

Health Data Poverty and Biases in Sub-Saharan African Community Health Settings: A Scoping Review Protocol

Barimwotubiri Ruyobeza

barimawotubiri.ruyobeza@uct.ac.za

University of Cape Town

Sudesh Sivarasu

University of Cape Town

Study protocol

Keywords: "sub-Saharan Africa", "SSA", "West Africa", "East Africa", "Southern Africa", "Central Africa", as well as specific leading digital health countries: "South Africa", "Kenya", "Nigeria", "Rwanda", "Ethiopia", "Ghana", "Zambia", "Uganda", "Tanzania"

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11233504/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

The rapid acceleration of artificial intelligence (AI) in healthcare holds transformative potential for sub-Saharan Africa (SSA). However, real-world deployment in primary healthcare (PHC) is severely constrained by '*health data poverty*'—the systemic under-representation and exclusion of marginalised populations. Furthermore, facility-centric data collection, primarily from urban tertiary health facilities, creates a '*pre-hospital*' data vacuum, systematically omitting early community-level symptom evolution and health-seeking behaviours. Compounding this, existing datasets suffer from '*gender blindness*' and structural fragmentation, while governance frameworks struggle to reconcile national data sovereignty mandates with cross-border, cloud-based AI processing. This scoping review aims to systematically map the existing literature on SSA community healthcare settings, dataset collection and curation limitations, data and algorithmic biases, and socio-governance challenges to establishing an empirical knowledge base for designing a responsible, gender-inclusive health dataset collection framework.

Methods

Guided by Joanna Briggs Institute (JBI) methodology, the Arksey and O'Malley framework, and PRISMA-ScR standards, a Population, Concept, Context (PCC) framework was formulated. Comprehensive searches across major bibliographic databases (PubMed/MEDLINE, Embase, Web of Science, Scopus, AJOL), preprint servers, trial registries, and grey literature will capture studies published in English or French between January 2019 and November 2026. Following dual independent title/abstract and full-text screening in Rayyan, extracted data will be appraised using JBI Critical Appraisal Tools. Synthesis will combine quantitative descriptive statistics with qualitative thematic analysis guided by the Consolidated Framework for Implementation Research (CFIR).

Discussion

By diagnosing frontline data-capture bottlenecks, gendered data gaps, and governance redlines, the findings will directly inform downstream digital tools—such as longitudinal cancer pre-screening datasets—advancing proactive, equitable, and data-driven community and primary care across sub-Saharan Africa.

Registration

The protocol is registered with the Open Science Framework (<https://osf.io/qdwr4>) to ensure methodological transparency and accountability.

I. BACKGROUND

Artificial intelligence (AI) and machine learning (ML) are likely to fundamentally reshape healthcare delivery across sub-Saharan Africa (SSA) [1]. By leveraging complex datasets—including social determinants of health, geographical risk factors, and electronic health records—algorithmic systems are already being piloted for disease outbreak prediction, resource allocation, and the personalisation of health interventions [2]. For instance, during the COVID-19 pandemic, algorithms were utilised to prioritise vaccine distribution and map social

vulnerabilities at the community level [3], [4]. However, despite this radical potential, the implementation of AI-driven health research in SSA remains uneven and fragmented [5], [6]. While high-resource settings have demonstrated strong AI performance, translating these advancements into the data-constrained and socio-culturally diverse environments of SSA is hindered by fragile digital infrastructure, limited technical capacity, and significant governance gaps [7].

The central barrier to realising responsible AI in the region lies in the datasets used to train these models [8]. Existing literature indicates that African health datasets are frequently scarce, of poor quality, non-representative, and highly fragmented [7], [2]. A critical concern is the '*facility-centric*' nature of existing records, which predominantly originate from tertiary urban hospitals. This reliance on urban, facility-based data may lead to limited representation of rural populations, informal settlements, and vulnerable communities. There is also a '*pre-hospital*' data vacuum in which early individual and community-level symptoms, health-seeking behaviours, and environmental risk exposures are not recorded [9], [10]. Without this foundational data, AI systems designed for early disease detection, such as cancer or cardiovascular pre-screening, are unable to learn from the very patterns they are intended to identify [11], [12].

Compounding these geographic disparities is the pervasive issue of '*health data poverty*,' defined as the inability of specific groups to benefit from innovations due to non-representative data [8], [13]. The growth of digital health data is uneven; most available data comes from wealthy regions with high adoption of electronic health records, widening the data divide [7]. In the context of SSA, existing datasets also often suffer from '*gender blindness*,' failing to capture sex-specific symptoms or the social determinants of health that uniquely affect women and girls [14], [15].

This lack of inclusive data risks reinforcing systemic inequalities through algorithmic bias, in which models trained on skewed data produce inaccurate generalisations about marginalised populations [16], [17]. Furthermore, the lack of transparent governance structures and unclear mechanisms for community participation interferes with the development of '*naturally responsible*' and inclusive AI [18], [19]. While the literature on AI in global health is expanding, existing reviews fail to address the specific needs of the proposed data collection framework. A scoping review by Al-Ganad et al. [1] focused on the deployment of medical AI in low-resource settings, but its scope was primarily implementation-focused barriers, such as infrastructure and workforce capacity, rather than the granular mapping of community-level dataset limitations. Similarly, Baichoo et al. [20] conducted a scoping review of AI in Africa but specifically limited their scope to *mobile technology and social media* tools, omitting the broader governance and dataset design challenges required for integrated community and primary healthcare. Yet, a global review of '*health data poverty*' found that biases are driven by data gaps, including limited or non-participation by countries or authors from the African continent, which, in turn, limits the data's contextual relevance to the SSA landscape [13].

Moreover, recent reports on AI governance in Africa, such as the *Science Policy Engagement with Africa's Research (SPEAR)*'s comprehensive review [21], indicated that while health is a priority area, there is a distinct *absence of health-specific AI governance* frameworks. Most existing national strategies are generic and lack the '*redlines*' necessary to balance innovation with ethical patient safety and data protection in the health sector [7], [14]. Consequently, to our knowledge, there is no empirically validated framework to support the design and collection of high-quality, representative, and ethically labelled health datasets for AI/ML training, specifically tailored to the diverse community contexts of SSA [22], [23].

We therefore argue that a scoping review is the most appropriate research design for this study for several logical reasons. First, the existing literature on the intersection of SSA community health settings, dataset biases, and AI governance is *limited* and *fragmented*, making a full systematic review—which requires a more homogeneous body of evidence—unsuitable [24], [25]. Second, the evidence base spans *diverse methodologies*, including qualitative descriptions, descriptive reports, pilot implementations, and expert commentaries, none of which lend themselves to meta-analysis [26], [27]. Third, the review seeks to map the ‘*extent and nature*’ of available evidence and identify conceptual gaps to inform the initial design of a new framework rather than to provide evidence or confirm some hypothesis [28], [29]. Fourth, scoping reviews are uniquely effective for *mapping complex* and *under-theorised concepts*—such as context-specific algorithmic governance and adequate data collection mechanisms—and for establishing a ‘knowledge base’ for future research. This approach aligns with the *Arksey and O'Malley framework* for identifying key concepts in emerging fields [30], [31].

The purpose of this protocol is therefore to detail the methodology for a scoping review that will systematically synthesise evidence on SSA community health settings and their associated dataset limitations, biases, and governance challenges. By mapping the "problem space" of African data poverty, this review will develop a reusable methodological framework for designing responsible AI data infrastructures, beginning with the collection of healthcare data. The specific objectives are to:

- *Map diverse African community healthcare settings* (e.g., rural, urban-poor, post-conflict) and identify their unique data collection challenges.
- *Synthesise evidence on dataset quality* and the specific biases identified in currently deployed AI models in SSA.
- *Identify governance and ethical hurdles* across different contexts, including issues of data sovereignty versus cloud dependency.
- *Highlight implications for gender-inclusive AI*, specifically focusing on the protection of vulnerable groups.

The review is expected to generate novel insights necessary to transition SSA's healthcare systems from reactive, curative models toward more *proactive, data-driven, and equitable care*, ensuring that African populations are not merely data sources but also the primary beneficiaries of the AI revolution.

II. METHODS/DESIGN

To guide this scoping review, we adopt the *Arksey and O'Malley methodological framework* [25], enhanced by the refined guidelines of the *Joanna Briggs Institute (JBI)* [27]. The final synthesis of evidence will adhere strictly to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR)* reporting standards [32]. For greater research transparency and accountability, this protocol will be registered on the *Open Science Framework (OSF)*: <https://osf.io/qdwr4> [33]. In the next section, we describe our core research questions, literature search strategies, eligibility criteria, study screening workflows, and quality assessment procedures, as well as our plans for data extraction, charting, and synthesis. We have provided the completed *PRISMA-P checklist* as a supplementary Table S1, and we conclude this protocol with a brief discussion addressing the broader implications of this research.

A. RESEARCH QUESTIONS

As outlined in the background, primary and community healthcare (PHC) systems in SSA face a critical challenge in establishing a resilient digital health foundation due to systemic *health data poverty* [8]. Despite the rapid global expansion of AI, most available datasets are facility-centric, skewed, and unrepresentative, leaving a profound gap in our understanding of community-level data limitations. This scoping review is designed to systematically map these community settings and their associated dataset limitations, biases, and governance hurdles to establish the foundational knowledge base for our wider research project. As shown in Table 1, we applied the *Population, Concept, Context* (PCC) framework to formulate a research question aligned with the objective of our study [29]:

· **What health dataset collection and curation limitations and challenges exist within the sub-Saharan African community and primary healthcare settings?**

The above question is structured to systematically address distinct operational and conceptual domains. Within the above question, *collection* refers to the physical and digital processes of gathering clinical, demographic, or self-reported patient information from the field [13]. *Curation* includes the subsequent cleaning, structuring, annotating, and labelling of this raw data to make it programmatically useful [34]. Related sub-research questions to address include:

- *What* distinct community healthcare settings (e.g., urban, rural, underserved, urban-poor informal settlements) are documented in SSA, and *what* are their unique data collection environments?
- *What* dataset limitations (e.g., scarcity, incompleteness, fragmentation) and data biases (e.g., geographic exclusion, gender blindness) have been documented in currently deployed or piloted health AI models in each of these contexts?
- *What* ethical, legal, and data governance challenges (specifically regarding *data sovereignty versus cloud dependency*) are reported when managing health data in these community settings?
- *What* design requirements, operational standards, or data collection practices are identified in the literature as necessary for building representative, ethically governed, and inclusive health datasets in SSA?

Table 1: PCC-Informed Scoping Review Question Formulation

PCC Element	Description for the Scoping Review	Justification and Linkage to the Study Aim
Population (P)	All SSA citizens of all ages, including underserved, marginalised, or rural populations and patients within the SSA community and primary healthcare (PHC) contexts.	Expand the review to the previously excluded human populations from urban, facility-centric datasets, directly addressing "health data poverty".
Concept (C)	Health dataset collection and curation practices and mechanisms, technical limitations (scarcity, quality, fragmentation, lack of labelling), biases (geographic exclusion, gender blindness), and socio-governance challenges (data sovereignty vs cloud dependency, consent gaps, regulatory deficits).	Maps how community-level health data is gathered, cleaned, structured, and governed, establishing empirical parameters needed to design a responsible, gender-inclusive AI dataset collection framework.
Context (C)	SSA primary and community healthcare settings (e.g., rural clinics, urban-poor informal settlements, underserved regions, crisis/post-conflict zones, and community-oriented care facilities).	Explicitly bounds the review to the frontline primary healthcare environments where early screening, risk factor recording, and community data capture must occur.

Limitations include documented technical gaps in accuracy, completeness, consistency, uniqueness, timeliness, and validity, as well as overall dataset scarcity, extreme fragmentation across clinics, paper-based dependencies, non-interoperable formatting, and a lack of clinically meaningful labels or annotations for robust AI/ML training. *Challenges* encompass the broader non-technical barriers, including *ethical, legal, social, and institutional* hurdles. In the SSA context, this also includes *weak personal data protection frameworks, limited local data sovereignty safeguards, unclear consent models, and a lack of transparent data governance mechanisms* that allow for active community participation.

The term '*sub-Saharan African community health settings*' explicitly defines the geographical and institutional context of the inquiry, focusing on community and primary healthcare (PHC) and related facilities in SSA, as well as localised settings (e.g., rural clinics, underserved regions, and urban-poor informal settlements), as opposed to urban tertiary facilities.

B. SEARCH STRATEGY AND IDENTIFICATION OF RELEVANT STUDIES

A comprehensive, reproducible, and multi-phased search strategy will be employed to map the existing literature on community health data collection, dataset curation limitations, and AI governance challenges across SSA. The study has enlisted the services of a qualified librarian to assist in designing the search strategy. The search will target literature published from *January 2019 to November 2026*. This date range is selected because 2019 marks a major turning point in African digital health and data governance, characterised by the implementation of key national digital strategies (e.g., South Africa's National Digital Health Strategy 2019–2024 and Kenya's Data Protection Act of 2019) and a subsequent significant global acceleration in health-related AI/ML applications.

B1. Key Sources of Literature

To capture a highly interdisciplinary body of clinical, computational, legal, and ethical evidence, we will query three distinct bibliographic categories:

1. Bibliographic Databases:

- o *Global Medicine & Public Health*: PubMed/MEDLINE, Embase, and Web of Science (Core Collection).
- o *Nursing & Allied Health*: EBSCO (CINAHL).
- o *African Region & ICT Research*: Scopus, African Journals Online (AJOL), and EBSCO (Africa-Wide Information).

2. Preprint Servers and Trials Registries:

- o *Emerging Data Science*: medRxiv and JMIR Preprints will be searched to capture rapid developments in digital health before formal peer review.
- o *Clinical Registries*: The Pan African Clinical Trials Registry (PACTR) and ClinicalTrials.gov will be searched to identify clinical trials utilising or generating community-level datasets across SSA.

3. Grey Literature and Supplemental Sources:

- o High-impact digital health and global policy databases will also be queried, including the World Health Organisation (WHO) Digital Health Atlas, the Science for Africa (SFA) Foundation repository, national Ministry of Health portals of leading SSA digital innovators, the African Union (AU) Commission, and the GIZ FAIR Forward Initiative.
- o Targeted hand-searching will, in addition, be conducted in high-impact journals, including *The Lancet Digital Health*, *BMC Digital Health*, *Journal of Medical Internet Research*, and *BMJ Global Health*.

B2. Eligibility Criteria (PICOS Framework)

Study eligibility is defined using the PICOS (*Population, Intervention, Comparison, Outcomes, Study Design*) framework. This structure aims to maintain scientific objectivity and replication.

Table 2: Inclusion Criteria (Based on PICOS Framework)

PICOS Element	Inclusion Criteria	Justification and Linkage to Study Sub-Questions
Population (P)	All citizens, including underserved, marginalised, or rural populations and patients within the SSA community and primary healthcare (PHC) contexts.	Expand the review to include previously excluded individuals from urban, tertiary "facility-centric" datasets, addressing the "pre-hospital" data vacuum.
Intervention / Concept (I)	Existing (or non-existing) health dataset collection and curation practices and mechanisms (including clinical recording, electronic health records, paper-to-digital data capture, symptom classification, and consent schemas).	Maps how community-level patient health data is currently gathered, cleaned, structured, and labelled within primary care workflows to identify technical gaps.
Comparison (C)	Not applicable.	N/A
Outcome(s) (O)	Documented <i>dataset limitations</i> and collection barriers (scarcity, completeness, quality, fragmentation, lack of labelling) and <i>socio-governance challenges</i> (ethical consent gaps, genomic under-representation, gender blindness, and data sovereignty versus cloud dependency).	Identifies the empirical limitations and real-world governance hurdles involved in designing a responsible, representative, and ethically governed dataset collection framework.
Study Design (S)	Peer-reviewed qualitative, quantitative, or mixed-methods studies; descriptive pilots; case studies; policy reports; regulatory guidelines; and intergovernmental charters.	Captures both structured scientific studies and real-world regulatory implementations across Africa.
Language & Period	Studies published in English or French between January 2019 and November 2026.	Aligns with the post-2019 digital health policy landscape in SSA while matching reviewer linguistic capacities.

Exclusion Criteria

Studies will be excluded if they:

1. Are conducted entirely outside of sub-Saharan Africa.

2. Do not address data collection, curation, structuring, labelling, or algorithmic governance issues (e.g., purely laboratory-based computer science papers testing an algorithm on standard global benchmark datasets without an African system context).
3. Focus solely on clinical efficacy trials without discussing data infrastructure, EHR characteristics, or database curation challenges.
4. Were published before January 2019.
5. Study protocols or those lacking sufficient primary or secondary data to address at least one of our sub-research questions.

B3. Search Terms and Controlled Vocabulary

To construct a comprehensive search string, keywords will be mapped to database-specific controlled vocabularies—specifically, MeSH (Medical Subject Headings) in PubMed/MEDLINE and Emtree in Embase—and combined with natural-language search terms.

Key Search Terms by Concept Block

· Block 1: Geographic Focus (Sub-Saharan Africa)

o *MeSH*: "Africa South of the Sahara"[Mesh], "Developing Countries"[Mesh]

o *Keywords*: "sub-Saharan Africa", "SSA", "West Africa", "East Africa", "Southern Africa", "Central Africa", as well as specific leading digital health countries: "South Africa", "Kenya", "Nigeria", "Rwanda", "Ethiopia", "Ghana", "Zambia", "Uganda", "Tanzania".

· Block 2: Population / Primary Care Context

o *MeSH*: "Primary Health Care"[Mesh], "Community Health Services"[Mesh], "Rural Population"[Mesh], "Vulnerable Populations"[Mesh]

o *Keywords*: "primary care", "primary healthcare", "community health", "rural health", "underserved clinics", "crisis", "post-conflict", "conflict zone", "informal settlements", "urban-poor", "pre-hospital".

· Block 3: Concept (AI Data Science, Health Informatics & Dataset Curation)

o *MeSH*: "Artificial Intelligence"[Mesh], "Machine Learning"[Mesh], "Datasets as Topic"[Mesh], "Electronic Health Records"[Mesh], "Information Storage and Retrieval"[Mesh]

o *Keywords*: "artificial intelligence", "AI", "machine learning", "deep learning", "natural language processing", "NLP", "algorithm", "algorithmic", "predictive model", "electronic health record", "electronic health records", "EHR", "EHRs", "electronic medical record", "EMR", "EMRs", "dataset", "database", "data collection", "data curation", "labeling", "labelling", "annotation", "data quality", "data storage", "data retrieval", "information storage", "information retrieval", "registries", "incompleteness", "fragmentation", "data poverty", "data sharing", "representativeness", "biomedical registry".

· Block 4: Outcomes (Limitations, Biases, and Governance & Security Tensions)

o MeSH: "Bias"[Mesh], "Information Governance"[Mesh], "Computer Security"[Mesh], "Ethics, Medical"[Mesh]

o Keywords: "bias", "biases", "biased", "algorithmic bias", "gender bias", "dataset bias", "sampling bias", "racial bias", "underrepresentation", "disparities", "algorithmic bias", "systemic exclusion", "gender blindness", "data sovereignty", "cloud dependency", "information governance", "data governance", "computer security", "data security", "cybersecurity", "data residency", "NDPR", "POPIA", "consent", "informed consent", "operational standards", "data standards", "regulation", "ethics", "ethical", "medical ethics", "bioethics", "regulatory", "redlines", "infrastructure constraints".

PubMed / MEDLINE Search String Example

An example of a PICO-guided PubMed search string is provided below and will be adapted syntactically for other databases. The search string is formulated using Boolean operators (AND, OR) and exact wildcards:

```
((("Bias"[Mesh] OR "Computer Security"[Mesh] OR "Ethics, Medical"[Mesh]) OR ("algorithmic bias" OR "systemic exclusion" OR "gender blindness" OR "data sovereignty" OR "cloud dependency" OR POPIA OR consent OR "operational standards" OR "data standards" OR regulation OR ethics OR "infrastructure constraints")) AND (("sub-Saharan Africa" OR "SSA" OR "West Africa" OR "East Africa" OR "Southern Africa" OR "Central Africa" OR "South Africa" OR Kenya OR Nigeria OR Rwanda OR Ethiopia OR Ghana OR Zambia OR Uganda OR Tanzania) OR ("Africa South of the Sahara"[Mesh]))) AND (((((primary care OR primary healthcare OR primary health-care OR community health OR rural health OR "underserved clinics" OR crisis OR "post-conflict" OR "conflict zone" OR "informal settlements" OR "urban-poor" OR pre-hospital OR prehospital) OR ("Primary Health Care"[Mesh])) OR ("Community Health Services"[Mesh])) OR ("Rural Population"[Mesh])) OR ("Vulnerable Populations"[Mesh]))) AND ((dataset OR database OR data collection OR data curation OR labeling OR labelling OR annotation OR data quality OR incomplete OR fragmentation OR data quality OR data poverty OR data sharing OR representative OR "biomedical registry") OR ("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Datasets as Topic"[Mesh] OR "Electronic Health Records"[Mesh] OR "Information Storage and Retrieval"[Mesh]))
```

Filters Applied: Publication date from 2019/01/01 to 2026/08/31; Language: English, French.

B4. Study Selection and Screening Process

All records identified through the electronic database and grey literature searches will be exported to *Rayyan* for automated and manual deduplication. The deduplicated list will then be moved into the Rayyan screening section for screening. The screening process will be executed in two distinct, sequential stages to ensure methodological rigour and minimise bias:

1. **Title and Abstract Screening:** Two authors, Mary Kama (MK) and Carlin Foka (CK), will independently screen all titles and abstracts against the predefined inclusion and exclusion criteria to filter out clearly irrelevant studies (e.g., studies on non-medical AI or studies conducted exclusively in high-income settings).

2. **Full-Text Screening:** Next, a different set of two independent reviewers, Veronica Kariuki (VK) and Richard Wairagu (RW), will independently review the full text of all potentially eligible articles. They will evaluate each paper using the PICOS criteria detailed in Table 2.

Any disagreements between the two reviewers regarding a study's eligibility will first be resolved through a structured discussion. If consensus cannot be reached, the lead author, BR, will serve as a third-party arbitrator to

make the final decision after consultations. The results of the search, screening, and selection process will be reported in detail and illustrated using a standardised *PRISMA-ScR Flow Diagram* in the final manuscript.

C. CHARTING THE DATA

To comprehensively address the *Main Research Question (MRQ)* and related *Sub-Research Questions (SRQs)* set out in this protocol, data extraction will follow a structured charting form adapted from the Joanna Briggs Institute (JBI) guidelines and PRISMA-ScR standards [35]. Table 3 below outlines the specific data elements to be extracted, their detailed descriptions, and their explicit justification for answering each research question.

Table 3: Charted Data Elements and Justification

Data Element	Description of the Data to be Extracted	Justification (How this information helps answer the MRQ and/or SRQs)
Bibliographic Information	Article title, lead author, co-authors, publication year, publication type (peer-reviewed journal, policy report, regulatory draft, grey literature), study region/country in sub-Saharan Africa (SSA), and country of corresponding author.	Identifies the origin and publication timeline, maps the geographical distribution across SSA regions, and evaluates the representation of local versus external authors to assess regional research capacity. <i>Supports general contextual mapping for the MRQ and SRQ1.</i>
Study Design and Methodology	Research design (e.g., qualitative interviews, quantitative surveys, mixed-methods, pilot evaluation, case study, policy analysis, regulatory review, or expert commentary) and analytical framework applied.	Evaluates the methodological diversity and quality profile of the literature, distinguishing between empirical field implementations, regulatory analyses, and theoretical models. <i>Supports the methodological appraisal for MRQ and SRQ1.</i>
Community Healthcare Setting and Context	Specific SSA healthcare environment (e.g., rural health clinics, urban-poor informal settlements, underserved regions, crisis/post-conflict zones, or community-oriented primary care) and primary care facility tier.	Addresses the ‘facility-centric’ data gap by evaluating whether datasets originate from primary, community settings or tertiary urban hospitals, addressing the ‘pre-hospital’ data vacuum. <i>Directly answers SRQ1.</i>
Health Data Collection Mechanisms	Primary mechanisms and tools used to gather clinical, demographic, or self-reported patient information (e.g., paper logs, mobile applications, community health worker [CHW] digital tools, electronic health records [EHRs], or community registries).	Maps how community-level patient health data is currently collected and recorded within primary care workflows across SSA, highlighting gaps in the paper-to-digital transition. <i>Directly answers SRQ1.</i>
Technical Curation Limitations and Dataset Quality	Documented dataset defects and curation gaps, including overall scarcity, structural fragmentation across clinics, paper dependencies, missing values, lack of interoperability (e.g., HL7/FHIR compliance), and absence of clinical labelling or machine-learning annotations.	Synthesises empirical evidence on dataset quality, curation limitations, and formatting deficits that constrain machine learning model training in African health systems. <i>Directly answers the MRQ and SRQ2.</i>
Algorithmic and Data Biases and Demographic Exclusion	Specific biases reported in datasets or deployed AI models, including geographic exclusion (urban vs. rural), genomic under-representation, ‘pre-hospital’ symptom data vacuums, and ‘gender blindness’ (failure to record sex-specific symptoms or gendered social determinants of health).	Identifies structural biases in existing health datasets that risk perpetuating health disparities through resultant algorithmic bias and defines necessary parameters for gender-inclusive data collection. <i>Directly answers the MRQ and SRQ2.</i>
Ethical, Legal, and Socio-Governance Challenges	Documented governance hurdles, including <i>data sovereignty versus cloud dependency</i> , personal data protection compliance (e.g., POPIA, NDPR, AU Malabo Convention), informed consent gaps, commercial exploitation risks, and lack of community participation mechanisms.	Identifies how cross-border data flows, cloud dependencies, and regulatory deficits impact patient privacy, data ownership, and ethical governance in SSA primary care. <i>Directly answers SRQ1 and SRQ3.</i>

Operational Principles and Framework Requirements	Recommended design requirements, operational practices, ethical safeguards, gender-sensitive data fields, governance structures, and alignment with national/global guidelines (e.g., SAMA Booklet 20, WHO, UNESCO guidelines).	Synthesises actionable design parameters, operational standards, and ethical safeguards required to construct a reusable, context-appropriate AI data-collection framework for SSA. <i>Directly answers SRQ2 and SRQ3.</i>
Implementation Feasibility and Acceptability	Reported evidence regarding technical feasibility, clinician/community acceptability, digital infrastructure constraints (power grid instability, low internet bandwidth), financial sustainability, workforce literacy, and pilot lessons learned.	<i>Determines real-world deployment viability</i> by evaluating operational enablers and systemic barriers that influence the adoption, usability, and scale of digital data collection systems in resource-constrained African settings. <i>Directly answers SRQ4.</i>
Design Requirements & Operational Standards	Recommended operational practices, data collection standards, ethical guidelines, gender-sensitive data fields, structured labelling/annotation protocols, and alignment with national/global frameworks (e.g., SAMA Booklet 20, WHO, UNESCO, AU guidelines).	Synthesises actionable design parameters, operational standards, and ethical principles for constructing a representative, ethically governed, and inclusive health data collection framework. <i>Directly answers SRQ4.</i>
Authors' Conclusions and Identified Evidence Gaps	Main takeaway messages, policy recommendations, strategic guidance for health system reorientation, and explicitly cited literature or evidence gaps.	Summarises strategic recommendations and clarifies unaddressed research areas, establishing the empirical knowledge base. <i>Supports the overarching synthesis for MRQ and SRQ1–SRQ4.</i>

The lead author, BR, will initially extract the above data elements from the included publications. Data extracted using the above standardised charting form will be managed in Rayyan and Microsoft Excel, where it will undergo iterative cross-verification by the review team to ensure completeness, accuracy, and methodological consistency.

D. COLLATING, SUMMARISING AND REPORTING THE RESULTS

Following data extraction from included publications, the review team will systematically synthesise, analyse, and report the findings of the study to comprehensively address the primary research question: ‘*What health dataset collection and curation limitations and challenges exist within sub-Saharan African community health settings?*’ A mixed-methods synthesis approach—combining quantitative descriptive mapping and qualitative thematic analysis—will be employed to summarise the findings, in line with the Joanna Briggs Institute (JBI) guidance for scoping reviews and the PRISMA-ScR reporting standards [36].

1. Quantitative Descriptive Synthesis

Quantitative analysis will be conducted to map the overall scope, volume, and distribution of the retrieved literature across SSA. Descriptive statistics (frequencies and percentages) will be calculated to present:

- *Bibliographic and Publication Trends*: Temporal distribution of studies published between January 2019 and November 2026, tracking post-2019 policy shifts and digital health acceleration in SSA.

- *Geographical Scope and Research Capacity*: Mapping study locations across West, East, Central, and Southern Africa, alongside an evaluation of lead author affiliations (local African vs. external international institutions) to assess regional research leadership.
- *Methodological Profile*: Distribution of study designs (qualitative, quantitative, mixed-methods, pilot evaluations, case studies, policy reviews, and regulatory charters).
- *Community Healthcare Tiers & Data Capture Tools*: Categorisation of documented primary healthcare settings (rural clinics, urban-poor informal settlements, underserved regions, and crisis/post-conflict zones) and the primary mechanisms used for patient data gathering (paper registers, mHealth apps, CHW digital tools, EHRs, and clinical registries).

The results of this quantitative mapping will be presented using summary tables, bar charts, and geographical maps of SSA to illustrate where evidence is concentrated and where critical data voids persist.

2. Qualitative Thematic Synthesis

To dissect the socio-technical, biopolitical, and legal complexities of African health data ecosystems, qualitative evidence will be analysed using thematic synthesis, guided by the Consolidated Framework for Implementation Research (CFIR) [37]. The qualitative synthesis will be organised into four core thematic domains that map directly to our sub-research questions (SRQs):

1. *Contextual Healthcare Environments & Capture Mechanisms (SRQ1)*: Synthesis of evidence detailing the physical, digital, and structural environments of primary healthcare facilities in SSA. This theme will analyse how data capture operates at the community level and evaluate the extent to which existing datasets suffer from an urban, tertiary-hospital '*facility-centricity*' that may exclude the '*pre-hospital*' clinical narrative.
2. *Technical Curation Deficits & Biases (SRQ2)*: Analysis of documented technical dataset defects, including overall scarcity, incompleteness, structural fragmentation across clinics, paper-based dependencies, non-interoperability, and the lack of clinical labels or machine-learning annotations. Furthermore, this theme will map structural biases in deployed AI models—specifically geographic exclusion, genomic under-representation, and '*gender blindness*' (the failure to record sex-specific symptoms and gendered social determinants of health).
3. *Ethical, Legal, & Socio-Governance Tensions (SRQ3)*: Examination of regulatory friction points and governance hurdles surrounding African health data. This includes analysing the tension between local *data sovereignty mandates* and *cloud-based AI dependencies*, personal data protection compliance (e.g., South Africa's POPIA, Nigeria's NDPR, Kenya's Data Protection Act, and the AU Malabo Convention), informed consent gaps, commercial exploitation risks, and the absence of transparent community participation mechanisms.
4. *Operational Design Requirements & Standards (SRQ4)*: Synthesis of actionable operational practices, data collection standards, ethical principles, and gender-sensitive data schemas documented in the literature. This theme will synthesise the best practices and regulatory guidelines (such as SAMA Booklet 20, WHO, UNESCO, and AU guidelines) required to construct representative, ethically governed, and inclusive health datasets.

3. Integration and Reporting

The final synthesis will merge these quantitative overviews and qualitative thematic insights into an integrated narrative report. Findings will be illustrated using a PRISMA-ScR Flow Diagram (detailing study identification, screening, eligibility, and inclusion) alongside structured summary tables. The synthesised results will define the operational requirements and governance principles for designing, piloting, and validating a reusable, gender-inclusive AI dataset collection framework and its associated cancer pre-screening tool in sub-Saharan Africa.

III. DISCUSSION AND SIGNIFICANCE

The rapid global acceleration of AI and ML in healthcare risks exacerbating existing health disparities in SSA if systems continue to be built on flawed, non-representative data [38]. While AI algorithms demonstrate high accuracy in high-resource settings, their deployment in African community and primary healthcare (PHC) is severely constrained by *'health data poverty'*—the systemic under-representation and exclusion of marginalised populations from digital health datasets [6], [8].

The primary value of this scoping review lies in its role as an *empirical mapping tool*. Currently, the literature on African health data, digital tools, and AI governance remains highly fragmented across clinical, technical, legal, and policy domains [1], [16]. By systematically searching, extracting, and synthesising this scattered evidence across SSA community settings, this review provides the essential evidence base needed *before* technological or policy solutions are designed. Building AI data collection tools or screening decision-support systems without first mapping the real-world limitations of frontline data capture risks embedding the very structural biases and infrastructure failures that cause health inequities in the first place [8], [6], [13], [38].

In addition to structural health data problems in SSA, *health data poverty* across the region manifests in three operational challenges that further undermine the quality, completeness, and reach of community primary healthcare (PHC) systems [8], [13]. Rather than existing as abstract theoretical hurdles, these problems actively compromise patient outcomes, erode clinical trust, and restrict equitable access to healthcare services across the SSA region [39], [13], [38].

One such problem is the pervasive issue of *gender blindness and demographic non-representativeness* in existing clinical records. Standard health data capture schemas routinely omit sex-specific symptom variations, maternal health indicators, and gendered social determinants of health [21], [38]. When digital health and AI models are trained on these skewed, incomplete datasets, they inherit and amplify systemic algorithmic biases [13]. Algorithms deployed at the primary care level produce inaccurate risk assessments, misclassify diagnostic priorities, and underdiagnose health conditions in women, girls, and marginalised demographic groups [38]. This lack of representative training data subjects vulnerable populations to misapplied generalisations, directly reducing the clinical accuracy of frontline care and widening health disparities between demographic groups [13], [38].

A second major operational problem is the acute friction between *national data sovereignty mandates and the functional reliance on cloud-based AI processing* [40], [6], [21]. National data privacy laws—such as South Africa's POPIA, Nigeria's NDPR, Kenya's Data Protection Act, and the AU Malabo Convention—mandate strict local data protection, data residency, and patient privacy safeguards [41],[6]. However, high-performance AI tools often rely on centralised, cross-border cloud infrastructures for machine learning training and processing [40], [21]. In the absence of context-specific governance frameworks and clear regulatory *'redlines,'* community and primary care health systems operate under significant legal and ethical ambiguity [39]. Unclear consent models and a lack of

transparent community participation leave patients vulnerable to privacy breaches and the exploitation of commercial data [6], [40], [13].

At the same time, this regulatory uncertainty prevents local PHC facilities from safely adopting cloud-enabled screening and screening decision-support tools. This governance deadlock paralyses digital health innovation, erodes community trust in electronic record systems, and restricts rural and primary care clinics' access to advanced screening and diagnostic capabilities [39], [38], [40].

Finally, there is a profound structural disconnect caused by *tertiary, facility-centric health data collection* [39], [38], [42], [43]. Existing health datasets in SSA are overwhelmingly collected in urban centres, with limited data from rural clinics, informal settlements, and PHC workflows [10], [38], [42]. This reliance on health facility-based data creates a '*pre-hospital*' data vacuum, wherein early community-level symptom evolution, environmental risk exposures, and initial primary care health-seeking behaviours remain completely unrecorded [10], [44], [9]. Because these early clinical trajectories are missing, community and primary healthcare systems remain locked in reactive, treatment-oriented models where patients are only captured upon reaching advanced, late-stage disease—such as Stage III and IV cancer presentations [11], [42], [9]. Furthermore, AI tools designed for early pre-screening or risk stratification (such as for breast, cervical, or prostate cancers) cannot learn from real-world community patterns when trained on retrospective hospital records, rendering community and primary care providers incapable of offering proactive, preventive care [45], [11], [13].

Systematically mapping current SSA community healthcare settings alongside their associated dataset limitations, biases, and governance hurdles brings these three interlocking operational problems into a clear view [1], [2], [8], [13]. By searching across diverse primary care contexts—including rural clinics, urban-poor informal settlements, underserved regions, and crisis zones—this scoping review visualises where data collection fails, where algorithmic biases originate, and where governance friction halts implementation [38], [6].

Beyond identifying gaps, this evidence mapping establishes the essential empirical *knowledge base* (Study 1 within a Design Science Research framework) required to construct a context-appropriate framework for collecting health datasets [46]. By synthesising real-world data-capture limitations and governance standards across SSA, the review defines the specific design requirements, gender-sensitive data schemas, and ethical parameters needed to design the proposed representative, ethically governed dataset-collection framework [15], [47], [41]. The above knowledge base will facilitate downstream digital tools—such as longitudinal cancer pre-screening dataset development tools—ensuring that early community symptom capture, primary clinic screening, and tertiary hospital diagnostic outcomes can be linked, ethically labelled, and subsequently, made ready for responsible AI training [10], [44], [42], [9].

The design of this protocol is particularly methodologically rigorous. It combines the *Arksey and O'Malley framework* with *Joanna Briggs Institute (JBI)* refinements and *PRISMA-ScR* reporting standards, thus establishing a transparent and reproducible mapping process. Key strengths include dual independent screening, an extensive multi-database and grey literature search strategy, and the application of standardised JBI critical appraisal tools to evaluate the quality of evidence across qualitative, quantitative, and policy literature.

We also anticipate several limitations during the course of the study, including restricting study selection to English and French, which excludes Portuguese or other local-language evidence from certain African contexts. Additionally, the heterogeneity of the included study designs precludes meta-analysis, necessitating a qualitative

thematic synthesis guided by the CFIR. However, these constraints are necessary to ensure a manageable, rigorous mapping of a highly diverse evidence base.

In the end, this scoping review will provide the missing link between *high-level ethical AI principles* and the *frontline realities* of the African community and primary healthcare. By systematically diagnosing *health data poverty*, dataset curation gaps, data biases, and governance hurdles at source, this review ensures that future digital and AI health tools in SSA are built on representative, gender-inclusive, and ethically sound foundations.

Abbreviations

- **AI**

Artificial Intelligence

- **AJOL**

African Journals Online

- **AU**

African Union

- **CFIR**

Consolidated Framework for Implementation Research

- **CHW**

Community Health Worker

- **CINAHL**

Cumulative Index to Nursing and Allied Health Literature

- **EBSCO**

Elton B. Stephens Company Information Services

- **EHR / EHRs**

Electronic Health Record / Electronic Health Records

- **EMR / EMRs**

Electronic Medical Record / Electronic Medical Records

- **FHIR**

Fast Healthcare Interoperability Resources

- **GIZ**

Deutsche Gesellschaft für Internationale Zusammenarbeit (German Agency for International Cooperation)

- **HL7**

Health Level Seven

- **ICT**

Information and Communications Technology

- **JBIR**

Joanna Briggs Institute

- **JMIR**

Journal of Medical Internet Research

- **MEDLINE**

Medical Literature Analysis and Retrieval System Online

- **MeSH**

Medical Subject Headings

- **mHealth**

Mobile Health

- **ML**

Machine Learning

- **MRQ**

Main Research Question

- **NDPR**

Nigeria Data Protection Regulation

- **NLP**

Natural Language Processing

- **OSF**

Open Science Framework

- **PACTR**

Pan African Clinical Trials Registry

- **PCC**

Population, Concept, Context

- **PHC**

Primary Healthcare

- **PICOS**

Population, Intervention, Comparison, Outcomes, Study Design

- **POPIA**

Protection of Personal Information Act

- **PRISMA-P**

Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols

- **PRISMA-ScR**

Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

- **SAMA**

South African Medical Association

- **SFA**

Science for Africa Foundation

- **SPEAR**

Science Policy Engagement with Africa's Research

- **SRQ / SRQs**

Sub-Research Question / Sub-Research Questions

- **SSA**

Sub-Saharan Africa

- **UNESCO**

United Nations Educational, Scientific and Cultural Organization

- **WHO**

World Health Organization

Declarations

Ethics approval and consent to participate

Not applicable. This manuscript is a scoping review protocol based entirely on published, secondary, and publicly available literature and grey data. It does not involve human participants, individual patient records, animal experiments, or the collection of confidential primary data.

Consent for publication

Not applicable. No individual personal data, clinical images, or identifiable participant details are contained within this protocol.

Availability of data and materials

The protocol is registered on the **Open Science Framework** (<https://osf.io/qdwr4>). All search strategies, eligibility criteria, charting forms, and analytical templates generated during the study will be reported, archived and made available on request.

Competing interests

The authors declare that **they have no competing or financial interests** that could inappropriately influence or bias the design, execution, or reporting of this protocol.

Funding

The authors would like to acknowledge the Andrew Carnegie Foundation, through the African Technology Policy Studies Network (ATPS) and research funding support from the National Research Foundation of South Africa and SARCHI in Biomedical Engineering and Innovation grant number SARCI150224114428 & Grant UID 98788.

Authors' contributions

- **BR** conceptualised the study, formulated the research questions, designed the methodology, and drafted the initial manuscript.
- **Sudesh Sivarasu** critically reviewed, revised, and approved the final manuscript for protocol submission.

Acknowledgements

The authors gratefully acknowledge the technical assistance and guidance provided by **Ms Gill Morgan**, a qualified academic librarian, in formulating, testing, and refining the multi-database search strings.

References

1. A. Al-Ganad, A. Al-Shahdhi, O. Al-Dhaifi, E. Hajeb, H. Hajeb, and A. Al-Motarreb, "Deploying medical AI in low-resource settings: a scoping review of challenges and strategies," 2026, *Frontiers Media SA*. doi: 10.3389/fdgth.2026.1743634.
2. E. Mbunge and J. Batani, "Application of deep learning and machine learning models to improve healthcare in sub-Saharan Africa: Emerging opportunities, trends and implications," *Telematics and Informatics Reports*, vol. 11, Sep. 2023, doi: 10.1016/j.teler.2023.100097.
3. Fredrick and Ishengoma, "Applying Machine Learning to Optimize Vaccine Distribution for COVID-19," *International Journal of Open Information Technologies*, vol. 11, no. 5, pp. 47–53, 2023.

4. C. Zuo, Z. Meng, F. Zhu, Y. Zheng, and Y. Ling, "Assessing Vaccination Prioritization Strategies for COVID-19 in South Africa Based on Age-Specific Compartment Model," *Front. Public Health*, vol. 10, Jun. 2022, doi: 10.3389/fpubh.2022.876551.
5. A. Cossy-Gantner, S. Germann, N. R. Schwalbe, and B. Wahl, "Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings?," *BMJ Glob. Health*, vol. 3, no. 4, Jul. 2018, doi: 10.1136/bmjgh-2018-000798.
6. A. Oluka, "Mitigating Biases in Training Data: Technical and Legal Challenges for Sub-Saharan Africa," *International Journal of Applied Research in Business and Management*, vol. 5, no. 1, Jan. 2024, doi: 10.51137/ijarbm.2024.5.1.10.
7. A. Victor, "Artificial intelligence in global health: An unfair future for health in Sub-Saharan Africa?," Feb. 01, 2025, *Oxford University Press*. doi: 10.1093/haschl/qxaf023.
8. H. Ibrahim, X. Liu, N. Zariffa, A. D. Morris, and A. K. Denniston, "Health data poverty: an assailable barrier to equitable digital health care," Apr. 01, 2021, *Elsevier Ltd*. doi: 10.1016/S2589-7500(20)30317-4.
9. S. Rayne, K. Schnippel, D. Kruger, C. A. Benn, and C. Firnhaber, "Delay to diagnosis and breast cancer stage in an urban south african breast clinic," *South African Medical Journal*, vol. 109, no. 3, pp. 159–163, Mar. 2019, doi: 10.7196/SAMJ.2019.v109i3.13283.
10. S. Day et al., "Challenges and facilitators in pathways to cancer diagnosis in Southern Africa: a qualitative study," *BMJ Open*, vol. 15, no. 6, Jun. 2025, doi: 10.1136/bmjopen-2025-099296.
11. O. A. Bamodu and C.-C. Chung, "Cancer Care Disparities: Overcoming Barriers to Cancer Control in Low- and Middle-Income Countries," *JCO Glob. Oncol.*, no. 10, Aug. 2024, doi: 10.1200/go.23.00439.
12. N. Sarmah, M. N. Sibiya, and T. E. Khoza, "Barriers and enablers to breast cancer screening in rural South Africa," *Curationis*, vol. 47, no. 1, 2024, doi: 10.4102/curationis.v47i1.2583.
13. K. E. Paik et al., "Digital Determinants of Health: Health data poverty amplifies existing health disparities-A scoping review," *PLOS Digital Health*, vol. 2, no. 10, Oct. 2023, doi: 10.1371/journal.pdig.0000313.
14. S. Araba and A. Shamira, "An African perspective on gender and artificial intelligence needs African data and research," Cape Town, Oct. 2020. [Online]. Available: <https://www.mcit.gov.sa/sites/default/files/examples-of-ai-national-policies.pdf>
15. D. Cirillo et al., "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," Dec. 01, 2020, *Nature Research*. doi: 10.1038/s41746-020-0288-5.
16. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science (1979)*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.
17. S. M. Musa et al., "Paucity of Health Data in Africa: An Obstacle to Digital Health Implementation and Evidence-Based Practice," *Public Health Rev.*, vol. 44, 2023, doi: 10.3389/phrs.2023.1605821.
18. N. Tiffin, A. George, and A. E. Lefevre, "How to use relevant data for maximal benefit with minimal risk: Digital health data governance to protect vulnerable populations in low-income and middle-income countries," *BMJ Glob. Health*, vol. 4, no. 2, Apr. 2019, doi: 10.1136/bmjgh-2019-001395.
19. S. L. M. Davis, "Towards digital justice: participatory action research in global digital health," May 03, 2022, *BMJ Publishing Group*. doi: 10.1136/bmjgh-2022-009351.
20. S. Baichoo et al., "Scoping review of artificial intelligence via mobile technology and social media for health in Africa," *Nature Communications*, vol. 16, no. 1, Dec. 2025, doi: 10.1038/s41467-025-64766-4.

21. Alam and Uzma, "GOVERNANCE OF ARTIFICIAL INTELLIGENCE FOR GLOBAL HEALTH IN AFRICA A Review of Policy and Regulatory Frameworks," Pretoria, Jun. 2023.
22. A. Nyamawe and D. Shao, "On the use of synthetic data for healthcare AI in Africa: Technical performance, governance challenges, and policy readiness," Jan. 01, 2026, *SAGE Publications Inc.* doi: 10.1177/20552076261418897.
23. L. F. Vibbi, "Poor Data Quality in Sub-Saharan Africa and Implications on Ethical AI Development," in *SpringerBriefs in Applied Sciences and Technology*, vol. Part F2336, Springer Science and Business Media Deutschland GmbH, 2024, pp. 83–92. doi: 10.1007/978-3-031-52962-7_7.
24. Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," *BMC Med. Res. Methodol.*, vol. 18, no. 1, Nov. 2018, doi: 10.1186/s12874-018-0611-x.
25. H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework," *The International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, 2005, [Online]. Available: <http://journalsonline.tandf.co.uk/OpenURLlinktothearticle:http://www.journalsonline.tandf.co.uk/openurl.asp?genre=article&eissn=1464-5300&volume=8&issue=1&spage=19>
26. S. N. H. Hadie, "ABC of a Scoping Review: A Simplified JBI Scoping Review Guideline," *Education in Medicine Journal*, vol. 16, no. 2, pp. 185–197, 2024, doi: 10.21315/eimj2024.16.2.14.
27. M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, "Guidance for conducting systematic scoping reviews," *Int. J. Evid. Based. Healthc.*, vol. 13, no. 3, pp. 141–146, Sep. 2015, doi: 10.1097/XEB.0000000000000050.
28. D. Levac, H. Colquhoun, and K. K. O'Brien, "Scoping studies: Advancing the methodology," *Implementation Science*, vol. 5, no. 1, Sep. 2010, doi: 10.1186/1748-5908-5-69.
29. The Joanna Briggs Institute, "Reviewers-Manual_Methodology-for-JBI-Scoping-Reviews_2015_v2," Adelaide, 2015.
30. M. D. J. Peters et al., "Updated methodological guidance for the conduct of scoping reviews," *JBI Evid. Synth.*, vol. 18, no. 10, pp. 2119–2126, Oct. 2020, doi: 10.11124/JBIES-20-00167.
31. H. M. L. Daudt, C. Van Mossel, and S. J. Scott, "Enhancing the scoping study methodology: A large, inter-professional team's experience with Arksey and O'Malley's framework," 2013. doi: 10.1186/1471-2288-13-48.
32. A. C. Tricco et al., "PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation," Oct. 02, 2018, *American College of Physicians*. doi: 10.7326/M18-0850.
33. E. D. Foster and A. Deardorff, "Open Science Framework (OSF)," *Journal of the Medical Library Association*, vol. 105, no. 2, Apr. 2017, doi: 10.5195/jmla.2017.88.
34. M. L. Brodie, "What is Data Science?," in *Applied Data Science – Lessons Learned for the Data-Driven Business*, Eds., Braschler, M, Stadelmann, T, Stockinger, and K, Eds., Cambridge: Springer, 2019, ch. 8, pp. 101–130. [Online]. Available: <https://www.insight-centre.org/>
35. Covidence, "A practical guide: Data Charting and Extraction for Scoping Reviews," 2026.
36. S. Mak and A. Thomas, "Steps for Conducting a Scoping Review," *J. Grad. Med. Educ.*, vol. 14, no. 5, pp. 565–567, Oct. 2022, doi: 10.4300/JGME-D-22-00621.1.
37. C. M. Reardon et al., "The Consolidated Framework for Implementation Research (CFIR) User Guide: a five-step guide for conducting implementation research using the framework," *Implementation Science*, vol. 20, no. 1, Dec. 2025, doi: 10.1186/s13012-025-01450-7.

38. S. Naidoo and R. Naidoo, "Algorithmic governance in artificial intelligence-driven health systems: A southern African perspective," *Southern African Journal of Public Health (incorporating Strengthening Health Systems 2312–9360)*, p. e3530, Mar. 2026, doi: 10.7196/shs.2026.v8i3.3530.
39. A. Al-Ganad, A. Al-Shahethi, O. Al-Dhaifi, E. Hajeb, H. Hajeb, and A. Al-Motarreb, "Deploying Medical AI in Low-Resource Settings: A Scoping Review of Challenges and Strategies," Nov. 10, 2025. doi: 10.21203/rs.3.rs-8051581/v1.
40. B. Mukundwi, "Data Sovereignty Versus Cloud Dependency: Governance Challenges for African Health AI," 2026.
41. S. Mahomed, M. V. Ncube, A. Dhali, W. Janneker, and M. Nodikida, "Ethical governance of artificial intelligence in healthcare: Critical reflections of the AI Task Team of the South African Medical Association on the Health Professions Council of South Africa guidelines on the ethical use of AI in healthcare (Booklet 20)," Jun. 02, 2026, *South African Medical Association*. doi: 10.7196/SAMJ.2026.v116i5.5072.
42. W. Ngwa *et al.*, "Cancer in sub-Saharan Africa: a Lancet Oncology Commission," Jun. 01, 2022, *Elsevier Ltd*. doi: 10.1016/S1470-2045(21)00720-8.
43. O. Omotoso *et al.*, "Addressing cancer care inequities in sub-Saharan Africa: current challenges and proposed solutions," Dec. 01, 2023, *BioMed Central Ltd*. doi: 10.1186/s12939-023-01962-y.
44. J. Moodley *et al.*, "Mapping local evidence on early recognition and management of people with potential cancer symptoms: a narrative review," *S. Afr. Health Rev.*, vol. 26, Oct. 2024, doi: 10.61473/001c.120468.
45. K. D. Arendse *et al.*, "Implementation of mHealth to support cancer diagnosis in Sub-Saharan Africa: A systematic review," *Afr. J. Prim. Health Care Fam. Med.*, vol. 17, no. 1, pp. 1–13, 2025, doi: 10.4102/phcfm.
46. K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.
47. M. S. Lee, L. N. Guo, and V. E. Nambudiri, "Towards gender equity in artificial intelligence and machine learning applications in dermatology," *Journal of the American Medical Informatics Association*, vol. 29, no. 2, pp. 400–403, Feb. 2022, doi: 10.1093/jamia/ocab113.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S1CompletedPRISMAPchecklist.docx](#)