms.	Google.
4	A1.1	What access policies and protocols are in place for this community?	Self-describing access restrictions, based on policy.	Hospital patient data is only accessed by the hospital data management personnel.

Continued on next page

Table 6.1 continued from previous page

#	Metadata Component	FAIR Guide-lines	Question for the domain community participating in the Metadata for Machines (M4M) workshop	Kampala International University Uganda FDP 1.1
5	A2	What is the persistence policy for the metadata?	Policy statement from the domain community regarding the requirements for metadata longevity.	In Uganda, archiving metadata to be saved permanently is done.
6	I2	How FAIR are the vocabularies?	Metadata linking to canonical descriptions of terminology systems, vocabularies, and ontologies.	Open Science Framework.
7	R1.1	What usage license(s) will be used?	Metadata linking to machine-actionable license.	CC License.
8	R1.2	What are the minimum provenance metadata needed to ensure reuse?	Examples: laboratory methods, analytical methods, computational platforms.	
9	R1.3	What FAIR Implementation Profile (FIP) is used to make data/metadata FAIR?	A machine-actionable list of the FAIR-enabling resources used to make the data and metadata FAIR.	VODAN FIP.

During a six-month series of Metadata for Machines (M4M) workshops, the VODAN-Africa team carefully selected nine customized metadata components for each Fair Data Point (FDP) in six African countries. These metadata components addressed crucial aspects such as the persistence policy for identifier systems (F1), the minimum profile for findability (F2), acceptable indexing requirements (F4), access policies (A1.1), metadata persistence policy (A2), FAIRness of vocabularies (I2), usage licenses (R1.1), minimum provenance metadata (R1.2), and the FAIR Implementation Profile (FIP) for making metadata FAIR (R1.3). Over time, additional metadata features have been refined and updated across different countries to enhance their applicability and alignment with local contexts.

The active participation of five data managers played a pivotal role in the six-month trainer

Proof of Concept

preparation program. These data managers were instrumental in the preparatory training process, actively collaborating with health agencies, regional health authorities, and university officials. Their responsibilities include setting up local servers, purchasing domain names, setting up domain names via DNS, installing ViB on Virtual Private Servers (VPS), analyzing ViB for data generation, populating COVID data from various sources, and integrating with the first SPARQL.

This period marked a critical initiation where data stewards, by engaging in data collection and FDP setup, retained ownership of data within their respective institutions. Through internal surveys, practical assignments, and discussions, it was observed that data stewards found the WHO standard COVID-19 eCRF Wizard user-friendly. Once COVID-19 data was entered into the eCRF, the template automatically converted the raw data into RDF format, storing it in the triple store as FAIR-compliant data. This entire process, from training to installation and query, spanned four months. Subsequently, a phase of documentation, assessment, and reflection ensued, preparing for the upcoming stages of the project.

6.6 Proof of Concept

6.6.1 Deployment in A Real-life Situation

For the proof of concept, deployment in a real-life situation was critical. The research team developed the following steps for the deployment: Step 1: Testing of the proposition in a real-life situation Step 2: Presentation of a clear proposition to stakeholders in universities, university hospitals, other hospitals, clinics, and ministries of health Step 3: Analysis to ensure compliance with relevant Regulatory Frameworks Step 4: Approval by stakeholders in all locations Step 5: Establishment of FAIR Data Points in two African and one European countries reachable over the Internet Step 6: Machine-actionable data production (test data and real data) Step 7: Running queries over the Internet across the FAIR Data Points that visit the machine-actionable data Step 8: Completion of proof of concept

This describes the initiation of the VODAN-Africa project, highlighting the crucial role of capacity building through the training of trainers. The objective was to empower African experts, including researchers and traditional data stewards from health facilities, with the skills needed to generate FAIR (Findable, Accessible, Interoperable, and Reusable) machine-actionable COVID-19 data. The training sessions focused on using the Electronic Case Report Form (eCRF) wizard, a component of the ViB, to facilitate the entry of structured and standardized data. Participants were trained not only in data entry but also in the subsequent steps of publishing metadata on FAIR Data Points (FDPs) and querying multiple triple stores across FDPs using SPARQL (SPARQL Protocol and RDF Query Language).

At the same time, Dockerized FDP and ViB were deployed together with training workshops. Technical dockerization was used for packaging applications and dependent containers for scalability and ease of use. On 22 July 2020, Kampala International University reached a milestone with the launch of the first FAIR Data Point for COVID-19 data called VODAN in a Box (ViB). This is a significant step being made towards realizing the VODAN-IN Africa initiative that provides a platform for collecting and disseminating data on COVID-19 that is suitable for device operations in line with FAIR principles.

6.6.2 Procedure for Proof of Concept

The proof of concept is inclusive of the World Health Organization (WHO) Electronic Case Report Form (eCRF), which can be read by humans and machines and distributed to partners in different countries. This Electronic Case Report Form is the basis for creating machine-readable metadata on FAIR Data Points (FDPs). The FDPs were configured to be viewable and accessible through the Internet. This allowed them to be discovered by algorithms through the VODAN FDP community.

The steps that are key in designing a proof of concept:

Preparing the WHO Electronic Case Report Form: A WHO Electronic Case Report Form that can be read by humans and machines was developed and distributed to partners in various locations.

Visibility and Accessibility of FDPs: FDPs were designed to be visible and accessible through the Internet. This, along with the appropriate licenses, ensured that algorithms running on the Internet could be accessed.

Automation of Data Exchange: The already installed FDPs supported the automated cross-country exchange of observational COVID-19 patient data. This facilitated querying of FAIR-compliant data by hospitals, facilities, and researchers across different countries while adhering to local data protection and governance laws.

Mock Data Entry: Mock data was entered into the ViB using the CRF Wizard tool, which adhered to the semantic model based on the WHO eCRF. This was achieved using SPARQL, an RDF query language.

Example Query: An example query was formulated to retrieve information from the triple store. The query was designed to present the names of healthcare facilities that admitted male patients with a fever (body temperature greater than or equal to 38 degrees). This query was executed through the AllegroGraph query interface.

SPARQL Query Execution: SPARQL queries were entered into the AllegroGraph query interface, and service call requests were made to the SPARQL endpoints of the three countries involved (Uganda, Kenya, and the Netherlands). The results displayed the facility names against the number of COVID-19 patients.

The proof of concept aimed to demonstrate the functionality of VODAN in a Box (ViB) in a real-world setting, showcasing the viability of querying FAIR-compliant data across different geographical locations.

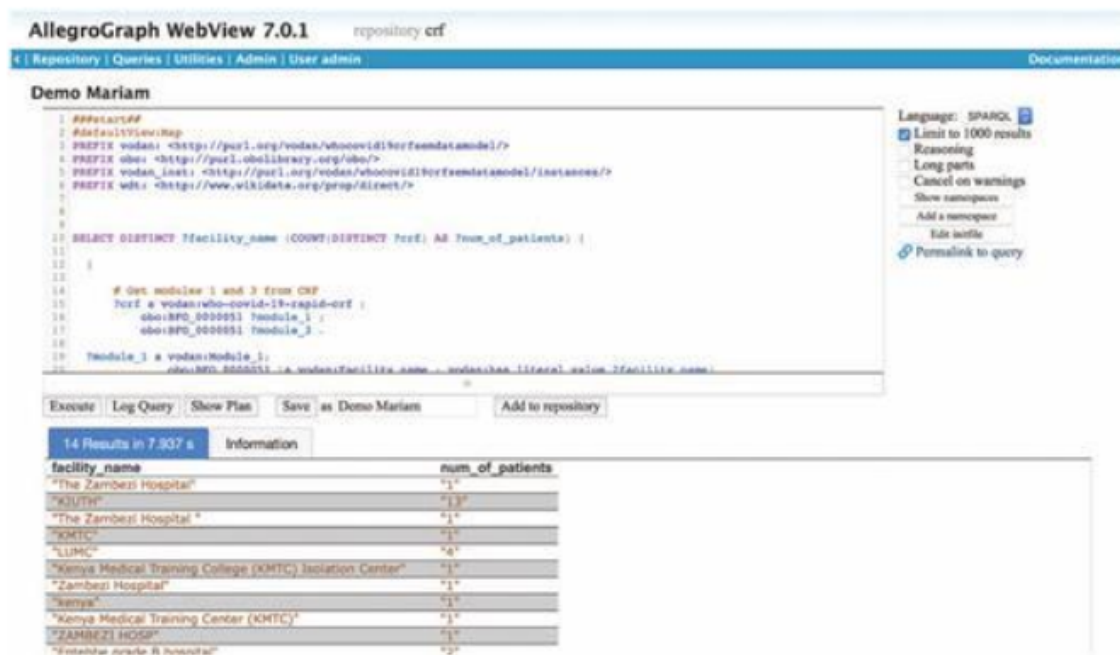


Figure 6.2: Output produced by the VODAN in a Box proof of concept demonstrating federated querying and retrieval of distributed health data from participating repositories

6.7 SWOT Analysis

The reflection period was used to reflect on the potential of the architecture, but also to identify limitations and alternatives. To assess the architecture, an analysis was carried out of strengths, weaknesses, opportunities, and threats (SWOT). The analysis was carried out as an internal evaluation by the data stewards involved in the research.

6.7.1 Strengths

The strength of the architecture was identified as its potential to organize data ownership and data visiting, and generate potential for the data within the health facilities to contribute to the quality of healthcare. Together, the FDPs form a distributed data network through which queries are able to visit the data and pose research queries on the data therein without the necessity of downloading or importing the data to another location. This distributed network of human and machine-readable data will constitute the Internet of FAIR Data and Services. The tools were deployed in diverse real-life African settings and generated interest.

Of the original requirements for the proof of concept, the following assessment was made by the resident data stewards within the hospitals and facilities. Data ownership and residency in health facilities were feasible, and secure data-visiting over the Internet was possible.

In terms of strengths identified, the data stewards found that the design and deployment of the ViB, together with the FDP project, uses some of the best practices in the modern computing world, such as distributed computing, software virtualization/containerization, and semantic

interoperability. Scientific fields that generate high-dimensional data sets have embraced the principle of data visitation (i.e., moving the code to data instead of data to code). This is because moving big data across networks takes massive computational resources and time. Besides, such data transfers incur myriad bottlenecks, including network issues, particularly with internet speed, especially in Africa. Code and data usually exist on different machines/servers, so either the code or the data has to be moved to execute the code on the data. As the code (usually contained in an executable file) is almost always smaller than the data, it is better to move the code to the data (or closer to the data). Apache Hadoop, the distributed computing framework for big data storage and processing, which consists of a storage layer (Hadoop Distributed File System) and a batch processing engine (MapReduce) [132], is based on the design principle that moving computation is cheaper than moving data [133]. In addition to this, further ViB development is planned to allow local customizations and other data types.

6.7.2 Weaknesses

A critical issue was the inflexibility of the DSW Wizard tool towards the FAIR metadata production of research data and health data types. The limited scope of the WHO eCRF within the DSW Wizard did not allow the production of any alternative data in a machine-readable format. The alternative data, including research data on COVID-19, could not be queried despite the availability of the FDP. This was a setback in the integration of both the patient data and research data. The focus on COVID-19 clinical patient data needed more flexible data entry templates. The DSW tool did not allow for bulk input, which would be a more efficient way of data production. The DW Wizard also did not allow seamless integration with the workflows and requirements in the health facilities, especially regarding the requirements to upload data in the HMISs (such as the District Health Information System, DHIS2). Moreover, the integration with COVID-19 research data on prevalence was limited by the inflexibility of the tools used to produce semantic data.

Subsequent to the proof of concept, and to overcome the limitations encountered, VODAN-Africa has extended collaboration with developers in the semantic world to enhance and develop education and development. Moreover, to create greater flexibility in the production of semantic languages, the VODAN-Africa group has engaged with the digital online learning courses developed to help advance these skills among computer science and data science graduates so that data managers will obtain the skills for the development of machine-readable semantic data in Africa [134].

Other weaknesses included a lack of flexibility in creating new templates and an inability to import bulk data when the immediate requirements go beyond the WHO eCRFs. The DSW toolset also had further weaknesses: the bulk inputting of data was not possible; it was not adaptable to seamlessly integrate into clinics and hospitals; it could not integrate different types of data on COVID-prevalence, including those derived from scientific projects; it could not solve the potential problems of data export from the Wizard into the DHIS; and it could not be easily extended to include the functionalities based on differentiated situations on the ground. Key considerations for local deployment include the need for the localization of user interfaces and the semantic content, and adaptation to accommodate languages and cultures within the system. Ease of use and flexibility, as well as ease of deployment through Dockerization or virtualization [135] for local deployment, with due consideration for infrastructure (such as bandwidth and storage), are also equally important. The system also needs to be customized

for it to seamlessly integrate with FPDs and to be able to add new features beyond eCRFs.

6.7.3 Opportunities

Opportunities were identified in response to the lack of flexibility in creating templates in existing tools. The Center for Expanded Data Annotation and Retrieval (CEDAR) Workbench [136] was approached as an alternative to the implementation for data entry. CEDAR provides an easy-to-use and intuitive template creation feature.

In addition, it was identified that both the CRF Wizard and the CEDAR Workbench data entry tools lacked the ability to import bulk data, as reported by a number of health facilities. To overcome the limitation on bulk data import within the CEDAR platform, the VODAN team developed a custom script for the two available solutions. The script allows for the upload of bulk comma-separated data to the CEDAR server using its API, as shown in Figures 3 and 4. While it is easy to install the ViB using Docker, it comes bundled with the DSW eCRF Wizard.



Figure 6.3: Schematic diagram illustrating Bulk data upload workflow for the CEDAR platform, depicting bulk data ingestion into the platform using a dedicated upload script

The construction of metadata templates and the filling in templates to generate high-quality metadata seems to be versatile in the CEDAR Workbench [136]. The Representational State Transfer (REST) based environment provides the CEDAR web-based tool to ease communication with other tools. Multiple file formats are available as JSON, JSON-LD, or RDF to produce machine-readable and actionable metadata and data.

Given the project’s interest in going beyond eCRF, there was a need for the localized deployment of the CEDAR Workbench, which is currently a cloud service. This is essential in the context of low-resource settings and the need for federated data stores deployed on-premises. The team is currently working on a locally deployable editor for CEDAR.

Further opportunities were identified in the possibility of using and visualizing data at the point of care through data dashboards. An aggregate dashboard based on data visualization of the locally deployed data was also conceivable, strengthening the possibility of monitoring the data for early health risk detection.

6.7.4 Threats

Threats identified pertained to the lack of ability to integrate the requirements of the ministries of health to deploy data in HMISs. The lack of an output component to the required systems would mean that data clerks would have to input the data twice, which would undermine the

Conclusion

objective intended: to edit data once for interoperable and reuse across different needs. The lack of integration with other systems was regarded as a serious threat that undermined the purpose of the innovation. In addition, training and capacity-building was identified as critical to support the system with knowledgeable data managers.

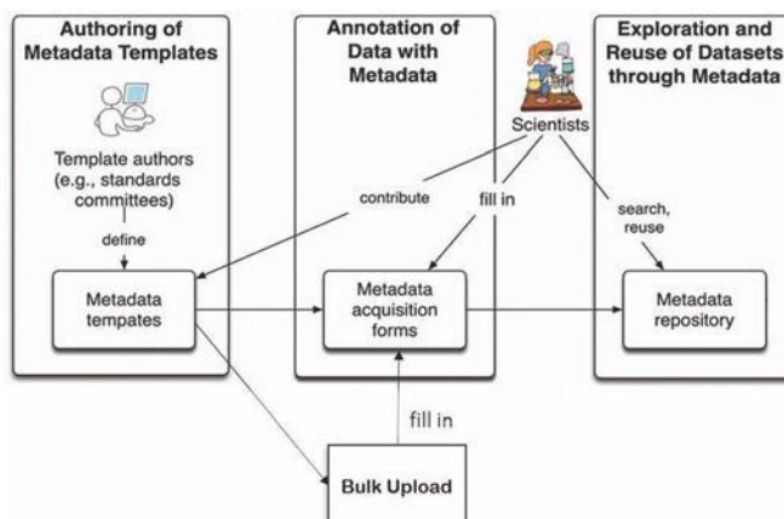


Figure 6.4: Schematic rendering of the CEDAR ecosystem supporting metadata management showing interaction with the bulk upload mechanisms

Source: [137]

6.8 Conclusion

During the ongoing COVID-19 pandemic, there remains a significant scarcity of publicly accessible data across Africa. It is challenging to ascertain whether sufficient data exists or if existing data is inaccessible due to limitations in availability and sharing mechanisms. The establishment of FAIR Data Points (FDPs), designed to support diverse data types, presents a viable solution to the challenges of making data Findable, Accessible, Interoperable, and Reusable (FAIR). This may enhance Africa's role in the fight against the coronavirus pandemic; ensure data ownership (i.e., that African data resides in Africa and is not removed to storage facilities elsewhere); and enable Africa to be a huge resource of verified data and strengthen data-informed health systems for Africa and the world.

This study explored the deployment of the VODAN in a Box (ViB) tool to enhance the interoperability of COVID-19 clinical patient data, ensuring that the data remains locally stored and accessible as machine-readable semantic metadata. The proof of concept generated a first query, which was implemented across two continents. These results are promising and highlight the potential for further development and refinement of the ViB tool. However, a SWOT analysis revealed several limitations associated with the ViB tool that require attention. The most significant limitation identified was the lack of flexibility in semantic data production. This could be solved by localizing existing platforms, such as CEDAR, to locally deployable versions.

Additionally, the study highlighted the necessity of enhancing the capacity of data stewards through targeted training in semantic data science to support the effective development and deployment of such tools.

6.9 Discussion

The task at hand was to FAIRify the COVID-19 data (i.e., make it Findable, Accessible, Interoperable, and Reusable), with the long-term vision of also making it federated and artificial intelligence (AI)-ready, in the sense that machine learning and AI approaches can discover meaningful patterns in data from future epidemic outbreaks [122].

The ViB-FDP was designed to host WHO electronic Case Report Forms (eCRFs) for patients, enabling data querying while ensuring the data remains localized within the jurisdiction of the hospital or clinic where it is generated. Data stewards identified a critical limitation in the WHO eCRF, noting its restricted capacity to handle diverse data types, such as FAIR metadata for both research and health data. The limited scope of the WHO eCRF within the DSW Wizard did not allow for the production of this data in a machine-readable format. Other data types collected within the different hospitals and clinics could not be queried despite the availability of the FDP. This limitation posed a significant challenge to the integration of patient and research data. It was concluded that the integration of the data on COVID-19 requires facilitation from semantic platforms, such as CEDAR, which supports the creation of customized metadata templates, with convergence developed to facilitate queries.

There also needs to be a change for the quick adoption of the system by health facility management procedures and HMIS requirements. Technical support should also be locally available for sustainability. Proper documentation and the availability of training and administration manuals are essential for the technical team, as well as end users. Governance documents also help in setting the rules of engagement within the context of the health system. As the end goal of a digital health system is to be able to use data for decision-making, there have to be features that facilitate data utilization, such as analytics tools, dashboards, and other reporting tools and formats.

Pandemics necessitate rapid responses in both health services and data management to support effective decision-making and strategic planning. While the WHO eCRF serves as an effective tool for collecting COVID-19 data, it lacks the additional HMIS elements required by routine health systems. This emphasises the need for a tool capable of integrating additional HMIS elements while adhering to FAIR principles. This would include new tools for creating new and adapting existing templates to accommodate the information needs of health systems. These tools need to work well with existing tools without creating much pressure on health workers and avoiding duplication of effort. Hence, other data production tools should be explored, and new integration tools created beyond the ViB, which is dependent on the WHO's COVID-19 eCRF. This requires the enhancement, customization, and localization of existing and new tools to accommodate demand. It was concluded that the DSW toolset was not the best match with the requirements and specifications, and alternative tools, such as the CEDAR Workbench, should be explored in further research.

Chapter 7

Privacy-Preserving Health Data Analytics in Uganda: A Comparative Study of DHIS2 and FAIR VODAN-Africa Architecture

This chapter is partially based on the following publication:

Basajja, M. and Paul-Chima, U. O. (2025) ‘Comparative Effectiveness of DHIS2 and FAIR Data Approaches for Privacy-Preserving Health Data Analytics in Uganda: A Systematic Review’, *ClinicoEconomics and Outcomes Research*, 17, pp. 849–864. doi: 10.2147/CEOR.S560265.

Abstract

The increasing use of digital technologies in health systems offers immense potential to improve the reuse of clinical and health data. In Uganda, the implementation of DHIS2 (District Health Information Software 2) and the adherence to FAIR(Findable, Accessible, Interoperable and Reusable) data principles represent significant steps toward improving health analytics. This chapter examines the comparative strengths and limitations of DHIS2 and FAIR data in ensuring data privacy while enabling accessibility and reusability for effective health analytics. Through a detailed critical analysis of their functionalities, data governance frameworks, and application contexts, the study identifies best practices and challenges to ensure privacy-preserving analytics. The chapter elaborates on the integration of VODAN-Africa architecture to enhance the findability, accessibility, integration, and reuse of clinical and research data in various healthcare facilities in Uganda.

Keywords: metadata templates, community standards, digital health data, Outpatient Register, FAIR principles, FAIR data, data life cycle, FAIR data management, data retrieval, data workflows, semantic web standards, ontology, Bio Portal

The source code for the scripts in this chapter can be found here: <https://github.com/mariambasajja/CH7-THESIS-CODE>

7.1 Introduction

The effective use of information at various levels within the health system is fundamental for the planning and delivery of health services. This includes its role in influencing policy decisions, programming actions, and ensuring evidence-based practices [138]. The District Health Information System version 2 (DHIS2), a flexible open source electronic information system, has been adopted in more than 60 countries, including Uganda, to manage and visualize routine health data [139]. Bhattacharya et al. demonstrated the feasibility of using DHIS2 for the local, real-time monitoring of maternal and newborn health services in low-resource settings. The Health Management Information System print-based forms and registers are compiled and submitted to the DHIS2 system for analysis. However, after submission, data analyses are not easily accessible at the point of care due to several factors. The DHIS2 system is often managed at a central or regional level, away from local health facilities, creating a physical disconnect. Many point-of-care locations, especially in rural or resource-constrained areas, may lack reliable Internet access to retrieve analyzed data from the DHIS2 system. The implementation and effective use of DHIS2 remain hindered by significant challenges, including limited interoperability, concerns about data privacy, and a lack of seamless integration with other systems. These issues underscore the need for innovative approaches, such as incorporating FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, to enhance data management and ensure privacy-preserving health analytics.

Using the VODAN-Africa architecture, the data was transformed into machine-actionable data that conforms to the FAIR principles. This supports machine-actionable data and increases health data interoperability, thus enabling efficient incorporation in other systems, analysis workflows, and future applications, including third-party applications that interface with DHIS2.

The primary challenge lies in the current state of health information management in Uganda and many African nations. The Health Management Information System (HMIS) exists primarily as paper forms compiled in registers [144], with some facilities implementing basic digital systems. These pre-printed paper formats, provided by the Ministry of Health to all health facilities, face multiple challenges: poor data quality, late submissions, incomplete capture of data from private health providers, inadequate data segregation, limited political support, and insufficient funding [141]. Additionally, limited access to computers and the reliable Internet impedes the full implementation of digital systems in areas limited by resources. The current health information system workflow consists of the steps, shown in Figure 7.1.

A Health Information System (HIS) follows a hierarchical data flow starting from primary data capture at healthcare facilities and progressing toward consolidated national-level reporting. At the initial stage, data is recorded manually using pre-printed paper forms at the point of care. This represents the foundational level of the system, where raw patient- or service-level information is collected directly within healthcare facilities. Following collection, the data is processed through two parallel pathways: data summarization and data aggregation. In the summarization pathway, raw entries are manually reviewed and converted into simplified indicators such as total counts of patients, cases, or treatments. In parallel, the aggregation pathway involves organizing individual records into structured datasets based on relevant dimensions such as time period, disease category, or healthcare facility. The outputs from both pathways are subsequently integrated into the District Health Information Software 2 (DHIS2) platform, which serves as the central digital repository for health information reporting. Within DHIS2, the data is digitized, standardized, and stored in a structured format, enabling consistent re-

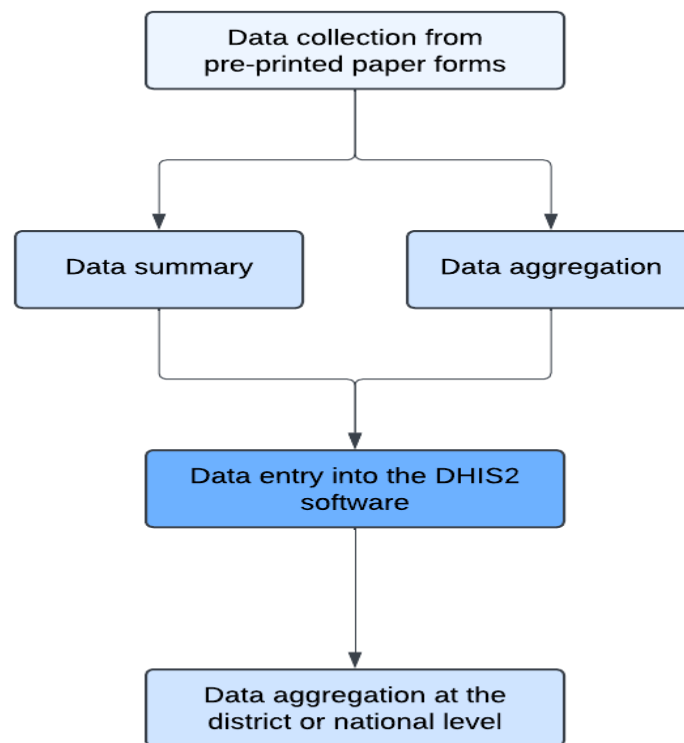


Figure 7.1: Workflow diagram illustrating the Health Information System data management process

porting across the health system. Finally, the integrated data is further aggregated at higher administrative levels, including district and national tiers. This hierarchical consolidation supports health authorities and policymakers in analyzing trends, monitoring disease patterns, and making informed, evidence-based decisions for healthcare planning and intervention.

In a review of the utilization of DHIS2 data for decision making in Ghana, Odei-Lartey et al. [143] identified critical barriers including the absence of guide manuals, infrequent training, and infrastructure limitations. Despite these challenges, their study demonstrated instances where DHIS2 data contributed to action-oriented decisions, emphasizing the system’s potential when properly supported and implemented. The terminology used by Uganda’s Health Management Information System(HMIS) registers, organized as tables with column headings, represents progress toward interoperability. However, the lack of standardized representation in semantic web formats such as SKOS, OWL, and RDF / XML limits the reusability of data and the interoperability with external datasets. While existing terms and codes facilitate local data usage, integration with broader semantic web standards requires not only adopting these formats, but also dedicated resources to create and maintain the actual ontologies and vocabularies within them.

A critical consideration in health data management is the preservation of privacy and access under well-defined terms whenever needed. Health facilities routinely generate large amounts of clinical data that have potential uses beyond patient diagnosis [147], including research, plan-

ning, decision-making, and improvement in service delivery. However, balancing data utility with patient privacy remains a pressing challenge. Although the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles promote openness and reuse, their implementation must address potential risks to patient confidentiality.

Issues of handling, sharing, and retrieving patient information in healthcare facilities are a challenge, especially in developing countries like Uganda [276]. Technology has advanced massively in the field of health, with digital health technologies promising to improve patient experiences, better resource use, and more effective policies for better health. However, these advancements come with a significant challenge [284]. The basic challenge of creating platforms for handling such big data results from the required balance between privacy protection for patients and making the related data easily accessible and analyzable [271]. This balance becomes even more important as health systems in developing nations turn to digital tools as a means to fix significant health challenges [285]. It also includes, but is not limited to, disease, surveillance, maternal and child health, as well as epidemics.

Digital health tools like DHIS2 (District Health Information Software 2) and the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles are pivotal in addressing these challenges. DHIS is an open-source health information management system implemented in Uganda for data collection, compilation, and reporting [310]. On the other hand, FAIR principles set a general practice applicable internationally as a data management practice focusing on the reusability of data, which is understood as machine readability and interoperability without compromising data confidentiality [269]. While both conceptually and operationally closely related, these approaches have different focus areas, methods of execution, and methods for maintaining user anonymity and privacy [308].

The VODAN-Africa project has taken steps to address these challenges by incorporating tools from the *Center for Expanded Data Annotation and Retrieval (CEDAR)* [136] to enhance metadata management and make health data FAIR-compliant (as explained in Chapter 6). The project has undertaken the conversion of Uganda's HMIS Registers to a standards-based knowledge representation [146], deploying a localized CEDAR-based system integrated with visualization and query tools [148].

This study aims to analyze the intersection of the FAIR principles with the implementation of DHIS2 in Uganda and broader Africa, focusing on examining how the FAIR principles can enhance the data management capabilities of DHIS2 while maintaining robust privacy preservation for health analytics and evaluating the practical challenges and opportunities to implement these principles within existing health information systems.

7.1.1 Motivation

The motivation for this research stems from the growing recognition of data as a strategic asset in healthcare delivery and public health decision-making. In Uganda, a country grappling with resource constraints and a high disease burden, robust health data systems can dramatically improve health outcomes. However, the sensitive nature of patient data necessitates a dual focus on usability and confidentiality [281]. Data breaches or mismanagement could undermine public trust, exacerbate health inequalities, and lead to serious ethical and legal ramifications [275].

Moreover, while DHIS2 has been successfully deployed in Uganda, there is limited evidence

Introduction

on how well it safeguards patient privacy in alignment with global data stewardship standards like FAIR principles [271]. By critically examining these two approaches, this research seeks to address a knowledge gap and provide actionable insights for policymakers, health administrators, and technologists in Uganda and similar countries.

7.1.2 Research Significance

This research holds substantial significance for several stakeholders:

- It highlights best practices for managing patient data securely while enabling operational efficiency in public health interventions.
- The findings will inform regulatory frameworks for data privacy, aligning with Uganda's Data Protection and Privacy Act (2019).
- By comparing DHIS2 and FAIR principles, this study identifies opportunities for technical innovation and system integration.

7.1.3 Research Objectives

The primary aim of this research is to evaluate the applicability of DHIS2 and FAIR data principles in fostering privacy-preserving health analytics for patient data in Uganda.

Specific objectives include:

- To explore and discover ways of facilitating the secure sharing of clinical health data in a Findable, Accessible, Interoperable, and Reusable way under well-defined terms using the VODAN-Africa architecture for data privacy-preserving analytics in Uganda.
- To evaluate and compare privacy-preserving health data analytics in aggregated, FAIR-ified DHIS2 datasets retrieved from the AllegroGraph database, which functions as a repository for linked data, specifically storing information in the RDF triple format [175].
- To analyze its mechanisms, such as role-based access control and encryption, for safeguarding patient data.
- To examine how these principles enhance data sharing and interoperability while maintaining privacy.
- To highlight barriers to implementing both systems effectively, including technical, legal, and organizational factors.
- To recommend strategies that leverage the strengths of both DHIS2 and FAIR principles to achieve optimal outcomes for Uganda's healthcare system.

7.1.4 Research Questions

This chapter explores the possibility of improving the analysis of health data at the point of care to propagate, integrate, and maintain it through FAIR data.

Research Questions

Main Research Question: Can implementing FAIR Data principles enhance the privacy-preserving accessibility of various types of health data and support the aggregation and analysis of these data?

Sub-Research Questions:

Question 1:

How does VODAN-Africa’s architecture ensure FAIR principles while preserving data privacy?

Question 2:

How can VODAN-Africa dashboards influence healthcare decisions across African countries based on FAIR data analytics?

7.2 Related Work

7.2.1 Overview of DHIS2

DHIS2 (District Health Information Software 2) is a widely used open-source software for health information management. It has been developed by the Health Information Systems Program (HISP) and provides a set of tools for data collection, validation, analysis, and visualization, which allows effective monitoring and reporting of the health system. In Uganda, DHIS2 plays an important role in health data compilation at the national and sub-national levels [264].

Implementation of DHIS2 in Uganda

DHIS2 implementation in Uganda started in 2009, encouraged by the government due to constraints with the paper-based health data management system. Before migrating to the DHIS2 system, the Ministry of Health (MoH) used EpiInfo, an Access-based system, and it proved to be especially challenging in the process of recording health facility data and aggregation on the national level [252]. These manual operations created problems such as time-consuming reporting, wrong information being documented, inaccuracies, and poor management of health information [291].

To overcome these challenges, the MoH, together with HISP Uganda and other partners, introduced and adopted the use of DHIS as the national electronic Health Management Information System (eHMIS), commonly known as DHIS2. Therefore, it was instrumental in realizing the reporting of health data in digital format, making it possible to enter the data on real-time basis, enhancing the data quality, and making the data easily accessible and usable for the decision-making processes. This flexibility made it possible to apply them in addressing the country’s particular health information requirements, including dimensions of the different health programs and indicators [254].

Impact & Significance

The strengthening of routine data collection and utilization in Uganda has immensely benefited from the implementation of DHIS2 [267]. In the health sector, it has enhanced the general health status and disease monitoring and thus informed decision-making [262]. The system can

Related Work

receive and analyze large amounts of data through real-time entry, which has improved the way health programs are supervised and resources are distributed and allocated.

Overall, Uganda's experience with DHIS2 underscores the platform's versatility and effectiveness in strengthening information systems, promoting transparency, and enhancing service delivery in the healthcare sector [267].

7.2.2 FAIR Data Principles

The FAIR data principles (Findability, Accessibility, Interoperability, and Reusability) provide a global framework to optimize the usability of data. In Uganda, the adoption of these principles is still growing more and more, especially within the fields of digital health and research data management [278].

The initiative of the Uganda National Council for Science and Technology (UNCST) and its international counterparts to advance the principles of FAIR data among Ugandan researchers is particularly noteworthy [263]. This initiative attempts to enhance the exchange and interoperability of data across multiple scientific domains to create a more collaborative scientific community.

Moreover, during the celebrations of Open Data Day 2020 in Uganda, stakeholders noted the aspect of FAIR data in enhancing the accountability of creating transparency and supporting monitoring, evaluation, and learning practices [268]. From the discussions, it was evident that government and district planning committees need to appreciate the importance of opening up more key data to assist in informed decision-making.

However, challenges related to the implementation of the FAIR principles in Uganda are encountered, such as low awareness, poor infrastructure support, and the absence of broad and harmonized data management policies. Solving these issues is essential to unlock the full potential of FAIR data to improve the quality of both research and public health [271].

Privacy Challenges in Uganda

Uganda's health sector encounters several significant challenges in managing patient data, particularly concerning privacy and data protection:

- **Inadequate Data Protection Laws:** Even though Uganda passed the Data Protection and Privacy Act in 2019, the enforcement currently remains limited and weak. A legal analysis conducted in 2015 pointed out that while Ugandans enjoy the protection of their right to privacy in the Constitution, several laws require one to disclose his/her health information in some or other circumstances, thereby eroding his/her privacy rights [264].
- **Resource Constraints:** The health sector experiences challenges of scarcity of healthcare human capital and restricted infrastructure to put in place elaborate data privacy systems. These constraints limit the ability to provide an effective data governance framework and to meet data privacy requirements [292].
- **Cultural and Ethical Considerations:** The urge to respect traditional cultural ways when managing personal health information, on the one hand, and the modern world's desire to keep health information private, on the other hand, complicates the process of managing

health data. These socio-cultural beliefs and practices can also determine perceptions of privacy in Ugandan society, affecting the acceptance and implementation of data protection measures. In addition, where the patient has AIDS or any other diseases that are socially unacceptable, patient information must be handled with a lot of care to avoid stigmatization of the patient [287].

Addressing these challenges requires a multifaceted approach, including strengthening legal frameworks, investing in capacity building, and fostering a cultural shift towards recognizing the importance of data privacy in healthcare.

7.3 Methodology

This chapter compares privacy-preserving health analytics between datasets generated from the existing data architecture within health facilities (DHIS2) and FAIR data generated after implementing the VODAN-Africa architecture. The chapter elaborates on the integration of VODAN-Africa and FIP tools to enhance the findability, accessibility, integration, and reuse of clinical and research data across various healthcare facilities in Uganda. The focus is on how these tools streamline the exchange of health data while ensuring patient privacy.

The development of the FAIR Implementation Plans (FIPs) was a collaborative effort, as detailed in the Paper "Basajja, M.: Comparing DHIS2 and FAIR data for privacy-preserving health analytics of patient data in Uganda" and further explained in Chapter 7 of this thesis. These FIPs were designed before deployment and specifically tailored to meet the needs of Ugandan health facilities. They provided a structured approach to digitizing and transforming routine Health Management Information System (HMIS) data collection registers into machine-actionable FAIR data using the VODAN-Africa toolset.

During deployment, clinical and research data were merged into the VODAN-Africa architecture following a set of procedures that prioritized the protection of patient privacy. This work is described in greater detail in Section 7.4 of this thesis. The methodology also details the processes for accessing, using, and analyzing the digitized data, ensuring compatibility and smooth operation across platforms, as elaborated in the data section of this chapter.

Data analysis was conducted on the digitized datasets aligned with FAIR principles. The study focused on how the collected data could be accessed, tested, and integrated for cross-continental comparisons and analytics. Specific methodologies for data collection, testing, and integration are described in the Data Section 7.4 of this chapter, where I will explicitly explain which data sets were used, how they were accessed, and the processes followed to ensure that research information flows smoothly between platforms. The results demonstrate the ability of this architecture to analyze and support decision-making across distributed datasets while maintaining privacy-preserving standards.

7.3.1 Data Privacy and Security Considerations and Identification of the Key Statistics of the Selected Datasets

The GDPR Law, introduced in 2018 [149], ensures that organizations and other bodies operating within the EU or collecting data from users within the EU must comply with certain policies detailed in this law. We are also dealing with data from Uganda; we must ensure compliance with

Methodology

local data protection and privacy laws, namely the Data Protection and Privacy Act introduced in 2019 [150, 151]. However, about this project, if we ensure compliance with GDPR along with the Ugandan Data Protection and Privacy Act, the security of health data information is enforced. The GDPR and Ugandan Data Protection and Privacy Act were highly considered and implemented during the implementation of the VODAN-Africa architecture within the different health facilities. Any sensitive data that could be traced back to a single person was excluded. The data subject decides whether the data should not be exposed. The following set of layers was identified as necessary for the VODAN-Africa community:

1. Data processing agreement among all data processing partners.
2. FAIR Equivalency tool based on the regulatory framework in the different contexts.
3. Agreement about machine-actionable templates based on the three Health Management Information System registers, outpatient, Maternal Child Care/Antenatal.
4. An agreement on the data repositories and security to keep the data secure within the residence where the data is stored.
5. An agreement on the data pipelines used for the dashboards providing real-time aggregate data analytics at the facility.
6. An agreement on the data pipelines used for the dashboards providing real-time aggregate data analytics at the community level.

7.3.2 Data Collection

For this preliminary study, four datasets were identified for analysis. These included data from the HMIS Outpatient Register 002 and the Antenatal Care Register (ANC), both of which were produced locally by VODAN-Africa and Asia health facilities. Additionally, two Point of Care datasets were included: one dataset containing data on HIV, schistosomiasis, and HPV for the period 2013-2022 (9 years) from the Ministry of Health (MOH) Uganda and three districts—Lira, Entebbe, and Hoima; and another dataset containing data on schistosomiasis cases and deaths from different districts in Uganda.

The data from the HMIS Outpatient Register 002 and ANC were originally collected on paper forms and submitted to the DHIS2 system daily, weekly, or monthly, as per the standard reporting requirements. Agreements were reached between health facilities and several key individuals involved in the initiative, such as the general coordinator, the executive director of VODAN Africa, and the national Country Coordinators for the use of these data. Specifically, the facilities provided the paper-based data to VODAN-Africa for digitization and integration into the VODAN-Africa system. For this study, anonymized data were used, with the necessary permissions obtained from the relevant Ugandan authorities.

To ensure that the data was accessible and reusable at the point of care, the VODAN-Africa architecture was employed to digitize and "FAIRify" the data. This process involved converting the paper-based registers into machine-actionable formats using locally installed CEDAR platforms within selected health facilities (as described in detail in Chapter 6 section 6.7.3).

Methodology

This digitization and transformation process was done through template creation and data production, ensuring the data adhered to the FAIR principles (Findable, Accessible, Interoperable, and Reusable).

Data were stored in the AllegroGraph triple store and queries were performed using the internal and community dashboards. These platforms allowed for easy retrieval and analysis of the data, facilitating improved access to point-of-care data. Integration of these data into the VODAN-Africa system demonstrates its potential to improve healthcare outcomes by enabling access to real-time data at the point of care.

In terms of privacy preservation, strict data protection measures were followed. The anonymized data was digitized, ensuring that personally identifiable information was not included. Additionally, data was only made accessible to authorized users within healthcare facilities, with encryption and other privacy-enhancing technologies employed to safeguard sensitive information.

Regarding data availability, these datasets remain part of the VODAN-Africa ecosystem and continue to be accessible for ongoing research and health analytics. In line with VODAN-Africa's decentralized architecture, the data are stored within individual institutions' local repositories, where they remain under the control of the respective health facilities. These datasets can be accessed and reused for future research through secure, distributed queries, provided that the necessary ethical and privacy guidelines are adhered to.

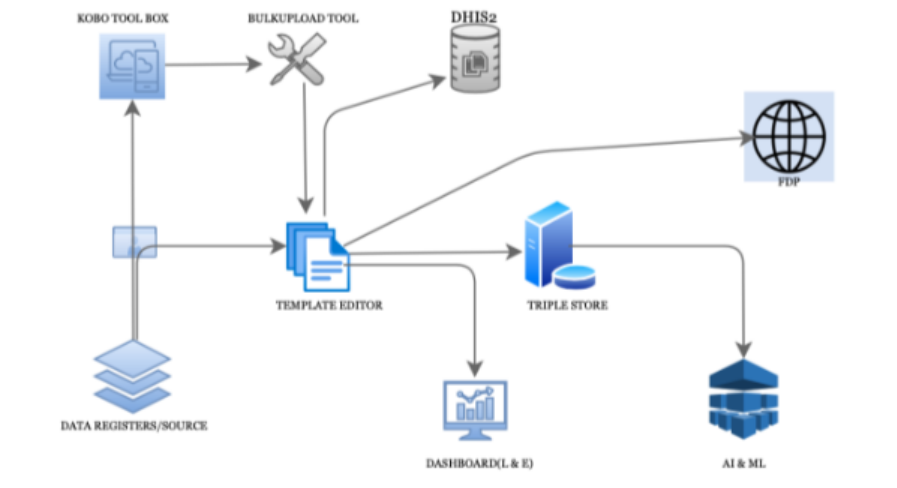


Figure 7.2: VODAN Africa internal and aggregate dashboard HIS integration scheme

Table 7.1 presents the data size produced within the health facilities, illustrating daily patient data collected from sample hospitals in Uganda.

Methodology

Table 7.1: Data size produced within the health facilities showing daily patient data from sample hospitals in Uganda

S.no.	Health facilities	Country	Daily patient data from previous survey	ANC (Monthly)	OPD (Monthly)	COVID (Monthly)
1	KIU Teaching Hospital	Uganda	130	685	1672	0
2	Ark Specialist Hospital	Uganda	± 50	1	0	0
3	Hoima referral Hospital	Uganda	± 170	62	118	0
4	Lira Referral Hospital	Uganda	1	1	0	0
5	Ayder Referral Center	Ethiopia	20	750	900	
6	Tukur Anbesa Specialized Hospital	Ethiopia	116	116	508	NA
7	Pumwani	Kenya	± 150	595	5203	6
8	Beacon of Hope	Kenya	62	54	1905	22
9	Railway clinic	Nigeria	+150	14	384	0
10	GH Lapai	Nigeria	± 130	132	1442	0
11	OOU	Nigeria	50	5	1656	15
12	FULL	Nigeria	100	23	819	NA
Total Uganda			909	1838		

7.3.3 Application of VODAN-Africa and Asia Health Management Information System Architecture

VODAN Africa developed an architecture to record clinical health data, making it accessible both to humans and machines in a distributed system that is locally governed and linked. This architecture supports data analytics directly at individual health facilities by enabling data visits for analytics. It includes the CEDAR system as a one-time data input editor, outpatient instances, ANC instances, and bulk upload of datasets to the CEDAR platform. The CEDAR platform, integrated within the VODAN-Africa architecture, facilitates the organization and storage of health data. The platform works with RDF and JSON-LD structured data. From these formats, template IDs (for Outpatient, ANC, and point of care datasets) and API keys are retrieved to configure the dashboard. This allows for analytics at the point of care using the internal health facility dashboard and the VODAN-Africa aggregate dashboard for real-time statistics. Data are stored in the AllegroGraph triple store, a specialized database designed to store and retrieve data as RDF triples through the running of dynamic SPARQL queries. In addition, it supports the import and export of data in the health facility HIS, such as DHIS2, in the form of reports.

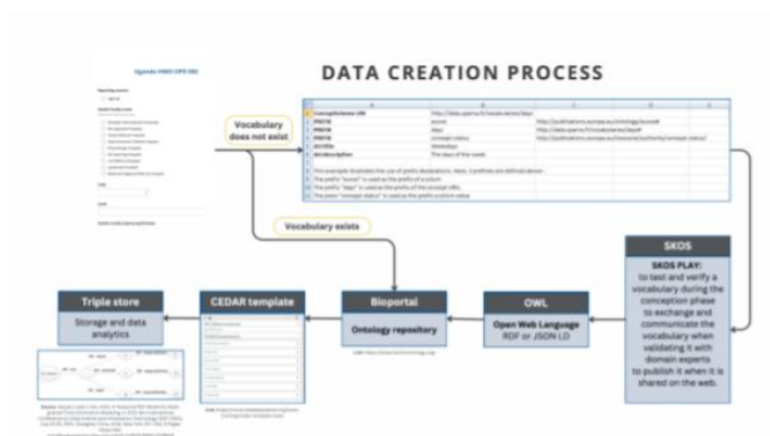


Figure 7.3: Data creation process using the VODAN Africa architecture

Why CEDAR?

The CEDAR is a flexible and agile workbench integrated with the bio portal to produce machine-readable data templates [290]. It was established in 2014 to create a computational ecosystem for the development, evaluation, use, and refinement of biomedical metadata. It was adapted to local data production and for the data capture of repositories controlled by the facility that produces the data by the VODAN-Africa technical team. Its approach is based on the use of metadata templates, which include controlled terms and semantic properties for specific data elements. The data were transformed into Resource Description Framework and JSON-LD format based on BioPortal ontologies (see Table 7.2). It also supports sharing and managing resources both locally and on the web. The CEDAR platform provides an API to support export/import and analysis of metadata, compatible with other tools, i.e., can be used for editing, creating, sharing, and accessing data.

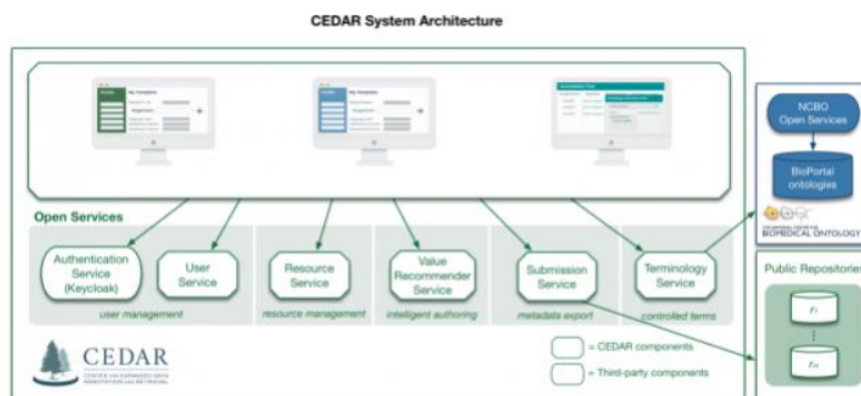


Figure 7.4: Cedar system architecture ¹

Allegro Graph Triple Store

FAIRification by design was successfully realized within the health facilities by digitizing and creating metadata templates based on the Health Management Information System paper registers and the DHIS2 system for aggregated data entry in residence. The data was generated and produced using the CEDAR system integrated into the locally deployed VODAN-Africa architecture within the different health facilities of Uganda. The CEDAR platform supports the creation of customized templates. The templates were then populated with patient data and automatically assigned unique identifiers(F1). The metadata templates are stored in JSON-LD format in MongoDB and the form of RDF triples in the Allegro graph triple store, respectively [152]. The populated templates can also be made accessible via the web under restricted measures, thereby fulfilling the Findable and Accessible aspect of the FAIR principles.

7.3.4 Standardization of the VODAN-Africa Controlled Vocabulary Terms and Semantic Properties

The VODAN-Africa controlled vocabularies and semantic properties were created using the Simple Knowledge Organization System (SKOS) framework, which is a standard for representing knowledge organization systems. These vocabularies are made available in multiple formats: CSV, RDF, and SKOS, to cater to different usage needs and tools.

The Cedar metadata templates (here) were used and examined iteratively to conform to the requirements of Resource Description Framework (RDF) SKOS conversion. Data were produced using the templates to ensure that all appropriate fields were adequately captured, as well as to ensure that the structure inherent in the templates allowed portability.

The conversion of tabular data for the controlled vocabulary terms and properties into SKOS RDF files was initially performed using the SKOS Play Convert tool [140]. However, it is important to note that fully automating the conversion process was not feasible due to the complexity of certain relationships between terms. Before conversion, several steps were taken to model the data to ensure their alignment with the controlled vocabulary. These steps involved assigning the controlled terms and semantic properties on the bioportal, such as [mapping

¹Source: <https://more.metadatascenter.org/tools-training/cedar-api>

Methodology

disease codes to standardized ICD-10 terminology’], to ensure consistency and coherence in the semantic structure, accounting for both quantitative and qualitative associations. This process accounted for both quantitative and qualitative associations, and the conversion scripts used for term mapping can be found on (here). These scripts facilitated the automation of the mapping process for commonly used terms, though some terms required manual intervention to ensure accuracy.

Despite these efforts, the SKOS Play Convert tool was able to automate only the initial stage of the conversion. Some operational associations between terms—particularly those involving more complex or nuanced relationships—were not accurately interpreted by the tool’s algorithms and required manual intervention. These issues were addressed through a process of semantic control, where expert review and refinement were necessary to ensure that the relationships between terms were correctly represented.

This limitation highlights an area for potential development in future iterations of the VODAN-Africa system. While some progress has been made in automating data conversion and semantic handling, further advancements are needed to enhance the accuracy and efficiency of the system, particularly in handling complex semantic relationships. Furthermore, flexibility became an issue due to the unfocused error handling of the tool: when input data was supplied in an incorrect format for conversion, the conversion stopped at some point, and incorrectly formatted data could lead to conversion failures. To address these challenges and improve the templates and workflows, the following enhancements were proposed:

- **Enhanced Template Design:** Validation rules and dropdown menus were incorporated into the templates to minimize user errors during data entry.
- **Automated Validation Scripts:** Conversion scripts (here) were introduced to validate data completeness and consistency before submission to the SKOS Play Convert tool.
- **Hybrid Conversion Approach:** A combination of automated conversion and manual review was adopted to mitigate the limitations of the SKOS Play Convert tool and ensure the semantic relationships and properties were accurately represented.
- **Continuous Feedback Loop:** User and stakeholders feedback was collected regularly to refine templates and address emerging requirements.

The current limitations primarily involve ensuring that non-technical users can effectively use the templates without introducing errors that complicate the conversion process. These enhancements require correcting the existing drawbacks that are related to the stability, simplicity, and applicability of the templates, providing for their wider utilization and constant repeated usage.

In addition, different versions of the VODAN-Africa vocabularies are managed using the GitHub version control system. This ensures that the latest versions of the vocabularies are available for download and use in various formats, including CSV and RDF, providing flexibility for different applications and integration scenarios.

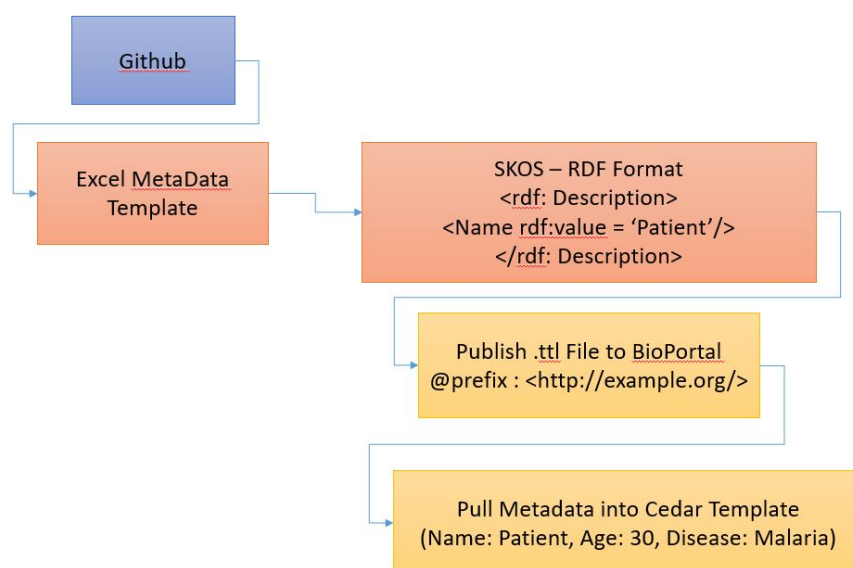


Figure 7.5: Workflow for the standardization of VODAN Africa controlled vocabularies and semantic metadata using RDF and BioPortal

Semantic interoperability is increased through importing and mapping concepts from semantic resources such as SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms), LOINC (Logical Observation Identifier Names and Codes) and ICD (International Classification of Diseases), and reusing definitions, for example aligning patient demographics and health indicators from these templates with VODAN-Africa data models. Within the MVP of VODAN-Africa, the reuse of suitable existing vocabulary terms and properties was implemented to avoid duplication and increase interoperability within machine-actionable semantic data templates in RDF and JSON-LD linked data formats. The templates contain fields for which each is assigned a controlled vocabulary term or semantic property stored in the BioPortal system. The controlled vocabulary terms and semantic properties created are used to provide values for the metadata attributes in these templates, representing community-endorsed standards and controlled terms from the VODAN-Africa ontologies hosted in Bio Portal (see Table 7.2). For instance, the term 'Hypertension' from SNOMED CT was used to harmonize disease classification across datasets, ensuring semantic alignment with the VODAN-Africa ontology.

Methodology

S.N	Description	Link
1	VODANA's General controlled vocabulary ontology and semantic properties that are used within Outpatient Department Registration and Antenatal Care Templates within the facilities.	https://bioportal.bioontology.org/ontologies/VODANA-GENERAL
2	VODANADISEASES is an ontology of the diseases whose data is collected within the VODANA facilities, especially.	https://bioportal.vodana.health/ontologies/VODANADISEASES
3	MCHVODANATERMS is an Ontology containing the controlled terms and semantic properties used in the Maternal Child Care generic template used within different facilities in VODANA	https://bioportal.vodana.health/ontologies/MCHVODANATERMS
4	VODANAMFLCODE is the ontology containing the health facility names along with the codes used within the VODANA community.	https://bioportal.bioontology.org/ontologies/VODANAMFLCODE

Methodology

S/N	Description	Link
5	VODANAUGANDA is an Ontology containing the controlled terms and semantic properties used in the OPD registers from different countries used within different facilities in VODANA	https://bioportal.bioontology.org/ontologies/VODANAUGANDA

Table 7.2: Overview of the Ontologies implemented within VODAN-Africa framework to facilitate standardized data representation and interoperability.

The ontologies in Table 7.2 were selected because they provide controlled vocabularies and semantic properties essential to ensure data interoperability and adherence to FAIR principles in the following contexts:

- **VODANA-GENERAL:** This controlled vocabulary offers a foundational framework and common properties used across various VODANA templates, including the Outpatient Department Registration and Antenatal Care templates.
- **VODANADISEASES:** Focuses specifically on diseases, enabling uniform data collection and reporting related to disease data within VODANA facilities, especially in Uganda.
- **MCHVODANATERMS:** Provides terms and properties tailored for Maternal Child Health care templates, ensuring consistency across facilities.
- **VODANAMFLCODE:** This ontology links health facility names with their standardized codes, promoting consistency when referring to facilities within the VODANA community.
- **COVIDTERMS:** A specialized ontology developed for managing COVID-19-related data, which is critical for pandemic response and monitoring.

Interrelation of Ontologies

These ontologies are interconnected, with the VODANA-GENERAL ontology serving as the overarching framework that supports and standardizes the specific vocabularies within the other ontologies. For example:

- **VODANADISEASES** and **COVIDTERMS** inherit the properties of **VODANA-GENERAL**, but apply them to disease-specific contexts.
- **MCHVODANA-TERMS** integrates terms from **VODANA-GENERAL** while focusing on maternal and child health.

Methodology

- VODANAMFLCODE ensures that the health facility codes used in the other ontologies are consistent and interoperable.

Five main source ontologies were used to constrain metadata fields in the CEDAR template, ensuring that users can enter attribute values correctly. Constraining metadata fields eliminates the possibility of entering syntactically different values or responses, such as misspellings or inconsistent formats. The four VODANA FAIR Vocabularies incorporated into the templates for HMIS OPD 002 REGISTER include VODANA-GENERAL, VODANAUGANDA, VODANA-COVID, VODANADiseases, and VODANAMFLCODE (VODANA Health Facility List).

For example, when entering patient attributes like "Health Facility," users must select from a predefined list provided by VODANAMFLCODE, ensuring uniformity (e.g., "Mulago National Referral Hospital" instead of variations like "Mulago Hosp." or "Mulago NRH"). This approach standardizes entries and supports interoperability in the datasets.

Interoperability of the VODAN-Africa Minimal Viable Product with HMIS systems

Within the VODAN-Africa minimal viable product, syntactic interoperability with point-to-point integration is increased through the support to compute and send aggregated monthly reports to hospital HMIS systems, as illustrated in Figure 7.6. The VODAN-Africa MVP is in compliance with the interoperability component of FAIR in that it guarantees that it integrates easily with other Health Management Information Systems (HMIS), including the DHIS2 system. It is done with the help of standardized data element report skeleton templates which use shared controlled vocabularies and semantic properties. These templates promote the flow of data between systems by responding to the concepts of interoperability described in the FAIR framework. The term interoperability here is understood as the capacity or capability of various systems to communicate with each other in a manner that is significant, advantageous, and without the necessity of putting much effort into it, in terms of manual labor. The MVP of VODAN-Africa has ensured that additional HMIS systems can easily understand and interpreted data through standardized data formats, such as JSON, which facilitates better data circulation and the usefulness of access to information by health systems. The aggregated indicators are then posted to DHIS2, which provides a web API to accept the data values. A separate enterprise service bus for authentication, management of interoperability services, logging, and mediation is envisioned as a microservice to ensure security and add flexibility during interoperability.

Challenges Encountered During VODAN-Africa MVP Deployment and Data Production.

The CEDAR platform was the main data entry tool used within the different VODAN-Africa health facilities. However, data acquisition posed significant challenges and massive data collection is vital for in-depth analysis. During the deployment and data production within the VODAN-Africa health facilities, several issues were identified and addressed through one-on-one interviews with data management personnel in Uganda, Kenya, Ethiopia and Nigeria. The CEDAR platform faced storage shortages and performance issues, leading to the use of alternative data collection tools. The locally installed Kobo toolkit was employed in some facilities, especially in Nigeria.

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

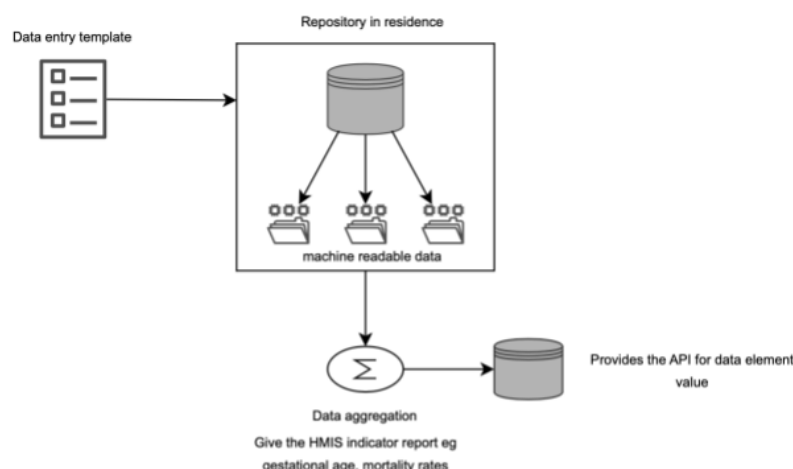


Figure 7.6: Interoperability workflow between VODAN-Africa and HMIS systems for machine-readable data aggregation and indicator reporting

Converting paper-format data into FAIR data required significant manpower. The health facilities in Uganda continued to report to the Ministry of Health (MOH) using more than 20 different paper registers, requiring the assignment of additional personnel to enter local health data. To FAIRify data collected over the past nine months, the bulk upload tool was adopted. Challenges such as internet connectivity and electricity scarcity in some regions of health facilities were mitigated by providing resources to purchase mobile data bundles for data production. Technical issues, such as resetting the admin password, were promptly resolved by the technical team.

To address the inadequate skillset between data management personnel and data stewards, comprehensive capacity-building initiatives were implemented. These included targeted training sessions and the distribution of shared materials aimed at improving competencies in data management and stewardship. Participants were trained on using the bulk upload tool, adhering to FAIR principles, and troubleshooting common issues. These initiatives not only enhanced the technical proficiency of the personnel, but also ensured the sustainability of FAIR data practices in health facilities.

7.4 Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

The usability of the dashboard is divided into two components:

- The aggregate dashboard of the VODAN-Africa community.
- The local dashboard (within the health facility).

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

The aggregate dashboard provides a summary of participating countries and the total number of records or instances of the Outpatient 002, COVID-19, and Antenatal Care registers. On the other hand, health facilities can visualize their data through the local dashboard. For example, a health facility can view regular statistics such as the number of patients registered per day, trends in antenatal visits over a specific period, or the distribution of COVID-19 cases within their facility. These insights help health facilities monitor their operations and make data-driven decisions.

The ownership of the data is determined by where the data or metadata is stored or shared. The domain of the local dashboard is attached to the local host. The database is set up through verification, the API key address is set up, admin, standard, user, and password are set up for security, and the update configuration is completed. The local dashboard automatically fetches data from the Cedar platform. The high frequency is attributed to the collection of real-time data that will be reliable in obtaining realistic results. A provenance template on the health facility metadata (template 0) with fields: Country, region, latitude, longitude, address, facility type: government or private, email address, phone number, and facility size is populated and submitted. The VODAN-Africa community dashboard manually updates when the local dashboard is updated and hosted on a public domain ²

No	Statistics	Aggregate Frequency	Description
1	Patient visits (real-time)	Daily and Monthly	Aggregate of the number of patient visits
2	Patient gender percentage (real-time)	Monthly	Male patient and Female patient percentage
3	Frequency of Diagnosed Diseases (real-time)	Daily and Monthly	Aggregate of the selected diseases grouped by diseases
4	COVID-19 patient visits (real-time)	Daily and Monthly	Aggregate of the number of patient visits for COVID-19
5	Patient tested for COVID-19 (real-time)	Daily and Monthly	Aggregate of the number of patients tested for COVID-19
7	Patient tested positive on COVID-19 (real-time)	Daily and Monthly	Aggregate of the number of patients tested COVID-19 positive
8	Patient death from COVID-19 (real-time)	Daily and Monthly	Aggregate of the number of patients died because of COVID-19
9	ANC Patient Age (real-time)	Monthly	Aggregate of the number of patients categorized by Age (by age in the interval)
10	ANC Patient visits (real-time)	Daily and Monthly	Aggregate of the number of patient visits from ANC form
11	ANC Syphilis test result (real-time)	Daily and Monthly	Aggregate of the number of patient tested for Syphilis categorized by result from ANC form
12	ANC HIV test result (real-time)	Daily and Monthly	Aggregate of the number of patient tested for HIV categorized by result from ANC form
13	Gestational Age (real-time)	Monthly	Aggregate of the number of patient categorized by gestational age in weeks (by weeks in interval) from ANC form

Figure 7.7: Standardized fields selected for the key statistics for the VODAN-Africa local and community dashboards

7.4.1 Key statistics Data Sources for the VODAN-Africa Local and Community Dashboards

²Source: <https://dashboard.vodana.health/>

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

Template name	Attributes and terms selected
Template 0	Country, region, latitude, longitude, address, status: government or private, facility type: email address, phone number, number of patients, facility size :S, M, L
Outpatient register	Health facility name, attendance frequency, date of visit, diagnosis frequency (daily and monthly), sex (monthly gender composition)
Antenatal childcare register	Serial number, age (age groups: 18-25, 26-35, 36+), attendance frequency, syphilis & Hep Test Results with syphilis, gestational age (1-10, 11-20, 21-30, 31+ weeks)
COVID-19	Health facility name, count of number of visits, date of visit, patients tested for COVID (daily and monthly), mortality rate/death rates (daily and monthly)

Table 7.3: Attributes and terms selected for different templates

Use of FAIR data at the point of care Health Data Analysis was done with the data retrieved from the health facility AllegroGraph database using dynamic SPARQL queries. This included analysing the outcomes of grouping the Ugandan Outpatient Registry Data by aggregation, patient demographic analysis of the Ugandan Outpatient Registry Data, examining the temporal and geographical trends in Uganda, as well as comparing visualisations using FAIR Data.

7.4.2 Data Analytics on the AllegroGraph Repositories

Grouping the Ugandan Outpatient Registry Data by Aggregation

The SPARQL query below is structured to retrieve the number of diagnoses that are recorded under each health facility from the data in the OPDUganda repository. By designing the query to count occurrences of each diagnosis per facility, a clear summary of how often different diagnoses are documented across various health facilities is obtained. The results are ordered in descending order by the count of each diagnosis to highlight the most frequently occurring diagnoses first.

SPARQL Query 1

```
SELECT
?healthFacilityName ?diagnosis (COUNT(?diagnosis) AS ?diagnosisCount)
WHERE {
  ?s <http://dashboard.vodana.health/Diagnosis> ?diagnosis.
  ?s <http://dashboard.vodana.health/HealthFacilityName> ?healthFacilityName.
}
GROUP BY ?healthFacilityName ?diagnosis
ORDER BY DESC(?diagnosisCount)
```

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

healthFacilityName	diagnosis	diagnosisCount
"KIU teaching Hospital"	"UTI"	"35"
"KIU teaching Hospital"	"Spondylolysis lumbar region"	"25"
"KIU teaching Hospital"	"Peptic ulcer disease"	"23"
"KIU teaching Hospital"	"Dental carries"	"22"
"KIU teaching Hospital"	"Peripheral neuropathy"	"21"
"KIU teaching Hospital"	"Gastritis"	"19"
"KIU teaching Hospital"	"Fibromyalgia"	"18"
"KIU teaching Hospital"	"Irreversible pulpitis"	"18"
"KIU teaching Hospital"	"Spondylolysis"	"17"
"KIU teaching Hospital"	"Malaria"	"16"
"KIU teaching Hospital"	"Dental abscess"	"15"
"KIU teaching Hospital"	"Dog bite"	"14"
"KIU teaching Hospital"	"Acute nasopharyngitis"	"14"
"Hoima Referral Hospital"	"MALARIA"	"14"

Figure 7.8: Output produced from running the SPARQL query to retrieve the number of patients per diagnosis in each health facility from the OPDUganda Repository

The results from running this query can be analyzed to understand how the occurrence of various diagnoses is distributed across the various Ugandan health facilities in the dataset. This information can be used by stakeholders in healthcare management to efficiently allocate resources and implement targeted interventions to hospitals that have a significantly large number of cases compared to the resources that are available for managing the diseases. Consequently, this would potentially improve the effectiveness of disease management in these institutions.

Patient Demographic Analysis in the Outpatient Registry Uganda Data

Patient Demographics Aggregation

The results from running the SPARQL query below can be used to analyze the information on the demographics of patients with certain diagnoses. The query retrieves the number of diagnoses from the OPDUganda repository. The diagnoses are then grouped by age groups of under 18 years, 18 to 25 years, 26 to 35 years, 36 to 45 years, 46 to 55 years, and over 55 years. The diagnoses are also grouped by gender. Finally, the query orders the results by the number of diagnoses in each group in descending order. The pattern of the query is that the subject is a diagnosis that was made, and the query retrieves the age and gender attributes of the patient that are associated with each diagnosis. The findings from analyzing the data on patient demographics can then be used in strategic planning for implementing interventions and allocating resources to effectively manage the spread and fatality of the diseases. By understanding which age groups and genders are most affected by particular diagnoses, healthcare providers and policymakers can make decisions that are better informed to improve the outcomes in public health in Uganda.

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

SPARQL Query 2

```
PREFIX vodan: <http://dashboard.vodana.health/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?diagnosis ?ageGroup ?gender (COUNT(?diagnosis) AS ?numDiagnoses)
WHERE {
  ?record vodan:Age ?age.
  ?record vodan:Sex ?gender .
  ?record vodan:Diagnosis ?diagnosis .
  BIND (
    IF(xsd:integer(?age) < 18, "< 18",
      IF(xsd:integer(?age) >= 18 && xsd:integer(?age) <= 25, "18-25",
        IF(xsd:integer(?age) >= 26 && xsd:integer(?age) <= 35, "26-35",
          IF(xsd:integer(?age) >= 36 && xsd:integer(?age) <= 45, "36-45",
            IF(xsd:integer(?age) >= 46 && xsd:integer(?age) <= 55, "46-55",
              IF(xsd:integer(?age) > 55, "> 55", "Unknown")))))
    AS ?ageGroup)
}
GROUP BY ?diagnosis ?ageGroup ?gender
ORDER BY DESC(?numDiagnoses)
```

diagnosis	ageGroup	gender	numDiagnoses
"UTI"	"18-25"	"Female"	"9"
"Gastritis"	"< 18"	"Female"	"8"
"Peripheral neuropathy"	"> 55"	"Female"	"7"
"PUD"	"18-25"	"Female"	"7"
"Spondylolysis lumbar region"	"> 55"	"Female"	"7"
"Late syphilis"	"> 55"	"Female"	"6"
"UTI"	"46-55"	"Female"	"6"
"Dental carries"	"18-25"	"Female"	"6"
"Hypertension"	"> 55"	"Female"	"6"
"UTI"	"> 55"	"Male"	"6"
"Common cold"	"< 18"	"Female"	"6"
"Malaria"	"< 18"	"Female"	"5"
"Acute nasopharyngitis"	"< 18"	"Female"	"5"
"STI"	"18-25"	"Female"	"5"

Figure 7.9: Output produced from running the SPARQL query to retrieve and count the number of specific diagnoses grouped by age groups and gender from the OPDUganda Repository

Patient Demographic Analysis in the Outpatient Registry Uganda Data

Various visualizations were created to analyze the patterns in the OPD Uganda patient data. The age distribution of the top 10 highly diagnosed diseases and infections was plotted using bar graphs. Age groups from the OPD Uganda dataset were in intervals of 4 years between

Data Analytics Using the Local Health Facility Dashboard and VODAN-Africa Aggregate Dashboard

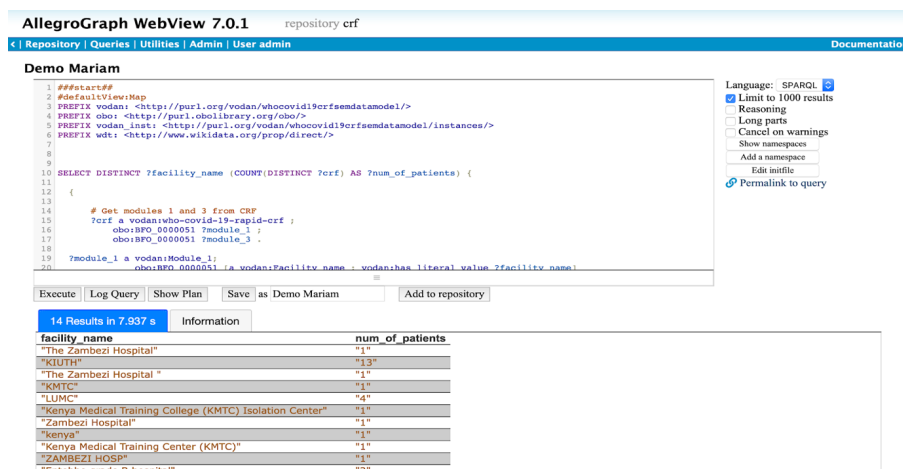


Figure 7.10: Output produced from running the SPARQL query to retrieve and count the number of patients from the OPD Uganda Repository in each hospital

0 and 80+ years. These visualizations can be used to diagnose diseases associated with age, to determine the appropriateness of health services, and to distribute these services between certain ages [214]. In the same manner, bar graphs were employed to indicate the distribution of the number of the ten most frequently diagnosed diseases and infections by sex. It is useful for studying gender-specific health concerns, for developing specific health initiatives, and for assessing potential discrepancies in disease rates or in access to medical services between male and female populations [214]. A word cloud image was generated to visually represent the diagnoses and prescriptions given. This visualization technique allows for quick identification of the most frequent diagnoses and prescriptions by displaying them in larger font sizes. It provides an intuitive and engaging way to summarize the prevalence of various health conditions in the dataset, helping to identify key areas of concern at a glance [229]. A histogram was created to represent the age distribution of the entire group of data, the mean age, and the standard deviation of the sample. This provides an overall picture of the age distribution of the population and allows the determination of age groups of the population that are diagnosed most frequently to plan the availability of medical services and healthcare for the prevailing age groups [215].

Data Acquisition

Data was retrieved from the AllegroGraph repository named OPD Uganda using SPARQL queries. The following query was run on the OPD Uganda repository to retrieve the age, sex, diagnosis, prescription, district of residence, blood pressure, and blood sugar levels from the OPD Uganda repository.

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

SPARQL Query 3

```
PREFIX opd: <http://dashboard.vodana.health/>
SELECT ?s ?age ?sex ?diagnosis ?prescription ?district
?blood_pressure ?blood_sugar
WHERE {
    ?s opd:Age ?age ;
        opd:Sex ?sex ;
        opd:Diagnosis ?diagnosis ;
        opd:Prescription ?prescription ;
        opd:District ?district ;
        opd:BloodPressureMmhg ?blood_pressure ;
        opd:BloodSugarMmL ?blood_sugar .
}
```

General Diagnoses and prescriptions from Uganda Outpatient data

Figure 7.11 illustrates the diagnoses and shows their frequency using the font size. Figure 7.12 shows the top prescriptions given using the font size.

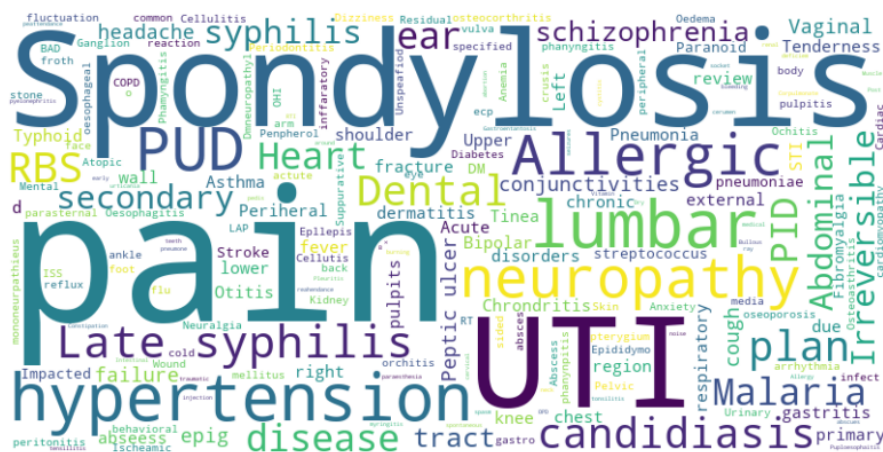


Figure 7.11: Word cloud visualization of the most frequently reported diagnoses in Uganda outpatient dataset, where word size corresponds to the relative frequency of each diagnosis

7.5 Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

In this section, we address the DHIS2’s role in transforming Uganda’s health Information systems, FAIR’s role in transforming Uganda’s Health Information systems, the strengths and limitations of DHIS2 and FAIR data principles.

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

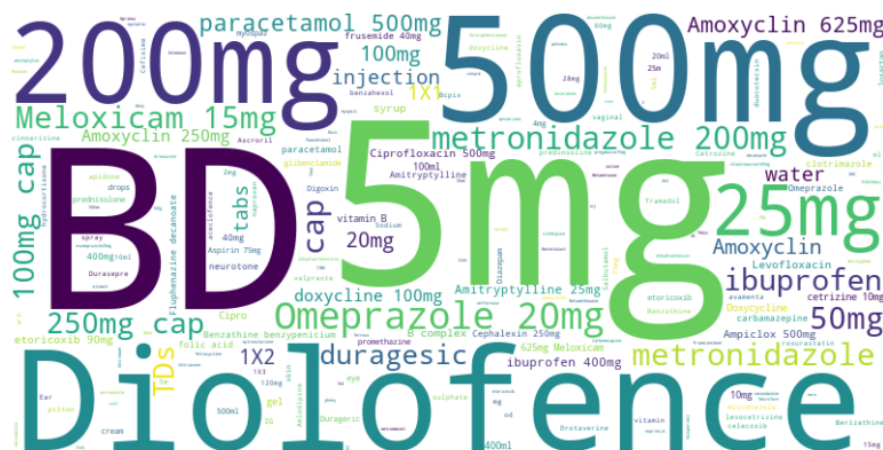


Figure 7.12: Word cloud visualization of the most frequently reported prescriptions in Uganda outpatient dataset, where word size corresponds to the relative frequency of each prescription

7.5.1 DHIS2's Role in Transforming Uganda's Health Information Systems

By implementing DHIS (District Health Information Software 2), Uganda has brought revolution in its health information management system via strong electronic support. DHIS2 has incorporated itself into healthcare practices, which is fundamental to disease immunization, surveillance, and maternal health programs within the country [283].

DHIS2 was introduced and has helped improve disease immunization, surveillance, and maternal health programs. DHIS2 has also made it possible to improve monitoring of vaccination rates, due to its real-time entry of data and reporting, which has resulted in more specific and focused immunization initiatives and used resources more efficiently. This has been exceptionally critical towards the containment of preventable diseases, particularly in the rural population.

In addition, DHIS2 has been instrumental in health surveillance programs in Uganda. Due to the union of data carried out by healthcare facilities nationwide, DHIS2 enables tracking disease outbreaks, enabling respondents in the field of public health to respond fast to emerging threats. It has also been useful for tracking infectious diseases such as malaria, cholera, and Ebola, which can be detected and contained in less time.

Through DHIS2, health care providers have enjoyed real-time access to vital information on pregnancy, childbirth, and post-delivery care in the maternal health sector. This has improved maternal services, as the required medical services are done at the expected time and pregnant women are well monitored during their pregnancy periods. The capabilities of the software to monitor the data related to antenatal visits, delivery results, and complications have helped to reduce the number of maternal and infant mortality rates.

In addition, the user-friendly nature of DHIS2 and its ability to integrate with any other health system have made its explosion to be employed by health workers of different levels. This has helped to create a strong healthcare data ecosystem in Uganda, which has ensured the sustainability of improving the delivery of data-driven health services.

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

Together, the functions of DHIS2 in Uganda's health information system serve as a good example of the benefits of digital technologies in healthcare by facilitating a change in workflow, efficient use of funds, and ultimately the health outcomes of low-resource countries.

Privacy Mechanism in DHIS2

DHIS2 incorporates several privacy mechanisms designed to protect patient data and ensure system security [264]. One of the main mechanisms is Role-based Access Control (RBAC), where users are granted access only according to their assigned roles in the healthcare organization. This control mechanism assists in discouraging unauthorized data access in the system. However, there is a problem with using RBAC, more specifically in the case of access control, where the controls are not sophisticated enough. This becomes problematic in settings with a diverse range of users, as the system may not fully account for the varying levels of access required by different roles within the healthcare environment [274].

Besides RBAC, DHIS2 uses data encryption to ensure the data is secure while transferring between the devices and the server to meet the organization's data confidentiality and integrity standards. This will reduce interceptions because data is encrypted while passing through a given network. However, there is still an important issue with the encryption scheme, more specifically, because the current system does not offer full end-to-end encryption. This constraint poses the data at risk both in the storage and processing phases in case of vulnerability in the system along the data journey [306].

In addition, the system has a feature of a user activity log, which records the activities of the users on the DHIS2 system. This monitoring assists in keeping a check on who is doing what, and who is up to which kind of something, to hold an individual or a party answerable, and to look out for signs of unauthorized activities. Since the user activities can be logged in DHIS2, the system can reveal the usage patterns and identify any abnormal behavior [50]. However, the system is limited in its technical capacity to analyze user logs adequately. This constraint hinders the possibility of identifying security breaches and responding to them promptly, with the result that the system is vulnerable to security threats. These challenges call for an even more robust strengthening of the DHIS2 privacy and security features in an environment that demands the highest degrees of health data confidentiality.

Case Study: DHIS2 and Immunization Campaigns in Uganda

In Uganda, the use of DHIS2 greatly helped a national polio immunization campaign because it provided real-time data tracking. Health workers in facilities across the country, including remote areas, used mobile devices to enter immunization data while the data were aggregated at the district and national levels. This was made possible because health officials were able to identify districts with low immunization rates and post additional hands, leading to a 92percent drop in polio cases [256].

However, problems were encountered, for example, weak Internet connection, especially in rural areas, resulting in the inability to complete real-time reporting and privacy issues. In identifying privacy issues, basic privacy measures such as encryption and access control were implemented in DHIS2, but end-to-end encryption and user authentication issues highlighted the need for stronger security protocols. Solving these connectivity and privacy problems is important to

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

enhance the functional performance of DHIS2 and protect health data in future future immunization campaigns [256].

Challenges in DHIS2 Implementation in Uganda

Despite the widespread adoption of DHIS2 in Uganda’s healthcare system, its implementation faces several challenges that hinder its full potential. These challenges span technical, human resource, and privacy/security aspects, all of which significantly impact the effectiveness of the system, particularly in rural and remote areas.

Technological Challenges

One of the limitations observed in the course of the implementation of DHIS2 in Uganda is the connectivity problem, characteristically marked by intermittent poor internet, especially in rural areas, which results in delayed data submission and synchronization. This impacts the efficiency and accuracy of the flow of health information. However, poor ICT infrastructure leads to slow processing and data errors. These challenges devalue the effectiveness of DHIS2 in generating real-time data for analysis and decision-making as well as a tool for healthcare enhancements [305].

Human Resource Constraints

Some of the limitations include Human resources and limited-trained staff to operationalize DHIS2, especially in rural areas of Uganda. The targeted groups of people are usually trained at the beginning of the project and do not receive consistent support later, which means they cannot make the most of the system’s tools, incorporating the capabilities of data analysis and report generation of reports. Further, high turnover of staff amplifies these difficulties, alongside interrupting data acquisition and integrating recurrent training [309]. It results in poor use of DHIS2 and poor standardization in the reporting of such data, bringing a lack of harmony in policies formulated and criteria set, in decision-making, and the effectiveness of the system.

Privacy and Security Concerns

While DHIS2 includes privacy measures like encryption and role-based access control (RBAC), concerns remain about data security. The lack of robust end-to-end encryption leaves patient data vulnerable during storage and processing. Weak implementation of RBAC, with insufficient granularity, risks unauthorized access to sensitive information. Additionally, limited customization options hinder the compliance with the Uganda Data Protection and Privacy Act (2019), making it difficult to tailor the system to local healthcare and legal requirements, thus increasing the risks of non-compliance and potential legal liabilities [294][277].

7.5.2 FAIR’s Role in Transforming Uganda’s Health Information Systems

Global Relevance of FAIR Data Principles in Uganda’s Context

The FAIR data (Findable, Accessible, Interoperable, and Reusable) are global principles guiding the management and sharing of data. Optimize data value through better search, availability, utilization, and recurrence, which is vital for resource-constrained health systems such as those

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

in Uganda. The application of these principles improves access, improves the decision-making process, and possibly improves interdisciplinary collaboration nationally and internationally [145]. However, for their applications in Uganda, a lot of effort must be put into making it understandable within the context of Ugandan health, laws, and technology.

Findable: Facilitating Access to Health Data

The “Findable” part of the FAIR concept focuses on the set of rules according to which the data source must be discoverable just by using the proper metadata and structured indexes. Currently in Uganda, there are numerous scattered data sources and no consolidated database/website of health information. Its use would set up a single database in which all the report’s different components could be traced [272]. For instance, DHIS2 has already assisted in exercising the centralization of data in health; the implementation of FAIR will likewise extend extensive metadata intake into the data-indexing process, supporting the decisions of involved health practitioners, policymakers, and international organizations [276].

Accessible: Breaking Down Barriers to Health Data Access

The “Accessible” principle allows data to be organized and to be accessed by identified and approved users only [290]. The flow of health data in Uganda is still faulty, meaning that different parties have unreliable means of getting the necessary information. Implementing FAIR principles would improve data availability for health officials, researchers, and policymakers. For instance, though Uganda’s Ministry of Health utilizes DHIS2, data sharing is often limited. Adopting FAIR principles would standardize data sharing, making it accessible for researchers, NGOs, and international organizations. It would foster cooperation and raise productivity in responding to health crises such as the coronavirus pandemic [258].

Interoperable: Enhancing Collaboration and Integration of Health Data Systems

The principle “Interoperable” means that data must be formatted to fit and integrate with other systems. In Uganda, data integration problems arise from format, platform, and standards. These barriers would be solved in case Uganda embraced FAIR principles so that its data in the health sector would correlate with international standards, as well as be published to international organizations. This would give better integration with other systems such as the WHO Global Health Observatory, adding value to global disease surveillance and research. The interoperability would also enhance the movement of information between the health departments in Uganda, the various Governments, and other stakeholders in the health sector to boost healthcare coordination.

Reusable: Promoting Long-Term Value from Health Data

The “Reusable” principle makes it possible for data to be useful for multiple purposes, for example, for research, policy, and program purposes [290]. Most healthcare data in Uganda is still fragmented, and the information can only be used for reporting without extension. With FAIR principles in mind, data would be well described with comprehensible metadata and stored

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

as usable formats for reuse. For instance, data collected in immunization campaigns could be used for vaccination analysis or monitoring of health status, to name but a few. This would improve data value; provide research-based decisions and popularity of Uganda in international collaboration and funding, of global health research [289].

Case Study Analysis: Implementation of FAIR Data Principles in Uganda through Open Science Framework (OSF)

The implementation of FAIR data principles has had a significant influence on the management of global health data and learning from such practices is valuable to Uganda's health data frameworks. An example of this is the Open Science Framework (OSF) [311], the core objective of which is related to the management and sharing of materials according to the FAIR principles. Such information is useful to understand how Uganda needs to apply these principles, specifically in connection with its current health data platforms comprising DHIS2.

The OSF is an all-encompassing platform that handles the life cycle of research work from data collection to sharing of data while at the same time keeping the data secure and making it readily available [286]. By incorporating the use of sophisticated encryption techniques and adopting RBAC methodology, OSF guarantees that research data of the sensitive kind can be shared safely and requires access control by the roles of users. This feature of OSF is especially valuable for Uganda because privacy considerations are important for handling often-sensitive data related to patients and public health in the country. Existing systems such as DHIS2, while indispensable in health data management, can present significant difficulties in guaranteeing strong privacy safeguards, including poor encryption mechanisms and user identification. Currently, OSF's strategy of encryption, alongside confined access as a method of protecting data, can be used by Uganda to enhance the security and privacy of health information [279].

Additionally, the scalability of the OSF is another key takeaway for Uganda. The OSF is capable of managing big databases in several research disciplines [286]. This scalability makes the platform suitable for managing the rising volumes of information that emerge in global health research, particularly in outbreaks or emergencies. The challenges Uganda has been experiencing when it comes to data aggregation, reporting, and analysis can learn from OSF architecture so that it can find a way of handling the big data that is produced by Uganda's health system. If the Ugandan government widens the use of DHS2 or adjusts it to other local health information systems through integration of other scalable data management solutions, then undoubtedly health analytics of Uganda could be enhanced for better and faster comprehensive illness surveillance, enabling more effective disease monitoring and faster decision-making [276].

The FAIR principles' application to the management of health data for Uganda could help increase openness, strengthen multipronged engagements between local and global health agencies, and generally boost health system outcomes [312]. So, following the example of the OSF and the lessons learned on privacy-preserving data sharing and scalability, Uganda could start the path toward improving data management to protect sensitive health data, but also make it available to conduct important biomedical and operational research as well as inform decision-making [293]. Although these FAIR principles can be applied in any country's health informa-

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

tion systems, their integration into Uganda's systems can go a long way in improving health analytics, better data-informed policies, and a stronger health system in general.

Privacy Features of FAIR Data Principles

The principles of FAIR data are intended to facilitate the management and utilization of research findings and data in any field of science. Of all the principles, they have significant consequences in the aspect of privacy, especially within the healthcare setting, which involves information from patient records [259]. Among the principal concerns of the FAIR principles, privacy is equally important, and here it is accentuated by many methods that can provide data privacy and increase the general benefits derived from data at the same time. Such privacy measures include anonymization, consent management, and proper execution of data-sharing techniques and procedures [279].

Data Anonymization

Anonymization is among the most important privacy operations recommended by the FAIR principles. In health data management, particularly in developing countries like Uganda, handling and sharing of patient data requires anonymization to avoid revealing people's identities. Anonymization is the process of excluding or masking nominated personally identifiable information (PII) from data sets so that even if unauthorized parties access the data, the data cannot be used to identify specific individuals or business entities [280]. For patient anonymity, this follows both global best practices and the Ugandan Data Protection and Privacy Act enacted in 2019.

However, compared to Health Information Systems (HIS) such as DHIS2, which have comparatively few integrated anonymization options, the FAIR principles aim to generalize the anonymization of data [266]. The authors do not recommend particular software methods but stress that anonymization should be a component of any data management strategy. This position is rather beneficial, especially for developing countries such as Uganda, since different health systems that may require the implementation of anonymization may do it differently based on the nation's needs and its technological capabilities.

Consent Management

Consent management is another privacy aspect that the FAIR principles encourage. Consent involves actions of people (including patients) that authorize access to their data, use, and sharing. The principles of FAIR support the proper consent procedures in a way in which people's data are to be used, explained, and consent is sought before using those data. This is especially true in health facilities where access, storage, transmission, and use of data are accompanied by the highest levels of security [270].

Like most systems, DHIS2 is implemented within a certain architecture that can integrate or lack the broad option of consent management solutions, but the guidelines of FAIR data stress the need for consent to be included in the process of data sharing. This means that each time data is shared, especially for research purposes, consent must be verified and documented [268]. The framework provides versatility that facilitates the incorporation of multiple consent man-

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

agement solutions, which are legal in global standards, in addition to Uganda's data protection laws [307].

Secure Data Sharing Protocols

The FAIR principles focus on how data is shared, whether between data repositories, platforms, or between different organizations, to reduce the ability of unauthorized persons to gain access to the data. Data sharing poses a very sensitive issue in healthcare systems, as the violation of the Patient's Information Act has legal consequences, and ethical and financial implications are involved. To communicate data safely, the FAIR principles recommend the utilization of encryption and access protocols and data integrity verification processes [288]. Unlike DHIS2, which uses a set of security measures (such as role-based access controls and encryption), the FAIR principles do not impose the necessity of any particular software solution for the secure sharing of data. Researchers have emphasized the right systems to ensure the best practices in data security that can be developed in different ways according to the technology base within respective regions. In this regard, for Uganda, this flexibility provides an opportunity for local systems to incorporate data preservation methods that can be shared securely in conjunction with the FAIR principles that are compatible with Ugandan challenges such as weak internet connection, resource constraints, and infrastructure constraints [145].

Adaptability and Flexibility of the FAIR Framework

One of the most compelling aspects of the FAIR principles, especially in comparison to systems like DHIS2, is their adaptability. The FAIR principles do not bring recommendations on the software or tools that should be used for privacy [255]. They do this without prescribing the specific contingency, which enables healthcare providers and countries to implement privacy practices that exist locally and yet meet international benchmarks. Such flexibility is desirable, more so in developing countries such as Uganda, where resources are relatively scarce. Although DHIS2 is a recognized platform in Uganda's health sector; though, it may lack the privacy required by the FAIR principles. On the other hand, with the help of the FAIR framework, Uganda can choose privacy solutions tailored to its conditions: from the deployment of other types of encryption to the more liberal concepts of consent and improved data masking methods. This flexibility enables Uganda to maintain patient confidentiality while keeping data open for use in public health [294].

Findings: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics of Patient Data in Uganda

Table 7.4: Comparison of DHIS2 and FAIR Data for Privacy-Preserving Health Analytics in Uganda.

Aspect	DHIS2	FAIR Data
Strengths	• User-friendly interface enabling widespread adoption in many health facilities. • Provides key performance metrics such as data completeness and reporting timeliness. • Facilitates evidence-based decision-making, even in rural areas with limited technical expertise.	• Superior interoperability for seamless data sharing across platforms and domains. • Promotes data reusability for research, policy development, and public health interventions. • Enhances global and regional collaboration through standardized data-sharing protocols.
Limitations	• Limited flexibility in accommodating diverse privacy protocols. • Weak role-based access control (RBAC) and insufficient encryption mechanisms expose data to breaches. • Rigid adaptation to dynamic local contexts and emerging privacy concerns.	• Requires substantial technical expertise, infrastructure, and resources for effective implementation. • Demands advanced tools for secure data sharing, consent management, and data governance. • Significant investments are needed for training, infrastructure upgrades, and alignment with international data standards.

Continued on next page

Table 7.4 continued from previous page

Aspect	DHIS2	FAIR Data
Implications for Uganda	Operational Impact • Effective in operational decision-making but constrained by privacy vulnerabilities and rigid structure. Integration Potential • Integrating FAIR principles with DHIS2 could address privacy shortcomings, including weak encryption and access controls. Challenges • Uganda faces resource constraints, outdated infrastructure, and a shortage of skilled personnel required for integration. Recommendations • Pilot projects in selected regions could evaluate the feasibility of integrating FAIR principles into DHIS2 for applications such as immunization tracking.	Operational Impact • Highly beneficial for data sharing and reuse but challenging to implement in resource-constrained settings such as Uganda. Integration Potential • FAIR-compliant systems could facilitate improved collaboration within Uganda and with international health organizations, strengthening health analytics and decision-making. Challenges • Implementation requires substantial investment in capacity building, regulatory frameworks, and international partnerships for technical and financial support. Recommendations • Collaboration with international partners and donor agencies could provide the expertise and funding needed to overcome implementation barriers.

7.6 Conclusion

This study demonstrated the critical need to convert Health Management Information System (HMIS) data from paper-based formats into machine-actionable, interoperable formats to enhance data reuse, accessibility, and real-time analytics. Through the implementation of FAIR (Findable, Accessible, Interoperable and Reusable) principles, data from various health facilities in Uganda—and across eight VODAN-Africa countries—was digitized using tools like the

Conclusion

DHIS2 and FAIRification platforms. In the FAIRification process, the data were transformed into standardized knowledge representations using semantic web technologies (e.g., RDF) and integrated into semantic models supported by existing and custom ontologies. These semantic models enhanced data interoperability and contextual understanding, allowing for more meaningful analytics.

The FAIR data terms were assigned persistent and resolvable web identifiers, enabling human and machine readability. The controlled vocabularies—originally legacy lists in printed HMIS registers—were published and mapped via Bio-portal using the LOOM algorithm, aligning them with international standards and enhancing their findability and reuse.

In contrast, while DHIS2 enables standardized data entry and has improved data comprehensiveness and decision-making within Uganda’s healthcare system, it still suffers from limitations. Its reliance on aggregated data restricts granularity, limits the discovery of deeper insights (as evidenced in the less detailed heat maps), and poses challenges related to privacy, especially in rural and resource-constrained areas.

The deployment of FAIR-compliant tools within the VODAN-Africa infrastructure, including knowledge graphs and localized CEDAR platforms, enabled federated querying and dynamic visualization through dashboards. These tools illustrated the capability of health facilities to view real-time statistics (e.g., Outpatient 002, COVID-19, maternal care) while maintaining patient privacy and data security. This hybrid approach—integrating FAIR principles with DHIS2—presents a viable and promising pathway for Uganda to achieve more robust, privacy-preserving health data analytics that comply with the Data Protection and Privacy Act, 2019.

However, scaling this approach requires overcoming substantial barriers, including limited technical capacity, inadequate infrastructure, and weak regulatory enforcement. Further work is needed to explore advanced uses of FAIR data, such as machine learning for predictive modeling or disease co-occurrence analysis, which could further unlock the value of semantic, linked health data in Uganda and similar contexts. Cost and programmatic effectiveness plans to scale up the health information systems will be helpful for the successful development of the Ugandan Health Information System. It has also been observed that FAIR Data are granular compared to non-FAIR data (see Appendix 9.6). In conclusion, the integration of FAIR principles into DHIS2 represents a significant opportunity for Uganda to enhance its healthcare data ecosystem. By addressing existing challenges and leveraging international collaborations, Uganda can establish a robust, privacy-preserving, and interoperable health information system that sets a benchmark for other developing nations.

Chapter 8

Machine Learning Models in FAIR Health Data from Uganda

This chapter is based on the following publication:

Mariam Basajja, Wolstencroft Katy, Fons Verbeek; A comparative analysis of the performance of Machine Learning Models in FAIR Health Data from Uganda, Africa. [Accepted, publication pending]

Abstract

In this chapter, we discuss the application of machine learning (ML) and the FAIR data principles - Findable, Accessible, Interoperable, and Reusable - within the healthcare sector. It utilizes anonymized datasets from VODAN-Africa, specifically from Uganda. The data include Uganda's Outpatient Department (OPD) records and Uganda's integrated antenatal care (ANC) information. These data sets are structured and standardized according to the FAIR principles and are used to train various machine learning models for tasks such as predicting pregnancy risk factors, assessing fetal conditions, detecting anomalies, and performing clustering analyses. This chapter highlights the effective integration of Machine Learning with FAIR data in a federated architecture in healthcare analytics and underscores the significance of high-quality, accessible, and reusable data to support collaborative research and enhance public health outcomes.

Keywords: FAIR data, machine learning, federated learning, federated architecture, outpatient register, antenatal care register, Allegrograph triple store.

The source code for the scripts in this chapter can be found here: <https://github.com/mariambasajja/CH8-THESIS-CODE>

8.1 Introduction

In recent years, the combination of machine learning (ML) and advanced data science techniques in digital health research has revolutionized the field. ML provides powerful tools for predictive analytics that enable the extraction of meaningful patterns from complex data. The patterns provide insights that inform the strategies used in the implementation of health interventions and the allocation of resources to reduce the prevalence and death of diseases, thus saving lives [153]. Despite the magnitude of this opportunity to improve healthcare systems, there are existing concerns about data privacy, data quality, and ethical issues that arise when applying machine learning to solve the problems faced in healthcare. Since the application of data science driven by advancements in big data and artificial intelligence (AI) has increased, the need to implement robust and ethical practices in data management has also grown [154]. Strict ethical and legal administration must be put in place to ensure that sensitive data in healthcare is protected. This will reduce the negative impact of malpractices in the management of health data [155].

An advantage of ML in healthcare is its utility in analyzing information in multiple domains, as valuable patterns often require large amounts of diverse data to emerge. A single source of data, such as from individual healthcare organizations or specific industries, is rarely adequate in scale or diversity to reveal robust and generalizable insights. However, integrating data from different sources presents challenges in data integration, standardization, and collaboration. In addition, it increases concerns about data privacy and security, as sensitive health information must now be shared or accessed across institutional and sometimes national boundaries. These challenges present the need to utilize an approach that supports collaborative data analysis in healthcare. The FAIR data principles provide disruptive potential in the biomedical sector since they critically improve data accessibility, interoperability, and reusability, which are determinants in the improvement of healthcare systems. Biomedical research and the digital health framework apply to FAIR (Findable, Accessible, Interoperable, and Reusable) data, which introduces a standardized approach to managing and sharing demanding data. Findability is the principle that ensures that the search for metadata-tagged data is simple, so that researchers and healthcare facilities can identify data that applies to their field on various platforms. Accessibility enables relevant users to access information, and various stakeholders in healthcare research and organizations can effectively access data in a manner that captures the provisions of security and privacy issues.

The interoperability aspect of FAIR data ensures that there is no interruption when integrating data on different systems and platforms. This is particularly critical in the healthcare sector because patient information may come from a large number of sources, including hospitals, clinics, mobile health applications, and research studies. FAIR data allows these disparate sources to interoperate with each other through standardisation of data formats and protocols, hence allowing collaborative efforts across institutions and even countries. This provides medical personnel with access to the detailed picture of the past of a patient, allowing them to make more reasonable clinical decisions and positively affect patient outcomes.

The reusability principle makes it possible to repurpose datasets and use them in new research and analyses rather than collecting the same data repeatedly. With respect to the biomedical industry, this is particularly useful when funding for data acquisition is scarce. The FAIR data allow the utilization of existing datasets across several studies, empowers researchers to perform Longitudinal analyzes, enhances predictive model accuracy, and finds new ways to track

Introduction

patterns of diseases, treatment reactions, and drugs.

Practically, based on FAIR data principles, data silos, which commonly occur in healthcare systems, are broken down, and different institutions, including local hospitals, global research organizations, and others, can share their experiences and collaborate more efficiently. To give an example, a rural hospital might not have easy access to the data of different facilities, but observing the principles of FAIR data accessibility, it would be able to safely share the anonymized data on patients and participate in collaborative research that would influence more general policies in healthcare. This introduces novel possibilities of using data in decision-making, which results in more effective treatment plans, rational use of resources, and healthier results among populations. In addition, privacy and data security are factors in the design of FAIR data. Improves the flow of information, but at the same time ensures that confidential data concerning patients remains secure. Access controls are clearly established within the system so that no unauthorized people or institutions can access specific datasets. This is especially decisive in the health care industry, where the issue of security of personal health information is not just mandated by the law, but is also a prime example of protecting health information on an ethical front. Its main strengths are that it allows sharing data, but at the same time ensures high control over access.

In general, the principles of FAIR data are critical in improving collaborations in healthcare systems at the local and global scales. These principles can advance the creation of more valid predictive models, enable collaborative work, and promote better decision-making and higher-quality care because they assist in the safe exchange of high-quality, standardized data. FAIR data will be essential in establishing a more connected, transparent, and efficient healthcare ecosystem as more healthcare systems move toward digital health technologies, where much more could be done with data to advance patient care, unlock operations efficiencies, and spur innovation. The implementation of FAIR data principles alongside machine learning emerges as a potential solution to these challenges [176]. Data can be FAIR, without being open. Strict accessibility criteria can be applied to individuals or institutions with permission to use them. Alternatively, Federated Machine Learning (FML) approaches can be applied more easily when data is FAIR and therefore standardised.

Federated Machine Learning facilitates the development of ML models using decentralized data sources without compromising data privacy [156]. This approach addresses ethical concerns in data management and security while simultaneously improving the accuracy of prediction and generalizability of learning models [157]. The application of federated learning and the adoption of the FAIR data principles guaranty that smaller healthcare facilities, including resource-constrained environments, can easily contribute to participating in data-driven patient health research without any significant emphasis on expenditure-heavy infrastructure support. The architecture can be scaled since health facilities only have to store and process local datasets, so they are easy to use. Moving toward the implementation of this architecture, hospitals and clinics will be able to collaborate safely and ensure privacy about their patients, therefore, enjoying the potential of machine learning systems, which are not only highly efficient but also safe to use. By enabling the construction of ML models using decentralized sources of data, Federated Machine Learning(FML) offers a promising path forward to harness the power of diverse datasets while maintaining stringent privacy standards. [315]

The emergence of federated machine learning allows models to be trained together across multiple organizations without sharing raw data, thus addressing some of the privacy concerns of centralized data processing [160]. If decentralized data sources are used, federated machine

Introduction

learning reduces the risks associated with data breaches and facilitates secure, scalable, and privacy-preserving machine learning [163]. FML provides an outline consistent with the FAIR principles to ensure that data is interoperable and reusable while maintaining data privacy [164].

Furthermore, federated machine learning frameworks have demonstrated their versatility and robustness in a variety of applications, such as real-time edge intelligence and mobile packet categorization [165], [166]. Through the combination of machine learning and federated learning, analytical capabilities continue to evolve, allowing sophisticated models to tackle challenges in healthcare, such as the complexity of biomedical image reconstruction [167].

In this study, we show the feasibility of using FAIR data resources in Ugandan Digital Health by training several traditional machine learning approaches on different datasets. AllegroGraph is a semantic graph database and RDF triple-store designed to store, query, and integrate complex linked data. In this study, it serves as the back end for managing anonymized VODAN-Africa health datasets, enabling structured storage and efficient retrieval of patient and maternal health information. FAIR data are organized and standardized, making it easier to work with. We are able to demonstrate different use cases that are possible in a federated architecture. The use cases include: FAIR Outpatient department diagnoses, clustering of K-means, detection of anomalies in FAIR Integrated Antenatal Care Uganda, clustering of pregnANCy risk factors and determinants of fetal states, and modeling of FAIR ANCfetal state prediction. For the FAIR Uganda ANC data, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Gradient Boosting models are employed for both pregnancy risk factor prediction and fetal state prediction. The performance of these algorithms is evaluated after training to determine the effectiveness of the implementation.

This chapter also presents a federated learning implementation for fetal state prediction in a simulated environment using ANC Uganda data. This approach demonstrates how federated learning can be applied to train models across geographically distributed datasets without sharing raw patient information. The simulated implementation mimics a decentralized setting in which data from different health facilities are represented as local data sources. Local models are trained independently on these subsets, which were generated by splitting the Ugandan ANC data set. A central server coordinates the training process and aggregates the locally trained models, ensuring data privacy throughout. Together, these implementations demonstrate the potential of applying FAIR data principles and machine learning in healthcare to enable collaborative model development while protecting sensitive patient data.

Among the primary findings of this study is the idea that the concepts of FAIR data can be implemented in the real world, especially in resource-limited environments, in healthcare settings. Most low-resource health care environments (e.g., rural hospitals, developing world clinics) have reduced access to sophisticated data infrastructure and sophisticated technologies. Such environments also find it difficult to embrace the full potential of digital health innovations due to the fact that they may not have the funds to invest in complicated systems of data management or have the financial capability to acquire costly infrastructure.

Nevertheless, FAIR data principles present a chance to transform this dynamic and introduce a case-flexible, scalable means to data management by providing a consistent way of managing data. FAIR data makes data access and reuse among various institutions easy without regard to the technical systems available at each institution. For example, health data in a rural hospital can be represented on a simple electronic health record (EHR) system that lacks advanced data-sharing capabilities; however, following the FAIR principles, the hospital would still be able to render its data interoperable with the other healthcare institutions, opening new possibilities

Introduction

of collaboration and research.

Our suggested architecture is an extension of these FAIR data concepts, integrated with machine learning, which enables multiple groups to build end-to-end models together without sharing individual raw patient data. Federated learning enables hospitals to train their own machine learning models on their locally stored sensitive data without outsiders learning more about said data. In so doing, by sharing sensitive information, hospitals have had an opportunity to utilize shared insights to develop predictive models collectively; moreover, quality of care also changes positively, and this helps in early detection and evaluation of disease risks. In particular, federated learning also maintains privacy-preserving access to patient data since model parameters (weights and gradients) are exchanged and not raw data.

The proposed architecture is expected to be scalable, i.e., applicable in a great number of care facilities, including small local hospitals and larger regional healthcare facilities. The secret behind such scalability is that the system does not require enormous infrastructure upheavals. Local hospitals are not required to invest in costly computing hardware and software to integrate new systems with existing ones and data sources, and can utilize their current systems without having to alter them. The federated learning strategy allows data access and cooperation between hospitals without the need to physically combine or transfer patient information. This is more so in areas where resources are sparse, since the process of reworking the whole data process might be costly and logistically impossible.

The simplicity in implementation means that even health facilities in other regions, such as those located remotely or in developing nations, can easily employ the architecture. With the necessary expertise and infrastructure in place, the facilities will easily implement federated learning and machine learning models on their local data.

The architecture helps to decentralize the process of learning, allowing hospitals to build predictive models based on their own history, taking the specifics of the health problems relevant to the area into account (e.g., outbreaks of a disease affecting the region, health trends unique to the region). With the help of this initiative, hospitals can exchange their experience, but do not infringe on the privacy of their patients. In contrast, this collaborative practice helps to improve the level of accuracy and generalizability of predictive models, since they are constructed on data sets of various environments. Furthermore, through such a collection of data, hospitals can learn about each other, have access to new knowledge, and improve healthcare outcomes in all regions.

FAIR data are also based on the concept of collaborative access to data without compromising patient privacy. Hospitals and clinics can use the force of multilateral information; however, it should be implemented while complying with ethical principles and privacy regulations. Under a federated learning approach to computation, sensitive patient information is not shared directly. Only the local model training results, e.g., updated parameters, are aggregated and shared. This gives institutions the opportunity to be involved in global machine learning activities and still retain data protection regulations (e.g., HIPAA or GDPR).

Finally, FAIR data allow hospitals in resource-restricted environments to take advantage of the combined potential of machine learning and big data analytics without having to constantly invest heavily in the infrastructure or in their data privacy. It offers an implementable, scalable approach to the application of machine learning to the delivery and outcomes of healthcare within low-resource settings.

8.1.1 Research Questions and Objectives

The research questions and sub-questions that we consider most relevant to the evaluation of the ease of use of FAIR data in digital healthcare in Africa are presented in this section. The main objective of this study is to assess the ease of use of FAIR data in VODAN-Africa by providing analytical information using machine learning to help drive operational decision-making and strategic planning. The study also aims to show the possibility of investigating the patterns revealed by applying ML prediction and clustering algorithms to the FAIR VODAN-Africa data. This will inform policies and other administrative decisions in the development, implementation, and maintenance of health interventions. These objectives will be achieved using data from the Outpatient Department Uganda, Antenatal Care Register Uganda, and data made available for retrieval from Allegrograph triple-store databases according to the signed data agreements between VODAN-Africa and the different health facilities. OPD Uganda data contain anonymized information on patients, including their age, sex, district of residence, blood pressure and blood sugar levels, body mass index (BMI), alcohol use, date of visit, and diagnosis. The FAIR ANC data contains anonymized data on pregnant women. The ANC Uganda dataset includes the woman's age, gestational age, height, weight, and mid-upper arm circumference (MUAC) measurement.

Research Questions

Main Research Question: Does the use of FAIR data support and strengthen the accessibility of different sources of health data in machine learning?

Sub-Research Questions:

Sub-Research Question 1:

what advantages does FAIR data provide for using ML algorithms for clustering and prediction? What patterns emerge when clustering the OPD Uganda Dataset, and how can these patterns inform healthcare practices?

Sub-Research Question 2:

Is the FAIR data sufficiently rich to predict pregnancy risks and fetal states across different age groups? How do pregnancy risk factors and fetal state correlate with the data obtained from the ANC Uganda datasets in predictive modeling? How can we apply ML in a federated architecture?

8.1.2 Our contributions

This research assesses the use of FAIR Data to support and strengthen privacy-preserving accessibility of different types of health data for machine learning. It investigates the use of FAIR data with machine learning to offer new insights into geographic patterns in disease and infection. This is achieved using FAIRified health data from VODAN-Africa, governed by shared Data Use Agreements between VODAN-Africa and health facilities in Uganda. Data anonymisation and these agreements ensure ethical data access to key health statistics, enabling the training of machine learning models for predictions and cluster analysis. The standardized

Related Work

and interoperable nature of the data within the VODAN network further enhances the quality and reliability of the models. These include:

1. Generation of visualizations that reveal patterns in FAIR OPD Uganda data patterns using K-Means clustering analyses.
2. Anomaly detection among pregnant women using the traditional isolation forest using FAIR Antenatal care(ANC) from Uganda.
3. Clustering analyzes on determinants of pregnancy risk factors and fetal states in FAIR ANC data from Uganda using K-Means clustering algorithms.
4. FAIR ANC Uganda data fetal state prediction modelling using traditional Machine learning and federated learning with a simulated dataset split.

Ethical clearance was obtained from Uganda’s Ministry of Health, health facility directors, and other key stakeholders to access the healthcare data used. The paper registers acquired containing OPD Uganda, ANC Uganda data were FAIRified using the VODAN-Africa architecture. [162] This makes data more accessible and reusable by researchers under well-defined terms, and enhances its integration capabilities, facilitating the application of machine learning algorithms. The use cases aim to show the possibility of providing analytical information showing trends that can be used for operational decision-making and strategic planning in healthcare settings.

This chapter explores related work in machine learning, federated machine learning, data science, and big data in health under Section 8.2. This is followed by the methodology implemented to train the machine learning models for the clustering of patient data, prediction of pregnancy risk factors, and fetal states in Section 8.3. The results and consequent recommendations are discussed in the same section. The limitations and future outlook are outlined in Section 8.4. The chapter’s conclusion details how the chapter answers the research questions and achieves the objectives set in Section 8.5.

8.2 Related Work

In this section, we discuss the application of machine learning in healthcare.

8.2.1 Machine Learning in Healthcare

Machine learning (ML), a subset of artificial intelligence, has become increasingly integral to healthcare systems, offering innovative solutions to various challenges in the sector. ML involves the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience, without being explicitly programmed [202]. In healthcare, ML has found applications in diagnosis, treatment planning, patient monitoring, and drug development, among others.

Machine learning can be broadly categorized into three types: Supervised Learning: The algorithm learns from labeled training data to predict outcomes for unforeseen data. In healthcare, this is often used to predict diagnoses based on patient data. Unsupervised learning: The algorithm finds hidden patterns or intrinsic structures in the input data without labels. This is

Related Work

useful in healthcare to segment patients and identify unknown disease subtypes. Reinforcement Learning: The algorithm learns to make decisions by interacting with an environment to achieve a goal. In healthcare, this can be applied to optimize treatment strategies over time.

Several machine learning (ML) models have found significant applications in healthcare. Neural networks, inspired by biological neural networks, are utilized for complex pattern recognition tasks such as image analysis in radiology and pathology, and prediction of patient outcomes based on electronic health records (EHRs) [203]. K-Nearest Neighbors (KNN), a simple, instance-based learning algorithm, is used for classification and regression and has been applied in cancer detection and heart disease risk classification [204]. Support Vector Machines (SVM), effective in both classification and regression tasks, especially in high-dimensional spaces, have been used in cancer classification, neuroimaging for Alzheimer's disease, and prediction of hospital readmissions [205].

K-means Clustering, an unsupervised learning algorithm, is used to partition data into clusters and has applications in patient segmentation, identifying disease subtypes, and analyzing gene expression data [206]. Hierarchical clustering creates a tree-like structure of clusters without requiring a prespecified number of clusters. In healthcare, it's used for patient stratification, gene expression analysis, and disease subtype identification. This method is valuable when the data structure is unknown or when exploring different levels of clustering granularity [207].

Gradient Boosting, an ensemble technique that combines weak learners, is applied to predict disease progression, patient risk stratification, and drug response prediction [208].

The Isolation Forest algorithm is an unsupervised learning method used for anomaly detection and can help identify unusual patient cases, rare diseases, or flag potential errors in medical records [197]. Lastly, Decision Trees, a non-parametric supervised learning method, are valued for their interpretability and are used in clinical decision support, diagnosis, and treatment planning [209].

These models are often combined or used in ensemble methods to improve performance. For example, random forests, which combine multiple decision trees, have been successful in various healthcare applications, including the prediction of patient survival, disease diagnosis, and drug response [211].

Moreover, deep learning, a subset of machine learning that uses artificial neural networks with multiple layers, has shown remarkable success in healthcare, particularly in medical imaging. Convolutional neural networks (CNN) have been used for tasks such as detecting diabetic retinopathy on retinal images or identifying cancerous lesions on radiological images [212].

The choice of ML model depends on the specific healthcare problem, the nature and amount of available data, the need for model interpretability, and computational resources. As healthcare data continues to grow in volume and complexity, the application of these ML models is expected to expand, potentially revolutionizing various aspects of healthcare delivery and medical research.

Applications of Machine Learning in Healthcare

The adoption of ML in healthcare systems has had positive impacts in solving different problems in the healthcare industry. ML has been most effective in cases in predicting disease occurrence, probable outbreaks, and patient outcomes. This capability has enabled a clearer definition of best practices in the treatment of diseases, increasing patient health status through earlier and more effective diagnosis of diseases.

Related Work

Epidemic Forecasting

Forecasting the epidemiological patterns of infections like the flu or COVID-19 using ML models has been proven effective in providing more insights into preventing and controlling the diseases. A recent study by Qian et al. (2020) presents an example of an ML model used for the prediction of the daily number of new cases of COVID-19 in China, demonstrating the potential of ML in the prediction of epidemics [195]. Estimates of the size, the peak, and the duration of outbreaks have also been done using machine learning techniques. Epidemic forecasting approaches include Artificial Neural Networks (ANN), time series, hybrid, and data mining to predict epidemic outbreaks. This is demonstrated in a study by Datilo et al. (2019) that focused on the utilization of ANNs to forecast the emergence of diseases, including influenza, SARS, HIV / AIDS, dengue, tuberculosis, malaria, and cardiovascular diseases [196].

COVID-19 Management

A review by Lalmuanawma et al. (2020) investigated the real-world use of various ML technologies to combat the SARS-CoV-2 outbreak. The study showed the results of using various machine learning models for the prediction of SARS-CoV-2:

1. The support vector regression and stacking assembly models were used for clinical data forecasting, showing an error range of 0.87%-3.51% for one day ahead, 1.02%-5.63% for three days ahead, and 0.95% - 6.90% for six days ahead, indicating strong predictive capabilities.
2. XGBoost Classifier achieved an accuracy of 90% when using clinical data and blood samples with 75 features, demonstrating its potential for accurate prediction.
3. Deep learning using the LSTM network was used for demographic data to predict the end point of the pandemic in Canada.
4. The Hybrid Wavelet-Autoregressive Integrated Moving Average Model and Regression Tree analysed demographic data from various countries to provide real-time forecasts and forecasts for 10 days, identifying seven key features associated with COVID-19-related deaths.

The results show the applicability of using machine learning as a solution to provide insights into how to manage widespread outbreaks. Machine learning technologies provide the alternative methods needed for rapid screening and prediction, as well as contact tracking, forecasting, and developing vaccines or drugs with more accurate and reliable operation [201].

Clinical Decision Support

ML algorithms have been applied to Electronic Health Records (EHR) to predict the probability that a patient's condition worsens, thus improving clinical decision making and diagnosis [210]. A study by Tomaev et al. (2019) showed how deep learning models applied to EHRs could predict acute kidney injury up to 48 hours in advance, potentially allowing earlier interventions [213].

The examples explored in this section demonstrate that applying machine learning is feasible for improving the strategies used to prevent, manage, and monitor diseases, particularly in areas with low human resources compared to the diseases' prevalence. By applying machine

learning methodologies, diagnostic activities can be conducted automatically in healthcare systems, treatment plans can be enhanced, and general patient control can be improved. It enables early intervention and accurate resource allocation for areas struggling with a lack of healthcare workers.

8.2.2 Federated Machine Learning

Security and privacy issues related to centralized data storage are addressed by federated machine learning, as it ensures that sensitive information remains locally available. This reduces the risk of data breaches and increases compliance with data privacy laws by enabling ML models to train ML models on decentralized data [160]. Various healthcare-related applications, including individual therapy and collaborative research between organizations, have demonstrated the feasibility of this approach. For example, FML could support collaborative research in multiple hospitals without jeopardizing patient confidentiality, as Rieke et al. mention [171]. Feki et al. (2021) demonstrated the use of FAIR principles in conjunction with federated learning for COVID-19 chest X-ray classification, showing how this combination can facilitate collaborative research while preserving data sovereignty [172].

8.2.3 Federated Architecture in Healthcare Application

Federated architecture has been extensively studied in the healthcare sector due to the sensitive nature of health data and the growing need for privacy-preserving machine learning models. Several studies have explored how federated learning (FL) and federated architectures can be employed to develop machine learning models that preserve the privacy of patient data while leveraging data across multiple institutions. These efforts are particularly important in developing countries, where healthcare data is often fragmented, and local hospitals or clinics may not have the computational resources to run advanced machine-learning models.

A notable example is the Federated Learning for Health (FL4H) initiative, which seeks to enable collaboration across different healthcare institutions without transferring sensitive patient data. In their research, [295] proposed a federated learning framework where data remains on local servers, and only model updates are shared with a central aggregator. This approach was tested using data from multiple hospitals in different regions, showing that federated learning can improve model accuracy while maintaining data privacy and confidentiality.

Another significant project, authored in [296], focuses on developing a federated architecture specifically for healthcare, integrating multiple data sources and machine learning models across hospitals in Europe. This project demonstrated the scalability and flexibility of federated learning in real-world healthcare scenarios, particularly in predicting patient outcomes using various forms of clinical data. It highlighted that federated learning's ability to collaborate across institutions, while ensuring data remains localized, provides significant benefits in terms of data diversity and model robustness.

In Africa, the application of federated learning in healthcare has started to gain traction, especially in the context of improving access to healthcare services and enhancing decision-making processes. Teo et al. [297] explored the feasibility of applying federated learning in resource-limited settings in Africa, with a focus on countries like Uganda. Their work emphasized the challenges posed by limited computational infrastructure and the need for privacy-preserving

Related Work

models when dealing with sensitive health data. They found that federated architecture, which allows hospitals and clinics to collaborate without sharing patient data, offers a promising solution to these challenges.

In Uganda, a study [298] was conducted on the implementation of federated learning within the context of maternal health data, particularly for predicting pregnancy complications. This study focused on how federated learning could be used to pool data from rural health facilities without transferring patient data across boundaries. It demonstrated that federated learning could provide accurate predictions of pregnancy outcomes even with the limited data available in rural areas. The study also noted the benefit of federated learning in improving model accuracy through the aggregation of data from multiple, geographically diverse healthcare centers, helping overcome the issue of data sparsity.

While federated learning offers significant benefits, several challenges persist in its implementation in countries like Uganda. One major challenge is the technological infrastructure of healthcare institutions. Most hospitals in Uganda, particularly in rural areas, face constraints such as unreliable internet connectivity and limited access to high-performance computing resources. These challenges were highlighted in [299], where they noted that while federated learning could be beneficial, the lack of reliable data transmission and storage capabilities would likely hinder its widespread adoption in Uganda's healthcare system.

Another study [300] focused on Uganda's health data interoperability challenges, suggesting that the FAIR data principles—Findable, Accessible, Interoperable, and Reusable—are fundamental in addressing issues related to data fragmentation and standardization. The study noted that federated learning could only be effective in Uganda if data from multiple institutions could be integrated and processed. This highlighted the need for both technical and organizational alignment to ensure that health data from disparate systems could be combined for effective federated learning without compromising privacy or security.

Despite these challenges, there have been several successful implementations of federated architecture in Uganda, particularly in maternal and child health data. For example, [301] explored a pilot project that implemented a federated learning model for predicting maternal health risks, using data from a network of Ugandan hospitals. They found that federated learning enabled the creation of a more accurate and generalized model by pooling data from different regions of Uganda, which otherwise would not have been possible due to the fragmented nature of the data. This project also demonstrated how federated architecture can facilitate collaboration between institutions while ensuring patient privacy through secure data-access protocols.

The integration of federated learning in healthcare requires that the data used across institutions be standardized and interoperable, which ties directly into the FAIR principles. Federated architecture is highly effective when data is structured in a way that makes it findable, accessible, interoperable, and reusable across different healthcare institutions. In this regard, [302] emphasized the role of standardized healthcare data formats like HL7 and FHIR to ensure that data from different sources can be integrated effectively.

A key challenge in the African context, especially in Uganda, is that many healthcare institutions use proprietary or non-standardized data formats. This lack of interoperability can

hinder the success of federated learning models, as data from different sources cannot be easily merged. For federated architecture to work effectively, data standards must be established, and healthcare institutions must collaborate on data integration efforts. This need for standardization has been highlighted in several studies, such as [303], which recommend that healthcare institutions in Africa move towards adopting global interoperability standards to facilitate federated data access.

Moreover, the use of federated learning models with FAIR health data allows the pooling of diverse health data across various African countries, facilitating cross-border health collaborations. For instance, [304] found that federated learning could enhance predictive models for disease outbreaks across multiple African regions by leveraging data from different healthcare systems without compromising patient privacy.

8.2.4 Data Science and Big Data in Health

The rise of data science and big data has enabled the processing and analysis of large volumes of health data, completely transforming health research. These developments have improved patient care and resource utilization by facilitating faster and more accurate decision-making in the healthcare industry [162]. On the other hand, FAIR principles and FML aim to solve the issues of privacy-preserving accessibility, quality, standardisation, and interoperability that are brought up by the integration of big data into healthcare, as maintaining high standards in managing data is important to ensure success in applying big data analytics [174].

8.3 Use cases, Data and Methods

This study demonstrates the use of FAIRified data in practical applications in machine learning modelling for data analysis and prediction. The use cases include:

1. **Use case 1:** FAIR Outpatient Department Uganda diagnoses with K-means clustering analysis.
2. **Use case 2:** FAIR Integrated Antenatal Care (ANC) Uganda Data Anomaly Detection.
3. **Use case 3:** FAIR ANC Uganda Pregnancy risk factors and fetal state determinants' clustering analysis.
4. **Use case 4:** FAIR ANC Uganda Data Fetal State Prediction Modelling using traditional ML and federated learning algorithms.

This implementation employs different data sets for each of the use cases mentioned to evaluate the performance of the models trained on Outpatient Registry and Antenatal Care register data from Uganda.

8.3.1 Data Sources and Code

The FAIR data used for the clustering and training of machine learning models were retrieved from the VODAN Africa AllegroGraph triple store databases in Uganda. These databases pro-

vide access to data governed by shared Data Use Agreements, ensuring ethical and compliant data usage. AllegroGraph, a graph database, specializes in managing RDF triple data structures for Linked Data applications [175]. Data retrieval utilized AllegroGraph version [7.3.1], available at ¹. Access to the repositories was established after setting up SPARQL queries using the SPARQLWrapper library to communicate with the Allegrograph SPARQL endpoint. Authentication credentials were provided via the setCredentials method, enabling secure access to the repositories. These data are not open to the public, but are accessible to authorized members of the VODAN-Africa network, including researchers affiliated with participating health facilities and other approved collaborators. The queries retrieved relevant data points essential for building and validating the models, while maintaining strict access control to protect sensitive health information.

These data are stored and retrieved from Allegrograph triple-store database repositories with clear protocols and usage rights, allowing authorized users to easily access, query, and analyse whenever needed under well-defined terms. The data templates and graph models used to structure the FAIR data across all repositories and the code used to train different machine learning models on the FAIR data are available on this repository containing this work's documentation.

The Antenatal Care Register(ANC) dataset contains anonymized data on pregnant women. The ANC Uganda dataset comprises the woman's age, gestational age, height, weight, and mid-upper arm circumference (MUAC) measurement. The shape of the data in this dataset included 909 patient records.

To increase the diversity and volume of model training data, the FAIR OPD Uganda datasets (Outpatient Department), as well as the FAIR ANC Uganda datasets (Antenatal Care), were augmented using the synthetic minority oversampling technique (SMOTE). SMOTE is a data augmentation technique that is used to address class imbalance by generating synthetic samples for underrepresented classes. This method enhances the robustness of machine learning models by preventing bias toward majority classes, ensuring better generalization and predictive performance. The FAIR OPD Uganda datasets were expanded from 200 instances each to 2,114 instances.

8.3.2 Federated Learning Architecture for Local Hospital Systems

In a Federated Learning (FL) architecture for healthcare, the data remain decentralized and stay at local sources, such as individual hospital databases, while the model updates (not the raw data) are shared with a central server. This design preserves patient privacy and ensures compliance with data protection regulations by eliminating the need to transfer sensitive medical data outside of the local institution.

How Hospitals Act as Local Clients

In this federated architecture, each hospital acts as a local client with its own dataset, often containing patient information, diagnostic records, or medical imaging data. The hospital does

¹<http://102.130.124.31:8081/>

Use cases, Data and Methods

not share its data with other institutions or a central server. Instead, it runs machine learning (ML) tasks, such as training a model, on its local data set.

The following outlines the process in detail.

- **Local Model Training:** Each hospital trains a machine learning model using its local data. For example, hospitals might use ML algorithms to predict disease outcomes, identify patient risk factors, or cluster patients based on similar characteristics. The model is trained solely on the local data.
- **Model Updates:** After the local training, each hospital sends the model updates (such as the weights and parameters adjusted during the learning process) to a central server. These updates do not contain sensitive patient data, thus preserving privacy.
- **Global Model Aggregation:** The central server aggregates all the received model updates and combines them to form a global model. This is done using the Federated Averaging technique, which calculates the average of the model updates received from each hospital. The global model benefits from the collective learning of all participating hospitals without the need to share data.
- **Distributing the Global Model:** The global model, which now incorporates insights from multiple hospitals, is sent back to each local hospital. The hospitals can then apply the updated global model to make predictions on new patient data or improve their local model further.

This decentralized approach ensures that while the knowledge from multiple hospitals contributes to the overall model, each hospital retains control over its own data, which remain local and private.

Challenges of Implementing Federated Learning in Local Hospitals

1. Limited Computing Power:

- Many hospitals, especially in rural or resource-constrained settings, may not have the computational infrastructure necessary to run sophisticated ML algorithms.
- **Solution:** Light-weight models and edge computing can be used to process smaller datasets locally. Additionally, computation-intensive tasks could be offloaded to a central server or cloud system, especially during aggregation of the model updates.

2. Data Heterogeneity:

- Data across different hospitals may vary in format, structure, or quality, making it difficult to train uniform models.
- **Solution:** Adopting standardized data formats (such as those defined by FAIR data principles) can help ensure data interoperability and facilitate smooth integration across different systems. Hospitals might need to pre-process and clean the data to ensure that it can be used effectively for ML tasks.

3. Network and Connectivity Issues:

- In rural areas, internet connectivity can be unreliable, complicating the transmission of model updates.
- Solution: Implementing periodic synchronization could help mitigate this issue. Hospitals could collect model updates locally and then transmit them to the central server during stable connectivity windows.

4. Data Privacy and Security:

- Even though FL does not transfer raw data, transmitting model updates can still pose privacy risks if not properly encrypted.
- Solution: Encryption and secure multi-party computation (SMPC) methods can be used to ensure that model updates are transmitted securely. These techniques allow hospitals to collaborate on machine learning models while maintaining the confidentiality of patient data.

Transitioning from Centralized Machine Learning to Federated Systems

The transition from a centralized machine learning model, where all data is stored and processed in a central database, to a federated architecture involves several steps:

- **Data Preparation:** The first step is to ensure that the data from various hospitals is standardized, accessible, and compliant with FAIR principles. Hospitals must ensure that data are well structured, with clear metadata and standardized formats to enable easy aggregation and model training.
- **Model Development:** The hospitals can then apply machine learning techniques to local data. The ML models will initially be trained on centralized datasets as a proof of concept, but the final models could be adapted to work in a federated setting.
- **Infrastructure Setup:** The technical infrastructure supporting the federated architecture must be established. This includes setting up secure communication channels, cloud or edge computing resources for model aggregation, and ensuring that local hospitals have the necessary hardware and software to run ML algorithms.
- **Model Deployment and Monitoring:** After federated learning is implemented, the models must be continuously updated and monitored. Hospitals will need to participate in periodic model updates, and the central server must manage the aggregation of new model updates and ensure that the system remains secure and efficient.
- **Scaling and Support:** For a wider deployment, it is crucial to ensure that the federated system can scale to accommodate more hospitals. This requires careful planning around data integration, secure communication protocols, and cloud or edge computing resources to support computational demands.

8.3.3 Machine Learning Modeling Workflow

The machine learning models employed in these implementations primarily utilize supervised learning methods, such as KNN, MLP, and Gradient Boosting, driven by the availability of labeled datasets. The Isolation Forest and K-means algorithms, which are unsupervised learning techniques, are implemented for anomaly detection and patient data clustering tasks, respectively. The workflow used to demonstrate the application of FAIR data in machine learning is illustrated in Figure 8.1

1. **Data Pipeline** This step consists of data acquisition, data preparation, and data pre-processing. Data acquisition involves retrieving FAIR data from various sources, such as the Outpatient Department Register (OPD) Uganda and the Integrated Antenatal Care Register (ANC) Uganda Allelograph databases. The data preparation step cleans and organizes the data by removing duplicates and handling missing values by removing rows that have missing values or converting data formats [177]. Data pre-processing involves selecting features that are relevant to the intended purpose of the models, such as ‘age’, ‘blood pressure’, and ‘blood sugar’, to cluster patient pregnancy risk factors.[178]. This step also includes the scaling of features and the encoding of categorical variables [179].
2. **Model engineering** This step comprises the definition of the model, the training of the model, and the evaluation. The definition of a model involves choosing the type of model to use based on the problem and the available data. For example, a K-means model is selected and defined in the clustering of patient data. Multiple models can be chosen and trained so that the best-performing model is selected for deployment [180]. The chosen model is then trained on the preprocessed data. This involves feeding the data through the model, comparing the predictions with the actual values, and adjusting the model parameters to minimize error [217]. The performance of trained models is assessed using task-relevant metrics [220], such as silhouette scores for clustering algorithms and classification reports showing precision, recall, F1 score, and support for predictive models. In addition, accuracy scores are reported to provide an overall classification performance evaluation.

Validation Techniques for Overfitting: To prevent overfitting, the following validation techniques are implemented where appropriate, based on the model type:

- **Early stopping:** Monitoring validation loss during training of neural networks to stop before over-fitting occurs
- **Data augmentation:** SMOTE (Synthetic Minority Oversampling Technique) [319] was subsequently applied to address data imbalances and compensate for any loss of diversity in the dataset due to missing value exclusion. Adding controlled Gaussian noise to training data was implemented to improve model robustness
- **Regularization:** Applying L1 (Lasso) and L2 (Ridge) penalties to model parameters to prevent them from growing too large
- **Evaluation of the test set:** Assessing model performance in a previously separated section of the data set to measure generalization ability

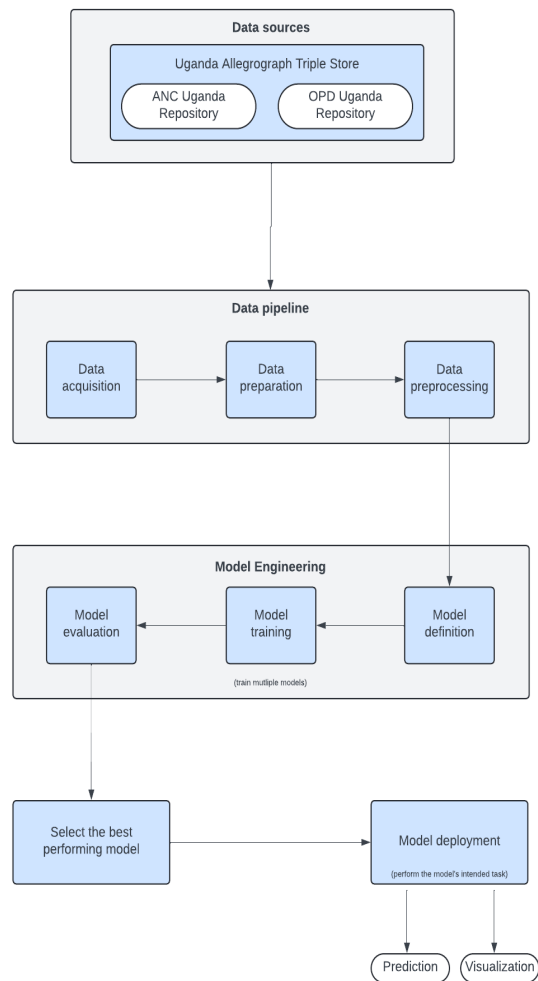


Figure 8.1: Implemented machine learning workflow

- Cross-validation: Evaluating model performance across multiple random train-test splits to ensure consistent behavior
3. Select the best performing model. Multiple models were trained, and the one with the best performance on evaluation metrics is selected. This decision also considers factors like model complexity and computational requirements
 4. Model deployment. The selected model is implemented in a production environment where it can perform its intended task on new, unseen data. In the machine learning implementation carried out for this study, models that are used for prediction are presented with user input to predict the pregnancy risk factors and fetal state of a hypothetical patient.
 5. Prediction and visualization. These are the outputs of the deployed model. Predictions are the model's results on new data, while visualization helps interpret results from the provided data.

The following sections discuss the implementation of this workflow that has been illustrated in the outlined case studies.

8.3.4 Use Case 1: FAIR Outpatient Department Uganda Diagnoses with K-Means Clustering Analyses

Data Acquisition and Preparation

The FAIR OPD data used to train the K-means clustering algorithms were acquired from the Allegrograph repository of the Ugandan Outpatient Registry Department, which contains information collected from hospital visits, including age, sex, district of residence, blood pressure, and blood sugar. The following query was run on the OPD Uganda repository to retrieve the required data:

SPARQL Query 4

```
PREFIX opd: <http://dashboard.vodana.health/>

SELECT ?s ?age ?sex ?diagnosis ?prescription ?district
?blood_pressure ?blood_sugar
WHERE {
    ?s opd:Age ?age ;
    opd:Sex ?sex ;
    opd:Diagnosis ?diagnosis ;
    opd:Prescription ?prescription ;
    opd:District ?district ;
    opd:BloodPressureMmhg ?blood_pressure ;
    opd:BloodSugarMmL ?blood_sugar .
}
```

The results of running this query were exported, yielding 168 instances and seven features. After confirming the absence of duplicates, missing values in the dataset were initially handled by list-wise deletion. While this method is simple to implement, as Kang [223] notes, it can reduce statistical power due to data loss. Given the potential limitations of list-wise deletion, SMOTE was subsequently applied to address data imbalances and compensate for any loss of diversity in the dataset due to missing value exclusion. By generating synthetic samples for minority classes, SMOTE ensured that the final data set was balanced and robust for model training and validation [223].

OPD Uganda K-Means Clustering Analysis based on Patient Age

Data Preprocessing

Three features, *age*, *blood pressure*, and *blood sugar*, were selected from the dataset for clustering. To ensure that the clustering algorithm treats all features equally, a `StandardScaler` was used to standardize the data. This step transforms the features such that they have a mean of zero and a standard deviation of one, which is crucial for algorithms like K-Means that rely on distance-based measurements.

The K-Means Clustering Algorithm

A clustering algorithm was implemented to group patients according to their age, blood pressure, and blood sugar levels using data from the FAIR Ugandan Outpatient Register. Clustering is useful for identifying patterns or subgroups within the data that might not be immediately apparent.

The K-Means algorithm was applied with a predefined number of clusters, in this case 3 and 5. The number of clusters were chosen arbitrarily, but can be adjusted based on domain knowledge. The `fit_predict` method assigned a cluster label to each patient, which was stored in a new column called *cluster* in the data set.

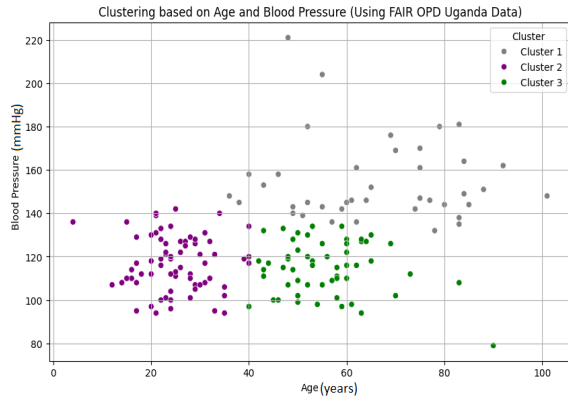
Visualizations were generated for each variation of the data set to examine how the clusters are formed based on the features; age, blood pressure, and blood sugar. Seaborn's `scatterplot` was used for these visualizations, with different colors representing different clusters. These graphs help us to understand the distribution of the groups and their relation to age, blood pressure, and blood sugar levels.

Cluster analysis was conducted using data derived from the FAIR Uganda Outpatient Register, which adheres to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles. The FAIR-compliant structure of the data set allowed efficient retrieval and integration of patient-level variables such as age, blood pressure, and blood sugar for analytical purposes. Through the use of standardized and semantically structured health records, the data could be transformed into machine-learning-ready formats while maintaining consistency and interoperability. This facilitated reproducible clustering analyzes and demonstrated how FAIR data infrastructures can support advanced data-driven healthcare research, knowledge discovery, and future artificial intelligence applications.

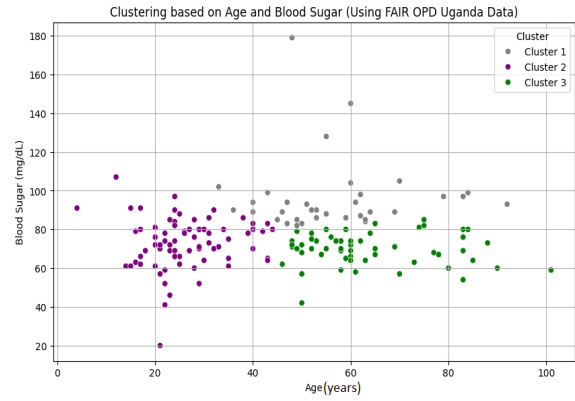
Cluster plots from clustering using FAIR OPD Uganda data (3 clusters):

K-Means Clustering Plots (no. of clusters = 3)

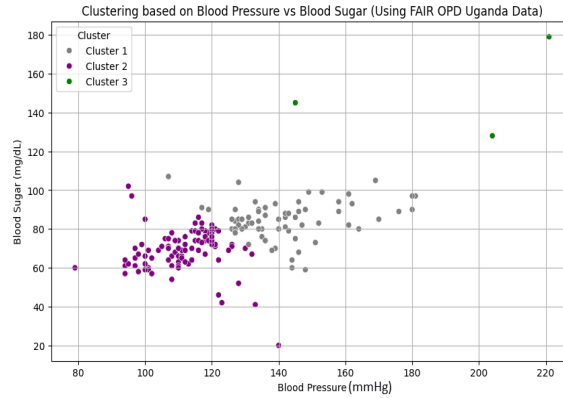
Figure 8.2 illustrates the K-Means clustering results obtained from the FAIR OPD Uganda dataset using three clusters ($K = 3$) and three different parameter pairings, namely age versus blood pressure, age versus blood sugar, and blood pressure versus blood sugar. The cluster-



(a) Clustering of OPD Uganda Patient Data Using K-Means: 3 Clusters Based on Age(yrs) and Blood Pressure(mmHg) with FAIR Data



(b) OPD Uganda Patient Data K-Means Clustering Plots, By 3 Clusters, Based on Age(yrs) and Blood Sugar(mg/dL) Using FAIR data



(c) OPD Uganda Patient Data K-Means Clustering Plots, By 3 Clusters, Based on Blood Pressure(mmHg) and Blood Sugar(mg/dL) Using FAIR data

Figure 8.2: K-Means clustering of OPD Uganda patient data with three different parameter pairings (where $K=3$, $n=168$)

ing process grouped patients according to similarities in their clinical measurements, enabling the identification of distinct patient profiles within the dataset. Prior to clustering, the data was standardized to ensure equal contribution of each feature during model training. After clustering, the cluster centers were inversely transformed back to the original scale to facilitate meaningful interpretation of the resulting patient groups.

Figure 8.2(a) presents the clustering distribution based on age and blood pressure measurements. The visualization demonstrates the presence of distinct patient groups characterized by varying blood pressure levels across different age ranges. Figure 8.2(b) illustrates the clustering pattern for age and blood sugar values, highlighting variations in glucose levels among patients belonging to different age categories. Similarly, Figure 8.3(c) shows the clustering re-

sults based on blood pressure and blood sugar, providing insights into the relationship between hypertension-related indicators and glucose measurements within the patient population.

The cluster centers provide a quantitative summary of the average characteristics of each identified group. Cluster 1 represents patients with moderate age and relatively high physiological measurements, while Cluster 2 corresponds to comparatively younger patients with lower average measurements. In contrast, Cluster 3 represents older patients with elevated physiological indicators, suggesting potentially higher health risks. These clustering results demonstrate the ability of unsupervised learning techniques to identify meaningful structures and hidden patterns within healthcare datasets, thus supporting patient stratification and data-driven clinical analysis.

From a FAIR data perspective, the clustering analysis demonstrates the value of using data that are Findable, Accessible, Interoperable, and Reusable for secondary analytical purposes. The FAIR-compliant OPD Uganda dataset enabled the integration and standardized processing of clinical variables such as age, blood pressure, and blood sugar, facilitating reliable unsupervised learning analysis. The interoperability of the dataset ensured that measurements were consistently represented and machine-readable, while its reusability supported the discovery of clinically meaningful patient groups without requiring extensive data restructuring. These findings illustrate how FAIR data principles can enhance healthcare analytics by enabling transparent, reproducible, and scalable data-driven investigations.

Table 8.1: Cluster descriptions for patient data based on different parameter pairings

Parameter Pairing	Cluster	Patients	Description
Age vs Blood Pressure	Cluster 1	41	Elderly individuals with high blood pressure, indicating potential risks of hypertension and diabetes.
	Cluster 2	71	Younger individuals with normal blood pressure, indicating lower risks of hypertension and diabetes.
	Cluster 3	56	Elderly individuals with normal blood pressure, indicating low risks of hypertension and diabetes.
	Total	168	
Age vs Blood Sugar	Cluster 1	34	Elderly individuals with normal blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 2	76	Younger individuals with normal blood sugar, indicating lower risks of hypertension and diabetes.

Continued on next page

Table 8.1 continued from previous page

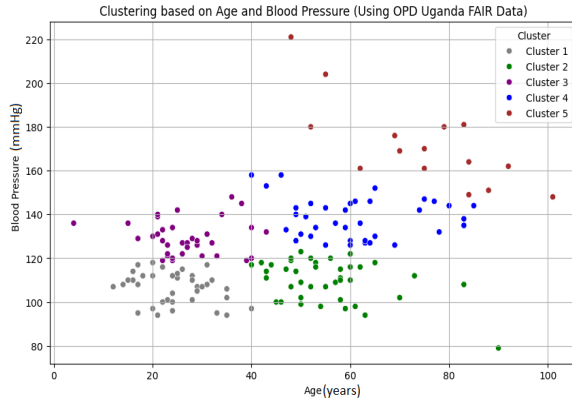
Parameter Pairing	Cluster	Patients	Description
Blood Pressure vs Blood Sugar	Cluster 3	58	Elderly individuals with normal blood sugar, indicating low risks of hypertension and diabetes.
	Total	168	
	Cluster 1	68	High blood pressure with normal blood sugar, indicating medium risks of hypertension.
	Cluster 2	97	Normal blood pressure with low blood sugar, indicating lower risks of hypertension and diabetes.
	Cluster 3	3	High blood pressure with high blood sugar, indicating high risks of hypertension and diabetes.
	Total	168	

Cluster Centers and Clustering Plots using FAIR OPD Uganda data (5 Clusters):

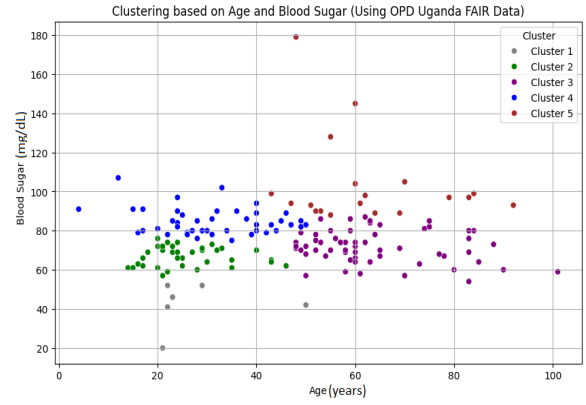
K-Means Clustering Plots (no. of clusters = 5)

Figure 8.3 presents the K-Means clustering results for the OPD Uganda patient dataset using FAIR data principles. The clustering analysis was performed using three different combinations of clinical parameters, namely age versus blood pressure, age versus blood sugar, and blood pressure versus blood sugar. These visualizations demonstrate how patient records can be grouped into distinct clusters based on similarities in physiological characteristics and health indicators. The clustering approach enables the identification of meaningful patterns within the dataset, supporting exploratory analysis and population stratification for healthcare decision-making.

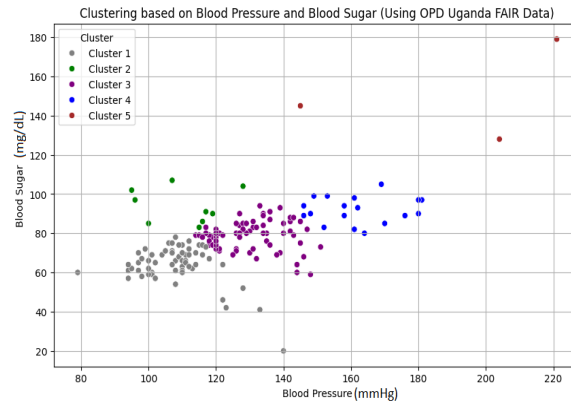
In Figure 8.3(a), patients are clustered according to age and blood pressure values, revealing distinct groups associated with varying blood pressure trends across different age ranges. The visualization indicates that certain patient groups exhibit elevated blood pressure levels at younger ages, while others maintain relatively normal blood pressure distributions. Figure 8.3(b) illustrates clustering based on age and blood sugar levels, highlighting variations in blood glucose patterns among different age groups. Similarly, Figure 8.3(c) shows clustering based on blood pressure and blood sugar measurements, demonstrating relationships between hypertension-related indicators and glucose levels. Collectively, these clustering results provide insights into patient health profiles and may support the early identification of high-risk patient groups for targeted healthcare interventions.



(a) OPD Uganda patient data K-Means clustering based on age(yrs) and blood pressure(mmHg) using FAIR data



(b) OPD Uganda patient data K-Means clustering based on age(yrs) and blood sugar(mg/dL) using FAIR data



(c) OPD Uganda patient data K-Means clustering based on blood pressure(mmHg) and blood sugar(mg/dL) using FAIR data

Figure 8.3: K-Means clustering plots for OPD Uganda patient data using FAIR data ($k = 5$, $n = 168$).

Table 8.2: Cluster descriptions for patient data based on different parameter pairings

Parameter Pairing	Cluster	Patients	Description
Age vs Blood Pressure	Cluster 1	38	Young with normal blood pressure, indicating low risks of hypertension.
	Cluster 2	41	Elderly with normal blood pressure, indicating lower risks of hypertension and diabetes.
	Cluster 3	36	Young with high blood pressure, indicating high risks of hypertension and diabetes.
	Cluster 4	38	Elderly with high blood pressure, indicating high risks of hypertension and diabetes.
	Cluster 5	15	Young with high blood pressure, indicating high risks of hypertension and diabetes.
Age vs Blood Sugar	Cluster 1	6	Young with low blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 2	38	Young with low blood sugar (range 26/67), indicating lower risks of hypertension and diabetes.
	Cluster 3	61	Elderly with normal blood sugar (range 64/72), indicating low risks of hypertension and diabetes.
	Cluster 4	44	Young with normal blood sugar, indicating low risks of hypertension and diabetes.
	Cluster 5	19	Elderly with normal blood sugar, indicating low risks of hypertension and diabetes.
Blood Pressure vs Blood Sugar	Cluster 1	65	Normal blood pressure with low blood sugar, indicating low risks of hypertension.

Continued on next page

Table 8.2 continued from previous page

Parameter Pairing	Cluster	Patients	Description
	Cluster 2	9	Normal blood pressure with normal blood sugar, indicating lower risks of hypertension and diabetes.
	Cluster 3	73	High blood pressure with normal blood sugar, indicating high risks of hypertension.
	Cluster 4	18	High blood pressure with normal blood sugar, indicating high risks of hypertension.
	Cluster 5	3	High blood pressure with high blood sugar, indicating high risks of hypertension and diabetes.

The differences between the 3 and 5 cluster analysis using the FAIR OPD Uganda data are:

Granularity of Age Groups: The 3-cluster approach produced broad age categories (younger 37, middle-aged 44, older 74), while the 5-cluster approach offered more nuanced age groupings (young 26, younger 29, middle-aged 51-58, older 81).

Health Risk Stratification: The 3-cluster approach provided general health categories (healthy, at-risk, high-risk), while the 5-group approach produced more detailed risk profiles, including a distinct group for very high-risk individuals.

Blood Pressure Patterns: The 3-cluster approach involved simple categorizations (normal, high), while the 5-cluster approach revealed more varied levels (healthy, slightly elevated, moderately high, very high).

Blood Sugar Patterns: The 3-cluster approach defined basic categories (normal, elevated, moderately elevated), while the 5-cluster approach identified finer distinctions in blood sugar levels. This includes three clusters for normal, high, and very high blood sugar levels to provide a better understanding of population glycemic control.

Insights into Young Adult Health: The 3-cluster approach identified one young group, while the 5-cluster approach provided more detailed insights by dividing the younger demographic into two groups. This differentiation enabled the identification of individuals with normal and those with slightly higher values, providing a better understanding of the variations in the health of young adults.

Insights into the Middle-Aged Population: The 3-cluster approach involved categorizing all middle-aged people into one cluster. However, the 5-cluster approach offered a more refined analysis by separating the middle-aged population into two distinct groups.

They are those who fall into the middle-risk category and those that fall into the high-risk category. This finer stratification makes it easier to conduct health risk assessments for the middle-aged population.

The 5-cluster approach provides a more detailed view of the health patterns of the population, potentially allowing for more targeted interventions and risk assessments. However, the 3-cluster approach offers a simpler, more general overview that might be easier to interpret and act upon in some contexts.

Normal Blood Pressure and Blood Sugar Levels by Age Group

Blood Pressure:

Normal blood pressure is typically considered to be around 120/80 mmHg.

The ideal range can vary by age. Older adults may have slightly higher normal levels (e.g., up to 140/90).

Blood Sugar Levels:

Fasting blood glucose levels for most people should be between 70-100 mg/dL.

Postprandial (after-meal) blood sugar levels are typically expected to stay below 140 mg/dL.

These ranges serve as general guidelines; individual health status or conditions (like diabetes) can influence what is considered "normal."

The following table 8.3 presents the normal ranges for blood pressure and blood sugar levels, categorized by age group. These values serve as a baseline to detect any anomalies or abnormalities in an individual's health data [318].

Table 8.3: Normal Blood Pressure and Blood Sugar Levels by Age Group

Age Group	Normal Blood Pressure (Systolic/Diastolic) (mmHg)	Normal Blood Sugar Level (Fasting) (mg/dL)	Normal Blood Sugar Level (Postprandial) (mg/dL)
0-9 years	90/60 to 110/75	70-100	90-140
10-19 years	90/60 to 120/80	70-100	90-140
20-29 years	90/60 to 120/80	70-100	90-140
30-39 years	90/60 to 120/80	70-100	90-140
40-49 years	90/60 to 130/85	70-100	90-140
50-59 years	90/60 to 130/85	70-100	90-140
60-69 years	90/60 to 140/90	70-100	90-140
70+ years	90/60 to 140/90	70-100	90-140

Table 8.3 presents the normal reference ranges for blood pressure and blood glucose levels across different age groups. These clinical thresholds serve as baseline indicators for identi-

fying physiological abnormalities within the ANC dataset. Blood glucose measurements are particularly important for detecting diabetes-related conditions, including gestational diabetes mellitus, while blood pressure measurements assist in identifying hypertensive disorders such as gestational hypertension and preeclampsia. By comparing patient observations against these established reference ranges, potential health anomalies can be identified and incorporated into subsequent machine learning analyses, including anomaly detection and patient clustering, thereby supporting early risk identification and improved maternal healthcare outcomes.

8.3.5 Use Case 2: FAIR Integrated Antenatal Care (ANC) Uganda Data Anomaly Detection and Patient Data Clustering Modelling

This section focuses on the process followed to train and apply the Isolation Forest and K-Means Clustering algorithms to analyze the Integrated Antenatal Care Data from Uganda.

FAIR ANC Uganda Acquisition and Preparation

The data used for training machine learning models to analyze patterns among pregnant women according to Integrated Antenatal Care Uganda were acquired from the repositories of the Uganda Allelograph databases that store data collected from hospital visits. The ANC Uganda dataset contains information on pregnant women, with features: the woman's age, gestational age, date of patient visit, weight and height measurements, and blood pressure. The following query was run on the mentioned repositories to retrieve the required data:

Query 5 on the ANC Uganda repository

SPARQL Query 5

```
PREFIX UgANC: <https://dashboard.vodana.health/>
SELECT ?s?Gestational_age ?Age ?Date_of_Patient_Visit
WHERE {
  ?s UgANC:Gestational_age ?Gestational_age .
  OPTIONAL { ?s UgANC:Age ?Age }
  OPTIONAL { ?s UgANC:Date_of_Patient_Visit ?Date_of_Patient_Visit }
}
```

FAIR ANC Uganda Dataset Preprocessing

After data cleaning through handling missing values and removing outliers, the cleaned ANC Uganda dataset had 794 instances and 7 features. A function was created to identify and encode categorical and numeric columns. One-hot encoding was used to convert categorical variables into a format that can be provided to ML algorithms to perform better. The numeric columns were concatenated with the categorical columns encoded into a single data frame. This was followed by dropping rows with missing values to ensure a complete dataset for verification. The “BinaryLabelDataset” from the “AIF360” library created a dataset with the specified

labels (“target”) and protected attributes [320]. In this case, the protected attribute is a specific age group. The privileged and unprivileged groups were then defined on the basis of the protected attribute. In this case, people aged 20-30 were considered privileged, while others were unprivileged. The “Reweighting algorithm” was used to adjust the weights of the instances in the data set to mitigate bias. It ensures that the learning algorithms trained on the data are less likely to inherit biases in the unaltered data. The transformed data set was converted back to a Pandas DataFrame. The one-hot encoded columns were then reversed to their original categorical form.

FAIR ANC Uganda Anomaly Detection and Clustering among Pregnant Women

The Isolation Forest algorithm was implemented to detect anomalies in the data, focusing on the features “gestationalAge”, “age”, and “visit_year” . The dataset contains first time visit instances of the pregnant women in Uganda. This algorithm is an unsupervised learning method for the detection of anomalies. It works by isolating anomalies in the data through random partitioning, exploiting the fact that anomalies are typically fewer and different from normal instances. This approach is particularly effective for high-dimensional data sets and is computationally efficient compared to many other anomaly detection techniques [197, 198, 199]. In practice, the model identifies approximately 10% of the data as anomalies and assigns an anomaly score to each data point. The results are visualized in a scatter plot showing maternal versus gestational age. The anomalies in the scatter plot were highlighted in different colors, allowing a clear identification of unusual patterns in the data set.

Recent advancements in anomaly detection, such as the MS2OD method [199], have demonstrated improved accuracy by combining minimum spanning trees with medoid-based selection strategies. These methods are particularly promising for complex medical datasets where traditional models may miss subtle irregularities. While Isolation Forest offers speed and generalizability, integrating ideas from structure-aware algorithms like MS2OD could enhance the detection of context-specific outliers in future work.

Findings from Anomaly Detection in ANC Uganda using Isolation Forest Algorithms

The plot 8.4 shows the relationship between the age of the women (x-axis) and the gestational age (y-axis). The ages of women range from about 15 to 45 years, while gestational ages range from approximately 5 to 40 weeks. This data distribution provides an overview of the range of ages and gestational periods observed in the study. Most data points are classified as normal (red dots). Normal data points indicate typical cases and help to identify general trends. The anomalies (blue dots) are scattered throughout the plot, but are more clearly concentrated at the edges of the data distribution.

The detected anomalies are primarily located at the boundaries of the maternal age and gestational age distributions, particularly among very young and older women, as well as cases with extremely low or high gestational ages. These observations deviate from the dense central cluster of normal cases identified by the isolation forest model. The highest density of normal data points appears to be for women aged 20-35 years with gestational ages between 15-35 weeks, which is the sample population’s most common age and gestational age range. This finding may be useful for targeting healthcare resources and interventions to the demographic that is most in need of care, such as pregnant women seeking clinic services for the first time.

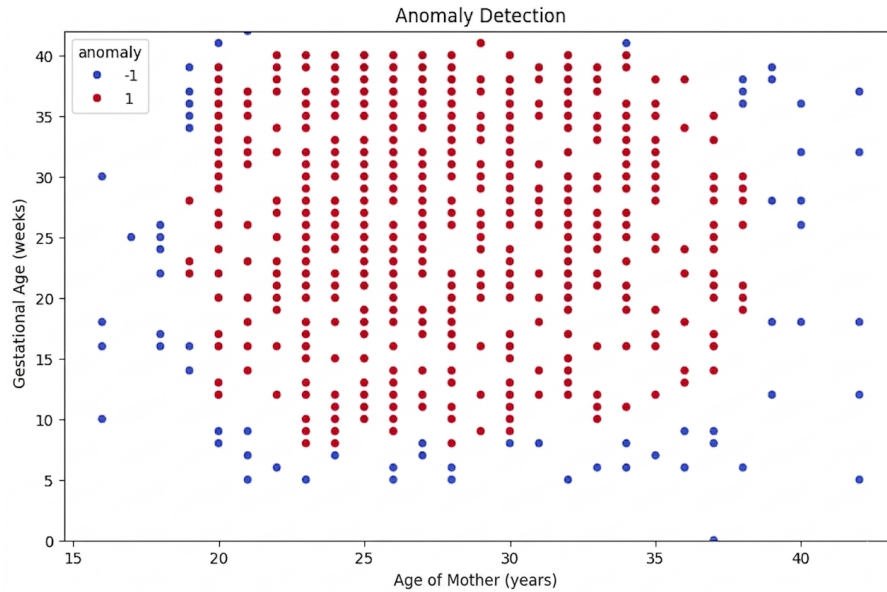


Figure 8.4: ANC Uganda Anomaly Detection Using Isolation Forest Algorithm by Age(yrs) and Gestational age(wks)

The anomalies observed at both extremes of the maternal age spectrum suggest that younger women may present at relatively higher gestational ages compared to other age groups. These age-related patterns suggest that specialized care and attention may be necessary for these groups due to their increased risk factors.

The wide range of gestational ages in all maternal ages indicates that women seek antenatal care at various stages of their pregnancy, irrespective of their age. This variation in gestational age highlights the diverse timing of antenatal visits and the need for flexible care strategies. Anomalies with very high or very low gestational ages can indicate high-risk pregnancies that require special monitoring and intervention within the Ugandan healthcare system. Identifying these anomalies can help prioritize and manage high-risk cases more effectively.

The data distribution suggests that most women are recorded in the dataset during mid-pregnancy. This pattern could be attributed to standard antenatal care schedules in Uganda or local healthcare seeker behaviors. Understanding these data collection patterns is crucial to improving the timing and effectiveness of antenatal care services.

The plot depicted in Figure 8.4 shows the outcome of the anomaly detection process. Using the Isolation Forest algorithm, patient records were classified as normal or anomalous, with anomalies representing observations that exhibit unusual characteristics of maternal age and gestational age relative to the overall population distribution.

8.3.6 Use Case 3: FAIR ANC Uganda Data Pregnancy Risk Factors and Fetal States Determinants' Clustering Analysis

FAIR ANC Uganda Data K-Means Clustering Analysis by Pregnancy Risk Factors

The data set consists of 909 records and 70 attributes. The numerical features that include “Weight(kg)”, “Height_(cm)”, and “Blood_pressure” were converted to appropriate data types, with non-numeric values being systematically replaced with NaN values. A Body Mass Index (BMI) feature was created using the following formula:

$$\text{BMI} = \frac{\text{Weight (kg)}}{\left(\frac{\text{Height (cm)}}{100}\right)^2} \quad (8.1)$$

Pregnant risk factors were determined using a function that classified patients based on multiple clinical parameters: blood pressure measurements and BMI thresholds. The classification included four risk categories: Hypertension, Obesity, Malnutrition, “no risk factor”, which was the default category for cases that did not meet any risk criteria. The two cluster analyzes for the age-risk and gestational age-risk relationships were performed simultaneously. Categorical risk factors were encoded using LabelEncoder and numerical features were scaled using RobustScaler to ensure comparable feature ranges. To address class imbalance issues, SMOTE was applied to generate synthetic samples for underrepresented classes. For gestational age analysis, domain-specific constraints were implemented to filter the data to clinically relevant ranges (8-42 weeks). The K-Means algorithm was configured to partition the data into three distinct clusters (n_clusters= 3) for both analyzes. The quality of the clustering was evaluated using silhouette scores, which measure the cohesion and separation of the resulting clusters [321]. To evaluate the quality and effectiveness of the K-Means clustering results, the silhouette score was used as an internal validation metric. The silhouette score measures how similar a data point is to its own cluster compared to other clusters. Provides an indication of the degree of separation and cohesion among the clusters, helping to determine whether the selected number of clusters is appropriate for the data set.

The silhouette score ranges from -1 to 1 . A score close to 1 indicates that the data points are well matched to their assigned clusters and clearly separated from neighboring clusters. A score around 0 suggests overlapping clusters or weak separation between groups, while negative values indicate that some data points may have been assigned to incorrect clusters.

In general, silhouette scores above 0.70 are considered strong and indicate highly distinct clusters. Scores between 0.50 and 0.70 represent reasonably good clustering performance with acceptable separation between groups. However, silhouette scores below 0.50 are typically considered weak or poor, suggesting that clusters are not clearly distinguishable and that there is significant overlap between data groups. Such low scores may indicate that the selected number of clusters is not optimal or that the dataset does not naturally form well-separated clusters.

Therefore, in this study, the silhouette score was used as an important criterion to evaluate the reliability of the cluster structure and to determine the suitability of the chosen value of K for patient grouping and exploratory healthcare analysis.

Visualization was accomplished through scatter plots using matplotlib, with the original (non-scaled) age and gestational age values plotted against risk factors. BMI and blood pressure variables were useful for the cluster analysis, but excluded from predictive modeling to avoid inflated classification performance. The plots were enhanced with appropriate labels, color-coding by cluster assignment, and gridlines to facilitate interpretation. The resulting visualizations demonstrate the distribution patterns of risk factors between different age groups and gestational periods in the Ugandan ANC population.

Model Performance Analysis: FAIR ANC Uganda Data K-Means Clustering by Pregnancy Risk Factors

The K-Means cluster analysis produced silhouette scores of 0.41 for age and blood pressure clustering and 0.48 for gestational age and blood pressure clustering. These scores indicate moderate cluster quality, suggesting that while meaningful grouping patterns exist within the FAIR ANC Uganda dataset, the clusters are not strongly separated. This is evident in Figure 8.5, where substantial overlap exists between cluster assignments, as shown by the intermixing of colors across the plots. The gestational age-based clustering demonstrates slightly better separation than age-based clustering, reflected by its higher silhouette score, suggesting that gestational age may be a more informative variable for identifying pregnancy risk-related patient groups. Nevertheless, the absence of clearly distinct cluster boundaries indicates that the underlying data contain gradual transitions rather than naturally well-separated groups.

Findings from FAIR ANC Uganda Data K-Means Clustering by factors Contributing to Risks in Pregnancy

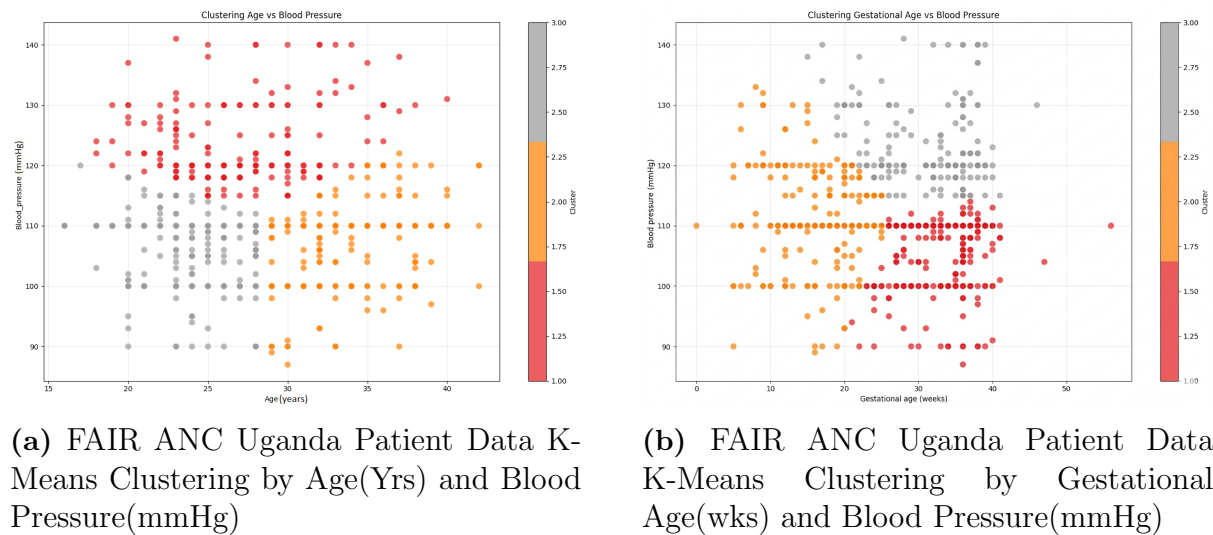


Figure 8.5: K-Means clustering plots for FAIR ANC Uganda patient data.

The K-Means clustering analysis revealed distinct patient groupings based on age and blood pressure patterns in plot 8.5(a) within the FAIR ANC Uganda dataset. The observed clusters suggest the presence of varying maternal health profiles and potential pregnancy risk categories associated with elevated blood pressure and demographic characteristics. Although clustering

itself does not measure FAIRness, the FAIR-compliant data infrastructure enabled efficient preprocessing, integration, and reuse of semantically structured clinical variables for advanced analytical tasks. The interoperability and reusability of the FAIR data framework facilitated machine learning-driven exploration of maternal risk patterns while preserving data consistency and accessibility across analytical workflows.

Table 8.4: Cluster Analysis for Age and Blood Pressure (BP)

Cluster	Sample Size	Age (Mean, Range)	Blood Pressure (Mean, Range)	Dominant Features
Young & High BP	203 (27.2%)	27.0 (18.0–40.0)	123.3 (11.5–141.0)	Fetal state: Non-reassuring (53.7%); Age group: 26–30 (39.9%)
Mid-age & Normal BP	232 (31.1%)	33.3 (22.9–42.0)	106.7 (87.0–122.0)	Fetal state: Reassuring (53.9%); Age group: 31–35 (48.3%)
Young & Normal BP	312 (41.8%)	23.8 (16.0–28.0)	106.6 (90.0–120.0)	Fetal state: Reassuring (43.6%); Age group: 20–25 (51.6%)
Total	747 (100%)	–	–	–

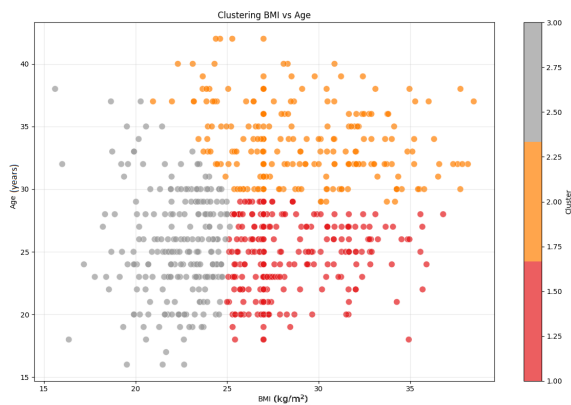
The K-Means cluster analysis revealed clear patient groups based on gestational age and blood pressure trends in plot 8.5(b) within the FAIR ANC Uganda dataset. These groups indicate different maternal health profiles and possible pregnancy risk categories related to variations in blood pressure at various stages of gestation. Although cluster analysis itself does not directly assess FAIRness, the FAIR-compliant data infrastructure supported the integration, preprocessing, and reuse of semantically organized clinical data for advanced analytical applications. The interoperability and reusability provided by the FAIR data framework enabled machine learning-based investigation of maternal risk patterns while promoting standardized and accessible healthcare data analysis.

A comparison of the four plots reveals notable differences in the clustering structure and the strength of the separation between physiological and demographic variables. The Age vs. blood pressure clustering plot shows moderate overlap between clusters, suggesting that although blood pressure varies between age groups, age alone may not strongly differentiate maternal risk categories. However, clusters with elevated blood pressure values appear to be concentrated between certain reproductive age ranges, indicating that blood pressure remains a clinically important indicator of maternal risk.

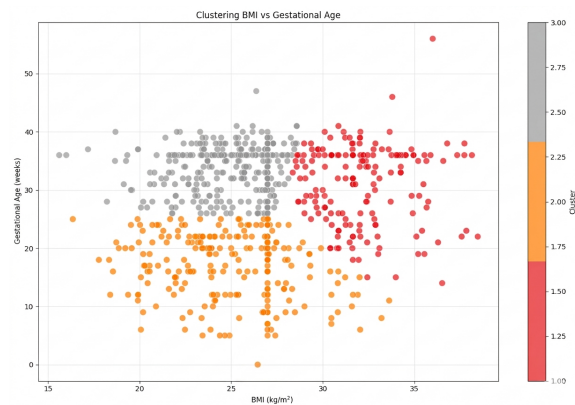
The Gestational Age vs Blood Pressure clustering plot demonstrates stronger and more structured clustering patterns than the Age vs Blood Pressure analysis. Different groups emerge in the gestational stages, particularly among patients who have elevated blood pressure during later periods of pregnancy. This suggests that gestational age may have a stronger relationship

Table 8.5: Cluster Analysis for Gestational Age and Blood Pressure (BP)

Cluster	Sample Size	Gestational Age (Mean, Range)	Blood Pressure (Mean, Range)	Dominant Features
Mid Term & Normal BP	339 (45.4%)	33.3 (21.0–56.0)	105.5 (87.0–114.0)	Fetal state: Reassuring (41.6%); Age group: 26–30 (33.0%)
Early Term & Normal BP	236 (31.6%)	16.6 (0.0–25.0)	110.5 (89.0–133.0)	Fetal state: Reassuring (55.5%); Age group: 20–25 (30.9%)
Early Term & High BP	172 (23.0%)	31.6 (15.0–46.0)	123.1 (11.5–141.0)	Fetal state: Non-reassuring (54.7%); Age group: 26–30 (33.1%)
Total	747 (100%)			



(a) FAIR ANC Uganda Patient Data K-Means Clustering by Age (Yrs) and BMI (kg/m^2)



(b) FAIR ANC Uganda Patient Data K-Means Clustering by Gestational Age (Wks) and BMI (kg/m^2)

Figure 8.6: K-Means clustering plots for FAIR ANC Uganda patient data.

with variability in blood pressure and progression of pregnancy-related risk than chronological age alone. Clearer separation of clusters implies that gestational progression plays a critical role in maternal cardiovascular changes during pregnancy.

The Age vs BMI clustering plot reveals a broader dispersion and partial overlap among clusters, indicating that BMI varies considerably across age groups. Nevertheless, several clusters containing moderate-to-high BMI values appear to be concentrated within the middle reproductive age ranges. This suggests the presence of heterogeneous nutritional and metabolic maternal profiles within the population. Compared to blood pressure clustering, BMI-based clusters appear less sharply separated, indicating weaker discrimination between maternal groups when age is used as the primary demographic variable.

In contrast, the Gestational Age vs BMI clustering plot demonstrates more distinct clustering structures than the Age vs BMI analysis. Patients with elevated BMI values appear to be concentrated within specific gestational stages, suggesting that gestational progression may influence maternal weight-related physiological patterns more strongly than age alone. The clearer cluster separation observed in this plot indicates that gestational age contributes substantially to differentiating maternal health profiles associated with BMI and pregnancy risk.

In general, a comparison of the four clustering analyzes suggests that gestational age produces stronger and more distinguishable clustering patterns than chronological age when combined with physiological indicators such as blood pressure and BMI. This finding implies that gestational progression may be a more influential factor in identifying maternal risk profiles within the FAIR ANC Uganda dataset. Furthermore, blood pressure-based clustering demonstrated stronger separation of potential risk categories compared to BMI-based clustering, highlighting blood pressure as a potentially more sensitive indicator of maternal health variation in pregnancy.

From a FAIR data perspective, these clustering analyzes demonstrate the practical analytical benefits of semantically structured and interoperable healthcare data. Although clustering itself does not directly measure FAIRness, the FAIR-compliant infrastructure enabled the integration, preprocessing, and reuse of standardized maternal health variables for advanced machine learning analysis. The interoperability and reusability of the FAIR ANC Uganda dataset facilitated the exploratory identification of maternal risk patterns while supporting accessible, machine-readable, and reusable healthcare analytics suitable for future federated learning and predictive modeling applications.

Tables 8.6 and 8.7 provide quantitative summaries of the clusters visualised in Figure 8.6. While Figure 8.6 illustrates the spatial distribution and separation of the K-Means clusters, the tables describe their demographic and clinical characteristics, including mean BMI, mean age or gestational age, cluster size, and dominant fetal state. Together, the visual and tabular analyses indicate that gestational age combined with BMI produces more distinct and clinically meaningful patient groupings than chronological age combined with BMI, as evidenced by the clearer cluster separation observed in Figure 8.6(b) and the well-defined cluster profiles reported in Table 8.7.

Table 8.6: Cluster Analysis for Age and BMI

Cluster	Sample Size	BMI (Mean, Range)	Age (Mean, Range)	Dominant Features
High BMI & Young	283 (37.9%)	28.2 (25.0–36.8)	24.6 (18.0–29.0)	Fetal state: Reassuring (47.7%); Age group: 20–25 (47.7%)
High BMI & Mid-age	227 (30.4%)	29.1 (21.0–38.5)	33.7 (29.0–42.0)	Fetal state: Reassuring (52.4%); Age group: 31–35 (54.6%)
Normal BMI & Young	237 (31.7%)	22.4 (15.6–25.1)	25.4 (16.0–38.0)	Fetal state: Distressed (38.4%); Age group: 20–25 (40.9%)
Total	747 (100%)			

Table 8.7: Cluster Analysis for Gestational Age and BMI

Cluster	Sample Size	BMI (Mean, Range)	Gestational Age (Mean, Range)	Dominant Features
High BMI & Early Term	176 (23.6%)	32.2 (28.4–38.5)	31.2 (14.0–56.0)	Fetal state: Non-reassuring (60.2%); Age group: 31–35 (33.0%)
Normal BMI & Early Term	245 (32.8%)	24.9 (16.4–33.6)	17.0 (0.0–25.0)	Fetal state: Reassuring (53.9%); Age group: 20–25 (32.7%)
Normal BMI & Mid Term	326 (43.6%)	24.9 (15.6–28.6)	33.8 (26.0–47.0)	Fetal state: Reassuring (36.2%); Age group: 20–25 (35.6%)
Total	747 (100%)			

8.3.7 Use Case 4: FAIR ANC Uganda Data Fetal State Prediction Tool Modelling

Sub-use Case 4.1: FAIR ANC Uganda Data Fetal State Prediction Tool Modelling

This involved data preprocessing, feature engineering, and fetal state determination. Multiple models were trained to predict the fetal state using ANC Uganda data. They include MLP, KNN, and Gradient Boosting models.

ANC Uganda Data Pre-processing

A search was conducted across all the data in the triplestore. The data set was then pivoted to convert the SPO structure into a more conventional DataFrame format, with subjects as rows and predicates as columns. This pivot operation significantly altered the shape of the dataset, organizing it into a suitable structure for machine learning tasks.

Feature Engineering and Transformation

In this step, numeric columns such as “Weight(kg)”, “Height_(cm)”, and “Blood_pressure” were converted to numeric data types to simplify data handling. Non-numeric values in the other columns were replaced with NaN to make sure that the data were consistent. This was followed by the calculation of the Body Mass Index (BMI) using the standard formula given in Equation 8.1. BMI was added as a new feature in the data set.

The missing values were imputed by replacing them with the most frequent value in the columns “Gestational age”, “Weight(kg)”, “Height (cm)”, “Blood pressure”, and “BMI” in the data set to ensure that the data was complete and ready for modelling.

Fetal State Determination

A custom function was designed to classify the fetus into one of the four categories according to gestational age, BMI, and blood pressure. This function applied logical conditions to assign one of four fetal states as follows: “Well”, “Distressed”, “Non-reassuring” or “Reassuring”.

‘Well’: gestational age ≥ 36 weeks, $23 \leq \text{BMI} \leq 28$, and blood pressure between 70 and 120 mmHg.

‘Distressed’: BMI < 22 or blood pressure > 140 mmHg.

‘Non-reassuring’: $22 \leq \text{BMI} < 23$ or $28 < \text{BMI} \leq 32$ or blood pressure between 120 and 140 mmHg.

‘Reassuring’: Other cases that do not fall into the above categories.

‘Unknown’: Cases in which data on gestational age, BMI, or blood pressure are missing.

Each row in the data set was evaluated, and the corresponding fetal state was assigned based on the criteria defined above.

Fetal State Prediction Tool Model Training and Evaluation

The processed data was split into features (X) and the target variable (y), with features including ‘Age’, ‘Gestational_age’, ‘Weight(kg)’, ‘Height_(cm)’, ‘BMI’, and ‘Blood_pressure’. Numeric features were identified and handled separately to ensure that they were appropriately imputed and scaled using the RobustScaler to mitigate the effects of outliers. To address class imbalance in the target variable, SMOTE was applied. This resulted in a balanced dataset for model training.

Table 8.8: Comparison of model performance for fetal state based on the FAIR ANC Uganda Data

Model	Fetal State	Precision	Recall	F1-Score
K-Nearest Neighbors	Distressed	0.9023 ± 0.0267	0.8332 ± 0.0492	0.8656 ± 0.0332
	Non-reassuring	0.6691 ± 0.0389	0.6610 ± 0.0324	0.6638 ± 0.0207
	Reassuring	0.7083 ± 0.0311	0.6143 ± 0.0616	0.6561 ± 0.0394
	Well	0.8105 ± 0.0159	0.9855 ± 0.0183	0.8892 ± 0.0088
MLP	Distressed	0.9431 ± 0.0231	0.9592 ± 0.0309	0.9507 ± 0.0170
	Non-reassuring	0.7084 ± 0.0521	0.6318 ± 0.0647	0.6659 ± 0.0461
	Reassuring	0.7300 ± 0.0212	0.6143 ± 0.0474	0.6659 ± 0.0284
	Well	0.7860 ± 0.0398	0.9766 ± 0.0116	0.8705 ± 0.0260
Gradient Boosting	Distressed	0.9971 ± 0.0058	0.9913 ± 0.0116	0.9941 ± 0.0055
	Non-reassuring	0.9822 ± 0.0218	0.9533 ± 0.0213	0.9674 ± 0.0184
	Reassuring	0.9659 ± 0.0189	0.9853 ± 0.0132	0.9754 ± 0.0108
	Well	0.9771 ± 0.0144	0.9913 ± 0.0116	0.9840 ± 0.0106

Three machine learning models were trained using stratified sampling: K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Gradient Boosting. The selected machine learning

algorithms were chosen to represent different learning paradigms and levels of model complexity. K-Nearest Neighbors (KNN) was selected as a distance-based, instance-learning algorithm that classifies observations according to the similarity of neighbouring data points. Decision Trees (DT) were chosen because they produce human-readable decision rules, making them particularly suitable for healthcare applications where model interpretability is important. Decision Trees can capture non-linear relationships between variables and provide clear explanations of the factors contributing to a prediction. Gradient Boosting was selected as an ensemble learning technique that combines multiple weak learners to improve predictive performance. Gradient Boosting is capable of modeling complex relationships within healthcare datasets and often achieves superior classification accuracy compared to individual classifiers. Each model was optimized with hyperparameters through . This was done to ensure that the best parameters were selected based on cross-validation performance. The training methodology used a 5-fold stratified cross-validation approach to ensure robust model evaluation. This involved 80 total fits for K-Nearest Neighbors with 16 candidate hyperparameter combinations multiplied by 5 folds. For the Multi-Layer Perceptron, 100 fits were conducted with 20 candidate hyperparameter combinations multiplied by 5 folds. The Gradient Boosting model involved 100 fits with 20 candidate hyperparameter combinations multiplied by 5 folds.

ANC Uganda Data Fetal State Prediction Tool Models' Performance Comparative Analysis

The Gradient Boosting model was selected to be used for prediction based on data input by users. This selection was due to its superior performance across all categories of fetal state compared to the other trained models. It outperformed both the K-Nearest Neighbors (mean accuracy score 0.7734 ± 0.0165) and the MLP (mean accuracy score 0.7953 ± 0.0181) models by achieving an overall mean accuracy of 0.7953 ± 0.0181 , making it the most reliable choice for predicting fetal states.

The resulting performance is shown in the table 8.9.

When comparing the initial results with the latest evaluations, the models showed improved mean accuracy, precision, recall, and F1 scores. The mean accuracy score range of the K-Nearest Neighbors (KNN) model improved from 0.7734 ± 0.0165 to 0.7909 ± 0.0109 , and the mean accuracy score range of the MLP model improved from 0.7953 ± 0.0181 to 0.7602 ± 0.0111 . These improvements suggest that data augmentation effectively improved the generalization capabilities of the models.

The Gradient Boosting model achieved an even better mean accuracy score range of 0.8622 ± 0.0161 , with precision, recall, and F1 scores that demonstrate improvement in all categories of fetal state. This shows its robustness in predicting fetal states for pregnant women in Uganda. These results underscore the role of data augmentation in improving model performance and reliability, particularly in medical applications where data may be limited or imbalanced.

Table 8.9: Comparison of Model Performance per Fetal State based on FAIR ANC Uganda data After Adding Noise

Model	Fetal State	Precision	Recall	F1-Score (%)
K-Nearest Neighbors	Distressed	0.9051 ± 0.0348	0.8743 ± 0.0154	0.8888 ± 0.0102
	Non-reassuring	0.6962 ± 0.0295	0.6491 ± 0.0305	0.6716 ± 0.0272
	Reassuring	0.7467 ± 0.0261	0.6944 ± 0.0473	0.7190 ± 0.0338
	Well	0.8089 ± 0.0165	0.9460 ± 0.0176	0.8720 ± 0.0164
MLP	Distressed	0.9133 ± 0.0213	0.9269 ± 0.0369	0.9193 ± 0.0130
	Non-reassuring	0.6631 ± 0.0206	0.6212 ± 0.0392	0.6404 ± 0.0209
	Reassuring	0.7746 ± 0.0256	0.5584 ± 0.0221	0.6484 ± 0.0134
	Well	0.7072 ± 0.0336	0.9342 ± 0.0146	0.8048 ± 0.0270
Gradient Boosting	Distressed	0.9493 ± 0.0101	0.9225 ± 0.0325	0.9352 ± 0.0128
	Non-reassuring	0.7716 ± 0.0283	0.7294 ± 0.0485	0.7492 ± 0.0335
	Reassuring	0.8251 ± 0.0282	0.8509 ± 0.0159	0.8376 ± 0.0194
	Well	0.9024 ± 0.0325	0.9459 ± 0.0120	0.9232 ± 0.0159

Fetal State Prediction Tool Based on User Input

A function was implemented to provide real-time predictions of the fetal state based on user input, including age, gestational age, weight, height, and blood pressure. The purpose of this feature is to help healthcare professionals and expectant women monitor fetal well-being. Once the user inputs the required information, the system first calculates the Body Mass Index (BMI) using the weight and height provided. The features are then scaled to match the format of the training data used by the Gradient Boosting model. This model is then used to predict the fetal state. This prediction can help identify potential risks and guide appropriate interventions. The tool is designed to be used in a clinical or home-care setting, enabling real-time decision-making during prenatal check-ups or consultations. The feature provides an immediate prediction after user input, which healthcare providers can use to adjust care plans or recommend additional tests if needed.

Sub-use Case 4.2: ANC Uganda Fetal State Prediction Tool Model Training using Federated Learning with a Simulated Dataset Split

Federated learning was implemented by simulating multiple clients (e.g., hospitals) that trained machine learning models on their respective local datasets. The ANC Uganda dataset was first transformed into a patient-level dataset and records with missing fetal state labels were excluded. To ensure a realistic evaluation and avoid information leakage, the variables BMI and Blood Pressure were removed from the predictor set for all models. BMI is directly derived from weight and height, while blood pressure is strongly associated with fetal health outcomes and could artificially inflate model performance if used as a predictor.

The final feature set consisted of Age, Weight, Height, Gestational Age, Parity, Hemoglobin (Hb), Mid-Upper Arm Circumference (MUAC), and ANC visit timing. The target variable was fetal state, consisting of four classes: Distressed, Non-reassuring State, Reassuring State, and Well.

The dataset was divided into training and testing subsets before model development. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training dataset. This ensured that information from the testing dataset did not influence the training process. After oversampling, the balanced training dataset contained an equal number of samples for each fetal state class.

The balanced training data were subsequently partitioned among three simulated federated clients representing independent healthcare facilities. Each client trained its own local model using only its assigned data partition. Model parameters or predictions were then combined to produce a global federated model. The overall federated learning workflow is illustrated in Figure 8.7.

Federated Dataset Distribution

This setup enables the development of collaborative models without requiring the direct sharing of sensitive patient records between participating institutions, thereby preserving privacy while benefiting from distributed learning. After removing records with missing fetal state labels, 710 patient records remained for model development. A stratified 80:20 train-test split was performed, resulting in 568 training records and 142 testing records. To address class imbalance, SMOTE was applied exclusively to the training set, increasing the training data size to 868 balanced samples while preserving the independence of the test set.

The balanced training dataset was subsequently divided among three federated clients. Client 1 contained 292 samples, while Clients 2 and 3 contained 288 samples each. The class distributions were approximately equal for all clients after SMOTE balancing. Each client independently trained its local model without sharing patient-level records, thereby preserving data privacy.

The performance of the local model was also evaluated in each federated node prior to aggregation. Across the ten experimental iterations, local test accuracies generally followed the same trend as the global model results. Gradient Boosting achieved the highest local accuracies, followed by Decision Tree and SVM models, while MLP produced the lowest local accuracies. Local accuracies for the Decision Tree and Gradient Boosting models typically ranged between 0.65 and 0.75 across the three simulated clients. The consistency of local performance across nodes indicates that the balanced data partitioning strategy produced comparable learning conditions for each participating client.

To evaluate local learning performance, each client model was assessed on its corresponding local validation subset. Across the ten experimental iterations, local accuracies ranged from

approximately 64.8% to 68.3% for Decision Tree models, 66.9% to 70.4% for Gradient Boosting models, 52.1% to 62.7% for SVM models, and 28.9% to 41.5% for MLP models. These results indicate that Gradient Boosting consistently achieved the strongest local performance among the evaluated algorithms.

During federated aggregation, each client model contributed to the global prediction according to the size of its local training set. The class probability outputs produced by each local model were aggregated using weighted averaging, where the contribution of each client was proportional to the number of samples used for local training. Consequently, Client 1 contributed a weight of 292 samples, while Clients 2 and 3 each contributed a weight of 288 samples. This approach more closely reflects real-world federated learning deployments, where healthcare facilities often possess different volumes of data. Model aggregation followed the Federated Averaging (FedAvg) principle for the MLP model. Instead of averaging predicted class labels, class probability outputs from local models were aggregated using weighted averaging based on the number of training samples available at each client. Let n_k denote the number of training samples at client k , and P_k denote the predicted probability distribution produced by that client model. The global probability estimate was computed as:

$$P_{global} = \frac{\sum_{k=1}^K n_k P_k}{\sum_{k=1}^K n_k} \quad (8.2)$$

where $K = 3$ represents the total number of clients. The final predicted class was obtained by selecting the class with the highest aggregated probability. This weighted aggregation strategy ensures that clients with larger datasets contribute proportionally more to the global model while maintaining complete data privacy.

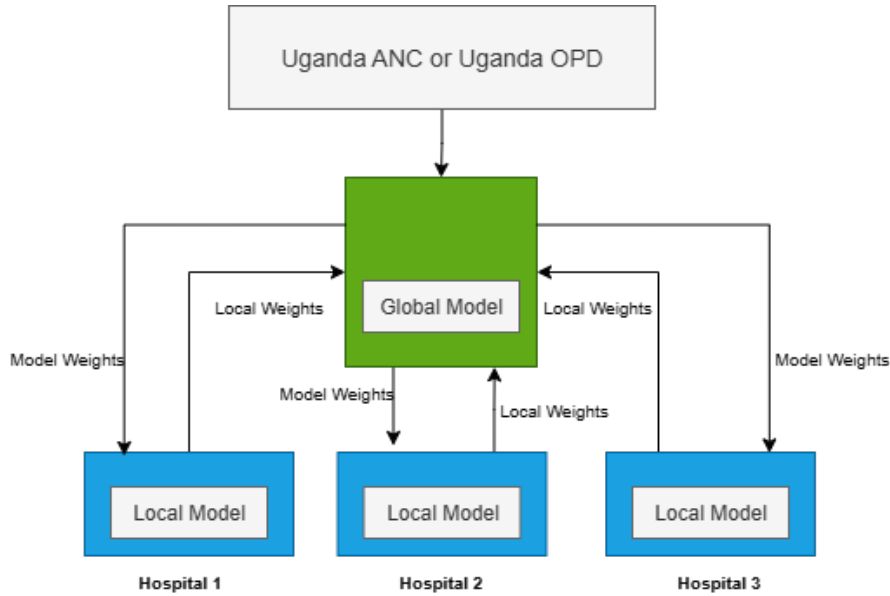


Figure 8.7: Simulated federated learning illustration

The fetal state prediction task was implemented using four machine learning models: a Multi-Layer Perceptron (MLP), Decision Tree, Support Vector Machine (SVM), and Gradient

Boosting model. These models were evaluated within the same federated learning environment to compare their effectiveness in predicting the fetal state.

Multi-Layer Perceptron Neural Network

A Multi-Layer Perceptron (MLP) classifier was implemented using Scikit-learn's `MLPClassifier`. The network consisted of two hidden layers containing 64 and 32 neurons, respectively. Data preprocessing included median imputation and robust scaling, while early stopping was employed to reduce overfitting and improve generalization performance. The model was trained within the federated learning framework and evaluated alongside the Decision Tree, Support Vector Machine (SVM), and Gradient Boosting models for fetal state prediction.

The MLP architecture consisted of:

- An input layer receiving the selected maternal health features.

- A first hidden layer containing 64 neurons.

- A second hidden layer containing 32 neurons.

- An output layer producing probabilities for the four fetal state classes.

- Early stopping to prevent overfitting during training.

Decision Tree Model

A Decision Tree classifier was implemented within the federated learning framework to predict the outcomes of the fetal state. Decision Trees are supervised machine learning algorithms that recursively partition data into homogeneous groups by selecting features that maximize class separation. The model was configured with a maximum tree depth of five levels to reduce overfitting and improve generalization, while balanced class weighting was applied to address class imbalance within the dataset. Each participating client trained a local Decision Tree model using its local data, and predictions were aggregated through the federated learning process.

The Decision Tree model consisted of:

- An input layer receiving the selected maternal health features.

- Recursive binary splits based on feature values that maximize class separation.

- A maximum tree depth of five levels.

- An output layer producing probabilities for the four fetal state classes.

- Leaf nodes representing the predicted fetal state classes.

- Balanced class weighting to mitigate class imbalance effects.

Support Vector Machine Model

The Federated Support Vector Machine employed a Radial Basis Function (RBF) kernel, making it a non-linear classification model. Unlike a linear SVM, which separates classes using a single hyperplane, the RBF kernel maps the data into a higher-dimensional feature space, enabling the model to capture complex, non-linear relationships among maternal health variables and fetal state outcomes. This makes the model particularly suitable for clinical datasets

where risk factors interact in a non-linear manner. The SVM algorithm identifies the optimal decision boundaries by maximizing the margin between classes in a transformed feature space. Balanced class weighting was applied to mitigate class imbalance, while probability estimation was enabled to support federated aggregation of local model outputs. Each participating client trained a local SVM model using its own data, and the resulting class probabilities were combined through weighted federated averaging to produce the global prediction.

The SVM model consisted of:

An input layer receiving the selected maternal health features.

Transformation of the feature space using an RBF kernel.

Construction of optimal separating hyperplanes that maximize class margins.

A maximum tree depth of five levels.

Probability estimation for multi-class fetal state prediction.

Leaf nodes representing the predicted classes of fetal state.

Balanced class weighting to address class imbalance.

Gradient boosting Model

A Gradient Boosting classifier was implemented within the federated learning framework to predict the outcomes of the fetal state. Gradient Boosting is an ensemble learning method that constructs a sequence of weak decision tree learners, where each subsequent tree is trained to correct the prediction errors of the previous trees. The model combines multiple decision trees to improve predictive accuracy and reduce bias. In this study, the model was configured with 150 estimators, a learning rate of 0.05, a maximum tree depth of three, and a sub-sampling rate of 0.9. Local models were trained independently at each federated client, and prediction probabilities were aggregated using weighted federated averaging to produce the final global prediction.

The Gradient Boosting model consisted of:

An input layer receiving the selected maternal health features.

A sequence of decision trees trained iteratively.

Error correction through gradient-based optimization, where each new tree focuses on reducing residual errors from previous trees.

Aggregation of predictions from all trees to generate the final fetal state classification.

Federated aggregation of local model probabilities using weighted averaging across participating clients.

Federated Learning Process

The federated learning process was simulated using three clients as illustrated in Figure 8.7. The implementation involved the following steps: The balanced training dataset was partitioned among three simulated clients. Each client independently trained a local model using only its assigned data. Local model information was communicated to a central aggregation server. The server combined client outputs to generate a global federated model. The global model was evaluated using a separate testing dataset that was not involved in training.

Federated MLP Neural Network Training

For the neural network implementation, federated learning followed the Federated Averaging (FedAvg) algorithm proposed by McMahan et al. [226]. During each federated round: The current global model was distributed to each client. Each client trained the model locally using its private dataset partition. The updated model weights were transmitted to the server. The server aggregated client weights by averaging them to produce an updated global model.

Federated Decision Tree, Gradient Boosting, and SVM Training

Unlike neural networks, traditional machine learning models such as Decision Tree, Gradient Boosting, and SVM models do not naturally support parameter averaging. Therefore, each client trained independently trained a local model using its private data partition. Instead of aggregating model parameters, predictions from locally trained models were combined during inference to generate a global prediction. This ensemble-based federated approach allowed the evaluation of traditional machine learning algorithms within a federated learning environment while preserving data locality and privacy.

K-Nearest Neighbours (KNN) was used only in the non-federated analysis and was not included in the federated learning experiments because it is an instance-based learning algorithm that does not produce a compact set of model parameters suitable for aggregation. Unlike Decision Trees, Support Vector Machines, Gradient Boosting, and Multi-Layer Perceptrons, which learn transferable model representations, KNN relies on retaining and accessing training instances during prediction. In a federated learning environment, this would require the exchange of patient records or local datasets between participating institutions, which is inconsistent with the privacy-preserving design of federated learning. Therefore, models supporting local training and parameter aggregation were selected for federated experiments.

ANC Uganda Data Fetal State Prediction Federated Learning Models' Performance Comparative Analysis

The ANC Uganda dataset was first transformed into a patient-level table and records with missing fetal state labels were excluded. To ensure a realistic evaluation and prevent information leakage, BMI and Blood Pressure were removed from the predictor variables because BMI is derived directly from weight and height, while blood pressure is strongly associated with fetal health outcomes and could artificially inflate model performance. The final set of characteristics consisted of Age, Weight, Height, Gestational Age, Parity, Hemoglobin (Hb), Mid-Upper Arm Circumference (MUAC) and ANC visit timing.

Table 8.10: Mean $\pm$ standard deviation of federated learning model performance across ten experimental iterations for fetal state prediction using FAIR ANC Uganda data.

Federated Model	Fetal State	Precision	Recall	F1-Score
Gradient Boosting	Distressed	0.8160 $\pm$ 0.0823	0.8600 $\pm$ 0.0583	0.8341 $\pm$ 0.0465
	Non-reassuring	0.6882 $\pm$ 0.0487	0.5220 $\pm$ 0.0817	0.5909 $\pm$ 0.0649
	Reassuring	0.6579 $\pm$ 0.0667	0.8500 $\pm$ 0.1026	0.7400 $\pm$ 0.0737
	Well	0.7730 $\pm$ 0.0367	0.8500 $\pm$ 0.0384	0.8089 $\pm$ 0.0282
Decision Tree	Distressed	0.7516 $\pm$ 0.1120	0.8250 $\pm$ 0.0873	0.7813 $\pm$ 0.0812
	Non-reassuring	0.6974 $\pm$ 0.0775	0.4700 $\pm$ 0.0744	0.5554 $\pm$ 0.0554
	Reassuring	0.5917 $\pm$ 0.0576	0.8278 $\pm$ 0.0911	0.6851 $\pm$ 0.0428
	Well	0.7743 $\pm$ 0.0425	0.8537 $\pm$ 0.0700	0.8090 $\pm$ 0.0270
SVM	Distressed	0.6643 $\pm$ 0.0644	0.7700 $\pm$ 0.1030	0.7084 $\pm$ 0.0612
	Non-reassuring	0.5702 $\pm$ 0.0417	0.4680 $\pm$ 0.0515	0.5128 $\pm$ 0.0410
	Reassuring	0.5632 $\pm$ 0.0673	0.7444 $\pm$ 0.1117	0.6382 $\pm$ 0.0713
	Well	0.7138 $\pm$ 0.0375	0.7074 $\pm$ 0.0452	0.7098 $\pm$ 0.0346
MLP	Distressed	0.3068 $\pm$ 0.1574	0.4750 $\pm$ 0.3010	0.3555 $\pm$ 0.1903
	Non-reassuring	0.3837 $\pm$ 0.1416	0.2860 $\pm$ 0.2193	0.2982 $\pm$ 0.1213
	Reassuring	0.3259 $\pm$ 0.1225	0.7333 $\pm$ 0.2639	0.4456 $\pm$ 0.1555
	Well	0.5598 $\pm$ 0.0899	0.4000 $\pm$ 0.1655	0.4458 $\pm$ 0.1250

The dataset was divided into training and testing subsets prior to model development. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training set. The balanced training data set was later partitioned and distributed among three simulated federated clients representing healthcare facilities. Each client trained local machine learning models using its own partition of the data without sharing raw patient records, thereby preserving data privacy and locality throughout the federated learning process.

Four machine learning models were evaluated within the federated learning environment: Multi-Layer Perceptron (MLP), Decision Tree, Support Vector Machine (SVM) and Gradient Boosting. The federated neural network used the Federated Average (FedAvg) approach, where the weights of the local model were aggregated to update the global model. For Decision Tree, Gradient Boosting, and SVM models, local models were trained independently on each client, and their predictions were aggregated to produce a global prediction. The models were evaluated on an independent test set using precision, recall, F1-score, and overall classification accuracy. The reported precision, recall, F1-score, and accuracy values represent the mean $\pm$ standard deviation obtained across ten independent experimental iterations. Reporting both statistics provides a more reliable assessment of model stability and robustness by accounting for variation arising from random train-test partitioning and federated client allocation. Results are reported as mean accuracy $\pm$ standard deviation across ten independent experimental iterations.

Table 8.11: Global accuracy of federated learning models across ten iterations using FAIR ANC Uganda data.

Federated Model	Global Accuracy
Decision Tree	0.7113 $\pm$ 0.0321
Gradient Boosting	0.7359 $\pm$ 0.0332
SVM	0.6366 $\pm$ 0.0294
MLP	0.4127 $\pm$ 0.0390

The overall classification accuracies obtained by the federated models were 71.83% for Federated Gradient Boosting, 71.13% for Federated Decision Tree, 66.20% for Federated SVM, and 40.85% for Federated MLP. Among the evaluated models, Federated Gradient Boosting achieved the highest classification accuracy and demonstrated the most balanced performance across the fetal state categories.

Federated Gradient Boosting achieved strong recall for the Distressed and Reassuring State classes while maintaining competitive precision and F1-scores across all categories. The Federated Decision Tree model produced comparable results and performed particularly well for the Well class. Federated SVM achieved moderate performance, but showed reduced effectiveness in distinguishing some categories of fetal state. The Federated MLP model produced the lowest accuracy, indicating that the available data set size and feature representation were less suitable for neural-network-based learning in the federated setting.

The performance obtained in this experiment is lower than that observed in earlier experiments because potential leakage variables were removed and the evaluation was conducted using a strict train-test separation strategy. Although the resulting accuracy values are lower, they provide a more realistic estimate of model performance on unseen patient data. Consequently,

Table 8.12: Average local client performance across ten federated iterations.

Federated Model	Client	Local Accuracy (Mean)
Decision Tree	Client 1 (n=292)	~0.65
Decision Tree	Client 2 (n=288)	~0.68
Decision Tree	Client 3 (n=288)	~0.65
Gradient Boosting	Client 1 (n=292)	~0.67
Gradient Boosting	Client 2 (n=288)	~0.70
Gradient Boosting	Client 3 (n=288)	~0.66
SVM	Client 1 (n=292)	~0.53
SVM	Client 2 (n=288)	~0.52
SVM	Client 3 (n=288)	~0.62
MLP	Client 1 (n=292)	~0.42
MLP	Client 2 (n=288)	~0.39
MLP	Client 3 (n=288)	~0.29

Federated Gradient Boosting was selected as the best-performing federated learning model for fetal state prediction based on the ANC Uganda dataset.

Prediction Based on User Input Using a Federated Learning Model

A function ‘predict_fetal_state’ was defined to make predictions using the best model. Processed user input, calculated BMI, and handled missing values. Predictions were made, and the result was printed with a formatted message showing the user’s input values.

Table 8.13: Comparison of Predictions Generated by the Traditional and Federated Gradient Boosting Models

Example	Age	Gest. Age	Weight	Height	BP	Traditional GB	Federated GB
1	25	36	70	165	110	Well	Well
2	50	40	90	165	110	Non-reassuring	Non-reassuring
3	24	28	50	170	100	Distressed	Distressed
4	30	34	70	165	150	Reassuring	Reassuring

Table 8.13: Comparison of the predictions of the fetal state produced by the Traditional Gradient Boosting and Federated Gradient Boosting models for four representative patient

cases from the Uganda dataset of the ANC. The table illustrates the consistency and differences between traditional and federated learning approaches when applied to identical patient inputs.

Performance Summary and Implications

The results obtained in this study demonstrate both the potential and current limitations of applying traditional and federated machine learning techniques to the prediction of fetal state using the FAIR ANC Uganda dataset. In all experimental settings, Gradient Boosting consistently achieved the highest performance, demonstrating superior precision, recall, F1-score, and overall classification accuracy compared to the other models evaluated. The model also exhibited relatively low standard deviations in repeated experimental runs, indicating stable and reliable performance. Although MLP produced competitive results in some categories, its overall performance remained lower than that of Gradient Boosting. Furthermore, the experiments involving noisy data showed that Gradient Boosting was comparatively robust, while other models exhibited greater sensitivity to data perturbations.

Federated learning experiments produced lower performance than traditional machine learning approaches, which was expected given the distributed nature of training and the challenges associated with client heterogeneity, model aggregation, and limited local data availability. Nevertheless, Federated Gradient Boosting achieved the highest global accuracy among the federated models and maintained balanced performance across the different fetal state categories. All federated models demonstrated a reasonable degree of stability across repeated iterations, although variations in performance were observed due to differences in data distribution and client-level training.

From a FAIR data perspective, this study demonstrates how FAIRification can transform maternal health records into machine-readable resources that support advanced analytics while enabling interoperability and data reuse. However, the available data set remains relatively small and does not fully capture the diversity of maternal healthcare conditions in Uganda. In addition, a substantial proportion of antenatal care records in Uganda are still maintained in paper-based systems, limiting the amount of digitized data available for analysis. Consequently, the results presented in this study should be viewed primarily as an indication of emerging trends rather than definitive clinical evidence.

The observed performance patterns are consistent with expectations for datasets of this size. The performance of traditional and federated models is expected to improve with access to larger, more diverse, and FAIR-compliant datasets collected from multiple healthcare facilities. Therefore, more testing, external validation, and large-scale evaluations are required to assess the generalizability and clinical applicability of the proposed approach.

The comparison between the predictions generated by the Traditional Gradient Boosting and Federated Gradient Boosting models demonstrated a high degree of agreement across representative patient cases, indicating that the federated approach was capable of capturing similar predictive patterns despite being trained in a distributed environment.

Despite the performance gap between traditional and federated learning, federated learning remains a highly promising direction for healthcare analytics. Its ability to train models across distributed datasets without sharing raw patient information provides significant privacy, security, and collaborative advantages. As healthcare institutions continue to digitize records and adopt FAIR data principles, federated learning offers a practical pathway to scalable, privacy-preserving, and data-driven maternal and fetal healthcare decision support systems. Although

further evaluation is required, the results of this study suggest that FAIR-enabled federated learning is an important and viable direction for future healthcare research and implementation.

8.4 Discussion

The results of this study demonstrate both the potential and limitations of applying federated learning and traditional machine learning techniques to maternal and fetal health prediction using the ANC Uganda dataset. While the initial centralized experiments showed strong performance, the federated learning setup produced more moderate and realistic results after applying strict preprocessing and leakage prevention measures.

A key improvement in this study was the removal of potential data leakage variables, specifically BMI and blood pressure. Although these features are clinically relevant, BMI is derived directly from weight and height, and blood pressure is strongly correlated with fetal health outcomes. Their exclusion ensured that the models were not indirectly exposed to target-related information, resulting in more reliable and generalizable performance estimates.

Furthermore, the application of SMOTE solely on the training dataset ensured that class balancing did not introduce bias into the evaluation process. This strict separation between training and testing data provides a more realistic assessment of model behavior in real-world clinical deployment scenarios, where unseen patient data must be classified without prior exposure.

The federated learning experiments showed that distributed model training is feasible for fetal state prediction; however, performance varies significantly across model types. Federated Gradient Boosting achieved the highest performance among all evaluated models, followed closely by the Federated Decision Tree model. Both models demonstrated relatively balanced precision and recall across fetal state classes. In contrast, the Federated SVM showed moderate performance, while the Federated MLP exhibited the weakest results, indicating that neural network models may require larger datasets or more optimized architectures in federated settings.

Compared to centralized machine learning approaches, federated learning produced lower accuracy values. However, this reduction is expected due to data partitioning across clients and the absence of centralized training. Despite this, federated learning provides a significant advantage in terms of privacy preservation, as sensitive patient data remains localized within each institution and is never directly shared.

The multilayer perceptron (MLP) model demonstrated the lowest performance among the evaluated algorithms. Neural network models generally require larger datasets to effectively learn complex feature representations and avoid overfitting. In this study, only 710 patient records were available after target generation and data cleaning, which limited the amount of information available for training. Tree-based models such as Gradient Boosting and Decision Trees are often more suitable for relatively small tabular healthcare datasets and therefore achieved superior performance. Future studies incorporating substantially larger datasets may allow deep learning models such as MLPs to realize their full predictive potential and potentially outperform traditional machine learning approaches.

Overall, the findings highlight a fundamental trade-off between predictive performance and data privacy. While centralized models achieve higher accuracy, federated learning offers a scalable and privacy-preserving alternative suitable for healthcare environments where data

Conclusion

sharing is restricted. This makes federated learning particularly valuable for multi-institutional collaboration in resource-constrained settings such as Uganda.

The machine learning and federated learning experiments conducted in this study demonstrated the potential value of FAIR-compliant health data for advanced healthcare analytics. Using FAIR ANC Uganda data, predictive models were successfully developed to classify fetal health states, with the federated Gradient Boosting model achieving the highest overall accuracy. These findings indicate that improved data management practices, standardized data collection, and FAIR data implementation can support the development of data-driven clinical decision support tools. Furthermore, the federated learning approach showed that predictive models can be trained collaboratively while preserving patient privacy, making it a promising strategy for future healthcare information systems in Uganda. The federated learning experiments demonstrated that FAIR-compliant health data can be effectively used within privacy-preserving machine learning architectures. Gradient Boosting achieved the highest overall performance, followed by Decision Tree and SVM models. The relatively lower performance of the MLP model is likely attributable to the limited size of the available dataset. Neural network-based approaches generally require substantially larger datasets to learn robust feature representations and fully exploit their modeling capacity. With larger FAIR-compliant datasets collected from multiple healthcare facilities, improved predictive performance would be expected across all evaluated models, particularly for deep learning approaches such as MLP.

Future improvements should focus on optimizing federated aggregation strategies, improving feature representation, and exploring advanced ensemble methods to further close the performance gap between centralized and federated approaches while maintaining strict privacy constraints.

8.5 Conclusion

This study investigated the integration of FAIR data principles with both traditional machine learning and federated learning approaches for fetal state prediction using the ANC Uganda dataset. The primary objective was to develop predictive models capable of classifying fetal health conditions while preserving patient privacy through distributed learning.

A key contribution of this work is the development of a leakage-free experimental pipeline. By removing derived and highly correlated variables such as BMI and blood pressure, and by applying SMOTE exclusively to the training dataset after a strict train-test split, the study ensured that evaluation was conducted under realistic and unbiased conditions. This improved the reliability and clinical relevance of the reported results.

Traditional machine learning models demonstrated strong predictive performance in earlier centralized experiments; however, the federated learning framework provided a more realistic assessment of model generalization in distributed healthcare environments. Among the federated models, Gradient Boosting achieved the highest performance with a global accuracy of 71.83%, followed closely by the Decision Tree model at 71.13%. The SVM model showed moderate performance, while the MLP model performed comparatively lower, indicating that neural networks may require larger datasets or further optimization in federated settings.

The results demonstrate that FAIR-compliant healthcare data can successfully support machine learning and federated learning workflows for maternal healthcare analytics. Among the evaluated models, the federated Gradient Boosting classifier achieved the highest overall accuracy of 73.59%, followed by the Decision Tree model with 71.13%. These findings indicate that

Conclusion

machine learning techniques can effectively utilize FAIR health datasets to predict fetal health states and support data-driven clinical decision making.

The findings of this chapter demonstrate that FAIR data support and strengthen the accessibility and usability of health data for machine learning applications. The transformation of health information into FAIR-compliant formats enabled the successful development of clustering, predictive analytics, and federated learning models using both the OPD Uganda and ANC Uganda datasets.

Regarding Sub-Research Question 1, FAIR data facilitated the application of machine learning algorithms for clustering and prediction by improving data interoperability, accessibility, and reusability. The clustering analyses revealed meaningful groupings of patients and diagnoses that can support healthcare planning and decision-making.

Regarding Sub-Research Question 2, the FAIR ANC Uganda dataset was sufficiently rich to support the prediction of pregnancy risks and fetal states. The predictive models successfully identified relationships between maternal characteristics and fetal outcomes, with Gradient Boosting achieving the strongest predictive performance among the evaluated models. The federated learning implementation further demonstrated that machine learning can be applied within a distributed healthcare environment while preserving patient privacy.

The federated learning experiments are particularly relevant to real-world healthcare systems because patient data remain within local institutions and only model parameters are exchanged during training. This approach reduces the need for centralized data sharing while still enabling collaborative model development across multiple healthcare facilities. The results therefore illustrate how FAIR data and federated learning can jointly support secure, scalable, and privacy-preserving health data analytics.

The federated learning framework further demonstrated that predictive models can be trained without sharing sensitive patient-level records between participating institutions. Instead, only model parameters and prediction probabilities were exchanged during aggregation, preserving patient privacy while enabling collaborative model development. This characteristic makes federated learning particularly suitable for healthcare environments where data sharing is restricted by privacy, ethical, or regulatory considerations.

Although promising results were obtained, the relatively limited dataset size constrained overall model performance, particularly for the MLP neural network model. Future work involving larger and more diverse datasets is expected to improve predictive accuracy, model generalization, and the effectiveness of federated learning systems in real-world healthcare settings. Overall, the study confirms that the combination of FAIR data principles, machine learning, and privacy-preserving federated learning provides a viable approach for developing scalable and trustworthy healthcare prediction systems.

The results highlight an important trade-off between predictive performance and data privacy. While centralized models typically achieve higher accuracy due to access to complete datasets, federated learning enables collaborative model training without sharing sensitive patient data. This makes it particularly suitable for healthcare systems where data privacy regulations and institutional constraints limit centralized data collection.

Overall, the study demonstrates that federated learning is a viable and privacy-preserving approach for fetal state prediction, especially when combined with careful preprocessing and leakage prevention strategies. Although there is a measurable reduction in performance compared to centralized learning, the approach provides a more realistic and ethically robust framework for multi-institutional healthcare analytics.

Conclusion

Future work should focus on improving federated optimization strategies, exploring advanced ensemble aggregation methods, and incorporating richer clinical features to further enhance predictive performance while maintaining strict privacy guarantees. Additionally, expanding the dataset across more healthcare institutions would improve model robustness and generalization across diverse populations.

Conclusion

Chapter 9

Discussion, Outlook and Conclusion

9.1 Observations Answering the Research Questions

The study sought to address the core question that revolved around whether FAIR data principles can overcome the issues of data non-interoperability and limited reuse within Africa. This question is vital to determine the viability of FAIR data to solve fundamental problems in digital health environments.

To answer this, the following specific research questions were defined:

To what extent is the lack of policy on data management, data sharing, and data analysis a contributing factor to the lack of use of clinical and health data in Uganda, Africa?

Can the introduction of systematic use of the FAIR principles help overcome the challenges of interoperability and reusability to the development and maintenance of digital health data in Uganda?

How can data analytics of health data be propagated, integrated, and maintained through FAIR data?

Can FAIR Data support and strengthen the accessibility of different types of health data, as well as aggregated health data analytics and machine learning models?

The unique contribution of this research lies in its demonstration of how the principles of FAIR data can be successfully implemented in resource-limited healthcare settings, such as Uganda. Specifically, this study shows that applying FAIR principles not only addresses data interoperability and reusability challenges but also facilitates privacy-preserving accessibility, hence opportunities like machine learning and federated learning improvements, which are crucial for advancing collaborative research among different institutions. By proving the feasibility of FAIR data in these settings, this research fills a critical gap in the existing literature.

The findings of this research affirm the enormous possibilities of implementing FAIR data principles in transforming digital health systems in Uganda. One of the findings is that the absence of well-coordinated policies on handling, sharing, and utilising big data makes their use nearly impossible. This gap poses difficulties in interoperability and data reuse, which are key factors in developing a sustainable framework of effective and efficient digital health systems. These are the problems that systematic deployment of FAIR principles can help solve, as shown by the experience of creating the VODAN (Virus Outbreak Data Network) in Africa and Asia. This endeavor embodies the virtues of the FAIR principles when it comes to handling and merging health data, especially after viral epidemics [30, 230].

The focus of this research is not on improving individual patient care but rather on using aggregated health data to identify population-level patterns. By analyzing large datasets, the research highlights key trends in disease outbreaks, health behaviors, and treatment outcomes, which can inform public health policy and resource allocation at the national level. This approach has the potential to improve healthcare systems by providing insights that go beyond individual treatment and contribute to the systemic management of health challenges.

Despite the promising outcomes of applying FAIR principles, several bottlenecks were encountered during the implementation phase, especially when transitioning from paper-based records to digital systems. Resistance from healthcare staff accustomed to paper-based processes was a significant hurdle. Furthermore, challenges related to data standardization and

Implications for Practice

ensuring interoperability among diverse health institutions were prevalent. However, by training personnel and emphasizing the benefits of FAIR data, these challenges were mitigated, and health facilities gradually embraced the digital shift. Moreover, some issues related to data security and privacy concerns arose, but the careful implementation of FAIR's access control mechanisms helped alleviate these concerns.

One of the main observations in this research is that the adoption of the FAIR principles has the potential to contribute significantly to the interoperability and reusability of data. It is easy to understand why the principles are advantageous after synthesizing the ideas from this thesis and especially the case studies where the FAIR principles were applied to existing data management architecture in Uganda. Generally, FAIR data have contributed to improving data access and reusability by researchers among multiple organisations and improving disease management [231].

In addition, the application of FAIR principles has shown promising outcomes regarding data analytics, as applying these principles ensures that data are easily accessible for analysis. Consequently, the analyzes gain an enhanced potential to provide the means for developing knowledge that is useful in informing the policies that are required and the practices in data management employed. This is especially the case in Uganda, where the proper use of data could significantly improve healthcare systems in the country. In addressing the research questions, the study found that the lack of a data management policy was a significant barrier to effective use of health data in Uganda. The introduction of FAIR principles helped overcome these barriers by establishing a framework for data interoperability, secure sharing, and reusability. As a result, machine learning models have been more effective in predicting disease outbreaks and assessing health risks. Furthermore, systematic use of FAIR data has improved the accessibility of various types of health data, improving the ability to track health trends and guide policy decisions.

The principles of FAIR have helped to maintain privacy-preserving accessibility in the management of health data. This has the potential to foster trust and increase data access and re-usability among different countries. The study identified several cases where FAIR data principles were applied to health data analysis to ensure access and reusability under well-defined terms. This is relevant as data privacy are increasingly becoming core issues in data handling [232]. The main contribution of this research lies in its demonstration that the application of FAIR data principles can significantly improve the accessibility of health data, the collaboration between institutions, and the reuse of health data for machine learning in a resource-limited setting. This work provides a conceptual framework and a practical example of integrating FAIR principles in digital health systems, paving the way for more effective disease management, health policy decisions, and predictive modeling in Uganda and similar settings across Africa.

9.2 Implications for Practice

The application of the FAIR principles provides a variety of solutions to the problems associated with the sharing and reuse of data in the Ugandan digital health system. Therefore, the application of the FAIR principles supports research efforts and data protection. The principles of FAIR raise the possibility that health professionals, policy makers, and data managers in Uganda should embrace the management of health data in a manner that would allow them to provide efficient public health services and promote health outcomes in Uganda.

Limitations of the Study

This thesis illustrates the practical implications of advancing patient care through the efficient handling of health information. When the FAIR principles are applied to healthcare practices, patient data can be accessed to perform data analytics whenever needed. The findings of this can be used to provide patients with better care promptly [233]. Moreover, applying FAIR data principles to the monitoring of public health challenges supports the management of such public health phenomena through improved data collection and analysis [234].

Long-term partnerships are advantageous because overall cost benefits will be realized through improved data management, as redundancy of original data structures will not be as much of an issue. Consequently, healthcare facilities can improve operational efficiency and reduce data storage and management costs [235]. In addition, the application of FAIR data principles can help develop innovative solutions for healthcare systems. For example, using high-quality, interoperable data can make machine learning and other advanced analytical techniques more productive. Contributes to the development of predictive machine learning models and other tools that can be used to improve health care services and outcomes [29].

One of the most significant benefits realized after FAIRifying health data in Uganda was the improved access of data between institutions, enabling better collaboration among healthcare providers. With improved interoperability and data sharing, healthcare professionals were able to access real-time analytics, leading to better-informed decision-making in disease management and resource allocation. Furthermore, the increased data reusability fostered innovation in predictive modeling for disease outbreaks, improving public health responses. [29]

9.3 Limitations of the Study

While this research provides various valuable insights into the adoption of FAIR principles, several limitations need to be acknowledged. This study was primarily focused on some of the health facilities in Uganda and, therefore has not been adopted throughout Uganda; therefore, there are limitations in the generalizability of this study in other related settings without further validation. Furthermore, the implementation of FAIR principles has been successfully implemented within the VODAN-Africa project, which comprises a variety of countries, thereby indicating a level of technological readiness within these countries. Although insights from the VODAN initiative form a rich case study, more data from different health facilities could strengthen the conclusions made. Moreover, the research did not dive fully into the economic effects of FAIR. The principles of FAIR have helped to maintain privacy-preserving accessibility in the management of health data. This has the potential to foster trust and increase data access and re-usability among different countries. The study identified several cases where FAIR data principles were applied to health data analysis to ensure access and re-usability under well-defined terms. This is relevant as the FAIR principles can potentially serve as a strength for decision-makers [236]. Adopting FAIR data principles involves changing the practices employed in managing, reproducing, and sharing data, which is difficult in environments where conventional practices have been established, for example, the use of paper registers. Mitigating this resistance will be a challenge that will require specific initiatives to sensitize and involve the various stakeholders at different levels [237].

9.4 Call to Action

To reach the tipping point of the application of the FAIR data principles, the combined efforts of stakeholders and policymakers to prioritize the development and implementation of supportive policies and frameworks are required. Health professionals and data managers must be trained on the relevance of FAIR principles to embrace them FAIR principles, and as a result, they would embrace these and build their capacity on the principles for better application. More research is needed to create and share appropriate guidelines and innovative methodological approaches for the process followed in implementing the FAIR data principles [238]. Moreover, it is now important for international organizations and donors to support the initiatives that promote the adoption of the FAIR principles in developing countries. Such organizations can contribute funding and technical assistance to establish the framework and expertise necessary for the implementation of FAIR data [239]. Stakeholders are encouraged to embrace FAIRification and data collaboration, as this will help improve the health of the population [240]. The sustainability of FAIR data practices must be checked regularly. This will help identify the areas that require more growth to improve the effective use of these principles. Therefore, the recommendations of stakeholders should be taken into account while the process of enhancing data management practices is ongoing [241].

The adoption of FAIR data principles offers substantial benefits, including streamlined data access, reduced duplication in data management, and enhanced data security. To realize these benefits on a larger scale, healthcare professionals and policymakers must prioritize the integration of the principles of FAIR into healthcare practices, ensuring that health data are not only secure but also useful for improving patient outcomes and public health management.

To fully realize the potential of FAIR data, health professionals, policy makers, and international organizations need to collaborate on the implementation of supportive policies and frameworks. Capacity building and training programs for healthcare professionals will be key to ensuring that FAIR data principles are adopted and integrated into routine healthcare practices, ultimately improving the quality of health systems.

9.5 Future Research Directions

Future studies should promote the development and enactment of political strategies that facilitate the use of FAIR data in Uganda. This encompasses policies and practices that meet the international standards of data collection, storage, utilization, and analysis thus encouraging policy development and implementation. Such guidelines should also incorporate the unique circumstances that may surround the handling of data [242]. Stakeholders should also be educated about the impacts that policy interventions have on information processing and the resulting health consequences of implementing them.

To support the principles of FAIR, educational courses for capacity building and training must be developed and launched to enhance the capacity of healthcare professionals and data managers to implement these principles. These courses should be able to provide information presentations that are aimed at ensuring the maximum level of awareness is achieved with the key competencies required to implement the FAIR data principles [243]. The interventions provided by these training programs should be assessed to understand the effectiveness of the implementation.

Further studies should focus on the development of the technical infrastructure and tools

Future Research Directions

that support FAIR principles. In this sense, such solutions and initiatives include the financial investment in digital solutions and platforms that aim to communicate and implement the principles of FAIR in data interoperability, access, and reuse [244]. Investigations should also factor in the costs and efficiencies of the options available in different industries and identify how different management strategies could be applied to those options. There is a need for advanced directions in applying the FAIR data principles to other progressive solutions, including artificial intelligence and blockchain. Studies in this field can find best practices that use these technologies for the FAIR data framework [248]. Researchers should also assess the preparedness of various health systems for these technologies and identify various hurdles to implementation.

Case studies should be conducted in areas or institutions that have successfully used the FAIR data principles, as this could reveal recommendations that can be used in different contexts. Relative research can be useful in defining the factors that can be influential in the effective implementation of the FAIR approach [245]. Recording and sharing such practices can help other countries and organizations to apply FAIR principles more effectively since they are based on the experience of other nations.

For future research, it is relevant to investigate the changes in health outcomes and efficiency of healthcare systems due to the implementation of the principles of FAIR data for impact assessment and evaluation. This involves evaluating how well FAIR data have improved data quality and decision-making while being of value to public health initiatives [246]. Actual patient health studies that monitor outcome indicators will be more informative about the practical effects of implementing FAIR data principles over time.

Future work should also analyze how to encourage stakeholder participation and collaboration, such as government bodies, healthcare organizations, and societies, to embrace the principles of FAIR data. It is crucial to have multi-stakeholder engagement approaches that can increase the sustainability and FAIRness of data [247]. Studies should be carried out on the patterns of interconnection between participants and knowledge of the approaches to cooperation.

Future research should investigate the evolving role of the GO FAIR initiative as a lightweight long-term coordination mechanism for the implementation of FAIR. As GO FAIR has transitioned from a centrally funded project into a loosely federated international network of offices, coordinated through the GO FAIR Foundation, further studies are needed to assess how such non-bureaucratic governance models can sustainably support FAIR adoption across diverse regions and disciplines. In particular, research should explore the effectiveness of GO FAIR's emerging *FAIR co-pilot* role, which focuses on guiding communities to implement FAIR principles properly rather than prescribing uniform solutions.

Additionally, future studies should examine how national and regional GO FAIR offices, operating with minimal resources, can maintain networking value, knowledge exchange, and interpretative guidance in the absence of central funding. Comparative analyzes between regions could provide insight into how GO FAIR coordination interacts with local policy environments, capacity-building initiatives, and technical infrastructures. Such research would contribute to understanding whether federated coordination models, supported through modest pooled resources and collaborative governance, can offer a scalable and resilient pathway for sustaining FAIR principles globally [117]. When it comes to health data, it is pertinent to highlight ethical and privacy considerations that are relevant to the discussion of the principles of FAIR data. It is also important that future studies emphasize best practices in data privacy and security

Conclusion

while at the same time encouraging data reuse [248]. This covers the issues of data accessibility, privacy protection, and the possibilities of finding a workable formula on how to best address these objectives.

Future work should also include highly specific, versioned analyzes of economic aspects of the adoption of the FAIR data principles. This would include a cost-benefit analysis, identifying the main financial implications and proposing financing models geared toward long-term implementation [249]. An accurate perception of the economic impact of FAIR data is another beneficial factor, as it would formulate a base to decide on how funds are to be spent.

Longitudinal patient health studies that include monitoring outcome indicators will be more informative on the practical effects of implementing the FAIR data principles over time. Such studies can help reveal trends, determine the long-term impact of data interoperability, and establish the new policies required [250]. Investigations should also be conducted to discover the determinant constructs that make up sustainable FAIR data.

With the ongoing development of digital health and biomedical research, the ability to access distributed data will become one of the major areas that will promote innovation. The emphasis should no longer be on having the data centralized at a single point but accessible in a distributed architecture. This distributed access is also key to facilitate collaboration and improve efficiency in health research. The fact that data need to be not only available but also findable and usable across geographies cannot be overstated, and this is particularly the case as the amount of health data continues to increase exponentially.

Researchers should be able to access various datasets distributed in various hospitals, health organizations, and even countries. It is imperative to ensure that even when data are made accessible, no security or privacy being jeopardized, and the new developments in technologies such as federated learning and data encryption may allow such data accessibility without privacy concern. Effective data sharing and access have been a major determinant of the future of digital health, and it will be critical to establish systems that enable researchers to access data stored in distributed sources to address health issues in the population and develop new scientific knowledge.

9.6 Conclusion

The use of FAIR data principles is a good solution to address the issues of data interoperability and reuse within the Ugandan context of digital health. These FAIR principles can improve the use of health data both in terms of efficiency and effectiveness by closing the gaps in the policy landscape, data integration, analysis, and hence data reuseability through privacy-preserving access. Therefore, the conclusions of this study serve as a basis for future projects that aim to improve data management approaches in Uganda and other countries. The directions for future research presented in this chapter could potentially enable new contributions to this body of knowledge and expand the knowledge and application of FAIR data principles in Uganda and the rest of Africa's health sector. These initiatives will also help promote better health and advance the availability of information to inform decision-making, as well as strengthen the development of the health system in the region. By embracing FAIR data principles, Uganda and other African nations can unlock the full potential of their health data, drive evidence-based policy, and decisively transform the future of healthcare delivery across the continent.

Bibliography

- [1] D. Le Blanc, "Towards integration at last? The sustainable development goals as a network of targets," *Sustainable Development*, vol. 23, no. 3, pp. 176-187, 2015.
- [2] G. Kobusinge, "Health Information Systems Interoperability: Towards a Managing as Designing Approach," 2021.
- [3] M. Basajja, M. Nambobi, and K. Wolstencroft, "Possibility of enhancing digital health interoperability in Uganda through FAIR data," *Data Intelligence*, vol. 4, no. 4, pp. 899-916, 2022.
- [4] B. Mittelstadt, "Designing the health-related internet of things: ethical principles and guidelines," *Information*, vol. 8, no. 3, p. 77, 2017.
- [5] J. S. Marwaha, A. B. Landman, G. A. Brat, T. Dunn, and W. J. Gordon, "Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation," *NPJ Digital Medicine*, vol. 5, no. 1, p. 13, 2022.
- [6] C. Alvarez-Romero et al., "FAIR4Health: findable, accessible, interoperable and reusable data to foster health research," *Open Research Europe*, vol. 2, 2022.
- [7] R. Plug et al., "FAIR and GDPR Compliant Population Health Data Generation, Processing and Analytics," in *SWAT4HCLS*, 2022, pp. 54-63.
- [8] A. B. Labrique, L. Vasudevan, E. Kochi, R. Fabricant, and G. Mehl, "mHealth innovations as health system strengthening tools: 12 common applications and a visual framework," *Global Health: Science and Practice*, vol. 1, no. 2, pp. 160-171, 2013.
- [9] C. B. Aranda-Jan, N. Mohutsiwa-Dibe, and S. Loukanova, "Systematic review on what works, what does not work and why of implementation of mobile health (mHealth) projects in Africa," *BMC Public Health*, vol. 14, no. 1, pp. 1-15, 2014.
- [10] M. Reglitz, "The socio-economic argument for the human right to internet access," *Politics, Philosophy Economics*, p. 1470594X231167597, 2023.
- [11] U. Nations, "Transforming our world: The 2030 agenda for sustainable development," New York: United Nations, Department of Economic and Social Affairs, 2015.
- [12] S. Goal, "9: Build Resilient Infrastructure, Promote Sustainable Industrialization and Foster Innovation," United Nations. Link available at:

Bibliography

- <https://www.un.org/sustainabledevelopment/infrastructureindustrialization/> (accessed 20.12.2019).
- [13] F. Griffiths et al., "Mobile consulting (mConsulting) and its potential for providing access to quality healthcare for populations living in low-resource settings of low-and middle-income countries," *Digital Health*, vol. 6, p. 2055207620919594, 2020.
 - [14] U. ESCAP, "SDG 17: Partnerships for the goals: strengthen the means of implementation and revitalize the global partnership for sustainable development," 2022.
 - [15] The Guardian, "'Gamechanger': Uganda launches drone delivering HIV drugs to remote islands," <https://www.theguardian.com/global-development/2021/may/04/gamechanger-uganda-launches-drone-delivering-hiv-drugs-to-remote-islands> (accessed 28 November, 2023).
 - [16] J. Driessen, D. Settle, D. Potenziani, K. Tulenko, T. Kabocho, and I. Wadembere, "Understanding and valuing the broader health system benefits of Uganda's national Human Resources for Health Information System investment," *Human Resources for Health*, vol. 13, no. 1, pp. 1-9, 2015.
 - [17] J. C. Spero, P. A. McQuide, and R. Matte, "Tracking and monitoring the health workforce: a new human resources information system (HRIS) in Uganda," *Human Resources for Health*, vol. 9, pp. 1-10, 2011.
 - [18] D. C. Mays et al., "Supporting and retaining Village Health Teams: an assessment of a community health worker program in two Ugandan districts," *International Journal for Equity in Health*, vol. 16, pp. 1-10, 2017.
 - [19] L. H. Kamulegeya, J. M. Bwanika, D. Musinguzi, and P. Bakibinga, "Continuity of health service delivery during the COVID-19 pandemic: the role of digital health technologies in Uganda," *The Pan African Medical Journal*, vol. 35, no. Suppl 2, 2020.
 - [20] V. M. Kiberu, R. E. Scott, and M. Mars, "Assessment of health provider readiness for telemedicine services in Uganda," *Health Information Management Journal*, vol. 48, no. 1, pp. 33-41, 2019.
 - [21] B. Heidt et al., "Point of care diagnostics in resource-limited settings: A review of the present and future of PoC in its most needed environment," *Biosensors*, vol. 10, no. 10, p. 133, 2020.
 - [22] D. Luyombya, "Framework for effective public digital records management in Uganda," UCL (University College London), 2010.
 - [23] W. M. Tierney et al., "Experience implementing electronic health records in three East African countries," in *MEDINFO 2010*, IOS press, 2010, pp. 371-375.
 - [24] L. Kamulegeya, C. Chikwata, W. Lubega, D. Musinguzi, and J. M. Bwanika, "Lessons Learnt from Initial Deployments of Rocket Health Telemedicine Service to Deliver Last Mile Medical Services in Uganda," in *2021 IST-Africa Conference (IST-Africa)*, 2021: IEEE, pp. 1-11.

Bibliography

- [25] R. B. Ness and J. P. Committee, "Influence of the HIPAA privacy rule on health research," *Jama*, vol. 298, no. 18, pp. 2164-2170, 2007.
- [26] J. Nagaba, "The Data Protection Officer within the context of Uganda, Kenya, Rwanda and the EU-GDPR."
- [27] M. E. Martone, "FORCE11: building the future for research communications and e-scholarship," *Bioscience*, vol. 65, no. 7, 2015.
- [28] Dutch Techcentre for Life Sciences, "The FAIR Data Principles explained." <https://www.dtls.nl/fair-data/fair-principles-explained/>.
- [29] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1-9, 2016.
- [30] M. van Reisen, M. Stokmans, M. Basajja, A. O. Ong'ayo, C. Kirkpatrick, and B. Mons, "Towards the Tipping Point for FAIR Implementation," *Data Intelligence*, vol. 2, no. 1-2, pp. 264-275, 2020.
- [31] M. van Reisen et al., "FAIR Practices in Africa," *Data Intelligence*, vol. 2, no. 1-2, pp. 246-256, 2020.
- [32] K. Wolstencroft et al., "SEEK: A Systems Biology Data and Model Management Platform," *BMC Systems Biology*, vol. 9, pp. 1-12, 2015.
- [33] S. Poux et al., "On Expert Curation and Scalability: UniProtKB/Swiss-Prot as a Case Study," *Bioinformatics*, vol. 33, no. 21, pp. 3454-3460, 2017.
- [34] A. Bernasconi, A. G. Simon, G. Guizzardi, L. O. B. d. S. Santos, and V. C. Storey, "Ontological Representation of FAIR Principles: A Blueprint for FAIRer Data Sources," in *International Conference on Advanced Information Systems Engineering*, 2023, Springer, pp. 261-277.
- [35] East African Health Research Commission Research, "DIGITAL REGIONAL EAST AFRICAN COMMUNITY HEALTH INITIATIVE." https://www.eahealth.org/sites/www.eahealth.org/files/content/attachments/2019-02-06/Digital-REACH-Initiative-Roadmap20171205_custom_size0.pdf(accessed).
- [36] M. van Reisen et al., "Design of a FAIR Digital Data Health Infrastructure in Africa for COVID-19 Reporting and Research," *Advanced Genetics*, vol. 2, no. 2, p. e10050, 2021.
- [37] A. Martínez-García, C. Alvarez-Romero, E. Román-Villarán, M. Bernabeu-Wittel, and C. Luis Parra-Calderón, "FAIR principles to improve the impact on health research management outcomes," *Heliyon*, vol. 9, no. 5, p. e15733, 2023, doi: <https://doi.org/10.1016/j.heliyon.2023.e15733>.
- [38] F. Adebessin, P. Kotze, R. Foster, and D. Van Greunen, "A Review of Interoperability Standards in E-health and Imperatives for their Adoption in Africa," *South African Computer Journal*, vol. 50, no. 1, pp. 55-72, 2013.

Bibliography

- [39] M. Mtey and M. Dida, "Towards interoperable e-Health system in Tanzania: analysis and evaluation of the current security trends and big data sharing dynamics," 2019.
- [40] A. B. Labrique et al., "Best practices in scaling digital health in low and middle income countries," *Globalization and health*, vol. 14, pp. 1-8, 2018.
- [41] B. Mons, E. Schultes, F. Liu, and A. Jacobsen, "The FAIR principles: First generation implementation choices and challenges," *MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info*, vol. 2, pp. 1-9, 2020.
- [42] B. Ali and P. Dahlhaus, "The role of FAIR data towards sustainable agricultural performance: A systematic literature review," *Agriculture*, vol. 12, no. 2, p. 309, 2022.
- [43] R. David et al., "“Be sustainable”: EOSC-Life recommendations for implementation of FAIR principles in life science data handling," *The EMBO Journal*, p. e115008, 2023.
- [44] T. Greenhalgh et al., "Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies," *Journal of medical Internet research*, vol. 19, no. 11, p. e8775, 2017.
- [45] C. Moucheraud et al., "Sustainability of health information systems: a three-country qualitative study in southern Africa," *BMC health services research*, vol. 17, pp. 1-11, 2017.
- [46] K. Mwantimwa, N. Ndege, J. Atela, and A. Hall, "Scaling innovation Hubs: impact on knowledge, innovation and entrepreneurial ecosystems in Tanzania," *Journal of Innovation Management*, vol. 9, no. 2, pp. 39-63, 2021.
- [47] M. Shimizu, "How Can a Resilience Approach Address SDGs?," in *A Resilience Approach to Acceleration of Sustainable Development Goals*, Springer, 2022, pp. 155-172.
- [48] R. Hoefer, "The Multiple Streams Framework: Understanding and Applying the Problems, Policies, and Politics Approach," *Journal of Policy Practice and Research*, vol. 3, no. 1, pp. 1-5, 2022, doi: 10.1007/s42972-022-00049-2.
- [49] J. W. Kingdon and E. Stano, *Agendas, alternatives, and public policies*, Little, Brown Boston, 1984.
- [50] M. Basajja, M. Van Reisen, and F. Oladipo, "FAIR Equivalency with Regulatory Framework for Digital Health in Uganda," *Data Intelligence*, vol. 4, no. 4, pp. 771-797, 2022, doi: 10.1162/dint.a_00170.
- [51] P. H. P. Jati, Y. Lin, S. Nodehi, D. B. Cahyono, and M. van Reisen, "FAIR Versus Open Data: A Comparison of Objectives and Principles," *Data Intelligence*, vol. 4, no. 4, pp. 867-881, 2022, doi: 10.1162/dint.a_00176.
- [52] E. A. Schultes, G. O. Strawn, and B. Mons, "Ready, Set, GO FAIR: Accelerating Convergence to an Internet of FAIR Data and Services," *DAMDID/RCDL*, vol. 19, p. 23, 2018.

Bibliography

- [53] G. FAIR, "GO FAIR Initiative." GO FAIR Foundation. <https://www.go-fair.org/go-fair-initiative/>.
- [54] J. Wise et al., "Implementation and relevance of FAIR data principles in biopharmaceutical RD," *Drug Discovery Today*, vol. 24, no. 4, pp. 933-938, 2019, doi: <https://doi.org/10.1016/j.drudis.2019.01.008>.
- [55] D. P. Marikyan, S. "Technology Acceptance Model: A review." <https://open.ncl.ac.uk/theories/1/technology-acceptance-model/>.
- [56] VODAN Africa, "About VODAN Africa." VODAN Africa. <https://www.vodan-totafrica.info/vodan-africa.php?i=1a=about-vodan-africa> (accessed Nov, 27th, 2023).
- [57] V. Moorthy, A. M. H. Restrepo, M.-P. Preziosi, and S. Swaminathan, "Data sharing for novel coronavirus (COVID-19)," *Bulletin of the World Health Organization*, vol. 98, no. 3, p. 150, 2020.
- [58] WHO, "Listings of WHO's response to COVID-19," WHO, 2021. [Online]. Available: <https://www.who.int/news/item/29-06-2020-covidtimeline>.
- [59] Telegraph, "Ebola's lost blood: row over samples flown out of Africa as 'big pharma' set to cash in," Telegraph Media Group Limited, 2019. [Online]. Available: <https://www.telegraph.co.uk/global-health/science-and-disease/ebolass-lost-blood-row-samples-flown-africa-big-pharma-set-cash/>.
- [60] Ministry of Health, "Ministry of Health U. The Second National Health Policy: Promoting people's health to enhance socio-economic development.," Ministry of Health, 2010. [Online]. Available: [http://library.health.go.ug/sites/default/files/resources/Second %20National%20Health%20Policy%202010.pdf](http://library.health.go.ug/sites/default/files/resources/Second%20National%20Health%20Policy%202010.pdf).
- [61] M. Segall, "District health systems in a neoliberal world: a review of five key policy areas," *The International journal of health planning and management*, vol. 18, no. S1, pp. S5-S26, 2003.
- [62] S. Nyakaisiki, "A context-aware m-health application: Towards a design model for developing rural areas," Delft University of Technology, 2016.
- [63] V. M. Kiberu, J. K. Matovu, F. Makumbi, C. Kyozira, E. Mukooyo, and R. K. Wanyenze, "Strengthening district-based health reporting through the district health management information software system: the Ugandan experience," *BMC medical informatics and decision making*, vol. 14, no. 1, pp. 1-9, 2014.
- [64] E. K. Wakida, C. Obua, G. Z. Rukundo, S. Maling, Z. M. Talib, and E. S. Okello, "Barriers and facilitators to the integration of mental health services into primary healthcare: a qualitative study among Ugandan primary care providers using the COM-B framework," *BMC health services research*, vol. 18, pp. 1-12, 2018.
- [65] Ministry Of Health. "Hospitals." <https://www.health.go.ug/hospitals/> (accessed 19th December 2023).

Bibliography

- [66] S. O. Wandera et al., "Facilitators, best practices and barriers to integrating family planning data in Uganda's health management information system," *BMC health services research*, vol. 19, pp. 1-13, 2019.
- [67] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," *Administration and policy in mental health and mental health services research*, vol. 42, pp. 533-544, 2015.
- [68] M. Q. Patton, *Qualitative evaluation and research methods*, SAGE Publications, inc, 1990.
- [69] J. Acheng, "The Health Management Information System Volume 1-Health Unit and Community Procedure Manual," Ministry of Health, Kampala, 2014.
- [70] M. Nambobi, M. S. H. Khan, and A. A. Alli, "Big Data: Prospects and Applications in the Technical and Vocational Education and Training Sector," *Data Analytics: Concepts, Techniques, and Applications*, CRC Press: Boca Raton, FL, USA, pp. 297-298, 2018.
- [71] L. Rainey, J. E. Lutomski, and M. J. Broeders, "FAIR Data Sharing: An International Perspective on Why Medical Researchers Are Lagging Behind," *Big Data Society*, vol. 10, no. 1, p. 20539517231171052, 2023.
- [72] G. FAIR, "Virus Outbreak Data Network (VODAN)." <https://www.go-fair.org/implementation-networks/overview/vodan/> (accessed 7/12/2023, 2023).
- [73] KIU, "Data Science through Go Fair in Africa: A New Generation Internet of Data Services." <https://kiu.ac.ug/engagements.php?i=data-science-through-go-fair-in-africa-a-new-generation-internet-of-data-and-services>.
- [74] S. Hurt, "Window of Opportunity: EU Development and Cooperation After the End of the Cold War," by Mirjam van Reisen, ed: Oxford University Press, 2010.
- [75] Ministry of Information and Communications Technology, "National Information and Communications Technology Policy for Uganda."
- [76] Ministry of Health, "Ministry of Health-Uganda National e-health Policy." <http://library.health.go.ug/download/file/fid/517> (accessed 11 February 2021).
- [77] Ministry of Health, "Ministry of Health-Uganda National e-health Strategy."
- [78] M. L. Ranney, Z. F. Meisel, E. K. Choo, A. C. Garro, C. Sasson, and K. M. Guthrie, "Interview-based Qualitative Research in Emergency Care Part II: Data Collection, Analysis and Results Reporting," *Academic Emergency Medicine*, vol. 22, no. 9, pp. 1103-1112, 2015.
- [79] N. Gale, G. Heath, E. Cameron, S. Rashid, and S. Redwood, "Using the Framework Method for the Analysis of Qualitative Data in Multi-Disciplinary Health Research," *BMC Medical Research Methodology*, vol. 13, p. 117, 2013.

Bibliography

- [80] L. P. Wong, "Data Analysis in Qualitative Research: A Brief Guide to Using NVivo," *Malaysian Family Physician: The Official Journal of the Academy of Family Physicians of Malaysia*, vol. 3, no. 1, p. 14, 2008.
- [81] B. A. Wolfe et al., "The OpenMRS System: Collaborating Toward an Open Source EMR for Developing Countries," in *AMIA Annual Symposium Proceedings*, 2006, vol. 2006: American Medical Informatics Association, p. 1146.
- [82] N. A. Mohammed-Rajput, D. C. Smith, B. Mamlin, P. Biondich, B. N. Doebbeling, and O. M. C. Investigators, "OpenMRS, a Global Medical Records System Collaborative: Factors Influencing Successful Implementation," in *AMIA Annual Symposium Proceedings*, 2011, vol. 2011: American Medical Informatics Association, p. 960.
- [83] Daily Monitor, "Daily Monitor-Government Embraces Electronic Medical Records System." <https://www.monitor.co.ug/News/National/Government-embraces-electronic-medical-records-system/688334-3481034-7qje6ez/index.html> (accessed 30 April 2020).
- [84] J. WHO, "UNF," The MAPS toolkit: mHealth assessment and planning for scale. Geneva: WHO Department of Reproductive Health and Research, 2015.
- [85] W. M. Tierney et al., "Experience implementing electronic health records in three East African countries," in *MEDINFO 2010*: IOS press, 2010, pp. 371-375.
- [86] Novartis. "Ghana telemedicine: Increasing access, efficiency and quality of care by centralizing expertise and empowering frontline healthcare workers." Novartis Foundation. Available: <https://www.novartisfoundation.org/past-programs/digital-health/ghana-telemedicine> (accessed 24th December 2023).
- [87] H. A. Singh, Digital health: A call for government leadership and cooperation between ICT and health. Broadband Commission for Sustainable Development, 2017.
- [88] E. F. Krah and J. G. de Kruijf, "Exploring the ambivalent evidence base of mobile health (mHealth): a systematic literature review on the use of mobile phones for the improvement of community health in Africa," *Digital health*, vol. 2, p. 2055207616679264, 2016.
- [89] EAHRC. "Digital REACH Initiative Roadmap." East African Health Research Commission Research (EAHRC). Available: <https://eahrc.org/research/digital-reach-initiative-roadmap> (accessed 24 December 2023).
- [90] Digital Curation Centre. "Fostering FAIR Data Practices in Europe (FAIRsFAIR)." Available: <https://www.dcc.ac.uk/projects> (accessed 24 December 2023).
- [91] WHO. "Digital Health Atlas." World Health Organization. Available: <https://digitalhealthatlas.org/en/-/> (accessed 24 December 2023).
- [92] HealthEnabled. "Uganda Digital Health Dashboard." Available: <http://healthenabled.org/wordpress/uganda-digital-health-dashboard/> (accessed 21 March 2021).

Bibliography

- [93] F. E. Services. "Resources and guidelines to assess the FAIRness of digital resources, Philosophy of FAIR testing." FAIR Evaluation Services. Available: <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/5121> (accessed 23 December 2021).
- [94] F. E. Services. "Resources and guidelines to assess the FAIRness of digital resources, UGANDA DH Dashboard." FAIR Evaluation Services. Available: <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/5121> (accessed 23 December 2021).
- [95] F. E. Services. "Resources and guidelines to assess the FAIRness of digital resources, DHA Uganda." FAIR Evaluation Services. Available: <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/5126> (accessed 23 December 2021).
- [96] F. M. Evaluation. "DHA Uganda; Tested identifier." FAIR Metrics Evaluation. Available: <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/5126> (accessed 2024).
- [97] F. E. Services. "Resources and guidelines to assess the FAIRness of digital resources, UDHD and DHA-Uganda." FAIR Evaluation Services. Available: <https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/!/evaluations/5213> (accessed 23 December 2021).
- [98] M. Tomlinson, M. J. Rotheram-Borus, L. Swartz, and A. C. Tsai, "Scaling up mHealth: where is the evidence?," *PLoS medicine*, vol. 10, no. 2, p. e1001382, 2013.
- [99] J. Lemaire, "Scaling up mobile health: elements necessary for the successful scale up of mHealth in developing countries," Actevis consulting group, 2011. [Online]. Available: <http://www.slideshare.net/jeanninelemaire/scaling-up-mhealth-in-developing-countries-white-paper-commissioned-by-advanced-development-for-africa>
- [100] E. Freudenthal, "Ebola's lost blood: row over samples flown out of Africa as 'big pharma' set to cash in," *The Telegraph*, vol. 6, 2019.
- [101] WHO, "WHO coronavirus (COVID-19) dashboard," World Health Organization. Available: <https://covid19.who.int/> (accessed).
- [102] B. Mons, "The VODAN IN: support of a FAIR-based infrastructure for COVID-19," *European Journal of Human Genetics*, vol. 28, no. 6, pp. 724-727, 2020.
- [103] F. Oladipo, "COVID-19 computer-readable observational data installed at Kampala International University," ed. Uganda: Kampala International University, 2020.
- [104] K. Modjarrad, V. S. Moorthy, P. Millett, P.-S. Gsell, C. Roth, and M.-P. Kieny, "Developing global norms for sharing data and results during public health emergencies," *PLoS Medicine*, vol. 13, no. 1, p. e1001935, 2016.
- [105] M. Van Reisen, *International cooperation in the digital era*, Universiteit Leiden, 2017.

Bibliography

- [106] R. D. Alliance, "Data together—RDA COVID-19 Working Group," Research Data Alliance. Available: <https://www.rd-alliance.org/groups/rda-covid19> (accessed).
- [107] S. Momtazmanesh et al., "All together to fight COVID-19," *The American Journal of Tropical Medicine and Hygiene*, vol. 102, no. 6, p. 1181, 2020.
- [108] P. Ozili, "COVID-19 in Africa: socio-economic impact, policy response and opportunities," *International Journal of Sociology and Social Policy*, vol. 42, no. 3/4, pp. 177-200, 2020.
- [109] I. Chitungo, M. Dzobo, M. Hlongwa, and T. Dzinamarira, "COVID-19: unpacking the low number of cases in Africa," *Public Health in Practice*, vol. 1, p. 100038, 2020.
- [110] WHO, "Building on Ebola response to tackle COVID-19 in DRC," World Health Organization. Available: <https://www.afro.who.int/news/building-ebola-response-tackle-covid-19-drc> (accessed 2 March 2021).
- [111] K. Owada, T. Eckmanns, K.-B. O. B. Kamara, and O. O. Olu, "Epidemiological data management during an outbreak of Ebola virus disease: key issues and observations from Sierra Leone," *Frontiers in Public Health*, vol. 4, p. 163, 2016.
- [112] African Academy of Sciences (AAS), "Recommendations for data and biospecimen governance in Africa," ASP Policy Paper 3, African Academy of Sciences (AAS). Available: <https://www.aasciences.africa/sites/default/files/Publications/Recommendations>
- [113] A. Jacobsen et al., "A generic workflow for the data FAIRification process," *Data Intelligence*, vol. 2, no. 1-2, pp. 56-65, 2020.
- [114] Y. Lin, P. H. Purnama Jati, A. Aktau, M. Ghardallou, S. Nodehi, and M. van Reisen, "Implementation of FAIR Guidelines in selected non-Western geographies," *Data Intelligence*, vol. 4, no. 4, pp. 747-770, 2022.
- [115] G. FAIR, "The Internet of FAIR Data Services," GO FAIR. Available: <https://www.go-fair.org/resources/internet-fair-data-services/> (accessed 3 March 2021).
- [116] O. NQd, V. Borges, H. Rodrigues, M. Campos, and G. Lopes, "A Practical Approach of Actions for FAIRification Workflows," 2022.
- [117] J. Wise et al., "Implementation and relevance of FAIR data principles in biopharmaceutical RD," *Drug Discovery Today*, vol. 24, no. 4, pp. 933-938, 2019.
- [118] R. D. Alliance, "Data together COVID-19 Appeal and Actions. Committee on Data for Science and Technology (CODATA), Research Data Alliance (RDA), World Data System (WDS) and GO FAIR." Available: <https://www.rd-alliance.org/sites/default/files/attachment/DataTogetherCOVID-19StatementFINAL.pdf> (accessed 21 May 2021).
- [119] B. Mons, *Data stewardship for open science: Implementing FAIR principles*, CRC Press, 2018.

Bibliography

- [120] WHO, "The Access to COVID-19 Tools (ACT) accelerator," World Health Organization. Available: <https://www.who.int/initiatives/act-accelerator> (accessed).
- [121] B. Mons, "FAIR science for social machines: Let's share metadata knowlets in the Internet of FAIR Data and Services," *Data Intelligence*, vol. 1, no. 1, pp. 22-42, 2019.
- [122] G. FAIR, "Data Together," GO FAIR. Available: <https://www.go-fair.org/go-fair-initiative/data-together> (accessed 20 May 2021).
- [123] M. Suchánek and M. Basajja, "VODAN in a Box: Proof of Concept," Zenodo, 2020.
- [124] L. O. Bonino, K. Burger, and R. Kaliyaperumal, Eds. "FAIR Data Point," [fairdatapoint.org](https://specs.fairdatapoint.org/fdp-specs-v1.2.html), 2023. [Online]. Available: <https://specs.fairdatapoint.org/fdp-specs-v1.2.html>.
- [125] G. FAIR, "VODAN in a Box: The all in one solution for easy instalment of VODAN FAIR Data Points," GO FAIR. Available: <https://www.gofairfoundation.org/vodan-in-a-box-the-all-in-one-solution-for-easy-instalment-of-vodan-fair-data-points/> (accessed 3 January 2024).
- [126] AllegroGraph, "AllegroGraph," ed, 2015.
- [127] O. M. Benhamed et al., "The FAIR data point: Interfaces and tooling," *Data Intelligence*, vol. 5, no. 1, pp. 184-201, 2023.
- [128] VODAN-Tech, "VODAN in the Box Production Deployment Configuration," GitHub. Available: <https://github.com/VOD>
- [129] VODAN, "VODAN in a Box documentation," Available: <https://docs.vodan.fairdatapoint.org/en/latest/> (accessed).
- [130] M. van Reisen et al., "Federated FAIR principles: Ownership, localisation and regulatory compliance (OLR)," *FAIR Connect*, vol. 1, no. 1, pp. 63-69, 2023.
- [131] FDP, "FAIR Data Point: Advanced configuration," Dutch Techcentre for Life Sciences Revision. Available: <https://fairdatapoint.readthedocs.io/en/latest/deployment/advanced-configuration.html> (accessed 20 May 2021).
- [132] D. Borthakur, "The hadoop distributed file system: Architecture and design," *Hadoop Project Website*, vol. 11, no. 2007, p. 21, 2007.
- [133] T. Sterling, M. Brodowicz, and M. Anderson, *High performance computing: modern systems and practices*, Morgan Kaufmann, 2017.
- [134] F. Oladipo, S. Folorunso, E. Ogundepo, O. Osigwe, and A. Akindele, "Curriculum development for FAIR data stewardship," *Data Intelligence*, vol. 4, no. 4, pp. 991-1012, 2022.
- [135] P. Kasireddy, "A beginner-friendly introduction to containers, vms and docker," *FreeCodeCamp*, Available: <https://medium.freecodecamp.org/a-beginner-friendly-introduction-to-containers-vms-and-docker-79a9e3e119b> (accessed 1 April 2018), 2016.

Bibliography

- [136] R. S. Gonçalves et al., "The CEDAR workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments," in *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II*, 16, 2017: Springer, pp. 103-110.
- [137] M. A. Musen et al., "The center for expanded data annotation and retrieval," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1148-1152, 2015.
- [138] E. V. Langlois, V. Becerril Montekio, T. Young, K. Song, J. Alcalde-Rabanal, and N. Tran, "Enhancing evidence-informed policymaking in complex health systems: lessons from multi-site collaborative approaches," *Health research policy and systems*, vol. 14, pp. 1-11, 2016.
- [139] A. A. Bhattacharya et al., "Quality of routine facility data for monitoring priority maternal and newborn indicators in DHIS2: A case study from Gombe State, Nigeria," *PloS one*, vol. 14, no. 1, p. e0211265, 2019.
- [140] Ameri, F., Yoder, R. & Zandbiglari, K., 2020. Skos tool: A tool for creating knowledge graphs to support semantic text classification. In *IFIP International Conference on Advances in Production Management Systems*, pp. 263–271. Springer.
- [141] P. Saunders-Hastings, "DHIS2 as a tool for health impact assessment in low-resource settings: examining opportunities for expanding use of routine health data," *Proceedings of the 38th Annual Conference of the International Association for Impact Assessment*, Durban, South Africa, 2018.
- [142] E. Byrne and J. I. Sæbø, "Routine use of DHIS2 data: a scoping review," *BMC Health Services Research*, vol. 22, no. 1, p. 1234, 2022.
- [143] E. O. Odei-Lartey et al., "Utilization of the national cluster of district health information system for health service decision-making at the district, sub-district, and community levels in selected districts of the Brong Ahafo region in Ghana," *BMC health services research*, vol. 20, pp. 1-15, 2020.
- [144] R. Dehnavieh et al., "The District Health Information System (DHIS2): A literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries," *Health Information Management Journal*, vol. 48, no. 2, pp. 62-75, 2019.
- [145] M. Basajja and M. Nambobi, "Information streams in health facilities: The case of Uganda," *Data Intelligence*, vol. 4, no. 4, pp. 882-898, 2022.
- [146] A. Kiwanuka, M. Amiyo, and J. Nabukenya, "Constraints to and enablers for contextualizing digital health terminology standards in Uganda's health services: a qualitative case study," *Oxford Open Digital Health*, vol. 1, p. oqad003, 2023.
- [147] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in *2012 Japan-Egypt Conference on Electronics, Communications and Computers*, 2012: IEEE, pp. 173-177.

Bibliography

- [148] R. Kievit, A. Abubakar Kawu, M. van Reisen, D. O’Sullivan, and L. Hederman, "Exploring the Design of Low-End Technology to Increase Patient Connectivity to Electronic Health Records," 2024.
- [149] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, "EU General Data Protection Regulation: Changes and implications for personal data collecting companies," *Computer Law Security Review*, vol. 34, no. 1, pp. 134-153, 2018.
- [150] K. Muhangi, "Overview of the data protection regime in Uganda," *Journal of Data Protection Privacy*, vol. 3, no. 1, pp. 82-92, 2019.
- [151] D. Nsobani, "A critical analysis of the Data Protection Act 2019 Viz-A-Viz the right to privacy under article 27 of the 1995 constitution," Kampala international University, School of Law, 2019.
- [152] E. Botoeva et al., "A Formal Presentation of MongoDB (Extended Version)," 2016, pp. 1-44.
- [153] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," 2020, arXiv:2001.01523. [Online]. Available: <http://arxiv.org/abs/2001.01523>
- [154] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, and Q. Yang, "A fairness-aware incentive scheme for federated learning," in Proc. AAAI/ACM Conf. AI, Ethics, Soc., Feb. 2020, pp. 393–399.
- [155] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in Proc. 12th ACM Workshop Artif. Intell. Secur. (AISec), 2019, pp. 1–11.
- [156] H. H. Zhuo, W. Feng, Y. Lin, Q. Xu, and Q. Yang, "Federated deep reinforcement learning," Feb. 2020, arXiv:1901.08277. [Online]. Available: <https://arxiv.org/abs/1901.08277>
- [157] S. Lin, G. Yang, and J. Zhang, "Real-time edge intelligence in the making: A collaborative learning framework via federated meta-learning," 2020, arXiv:2001.03229. [Online]. Available: <http://arxiv.org/abs/2001.03229>
- [158] S. R. Pandey, N. H. Tran, M. Bennis, Y. K. Tun, A. Manzoor, and C. S. Hong, "A crowd-sourcing framework for on-device federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3241–3256, May 2020.
- [159] E. Bakopoulou, B. Tillman, and A. Markopoulou, "A federated learning approach for mobile packet classification," 2019, arXiv:1907.13113. [Online]. Available: <http://arxiv.org/abs/1907.13113>
- [160] D. B. Larson, D. C. Magnus, M. P. Lungren, N. H. Shah, and C. P. Langlotz, "Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework," *Radiology*, vol. 295, no. 3, 2020, Art. no. 192536.

Bibliography

- [161] J. Konečný, H. Brendan McMahan, D. Ramage, and P. Richtarik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, arXiv:1610.02527. [Online]. Available: <http://arxiv.org/abs/1610.02527>
- [162] A. Alexander, A. Jiang, C. Ferreira, and D. Zurkiya, "An intelligent future for medical imaging: A market outlook on artificial intelligence for medical imaging," *J. Amer. College Radiol.*, vol. 17, no. 1, pp. 165–170, Jan. 2020.
- [163] C. Ilias and S. Georgios, "Machine learning for all: A more robust federated learning framework," in *Proc. 5th Int. Conf. Inf. Syst. Secur. Privacy*, 2019, pp. 544–551.
- [164] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, Oct. 2017, pp. 1175–1191, doi: 10.1145/3133956.3133982.
- [165] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, and A. Dubey, "No peek: A survey of private distributed deep learning," 2018, arXiv:1812.03288. [Online]. Available: <http://arxiv.org/abs/1812.03288>
- [166] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," 2019, arXiv:1905.10497. [Online]. Available: <http://arxiv.org/abs/1905.10497>
- [167] H. Ben Yedder, B. Cardoen, and G. Hamarneh, "Deep learning [162] W. Raghupathi and V. Raghupathi, 'Big data analytics in healthcare: promise and potential,' *Health Information Science and Systems*, vol. 2, p. 3, 2014.
- [168] E. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.
- [169] A. Verghese, N. H. Shah, and R. A. Harrington, "What this computer needs is a physician: humanism and artificial intelligence," *JAMA*, vol. 319, no. 1, pp. 19-20, 2018.
- [170] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, et al., "FAIR principles: Interpretations and implementation considerations," *Data Intelligence*, vol. 2, no. 1-2, pp. 10-29, 2020.
- [171] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, et al., "The future of digital health with federated learning," *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1-7, 2020.
- [172] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, "Federated learning for COVID-19 screening from Chest X-ray images," *Applied Soft Computing*, vol. 106, p. 107330, 2021.
- [173] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020.

Bibliography

- [174] S. Zillner, S. Neururer, and A. Hoppe, "Data quality in healthcare: A framework for data quality assessment, improvement and visualization," **Studies in Health Technology and Informatics**, vol. 228, pp. 265-269, 2016.
- [175] F. Stegmaier, U. Gröbner, M. Döller, H. Kosch, and G. Baese, "Evaluation of current RDF database solutions," in **Proceedings of the 10th International Workshop on Semantic Multimedia Database Technologies (SeMuDaTe), 4th International Conference on Semantics And Digital Media Technologies (SAMT)**, pp. 39-55, Dec. 2009.
- [176] Y. Mou, F. Guo, W. Lu, Y. Li, O. Beyan, T. Rose, and S. Decker, "Towards FAIR Data in Distributed Machine Learning Systems," in **GLOBECOM 2023-2023 IEEE Global Communications Conference**, pp. 6450-6455, Dec. 2023.
- [177] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.
- [178] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [179] J. Grus, *Data science from scratch: first principles with python*. O'Reilly Media, 2019.
- [180] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [181] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
- [182] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175(4), 7-9.
- [183] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.
- [184] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.
- [185] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).
- [186] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
- [187] M. G. Kahn, J. S. Brown, A. T. Chun, B. N. Davidson, D. Meeker, P. B. Ryan, et al., "Transparent reporting of data quality in distributed data networks," **EGEMS (Washington, DC)**, vol. 1, no. 1, article 1035, 2012.
- [188] E. V. Bernstam, J. W. Smith, and T. R. Johnson, "What is biomedical informatics?" **Journal of Biomedical Informatics**, vol. 43, no. 1, pp. 104-110, 2010.

Bibliography

- [189] S. G. Alonso, I. de la Torre Díez, and S. Hamrioui, "Data mining algorithms and techniques in mental health: A systematic review," **Journal of Medical Systems**, vol. 43, no. 5, pp. 1-11, 2019.
- [190] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," **The New England Journal of Medicine**, vol. 375, no. 13, pp. 1216, 2016.
- [191] D. Lewis, "Machine learning in healthcare: Applications, ethics, and challenges," **BMJ Health & Care Informatics**, vol. 27, no. 3, article e100236, 2020.
- [192] F. J. Mateen, J. Oh, A. I. Tergas, N. H. Bhayani, and B. B. Kamdar, "Titles versus titles and abstracts for initial screening of articles for systematic reviews," **Clinical Epidemiology**, vol. 7, pp. 253-257, 2015.
- [193] F. Dhombres, J. Charlet, and M. Fieschi, "The contribution of semantic web technologies to health knowledge representation for decision support and predictive analysis in oncology," **Journal of Biomedical Semantics**, vol. 7, no. 1, pp. 1-17, 2016.
- [194] K. Zheng, D. A. Hanauer, N. Weibel, Z. Agha, and D. Carrell, "Computational technology for clinical research informatics: Lessons learned through collaboration," **Journal of Biomedical Informatics**, vol. 57, pp. 3-10, 2015.
- [195] X. Qian *et al.*, "Fighting against the common enemy of COVID-19: a practice of building a community with a shared future for mankind," *Infectious Diseases of Poverty*, vol. 9, no. 1, pp. 1-6, 2020.
- [196] P. M. Datilo, Z. Ismail, and J. Dare, "A review of epidemic forecasting using artificial neural networks," *Epidemiology and Health System Journal*, vol. 6, no. 3, pp. 132-143, Sep. 2019.
- [197] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, 2008.
- [198] Li, J., Li, J., Wang, C., Verbeek, F. J., Schultz, T., and Liu, H. (2023). Outlier detection using iterative adaptive mini-minimum spanning tree generation with applications on medical data. *Frontiers in Physiology*, **14**, 1233341. <https://doi.org/10.3389/fphys.2023.1233341>
- [199] Li, J., Li, J., Wang, C., Verbeek, F. J., Schultz, T., and Liu, H. (2024). MS2OD: outlier detection using minimum spanning tree and medoid selection. *Machine Learning: Science and Technology*, **5**(1), 015025. <https://doi.org/10.1088/2632-2153/ad2492>
- [200] Md Zubair, S. Roy, and A. Rahman, "An improved K-means clustering algorithm towards an efficient data-driven modeling," *Annals of Data Science*, vol. 9, no. 1, pp. 1-20, Jan. 2022.
- [201] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," *Chaos, Solitons & Fractals*, vol. 139, p. 110059, Oct. 2020.

Bibliography

- [202] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
- [203] F. Amato, A. López, E. M. Peña-Méndez, P. Vañhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine*, vol. 11, no. 2, pp. 47–58, 2013.
- [204] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Image classification based on low-rank matrix recovery and Naive Bayes collaborative representation," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5248–5262, 2018.
- [205] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [206] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [207] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86-97.
- [208] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neuro-robotics*, vol. 7, p. 21, 2013.
- [209] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.
- [210] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1-10, 2018, doi: 10.1038/s41746-018-0029-1.
- [211] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [212] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.
- [213] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.
- [214] Regitz-Zagrosek, V. (2012). Sex and gender differences in health. *EMBO reports*, 13(7), 596–603.
- [215] Rechel, B., Doyle, Y., Grundy, E., McKee, M. (2009). How can health systems respond to population ageing?. World Health Organization Regional Office for Europe.
- [216] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, Feb. 2012.
- [217] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," MIT Press, 2016.

Bibliography

- [218] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.
- [219] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 1992.
- [220] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.
- [221] K. Sweatt, W. T. Garvey, and C. Martins, "Strengths and Limitations of BMI in the Diagnosis of Obesity: What is the Path Forward?," *Current Obesity Reports*, pp. 1-12, 2024.
- [222] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341-1390, 1996.
- [223] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, pp. 402-406, May 2013, doi: 10.4097/kjae.2013.64.5.402.
- [224] A. Kataria and M.D. Singh, "A review of data classification using k-nearest neighbour algorithm," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 6, pp. 354-360, Jun. 2013.
- [225] T.L. Athey, T. Liu, B.D. Pedigo, and J.T. Vogelstein, "Autogmm: Automatic and hierarchical gaussian mixture modeling in python," *arXiv preprint arXiv:1909.02688*, 2019.
- [226] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, vol. 54, pp. 1273-1282, PMLR, 2017.
- [227] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.
- [228] N. B. Kabatereine, S. Brooker, E. M. Tukahebwa, F. Kazibwe, and A. W. Onapa, "Epidemiology and geography of *Schistosoma mansoni* in Uganda: implications for planning control," *Tropical Medicine International Health*, vol. 9, no. 3, pp. 372-380, 2004.
- [229] Heimerl, F., Lohmann, S., Lange, S., Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences (pp. 1833-1842). IEEE.
- [230] Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A. O., Nakazibwe, P., Chindoza, K. (2020). FAIR Practices in Africa. *Data Intelligence*, 2(1-2), 246-256.
- [231] Basajja, M., Nambobi, M., Wolstencroft, K. (2022). Possibility of enhancing digital health interoperability in Uganda through FAIR data. *Data Intelligence*, 4(4), 899-916.

Bibliography

- [232] Plug, R., et al. (2022). FAIR and GDPR Compliant Population Health Data Generation, Processing and Analytics. SWAT4HCLS, 54-63.
- [233] Mtey, M., Dida, M. (2019). Towards interoperable e-Health system in Tanzania: analysis and evaluation of the current security trends and big data sharing dynamics.
- [234] Alvarez-Romero, C., et al. (2022). FAIR4Health: findable, accessible, interoperable and reusable data to foster health research. Open Research Europe, 2.
- [235] Greenhalgh, T., et al. (2017). Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. Journal of medical Internet research, 19(11), e8775.
- [236] Marwaha, J. S., Landman, A. B., Brat, G. A., Dunn, T., Gordon, W. J. (2022). Deploying digital health tools within large, complex health systems: key considerations for adoption and implementation. NPJ Digital Medicine, 5(1), 13.
- [237] Mittelstadt, B. (2017). Designing the health-related internet of things: ethical principles and guidelines. Information, 8(3), 77.
- [238] Jati, P. H. P., Lin, Y., Nodehi, S., Cahyono, D. B., Van Reisen, M. (2022). FAIR Versus Open Data: A Comparison of Objectives and Principles. Data Intelligence, 4(4), 867-881.
- [239] Schultes, E. A., Strawn, G. O., Mons, B. (2018). Ready, Set, GO FAIR: Accelerating Convergence to an Internet of FAIR Data and Services. DAMDID/RCDL, 19, 23.
- [240] GO FAIR, "GO FAIR Initiative." GO FAIR Foundation. <https://www.go-fair.org/go-fair-initiative/>
- [241] Wilkinson, M. D., et al. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific Data, 6(1), 174.
- [242] Spero, J. C., McQuide, P. A., Matte, R. (2011). Tracking and monitoring the health workforce: a new human resources information system (HRIS) in Uganda. Human Resources for Health, 9, 1-10.
- [243] Driessen, J., Settle, D., Potenziani, D., Tulenko, K., Kaboch, T., Wadembere, I. (2015). Understanding and valuing the broader health system benefits of Uganda's national Human Resources for Health Information System investment. Human Resources for Health, 13, 1-9.
- [244] Labrique, A. B., et al. (2013). mHealth innovations as health system strengthening tools: 12 common applications and a visual framework. Global Health: Science and Practice, 1(2), 160-171.
- [245] Shimizu, M. (2022). How Can a Resilience Approach Address SDGs? In A Resilience Approach to Acceleration of Sustainable Development Goals (pp. 155-172). Springer.

Bibliography

- [246] David, R., et al. (2023). Be sustainable: EOSC-Life recommendations for implementation of FAIR principles in life science data handling. *The EMBO Journal*, e115008.
- [247] Moucheraud, C., et al. (2017). Sustainability of health information systems: a three-country qualitative study in southern Africa. *BMC health services research*, 17, 1-11.
- [248] Kiberu, V. M., Scott, R. E., Mars, M. (2019). Assessment of health provider readiness for telemedicine services in Uganda. *Health Information Management Journal*, 48(1), 33-41.
- [249] Ali, B., Dahlhaus, P. (2022). The role of FAIR data towards sustainable agricultural performance: A systematic literature review. *Agriculture*, 12(2), 309.
- [250] Heidt, B., et al. (2020). Point of care diagnostics in resource-limited settings: A review of the present and future of PoC in its most needed environment. *Biosensors*, 10(10), 133.
- [251] Le Blanc, D. (2015). Towards integration at last? The sustainable development goals as a network of targets. *Sustainable Development*, 23(3), 176-187.
- [252] Adegoke, O.J. (2024) 'Migration from Epi Info to District Health Information Software 2 for Vaccine-Preventable Disease Surveillance — World Health Organization African Region, 2019–2023', *MMWR. Morbidity and Mortality Weekly Report*, 73. Available at: <https://doi.org/10.15585/mmwr.mm7323a2>.
- [253] Alzyoud, A.A.Y. (2018) 'The Influence of Human Resource Management Practices on Employee Work Engagement', *Foundations of Management*, 10(1), pp. 251–256. Available at: <https://doi.org/10.2478/fman-2018-0019>.
- [254] Azevedo, M.J. (2017) 'The State of Health System(s) in Africa: Challenges and Opportunities', *Historical Perspectives on the State of Health and Health Systems in Africa*, Volume II, p. 1. Available at: https://doi.org/10.1007/978-3-319-32564-4_1.
- [255] Baker, D.B., Kaye, J. and Terry, S.F. (2016) 'Governance Through Privacy, Fairness, and Respect for Individuals', *eGEMs*, 4(2), p. 1207. Available at: <https://doi.org/10.13063/2327-9214.1207>.
- [256] Bakkabulindi, P. et al. (2023) 'Engagement of community health workers to improve immunization coverage through addressing inequities and enhancing data quality and use is a feasible and effective approach: An implementation study in Uganda', *PLOS ONE*, 18(10), p. e0292053. Available at: <https://doi.org/10.1371/journal.pone.0292053>.
- [257] Barker, M. et al. (2022) 'Introducing the FAIR Principles for research software', *Scientific Data*, 9(1), pp. 1–6. Available at: <https://doi.org/10.1038/s41597-022-01710-x>.
- [258] Barugahare, J. et al. (2020) 'Ethical and human rights considerations in public health in low and middle-income countries: an assessment using the case of Uganda's responses to COVID-19 pandemic', *BMC Medical Ethics*, 21(1), pp. 1–12. Available at: <https://doi.org/10.1186/s12910-020-00523-0>.
- [259] Basajja, M. and Nambobi, M. (2022) 'Information Streams in Health Facilities: The Case of Uganda'. Available at: https://doi.org/10.1162/dint_a_00177.

Bibliography

- [260] Basajja, M., Nambobi, M. and Wolstencroft, K. (2022a) ‘Possibility of Enhancing Digital Health Interoperability in Uganda through FAIR Data’. Available at: https://doi.org/10.1162/dint_a_00178.
- [261] Basajja, M., Van Reisen, M. and Oladipo, F. (2022b) ‘FAIR Equivalency with Regulatory Framework for Digital Health in Uganda’. Available at: https://doi.org/10.1162/dint_a_00170.
- [262] Begum, T. et al. (2020) ‘Perceptions and experiences with district health information system software to collect and utilize health data in Bangladesh: a qualitative exploratory study’, *BMC Health Services Research*, 20, p. 465. Available at: <https://doi.org/10.1186/s12913-020-05322-2>.
- [263] Brar, S. et al. (2010) *Science, Technology and Innovation in Uganda: Recommendation for Policy and Action*. The World Bank. Available at: <https://doi.org/10.1596/978-0-8213-8672-9>.
- [264] Bryant, J. (2021) ‘Africa in the Information Age: Challenges, Opportunities, and Strategies for Data Protection and Digital Rights’, 24.
- [265] Byrne, E. and Sæbø, J.I. (2022) ‘Routine use of DHIS2 data: a scoping review’, *BMC Health Services Research*, 22, p. 1234. Available at: <https://doi.org/10.1186/s12913-022-08598-8>.
- [266] Chevrier, R. et al. (2019) ‘Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review’, *Journal of Medical Internet Research*, 21(5), p. e13484. Available at: <https://doi.org/10.2196/13484>.
- [267] Chikwado (2023) Strengthening routine data collection with DHIS2 to boost education data use in four Ugandan districts, DHIS2. Available at: <https://dhis2.org/uganda-education-data-use/> (Accessed: 19 January 2025).
- [268] Day, O.D. (2020) Embracing the FAIR data principles in Uganda: Open Data Day 2020 report, Open Knowledge Foundation blog. Available at: <https://blog.okfn.org/2020/04/08/embracing-the-fair-data-principles-in-uganda-open-data-day-2020-report/> (Accessed: 19 January 2025).
- [269] Heller, P.S. and Hsiao, W.C. (no date) ‘CHAPTER 4: Health Policy Challenges and Issues Confronting Nations’, in. Available at: <https://www.elibrary.imf.org/display/book/9781589066182/ch04.xml> (Accessed: 19 January 2025).
- [270] Hildt, E. and Laas, K. (2022) ‘Informed Consent in Digital Data Management’, in K. Laas, M. Davis, and E. Hildt (eds) *Codes of Ethics and Ethical Guidelines*. Cham: Springer International Publishing (The International Library of Ethics, Law and Technology), pp. 55–81. Available at: https://doi.org/10.1007/978-3-030-86201-5_4.
- [271] Jati, P.H.P. et al. (2022) ‘FAIR Versus Open Data: A Comparison of Objectives and Principles’, *Data Intelligence*, 4(4), pp. 867–881. Available at: https://doi.org/10.1162/dint_a_00176.

Bibliography

- [272] Kawu, A.A. et al. (2022) ‘FAIR Guidelines and Data Regulatory Framework for Digital Health in Nigeria’. Available at: https://doi.org/10.1162/dint_a_00174.
- [273] Kiberu, V. et al. (2014) ‘Strengthening district-based health reporting through the district health management information software system: the Ugandan experience’, BMC Medical Informatics and Decision Making [Preprint]. Available at: <https://www.semanticscholar.org/paper/Strengthening-district-based-health-reporting-the-Kiberu-Matovu/0f657f5667f03e88c6f112087eefa656e3b498a6> (Accessed: 19 January 2025).
- [274] Kinkade, C. et al. (2022) ‘Extending and Strengthening Routine DHIS2 Surveillance Systems for COVID-19 Responses in Sierra Leone, Sri Lanka, and Uganda’, Emerging Infectious Diseases, 28(Suppl 1), p. S42. Available at: <https://doi.org/10.3201/eid2813.220711>.
- [275] Kruk, M.E. et al. (2018) ‘High-quality health systems in the Sustainable Development Goals era: time for a revolution’, The Lancet. Global Health, 6, p. e1196. Available at: [https://doi.org/10.1016/S2214-109X\(18\)30386-3](https://doi.org/10.1016/S2214-109X(18)30386-3).
- [276] Lawrence, M.R. et al. (2024) ‘Health Data Reporting Challenges in the Age of DHIS2: a Mixed-Methods Review in the Teso Sub-Region of Uganda’. Available at: <https://doi.org/10.1101/2024.09.18.24313921>.
- [277] Li, M. (no date) ‘Utilizing the Space for User Participation: Experiences from a Public Health Project in Uganda’.
- [278] Lin, N. van et al. (2021) ‘How Patient Organizations Can Drive FAIR Data Efforts to Facilitate Research and Health Care: A Report of the Virtual Second International Meeting on Duchenne Data Sharing, March 3, 2021’, Journal of Neuromuscular Diseases, 8(6), p. 1097. Available at: <https://doi.org/10.3233/JND-210721>.
- [279] Llebot, C. (2022) ‘Encouraging adoption of FAIR principles via institutional research data policies’. Available at: <https://osf.io/vfajt/> (Accessed: 19 January 2025).
- [280] Majeed, A. and Lee, S. (2021) ‘Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey’, IEEE Access, 9, pp. 8512–8545. Available at: <https://doi.org/10.1109/ACCESS.2020.3045700>.
- [281] Miir, C. et al. (2022) ‘Achieving optimal health data impact in rural African healthcare settings: measures to barriers in Bukomansimbi District, Central Uganda’, International Journal for Equity in Health, 21(1), pp. 1–13. Available at: <https://doi.org/10.1186/s12939-022-01814-1>.
- [282] Mujuni, M. et al. (2024) ‘Beyond diagnostic connectivity: Leveraging digital health technology for the real-time collection and provision of high-quality actionable data on infectious diseases in Uganda’, PLOS digital health, 3(8). Available at: <https://doi.org/10.1371/journal.pdig.0000566>.
- [283] Munyambabazi, E. et al. (2024) ‘Digitalizing Monitoring of National Development Plans: A Case Study from Uganda’, in Implications of Information and Digital Technologies

Bibliography

- for Development. International Conference on Implications of Information and Digital Technologies for Development, Springer, Cham, pp. 33–47. Available at: https://doi.org/10.1007/978-3-031-66982-8_3.
- [284] Nansikombi, H.T. et al. (2022) ‘Timeliness and completeness of weekly surveillance data reporting on epidemic prone diseases in Uganda, 2020–2021’. Available at: <https://doi.org/10.21203/rs.3.rs-2189383/v1>.
- [285] Nsubuga, P. et al. (2006) ‘Public Health Surveillance: A Tool for Targeting and Monitoring Interventions’, in Disease Control Priorities in Developing Countries. 2nd edition. The International Bank for Reconstruction and Development / The World Bank. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK11770/> (Accessed: 19 January 2025).
- [286] OSF (no date) The Open Science Framework. Available at: <https://www.cos.io/products/osf> (Accessed: 19 January 2025).
- [287] Pool, J. et al. (2024) ‘A systematic analysis of failures in protecting personal health data: A scoping review’, International Journal of Information Management, 74, p. 102719. Available at: <https://doi.org/10.1016/j.ijinfomgt.2023.102719>.
- [288] Purnama Jati, P.H. et al. (no date) ‘Data Access, Control, and Privacy Protection in the VODAN-Africa Architecture’. Available at: https://doi.org/10.1162/dint_a_00180.
- [289] Queralt-Rosinach, N. et al. (2022) ‘Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic’, Journal of Biomedical Semantics, 13, p. 12. Available at: <https://doi.org/10.1186/s13326-022-00263-7>.
- [290] Reisen, M. van et al. (2021) ‘Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research’. Available at: <https://doi.org/10.1002/ggn2.10050>.
- [291] Reynolds, E. et al. (2022) ‘Implementation of DHIS2 for Disease Surveillance in Guinea: 2015–2020’, Frontiers in Public Health, 9, p. 761196. Available at: <https://doi.org/10.3389/fpubh.2021.761196>.
- [292] Shinta, S.O., Handayani, P.W. and Hidayanto, A.N. (2022) ‘Health Data Governance Issues in Healthcare Facilities: Perspective of Hospital Management’, in 2022 10th International Conference on Information and Communication Technology (ICoICT). 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia: IEEE, pp. 1–5. Available at: <https://doi.org/10.1109/ICoICT55009.2022.9914865>.
- [293] Spies, J.R. (no date) ‘The Open Science Framework’.
- [294] Ssali, M. (2024) ‘Uganda Data Protection and Privacy Act 2019: Impact on Local Web Hosting’, Othware Uganda, 5 September. Available at: <https://www.othware.co.ug/2024/09/05/hosting-in-uganda-why-it-is-essential-for-your-website/> (Accessed: 19 January 2025).

Bibliography

- [295] B. Yurdem, M. Kuzlu, M. K. Gullu, F. O. Catak, and M. Tabassum, “Federated learning: Overview, strategies, applications, tools and future directions,” *Heliyon*, vol. 10, no. 19, p. e38137, Oct. 2024, doi: 10.1016/j.heliyon.2024.e38137.
- [296] A. Karargyris et al., “Federated benchmarking of medical artificial intelligence with Med-Perf,” *Nat. Mach. Intell.*, vol. 5, no. 7, p. 799, Jul. 2023, doi: 10.1038/s42256-023-00652-2.
- [297] Z. L. Teo et al., “Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture,” *Cell Rep. Med.*, vol. 5, no. 2, p. 101419, Feb. 2024, doi: 10.1016/j.xcrm.2024.101419.
- [298] Q. Sserwanja et al., “Continuum of care for maternal health in Uganda: A national cross-sectional study,” *PLoS ONE*, vol. 17, no. 2, p. e0264190, Feb. 2022, doi: 10.1371/journal.pone.0264190.
- [299] W. K. Kansiime et al., “Barriers and benefits of mHealth for community health workers in integrated community case management of childhood diseases in Banda Parish, Kampala, Uganda: a cross-sectional study,” *BMC Prim. Care*, vol. 25, p. 173, May 2024, doi: 10.1186/s12875-024-02430-4.
- [300] A. A. Sinaci et al., “A Data Transformation Methodology to Create Findable, Accessible, Interoperable, and Reusable Health Data: Software Design, Development, and Evaluation Study,” *J. Med. Internet Res.*, vol. 25, p. e42822, Mar. 2023, doi: 10.2196/42822.
- [301] L. J. Wallace and L. Kapiriri, “Priority setting for maternal, newborn and child health in Uganda: a qualitative study evaluating actual practice,” *BMC Health Serv. Res.*, vol. 19, p. 465, Jul. 2019, doi: 10.1186/s12913-019-4170-6.
- [302] W. Oh and G. N. Nadkarni, “Federated Learning in Health care Using Structured Medical Data,” *Adv. Kidney Dis. Health*, vol. 30, no. 1, p. 4, Jan. 2023, doi: 10.1053/j.akdh.2022.11.007.
- [303] F. Adebessin and P. Kotz, “Barriers and Challenges to the Adoption of E-Health Standards in Africa”.
- [304] M. E. van Genderen, M. Cecconi, and C. Jung, “Federated data access and federated learning: improved data sharing, AI model development, and learning in intensive care,” *Intensive Care Med.*, vol. 50, no. 6, p. 974, Apr. 2024, doi: 10.1007/s00134-024-07408-5.
- [305] Ssegujja, E. et al. (2024) ‘Quality improvement interventions targeting immunization data from urban lower-level private-for-profit health service providers in Kampala Capital City: processes and implementation experiences’, *Discover Health Systems*, 3(1), pp. 1–15. Available at: <https://doi.org/10.1007/s44250-024-00109-2>.
- [306] Sujatmiko, N.F. (2015) Implementing DHIS2 Feedback via Short Message Service: A Case Study for Uganda WEMR Health Workers. Master thesis. Available at: <https://www.duo.uio.no/handle/10852/47635> (Accessed: 19 January 2025).

Bibliography

- [307] Uganda data protection laws (2022) ‘International & U.S. Data Privacy Laws and Regulations You Need to Know’. Available at: <https://cdp.com/basics/international-u-s-data-privacy-laws-and-regulations-you-need-to-know/> (Accessed: 19 January 2025).
- [308] Ugochukwu, A.I. and Phillips, P.W.B. (2024) ‘Open data ownership and sharing: Challenges and opportunities for application of FAIR principles and a checklist for data managers’, *Journal of Agriculture and Food Research*, 16, p. 101157. Available at: <https://doi.org/10.1016/j.jafr.2024.101157>.
- [309] Wandera, S.O. et al. (2019) ‘Facilitators, best practices and barriers to integrating family planning data in Uganda’s health management information system’, *BMC Health Services Research*, 19, p. 327. Available at: <https://doi.org/10.1186/s12913-019-4151-9>.
- [310] Wang, D. et al. (2022) ‘Demand sensing and digital tracking for maternal child health (MCH) in Uganda: a pilot study for “E+TRA health”’, *BMC Medical Informatics and Decision Making*, 22(1), pp. 1–14. Available at: <https://doi.org/10.1186/s12911-022-01982-8>.
- [311] Wang, H. et al. (2022) ‘Implementation and assessment of an end-to-end Open Science & Data Collaborations program’, *F1000Research*, 11, p. 501. Available at: <https://doi.org/10.12688/f1000research.110355.1>.
- [312] Zerfu, T.A. et al. (2024) ‘Unveiling the role of DHIS2 in enhancing data quality and accessibility in primary healthcare facilities: Evidence from Ethiopia’, *PLOS ONE*, 19(12), p. e0314505. Available at: <https://doi.org/10.1371/journal.pone.0314505>.
- [313] N. A. of Medicine et al., “WHY INTEROPERABILITY IS ESSENTIAL IN HEALTH CARE,” in *Procuring Interoperability: Achieving High-Quality, Connected, and Person-Centered Care*, National Academies Press (US), 2018. Accessed: Aug. 09, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK594855/>
- [314] E. T. Inau, J. Sack, D. Waltemath, and A. A. Zeleke, “Initiatives, Concepts, and Implementation Practices of the Findable, Accessible, Interoperable, and Reusable Data Principles in Health Data Stewardship: Scoping Review,” *J. Med. Internet Res.*, vol. 25, p. e45013, Aug. 2023, doi: 10.2196/45013.
- [315] L. D. Hughes et al., “Addressing barriers in FAIR data practices for biomedical data,” *Sci. Data*, vol. 10, p. 98, Feb. 2023, doi: 10.1038/s41597-023-01969-8.
- [316] S. N. Mitchell et al., “FAIR data pipeline: provenance-driven data management for traceable scientific workflows,” *Philos. Transact. A Math. Phys. Eng. Sci.*, vol. 380, no. 2233, p. 20210300, Aug. 2022, doi: 10.1098/rsta.2021.0300.
- [317] A. Carballo-Garcia and J.-J. Boté-Vericad, “Fair Data: History and Present Context,” *Cent. Eur. J. Educ. Res.*, vol. 4, no. 2, pp. 45–53, Dec. 2022, doi: 10.37441/cej-er/2022/4/2/11379.
- [318] Mathew, Thomas K., Muhammad Zubair, and Prasanna Tadi. ”Blood Ducrose Control.” (2024).

Bibliography

- [319] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [320] Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Natesan Ramamurthy K, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. In: Proceedings of the 29th International Conference on World Wide Web Companion (WWW '20). New York, NY, USA: Association for Computing Machinery; 2020. p. 1534–1537. doi:10.1145/3366424.3388817.
- [321] Y. Januzaj, E. Beqiri, and A. Luma, “Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique,” *International Journal of Online & Biomedical Engineering*, vol. 19, no. 4, 2023.

SUPPLEMENTARY MATERIAL: HMIS OPD 002 OUTPATIENT REGISTER

Figure 1: HMIS OPD 002 OUTPATIENT REGISTER.

227

```

    & first name in full and Tel. contact
    "Age",
    "Sex",
    "Client Category",
    "RESIDENCE",
    "Next Of Kin Name And Contact",
    "Mid-Upper Arm Circumference (MUAC)",
    "BMI (SM/MM/N/OW/O)",
    "BMI For Age",
    "Weight For Height Z Score",
    "Weight (Kg)",
    "Weight For Age Z Score",
    "Height/Length (cm)",
    "Height/Length for Age Z Score (N/S)",
    "Blood Pressure (mmHg)",
    "Blood Sugar (mm/l)",
    "Need For Palliative Care",
    "Tobacco Use",
    "Tobacco Exposure",
    "Alcohol Use",
    "Fever",
    "Tests Done",
    "Malaria Tests Results",
    "Presumed Tuberculosis Case",
    "Patient Sent To The Laboratory",
    "Laboratory Tuberculosis Result",
    "Linked To Tuberculosis Clinic",
    "Where referred to",
    "Unit Tuberculosis Number",
    "Is this a new case of illness? (Tick Classification)",
    "Diagnosis",
    "Prescription",
    "Disability (Write code of disability)",
    "ReferralInNumber",
    "ReferraloutNumber"
  }
}

```

Supplementary Material 2: UGANDA ANTENATAL CARE REGISTER TEMPLATE SCHEMA

"DESCRIPTION": "UGANDA ANTENATAL CARE REGISTER TEMPLATE SCHEMA GENERATED BY THE CEDAR TEMPLATE EDITOR 2.6.19"

JSON Schema

```

{
  "The date of the patient visit",

```

"Serial Number",
"PatientID",
"Ugandan National Identification Number",
"Name of Client",
"Age",
"Client Category",
"Address",
"ANC Visit",
"Gravida",
"Gestational age of the women in weeks (GA)",
"Parity",
"ANC 1 Timing",
"Expected Date of Delivery",
"Weight in kgs",
"Fundal Height in (cm)",
"Mid-Upper Arm Circumference (MUAC) of the pregnant mother using centimeter (cm)",
"INR Number",
"bloodPressure",
"eMTCT Status W Initial Result",
"eMTCT Status W STK Given",
"eMTCT Status W Age",
"eMTCT Status P Facility Result",
"eMTCT Status P Linked to ART",
"eMTCT Status P TFV Yes/No",
"eMTCT Status P Linked to ART Where",
"eMTCT Status P Linked to ART Pre-Art Number",
"CD4 count",
"Viral Load",
"has_clinical_stage",
"eMTCT Risk Assessment",
"ART Art Code",
"ART Art Number",
"Infant Arv Prophylaxis",
"Infant and Young Child Feeding Counselling and Maternal Nutrition Counselling",
"TB Status",
"WOA U/S Scan Done",
"GBV Risk Phy",
"GBV Risk Sx",
"GBV Risk Psy",
"Blood Screening",
"Hb",
"Blood Group",
"Rhesus",
"Sickle cell",
"Urine Protein",
"Urine glucose",

```

    "Syphilis & Hep B Test Results W: Syphilis",
    "Syphilis & Hep B Test Results W: Hep B",
    "Syphilis & Hep B Test Results P: Syphilis",
    "Syphilis & Hep B Test Results P: Hep B",
    "Family Planning Counselling",
    "Tetanus Diphtheria (Td)",
    "IPT/CTX",
    "LLIN",
    "Mabendazole",
    "Other Treatments",
    "Referral in/out",
    "Complications/Risk Factors",
    "Number of Pills Iron sulphate",
    "Number of Pills Folic Acid",
    "Number of Pills Iron sulphate/Folic Acid Combined",
    "Number of Pills Iron sulphate/Folic Acid Combined Misoprostol"
}

```

Supplementary Material 3: COVID 19 GENERIC REGISTER TEMPLATE SCHEMA

"DESCRIPTION": "COVID 19 GENERIC REGISTER TEMPLATE SCHEMA GENERATED BY THE CEDAR TEMPLATE EDITOR 2.6.19"

JSON Schema

```

{
  "Reporting country",
  "Health Facility name",
  "Why tested for COVID-19",
  "Serial Number",
  "Age",
  "Name",
  "Sex",
  "Date of Patient Visit",
  "Health status at time of reporting",
  "COVID-19 Vaccination",
  "Tested on COVID-19",
  "Date of most recent COVID-19 test",
  "City of COVID-19 test center",
  "COVID 19 Result",
  "Mortality Status",
  "Cause of death includes COVID-19",
  "Is the COVID-19 case a health care worker",
  "Country"
}

```

APPENDIX B

FIELDWORK REPORT: ASSESSING VODAN-AFRICA'S DATA PRODUCTION PROGRESS AND CHALLENGES IN UGANDA

The main goal of the field work was to assess and address the progress and challenges faced by the VODAN-Africa Uganda health facilities during the data production phase of the project. The main aim was to strengthen the ongoing collaboration with the different stakeholders involved in the project, especially the deployment and data production phase of the project. On 8 May 2023, I visited the Case clinic hospital in the city of Kampala, Uganda. We had a focus group meeting with the data management personnel team spearheaded by Norbert, a one-on-one with the health facility director and addressed the technical issues faced during the data production phase of the project.

On 18 May 2023, I traveled from Kampala district, Uganda, to Ark Specialist Hospital in Wakiso district, which was about 45 minutes drive. I had the opportunity to have a one-on-one interview with the Health Facility Manager Sheeba Nasuna who expressed great interest in the current and future coordination with VODAN-Africa. We discussed the different issues the health facility was experiencing with the VODAN-Africa system during the data production phase. She mentioned that she oversaw operations, administration, and supervision. She also pointed out the need for extensive training for data management personnel. This was carried out for a duration of three days. I also had a few more interviews with the health facility receptionist Bridget Namusisi, Opio Jared, the records personnel, and two nurses.

From Ark Specialist Hospital, for about seven hours, we drove to Lira referral Hospital on 22nd, May 2023. I managed to meet the Lira Senior CDO, Francis Ifeero, who referred me to Hejang Kuman Pollyn, Dennis Olo and Sharon Akello, the data management personnel of Lira Referral Hospital. I successfully addressed the issues pushed forward during my three days stay in Lira district, Uganda.

On 26 May 2023, I went to Hoima Referral Hospital in Hoima district, which was 4 hours and a half drive from Lira district. I had a very insightful interview with Idris Muyingo. Data management personnel Hoima referral hospital. The brief encounter I had with the health facility director's secretary, Maureen Atugonza, persuaded her to schedule an interview slot with the hospital administrator who then introduced me to Livingstone Bantu, IT management personnel, Dr. Said Abdi, HRH medical doctor and Ramula Musanya, HRH hospital administrator. During my three day stay in Hoima district, I experienced some of the limitations encountered by the data management personnel during the data production phase, for example, the limited internet connectivity, electricity shortages, and lack of relevant skill set.

We arrived at the Kampala International University Teaching hospital on May 30, 2023. I conducted interviews and discussions with the hospital director, two data management personnel, a nurse during the 5 days stay in Bushenyi district. The trip from Bushenyi to Kampala was a 6 hour drive. This strengthened collaboration and trust between stakeholders in the health facility and the VODAN-Africa initiative. Below are some of the pictures from the different locations of the fieldwork.

You can scan this QR code to access the field work presentation.



Appendix B



Figure 2: KIU Teaching Hospital records storage room



Figure 3: Hoima Regional Referral Hospital



Figure 4: Case Clinic Hospital

APPENDIX C

Coding Functions

Function: Create Decision Tree model

```
def create_keras_model(input_shape):
    num_classes = len(np.unique(y_train))
    model = tf.keras.Sequential([
        tf.keras.layers.Dense(64, activation='relu', input_shape=(input_shape,)),
        tf.keras.layers.Dropout(0.5),
        tf.keras.layers.Dense(32, activation='relu'),
        tf.keras.layers.Dense(num_classes, activation='softmax')
    ])
    model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.0001),
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    return model
```

Function: Create Decision Tree and SVM model

```
# Define Decision Tree and SVM model
def create_decision_tree_model():
    return DecisionTreeClassifier()

def create_svm_model():
    return SVC(probability=True)

# Function to train Decision Tree or SVM on a client's data
def train_client_sklearn_model(model, client_data, features_list):
    features = client_data[features_list].values.astype(np.float32)
    labels = client_data['fetalState'].values.astype(np.int32)

    # Handle missing values
    imputer = SimpleImputer(strategy='mean')
    features_imputed = imputer.fit_transform(features)

    model.fit(features_imputed, labels)
    return model
```

Function: Split training data into per-client datasets

```
pythonCopyclient_data = {
    client: pd.concat([
        X_train_df.sample(frac=1/len(clients),
                           random_state=42).reset_index(drop=True),
        y_train_df.sample(frac=1/len(clients),
                           random_state=42).reset_index(drop=True)
    ], axis=1)
    for client in clients
}
```

Function: Federated training loop for Keras model

```
for round_num in range(1, num_rounds + 1):
    print(f"Round {round_num}")

    client_weights = []
    for client in clients:
        client_model = create_keras_model(input_shape)
        client_model.set_weights(global_keras_model.get_weights())
        weights, history = train_client_keras_model(client_model,
                                                       client_data[client], features)
        client_weights.append(weights)

    aggregated_keras_weights = aggregate_keras_weights(client_weights)
    global_keras_model.set_weights(aggregated_keras_weights)
```

Function: Train Decision Tree and SVM for each client

```
client_decision_tree_models = []
client_svm_models = []

for client in clients:
    client_dt_model = create_decision_tree_model()
    client_svm_model = create_svm_model()

    client_dt_model = train_client_sklearn_model
```

```
(client_dt_model, client_data[client], features)
client_svm_model = train_client_sklearn_model
(client_svm_model, client_data[client], features)

client_decision_tree_models.append(client_dt_model)
client_svm_models.append(client_svm_model)
```

APPENDIX D

ETHICAL CLEARANCE AND APPROVAL AND DATA AGREEMENTS

Ethical approvals were provided by all organisations required. These are available on request.

- Ministry of Health Approval Letter
- Ethical Clearance Renewal of Approval 2019
- Ethical Clearance Renewal of Approval 2021
- Ethical Clearance Renewal of Approval 2023
- Ethical Clearance Renewal of Approval 2024
- Signed Hoima Regional Referral Hospital Uganda Data Use Agreement Microsoft VODANA 28.4.24
- Signed KIU Teaching Hospital Data Use Agreement microsoft VODANA 28.4.24
- Uganda LU Data Processing Agreement

You can scan this QR code to access the documents mentioned above.



APPENDIX E

Heat map generation using Python with FAIR Antenatal care health data

The analysis focused on examining the relationship between the Mebendazole proxy and HIV status to assess potential co-occurrence. More granular FAIR antenatal care register data was retrieved from the allegro graph triple store using the Python API. The Python API allowed us to query the data using SPARQL. The inspect triples functionality in AllegroGraph helped to determine the URIs of the patient properties we were interested in. Using these, we were able to use SPARQL to return data on the patient prescription, Mebendazole in our case, patient HIV status, and patient address, without the patient's name or identification number. Being able to retrieve this information without patient names or patient IDs was critical, as it shows that it is possible to obtain data for analysis without compromising patient data privacy. With more granular FAIR data, the workflow for generating heat maps is not so different from the aggregated data. However, since it's on the granular level, the patient addresses were used instead of district coordinates. The relevant rows and columns were used to plot the addresses, these were translated into relevant DD coordinates using the geopy library for folium to create maps. Out of 500 patient entries, only 40 entries had valid addresses and could be plotted into the map.

1. Isolate various areas of Uganda and medical info (Mebendazole)
2. Use Geopy to validate city addresses and plot them using Folium
3. View the HTML output

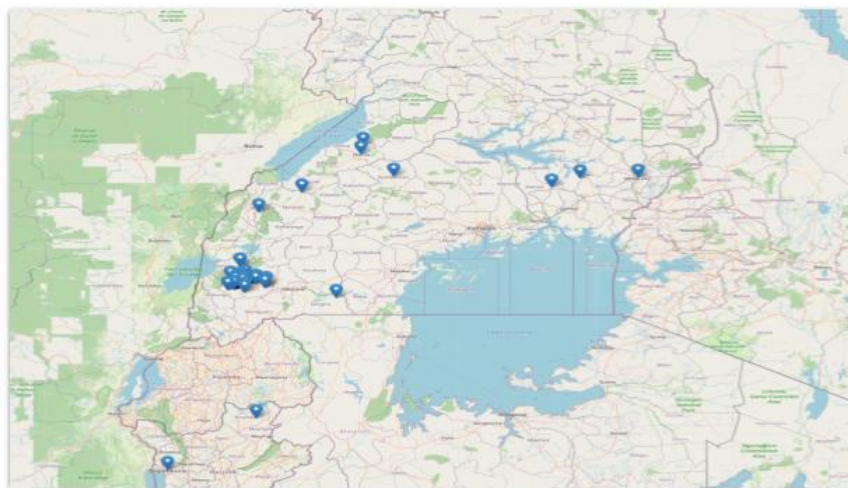


Figure 5: HTML output showing the use of Mebendacole - Uganda

Limitations of FAIR Antenatal Care Health data

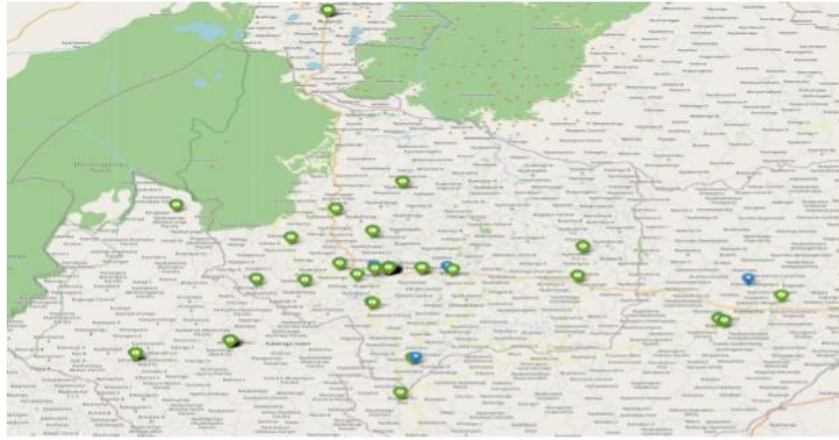


Figure 6: Bushenyi District addresses - Use of Mebendazole Yes (Green) or No (Blue)

The FAIR antenatal care dataset was more relevant to this analysis as data contained much more detailed information on HIV status and treatment than the DHIS2-generated aggregate dataset. However, data on Female genital schistosomiasis, HPV vaccination status, and treatment or diagnosis for schistosomiasis are not collected by the care providers.

Given the limitations of our data, we utilized antiparasitic Mebendazole treatment as a proxy for schistosomiasis infection, due to the unavailability of Praziquantel which is the preferred treatment for schistosomiasis. It is important to note that this approach was more a proof-of-concept than a reliably accurate method to determine schistosomiasis infection rates. The analysis focused on examining the relationship between the Mebendazole proxy and HIV status to assess potential co-occurrence.

FAIR data made it easy to extract the relevant information, so it is possible to run co-occurrence analysis with this type of FAIRified data. However, we ran into limitations of the second data set size; we had about 1000 total data points, 500 of which had any entry for both Mebendazole and HIV status, but there were no patients who were both HIV positive and being treated with Mebendazole. Unfortunately, this made it impossible to run any sensible co-occurrence test. Despite limitations, our efforts do demonstrate the potential of FAIRifying these data, and the viability of this type of co-occurrence analysis in the future, provided enough data. This illustrates the value of the FAIR principles for large data-driven projects. If the data was first, Findable, Accessible, Interoperable, and Reusable by design it would have mitigated the problems of our initial uncertainties and enabled us to plan effectively.

Appendix A

Abbreviations

- **eCRF** – electronic Case Report Form
- **AI** – Artificial Intelligence
- **API** – Application Programming Interface
- **ML** – Machine Learning
- **FAIR** – Findable, Accessible, Interoperable, and Reusable
- **SDG** – Sustainable Development Goal
- **ICT** – Information and Communication Technology
- **NITA** – National Information Technology Authority
- **MoH** – Ministry of Health
- **UCC** – Uganda Communications Commission
- **CHWs** – Community Health Workers
- **HRH** – Human Resources for Health
- **HIPAA** – Health Insurance Portability and Accountability Act
- **GDPR** – General Data Protection Regulation
- **UIRI** – Uganda Industrial Research Institute
- **EAC** – East African Community
- **BD2K** – US-based Big Data to Knowledge
- **EAOSCH** – East Africa Open Science Cloud for Health
- **EAHRC** – East African Health Research Commission

-
- **VODAN** – Virus Outbreak Data Network
 - **IFDS** – Internet-based FAIR Data and Services
 - **HIS** – Hospital Information System
 - **EMR** – Electronic Medical Record
 - **EHR** – Electronic Health Record
 - **PHR** – Patient Health Record
 - **SARS-CoV-2** – Severe Acute Respiratory Syndrome Coronavirus 2
 - **PNFPs** – Private-Not-For-Profit Health Providers
 - **PHPs** – Private Health Practitioners
 - **TCMPs** – Traditional and Complementary Medicine Practitioners
 - **VHTs** – Village Health Teams
 - **RRHs** – Regional Referral Hospitals
 - **HMIS** – Health Management Information System
 - **DHIS** – District Health Information Software
 - **KIU-TH** – Kampala International University Teaching Hospital
 - **HSDP** – Health Sector Development Plan
 - **NDP** – National Development Plan
 - **NIN** – National Identification Number
 - **OpenMRS** – Open Medical Records Systems
 - **DHA** – Digital Health Atlas
 - **MAPS** – mHealth Assessment and Planning for Scale
 - **HRHIS3** – Human Resource for Health Information Software 3
 - **UDHD** – Uganda Digital Health Dashboard
 - **ViB** – VODAN in a Box
 - **SWOT** – Strengths, Weaknesses, Opportunities, and Threats
 - **WHO** – World Health Organization
 - **IFDS** – Internet of FAIR Data and Services

-
- **CEDAR** – Center for Expanded Data Annotation and Retrieval
 - **REST** – Representational State Transfer
 - **SNOMED CT** – Systematized Nomenclature of Medicine – Clinical Terms
 - **LOINC** – Logical Observation Identifiers Names and Codes
 - **ICD** – International Classification of Diseases
 - **HIS** – Health Information System
 - **FAIRDOM-SEEK** – Findable, Accessible, Interoperable, Reusable DOMain

Appendix B

Summary

This thesis is about making health data in digital form accessible for Health Systems in Africa, and in particular in Uganda. In order for data to be computer readable and reusable, the data should be Findable, Accessible, Interoperable and Reusable (FAIR). This is commonly referred to as the FAIR-principles. This thesis presents the design and implementation of the FAIR principles to digital health data in Africa, with a specific focus on Uganda. Health data, when in digital form, remain reusable and beneficial to others. In Uganda and other African countries, health data are collected and stored in paper format using Health Management registers. For the Uganda case, the aggregated data are routinely submitted to the DHIS2 system as required by the Uganda Ministry of Health. Data are not easily accessible and reusable by researchers and health professionals at the point of care.

The Virus Outbreak Data Network (VODAN) has been a crucial part of the research presented in this thesis and as such our research became part of the VODAN-Africa and Asia implementation network in March 2020. This network which aims to address current and immediate challenges in the use and connectivity of digital health data. The VODAN-Africa and Asia initiative is central to an expanding network of equal partnerships between African and European universities. These partnerships produce or utilize high-quality, high-volume, machine-actionable data, bringing African health insights into the global medical knowledge base to improve health and healthcare in Africa. The VODAN Africa and Asia architecture is a set of integrated tools with triple-store databases where datasets are stored and made available for discovery, access, and reuse under clearly defined terms.

Our study is driven by the need to bridge the gaps in digital health systems in Africa by exploring the transformative potential of the principles of FAIR data. The purpose of FAIR data support design is to overcome prevalent concerns about the interoperability and reusability of digital health data by researchers and different stakeholders. This is achieved by implementing remote querying facilities in Uganda and other member countries of VODAN-Africa, ensuring that data remain within health facilities while enhancing the quality of care. Understanding the need for FAIR data requires recognizing its potential to transform the current data landscape. Adherence to FAIR principles ensures that data become more findable, accessible, interoperable, and reusable. This thesis is structured into nine chapters, beginning with an introduction in Chapter 1, which outlines the research problem, the objectives, and the significance of the principles of FAIR data. The general problem we are addressing is 'How can data integration, supported by the FAIR data principles, address the challenges of data interoperability and reuse

within the African digital health sector?’ In Chapter 2, we review the VODAN Africa and Asia initiative and the global implementation of FAIR principles. This chapter provides a global perspective on the transformative role of the FAIR principles in scientific data management and stewardship with a specific focus on Africa. This highlights the persistent challenges within African healthcare systems, as illustrated, underscoring the urgent need for inclusive and equitable FAIR practices in diverse scientific domains. We examine how the VODAN initiative addresses key challenges, including barriers with respect to collaboration, limitations in data reuse, sustainable development, and the co-creation of knowledge. According to the multiple stream framework of Kingdon, the corresponding theory of agenda setting is applied in the African context to explore the adoption of FAIR Data, highlighting the crucial role of policy actors in facilitating this transition.

In Chapter 3, we focus on investigating the data management practices and information flow in Ugandan health facilities through focus group discussions and interviews. We discuss the current challenges of health data management in Uganda, the limitations of paper-based systems, and the potential role of the FAIR architecture in improving data accessibility and reuse.

In Chapter 4, we assess Uganda’s regulatory framework for digital health and examine the feasibility of the adoption of FAIR principles. The critical challenges paper-based data systems pose in Ugandan health facilities are highlighted, especially amid health crises such as the COVID-19 pandemic. We argue that aligning Uganda’s digital health strategy with the FAIR principles could significantly improve data governance, interoperability, and reusability. To assure a good alignment, between the FAIR standards and Uganda’s health data regulations, we specifically address current gaps.

In Chapter 5, we explore how the FAIR principles can enhance interoperability in Uganda’s digital health initiatives. We assess the interoperability of two crucial digital health dashboards: the Digital Health Atlas Uganda (DHA-U) and the Uganda Digital Health Dashboard (UDHD), along with their associated vertical digital health tools within the Ugandan context. The initial phase involved assessing the level of FAIRness of DHA-U, the UDHD, and the vertical digital health solutions within these hubs using the FAIR Evaluation Services tool; an automated web-based service. Using a FAIR Data point, the level of FAIRness of the digital health hubs was enhanced. This facilitated integration and alignment with FAIR principles. Subsequently, the level of FAIRness was reassessed after the FAIRification process, using FAIR Evaluation Services.

In Chapter 6 we introduce the VODAN-in-a-Box (ViaB) system and demonstrate a proof-of-concept using the ViaB system for COVID-19 data transformation and access. To make this possible FAIR Data Points are used. These are distributed data repositories that host machine-actionable data and metadata that adhere to the FAIR Guidelines. In participating institutions, the FAIR Data Point is deployed through the ViaB that comprises a dockerised bundle of tools. ViaB is a set of multiple open-source and FAIR-enabling services with the goal to support the gathering of World Health Organization (WHO) electronic case report forms (eCRFs) as FAIR data in a machine-actionable way, but without exposing or transferring the data outside the facility. Following the execution of a proof-of-concept, ViaB was deployed in Uganda and at Leiden University.

In Uganda and other African counties, the DHIS2 platform is commonly used. In Chapter 7, we compare FAIR data with DHIS2 for privacy-preserving health analytics. The lack of flexibility in semantic data production was observed in the first proof-of-concept. This could be solved

by localizing existing platforms, such as the Center for Expanded Data Annotation and Retrieval (CEDAR), for locally deployable versions. We examined the comparative strengths and limitations of DHIS2, and FAIR data in ensuring data privacy while allowing accessibility and reusability for effective health analytics. We elaborated on the integration of the VODAN-Africa architecture to enhance the Findability, Accessibility, Interoperability, and reuse of clinical and research data across various healthcare facilities in Uganda. Through a detailed critical analysis of their functionalities, data governance frameworks, and application contexts, the study identifies best practices and challenges to ensure privacy-preserving analytics.

In Chapter 8 we examine the integration of machine learning models with FAIR health data for advanced analytics and decision making. Using anonymized datasets from VODAN-Africa, specifically from Uganda, i.e., Uganda's Outpatient Department (OPD) records and Uganda's integrated antenatal care (ANC) information. These data sets are structured and standardized according to the FAIR principles and are used to train various machine learning models for tasks such as predicting pregnancy risk factors, assessing fetal conditions, detecting anomalies and performing clustering analyzes. We have highlighted the effective integration of Machine Learning with FAIR data in a federated architecture in healthcare analytics and underscored the significance of high-quality, accessible, and reusable data to support collaborative research and enhance public health outcomes.

Finally, in Chapter 9, we enumerate the conclusions and future outlook of the research. The use of FAIR data principles provides a viable solution to improving data interoperability and reuse in Uganda's digital health sector. The FAIR principles can enhance the efficiency and effectiveness of health data use by addressing gaps in the policy landscape, strengthening data integration and analysis, and enabling data reusability through privacy-preserving access. Consequently, the findings of this study provide a foundation for future initiatives aimed at strengthening data management approaches in Uganda and comparable contexts.

Appendix C

Samenvatting

In dit proefschrift wordt beschreven hoe gezondheidszorg gegevens in digitale vorm toegankelijk kunnen worden gemaakt voor gezondheidszorgsystemen in Afrika, Oeganda in het bijzonder. Om data leesbaar en bruikbaar te laten zijn voor computersystemen, is het een vereiste dat de gegevens voldoen aan de principes van: Findable, Accessible, Interoperable and Reusable (FAIR). Deze principes zijn bekend als de zogenaamde FAIR-principes. Dit proefschrift presenteert het ontwerp en de implementatie van deze FAIR-principes voor digitale gezondheidszorg gegevens in Afrika, met Oeganda als specifieke casus. In een gedigitaliseerde vorm zijn gezondheidszorg gegevens herbruikbaar. In Oeganda en andere Afrikaanse landen worden gezondheidszorg gegevens verzameld en opgeslagen in papieren archieven waarbij gebruik gemaakt wordt van Gezondheid Management Registers. De verzamelde gegevens worden routinematig ingediend in het DHID2 systeem volgens de richtlijnen van het Oegandese Ministerie van Gezondheid. Deze gegevens zijn echter niet makkelijk toegankelijk en herbruikbaar voor onderzoekers en gezondheidsprofessionals bij specifieke zorglocaties.

Het Virus Outbreak Data Network (VODAN) is een cruciaal onderdeel van het onderzoek dat in dit proefschrift wordt gepresenteerd, daarmee is ons onderzoek onderdeel geworden van de implementatie van VODAN-Afrika/Azië netwerk in Maart 2020. Dit netwerk beoogt de huidige en onmiddellijke uitdagingen aan te pakken in het gebruik van en de verbindingen tussen digitale gezondheidszorg gegevens. Het VODAN-Afrika/Azië initiatief staat centraal in een zich uitbreidend netwerk van gelijkwaardige partners tussen Afrikaanse en Europese universiteiten. Deze partnerschappen produceren of gebruiken hoogkwalitatieve, groot volume, machine-uitvoerbare data, waarmee Afrika-gerelateerde inzichten in gezondheidszorg deel worden van de globale kennisbasis over gezondheid waardoor gezondheid en zorg in Afrika kunnen verbeteren. De architectuur in VODAN-Afrika/Azië bestaat uit een set van gereedschappen geïntegreerd in triple-store databases waarin datasets zijn opgeslagen en beschikbaar gemaakt voor bevinding, toegang en hergebruik door middel van duidelijk gedefinieerde termen.

Motivatie voor onze studie is de noodzaak de lacunes in de digitale gezondheidszorgsystemen in Afrika te overbruggen door het bestuderen van de potentiële hervorming van data door het toepassen van FAIR-principes. Het doel van het ontwerp van FAIR data ondersteuning is het overwinnen van bestaande zorgen over interoperabiliteit en herbruikbaarheid van digitale gezondheidszorg gegevens door onderzoekers en de verschillende belanghebbenden. Dit doel wordt gerealiseerd door het implementeren van query-faciliteiten op afstand in Oeganda en andere lidstaten binnen de VODAN-Afrika gemeenschap, met de waarborg dat de data binnen

de gezondheidsinstelling blijven terwijl de kwaliteit van de zorg vergroot. Het begrijpen van de noodzaak van FAIR data vereist erkenning dat dit potentieel het huidige datalandschap kan hervormen. Naleven van FAIR-principes zorgt ervoor dat de gegevens beter vindbaar, toegankelijk, interoperabel en herbruikbaar worden. Dit proefschrift is onderverdeeld in 9 hoofdstukken, waarbij in hoofdstuk 1 de probleemstelling, de doelstellingen en de significantie van het principe van FAIR data, worden beschreven. Het generieke probleem dat in dit proefschrift wordt geadresseerd is: “Hoe kan data integratie, ondersteund door FAIR-principes, worden aangewend bij uitdagingen van interoperabiliteit en herbruikbaarheid binnen de Afrikaanse sector van digitale platformen voor gezondheidszorg?”.

In hoofdstuk 2, geven we een overzicht van het VODAN-Afrika/Azië initiatief en de implementatie van FAIR-principes. Het hoofdstuk geeft een globaal perspectief van de transformatieve rol die FAIR-principes met betrekking tot wetenschappelijk gegevensbeheer en rentmeesterschap, specifiek gericht op Afrika. Hiermee worden de structurele uitdagingen binnen de Afrikaanse zorgsystemen benadrukt, en de urgentie voor een inclusieve en rechtvaardige FAIR-praktijk in de diverse wetenschappelijke domeinen onderstreept. We onderzoeken we hoe het VODAN-initiatief de belangrijkste uitdagingen aanpakt, inclusief de barrières met betrekking tot samenwerking, beperkingen voor hergebruik van data, duurzame ontwikkeling, en de co-creatie van kennis. Volgens het “multiple stream framework” van Kingdon, is de corresponderende theorie van agendering toegepast in de Afrikaanse context om de adoptie van FAIR data te onderzoeken, waarbij de cruciale rol van politieke spelers in het faciliteren van deze overgang wordt uitgelicht. In hoofdstuk 3, concentreren we ons op het bestuderen van de praktijk van het datamanagement en de informatie stroom in Oegandese zorg-faciliteiten. Dit wordt gerealiseerd door discussies en interviews met focus groepen. We bediscussiëren de recente uitdagingen binnen het management van gezondheidszorg gegevens in Oeganda, de beperkingen van systemen die op papier zijn gebaseerd, en de potentiële rol die de FAIR architectuur heeft in het verbeteren van toegankelijkheid en herbruikbaarheid van de data. In hoofdstuk 4, geven we een inzicht in de kaders van de Oegandese regelgeving voor digitale gezondheidssystemen en beoordelen daarmee de haalbaarheid van de adaptatie van FAIR-principes. De cruciale uitdaging van de papier-gebaseerde data registratie in Oeganda worden onder de loep genomen waarbij speciale aandacht wordt gegeven aan gezondheidscrisis zoals de COVID-19 pandemie. We beargumenteren dat een herschikking van de Oegandese strategie voor digitale gezondheidssystemen door gebruik van FAIR-principes een significante verbetering kan realiseren in gegevensbeheer, interoperabiliteit en herbruikbaarheid. Om een goede samenspraak tussen FAIR-standaarden en de Oegandese regelgeving voor gezondheidszorg gegevens te borgen, besteden we daarbij specifiek aandacht aan de bestaande lacunes. In hoofdstuk 5 gaan we na hoe de FAIR-principes interoperabiliteit in de initiatieven voor Oegandese digitale gezondheidssystemen kunnen vergroten. We doen dit door de interoperabiliteit van twee cruciale gezondheidsinstrumenten te onderzoeken: i.e. “Digital Health Atlas Uganda (DHA-U)” en “Uganda Digital Health Dashboard (UDHD)”, tezamen met de digitale zorgoplossingen die in Oeganda gebruikelijk zijn. De initiële fase betrof het maken van een inschatting van het niveau van FAIRness van DHA-U en UDHD en de digitale zorgoplossingen in deze instrumenten door middel van het FAIR-Evaluation Service instrument; een automatisch web-gebaseerde service. Door gebruik te maken van een FAIR Data Point wordt de mate van FAIRness van deze spil gezondheidssystemen vergroot. Dit faciliteerde de integratie en inrichting met FAIR principes. Vervolgens is de mate van FAIRness opnieuw onder de loep genomen aan hand van een FAIRification proces dat gebruik maakt van het FAIR-Evaluation Service instrument.

In hoofdstuk 6 introduceren we het VODAN-in-a-Box (ViaB) systeem en demonstreren we een proof-of-concept met het ViaB systeem voor transformatie en toegang van COVID-19 data. Om dit mogelijk te maken, wordt gebruik gemaakt van zogenaamde FAIR-Data Points. Dit zijn gedistribueerde data opslagplaatsen van waaruit machine-actionable data en metadata die voldoen aan de FAIR richtlijnen, beschikbaar worden gesteld. In de deelnemende instituten wordt het FAIR-Data point in omloop gebracht door de ViaB dat bestaat uit een zelfstandige gereedschapsbundel. ViaB is dus een verzameling open-source applicaties van FAIR services met als doel het verzamelen van elektronische patiëntenrapportageformulieren (eCRFs) van de Wereld Gezondheid Organisatie (WHO) als FAIR data in een machine-actionable formaat, maar zonder dat de data buiten het instituut beschikbaar komen of worden gemaakt. Volgend op het proof-of-concept is ViaB uitgezet in Oeganda en op de Universiteit Leiden. In Oeganda en andere Afrikaanse landen is het DHIS2 platform gangbaar voor gezondheids data. In hoofdstuk 7 vergelijken we FAIR-data met DHIS2 met betrekking tot privacybeschermende gezondheidssanalyses. Het gebrek aan flexibiliteit in de productie van semantische data is vastgesteld in een eerste proof-of-concept. Dit kon worden opgelost door bestaande platforms te gebruiken, i.e. het Center for Expanded Data Annotation and Retrieval (CEDAR), voor lokaal inzetbare versies. We hebben de sterke en zwakke punten van DHIS2 en FAIR-data onderzocht met betrekking tot privacybescherming onder voorwaarden van toegankelijkheid en herbruikbaarheid voor effectieve data-analyse. We hebben de integratie van de VODAN-Afrika architectuur uitgewerkt om de vindbaarheid, toegankelijkheid, interoperabiliteit en herbruikbaarheid (FAIR) van klinische en onderzoek data vanuit verschillende gezondheidszorg faciliteiten. Door een gedetailleerde kritische analyse van functionaliteiten, platforms voor databeheer, en toepassingsgebieden hebben we in deze studie de best-practices geïdentificeerd samen met uitdagingen om privacybeschermende analyse te garanderen.

In hoofdstuk 8 onderzoeken we de integratie van machine-learning modellen met FAIR gezondheidszorg data voor geavanceerde analyse en besluitvorming. Er wordt gebruik gemaakt van geanonimiseerde datasets uit VODAN-Afrika, specifiek uit Oeganda; i.e. Uganda's Outpatient Department (OPD) gegevens en Uganda's Integrated Antenatal Care (ANC) informatie. Deze datasets zijn gestructureerd volgens FAIR-principes en zijn gebruikt om verschillende machine-learning modellen te trainen. Dit is gedaan voor taken zoals voorspellen van zwangerschapsrisico factors, inschatten van conditie van de foetus, detecteren van anomalieën en verschillende vormen van clustering van de gegevens. We hebben de effectieve integratie van machine-learning van analyses voor gezondheidszorg in een federated architectuur en met FAIR data kunnen markeren en daarbij hebben we het belang onderstreept van data van goede kwaliteit die toegankelijk en herbruikbaar zijn voor het ondersteunen van onderzoekssamenwerking en het verbeteren van de uitkomsten van de gezondheidszorg.

Tenslotte, presenteren we in hoofdstuk 9 de conclusies van het onderzoek en een toekomstvisie. Het gebruik van FAIR data principes geeft een realistische oplossing voor het verbeteren van interoperabiliteit en herbruikbaarheid van data in de digitale gezondheidszorg in Oeganda. De FAIR principes kunnen de efficiëntie en effectiviteit van het gebruik van gegevens uit de gezondheidszorg door de lacunes in het beleid adresseren, de integratie en analyse te versterken, en het hergebruik mogelijk maken door privacybeschermende toegang. Dus, de bevindingen van deze studie geven een fundament voor toekomstige initiatieven die tot doel hebben aanpak van gegevensbeheer in Oeganda te versterken en verbeteren.

Appendix D

Curriculum Vitae

Mariam Basajja was born in Kampala, Uganda, in 1993. She completed her bachelor's degree in Applied Computer Technology with a focus on Software Engineering from the United States International University of Africa in 2012. In addition, she did a Master in Advanced Computing, Machine Learning, and Data Mining from the University of Bristol, UK, from 2016 to 2017 and a Master in Computational Intelligence from the University of Nairobi, Kenya, from 2016 to 2018. In both masters, her main focus was on machine learning models and health data. After the masters, Mariam did an internship at the GO-FAIR Foundation and started her PhD under the supervision of Prof. Dr. Ir. Fons J. Verbeek and Dr. Katy Wolstencroft in the Imaging and BioInformatics group at Leiden University in 2019. She participated in the VODAN-Africa project headed by Mirjam van Reisen. This comprised of software research, collaborations with different organizations and institutions. During her PhD studies, she enrolled for the following certifications; Academic writing, speed reading, effective communication, among others. She plans to pursue her career in health systems and FAIR data after her Ph.D.

Appendix E

Acknowledgements

On arriving to Leiden, The Netherlands, Prof. Mirjam Van Reisen welcomed me for a quiet dinner with a bowl of bitterballen, classic Dutch snacks, crispy on the outside, soft and savory on the inside, which automatically became my go-to snack since then. I would like to extend my deepest appreciation to Prof. Mirjam Van Reisen for the support and guidance during this study. Thanks to Mirjam and Barend Mons, I did an internship at the GO-FAIR Foundation, which led to the start of my PhD at Leiden Institute of Advanced Computer Science.

My heartfelt gratitude to my supervisors Prof.dr.ir. F.J. Verbeek and Dr. K.J. Wolstencroft for their continued support, encouragement, and unwavering guidance throughout this research. Their timely and insightful feedback, expertise, and mentorship have been invaluable in shaping this work. I couldn't have done it without you. My sincerest appreciation to Marloes van der Nat for the moral support and encouragement during this journey.

Profound appreciation goes to the GO-FAIR Foundation, Barend Mons, Erik Schultes, Bert Meerman, Marek Suchanek and the rest of the GO-FAIR Foundation team for providing technical guidance during the one on one interactions and FAIR data stewardship trainings carried out during the first phase of the proof of concept in this research.

My outmost gratitude to the CEDAR platform administration and technical support team for their prompt and effective support and expertise during the Training of Trainers data stewardship workshops. CEDAR was a key tool used to create machine-readable reusable meta-data in the second phase of the proof of concept. Their expertise in the design and implementation of FAIR principles in digital health greatly enriched this study, especially during the platform localization process.

My deepest gratitude to Kampala International University, Kampala Uganda, for making it possible for me to complete my studies at Leiden University. The author expresses sincere thanks to Leiden University for providing the necessary resources and academic environment that made this research possible. I am grateful to my colleagues and peers in the VODAN-Africa team; Dr. Reginald M. Nalugala, Samson Yohannes, Dr. Araya Abrha Medhanyie, Dr. Ibrahim Abdullahi, Getu Tadele, Ruduan Plug, Abdullahi Abubakar Kawu, Aliya Aktau, Tesfit Gebremeskel, and the LIACS imaging and Bioinformantics group; Danyi Liu, Shima Javanmardi, Solomiia Kurchaba, Lise Stork, Ana Arcos, Irene Martorelli and Sacha Goultiaev for their constructive discussions, motivation, and collaboration.

I also would like to acknowledge the Ministry of Health Uganda, the permanent secretary, Dr. Diana Atwine, district officers, hospital directors, statisticians, data management personnel,

KIU REC and other stakeholders for providing the necessary approvals and cooperation that made this work possible.

Special thanks goes to my family for their unconditional love, moral support, and understanding. My friends who stood by me through this journey, thank you for the belief in me and encouragement.

Finally, I dedicate this work to my father and mother, whose inspiration has been a guide for my journey.