Systematic Review

Mapping Routine and Novel Data Sources for Predictive HIV Surveillance in Sub-Saharan Africa: A Scoping Review Protocol

Vincent Samuel Phiri[1,2], Felix Chisoni[3], Jim Mtambo[3], Salule Masangwi[1], Atupele Mulaga[1], Mphatso Kamndaya[1]

1. School of Sciences and Technology, Malawi University of Business and Applied Sciences
2. School of Global and Public Health, Kamuzu University of Health Sciences
3. Library Department, Kamuzu University of Health Sciences

Corresponding Author: Vincent Samuel Phiri; **Email:** vsamuel@kuhes.ac.mw

Abstract

Background

HIV/AIDS remains a leading cause of morbidity and mortality in Sub-Saharan Africa (SSA), accounting for approximately two-thirds of global HIV burden. Effective epidemic control requires timely, accurate, and geographically granular surveillance data. While traditional surveillance systems such as antenatal care (ANC) sentinel surveillance and population-based surveys have been foundational, they are often subject to delays, geographic gaps and limited predictive capacity. Over the past decade, several routine health systems data and novel data sources (including mobile phone data, geospatial indicators, phylogenetic surveillance and social media signals) have been proposed as complements or alternatives to enhance predictive surveillance. However, no comprehensive synthesis of these sources applied to HIV in SSA currently exists.

Objective

This scoping review aims to systematically map the types of routine and novel data sources that have been used or proposed for predictive HIV surveillance in SSA, characterise their methodological applications, and identify evidence gaps and research priorities.

Methods

We will conduct a scoping review following the Arksey and O'Malley (2005) framework as updated by Levac, Colquhoun and O'Brien (2010) and reported according to the PRISMA extension for Scoping Reviews (PRISMA-ScR). We will search PubMed, EMBASE, Web of Science, Scopus, WHO IRIS repository and grey literature without language or date restrictions. Studies will be included if they describe the use or validation of any data source for predicting HIV incidence, prevalence, transmission dynamics, or service demand in any SSA country. Data will be charted on source type, country, prediction target, modelling approach, and key findings, followed by a thematic narrative synthesis.

Expected Outputs

The review will produce a conceptual map of data sources, a taxonomy of modelling approaches, and a prioritised research agenda to guide future investments in HIV surveillance infrastructure across SSA. Findings will be disseminated through peer-reviewed publication, a policy brief, and open-access data sharing.

Keywords: HIV surveillance; predictive modelling; Sub-Saharan Africa; routine data; novel data sources; scoping review; epidemic intelligence; geospatial; mobile data; phylogenetics

Introduction

Background and Rationale

Sub-Saharan Africa (SSA) carries a disproportionate share of the global HIV burden. According to UNAIDS, approximately 25.6 million people in the region were living with HIV in 2023, representing roughly 65% of all people living with HIV worldwide1. Although substantial progress has been achieved through expanded antiretroviral therapy (ART) programmes, voluntary medical male circumcision (VMMC) campaigns, and pre-exposure prophylaxis (PrEP) rollout, transmission remains unacceptably high in many settings, and the 2030 UNAIDS targets of 95-95-95 remain out of reach in numerous countries[2].

A critical enabler of effective epidemic response is surveillance, the continuous, systematic collection, analysis, and interpretation of health-related data for planning, implementation, and evaluation of public health programmes. HIV surveillance in SSA has historically relied on a number of complementary approaches. Antenatal care (ANC) sentinel surveillance, which monitors HIV prevalence among pregnant women attending selected facilities, provided the earliest epidemiological signal in most countries and continues to inform UNAIDS global estimates. Population-based surveys, including the Demographic and Health Surveys (DHS), Population-based HIV Impact Assessments (PHIA), and the AIDS Indicator Surveys (AIS), offer nationally representative seroprevalence estimates but are conducted infrequently and cannot capture rapid temporal changes. Routine programme data from health management information systems (HMIS), including facility-based testing registers, ART cohort records, and voluntary counselling and testing (VCT) data, offer more continuous coverage but suffer from incompleteness, underreporting, and health-

© 2026 Kamuzu University of Health Sciences. This work is licensed under the Creative Commons Attribution 4.0 International License.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

https://dx.doi.org/10.4314/mmj.v38i4.

facility access biases.

These traditional surveillance pillars, while invaluable, share a common limitation: they are predominantly retrospective, slow to generate actionable insights, and poorly suited for forward-looking, predictive epidemic modelling at subnational levels. The 2014 Ebola outbreak in West Africa and, more recently, the COVID-19 pandemic have accelerated global interest in real-time epidemic intelligence[3], spotlighting the limitations of conventional surveillance architectures and catalysing investment in novel data streams.

Over the past decade, an expanding portfolio of routine and novel data sources has been evaluated for their potential to augment or transform HIV surveillance. These include but are not limited to routine health facility data from electronic medical records (EMR) systems and HMIS platforms; mobile phone call detail records (CDRs) and mobility data as proxies for population movement and sexual network connectivity[4,5], geospatial and remote sensing data capturing environmental and socioeconomic determinants of HIV risk[6], phylogenetic and genomic surveillance data enabling reconstruction of transmission networks[7], social media and internet search data reflecting risk behaviour or care-seeking signals[3], and machine learning (ML) and artificial intelligence (AI) frameworks integrating multiple heterogeneous data streams for predictive modelling[8]. In parallel, mechanistic transmission models including deterministic compartmental and individual-based simulation approaches have been used to project HIV incidence and to evaluate the population-level impact of prevention and treatment strategies[9–11].

Despite a growing body of empirical work applying these sources to HIV prediction in SSA, no comprehensive synthesis exists. Studies are fragmented across disciplines such as infectious disease epidemiology, computational biology, geography, informatics, and public health and are published in diverse journals with varying methodological rigour. Policymakers, programme managers, and researchers lack a consolidated map of what data sources are available, what has been tested, what works in which settings, and where critical evidence gaps remain. This scoping review is designed to fill that gap.

Defining Predictive HIV Surveillance

Conventional HIV surveillance is largely descriptive: it reports what has already been observed, such as diagnosed cases, testing yield, or prevalence measured in a past survey round. In this review, predictive HIV surveillance refers to the systematic use of HIV-related data, combined with statistical, machine-learning, or mechanistic models, to generate estimates of HIV outcomes that have not yet been directly observed. This encompasses four types of prediction task: (i) forecasting future HIV incidence, prevalence, or service demand; (ii) nowcasting, which adjusts current data for reporting delays or incompleteness; (iii) spatial prediction for areas or sub-populations without direct measurement, including small-area estimation and hotspot identification; and (iv) population-level risk stratification to prioritise geographic areas or sub-populations for intervention. Its defining feature is the production of forward-looking, or out-of-sample information intended to inform public health action before, rather than after, changes in the epidemic are observed.

Drawing on established guidance for evaluating public health surveillance systems[15] and for developing and reporting prediction models[16,17], we consider effective predictive surveillance to require five interrelated components:

• Data: sources that are timely, complete, representative of the target population, available at adequate spatial and temporal resolution, of sufficient time-series length, and capable of being linked or harmonised with other sources.

• Analytic methods: modelling approaches appropriate to the data structure, including the handling of spatial and temporal correlation, reporting delays, and missing data, with explicit quantification of uncertainty.

• Validation: assessment of predictive performance against independent or out-of-sample data, using measures such as calibration, discrimination, forecast error, and prediction-interval coverage.

• Actionability: prediction horizons and geographic units aligned with programme planning and decision-making cycles, with outputs that are communicated to and used by decision-makers.

• Governance and sustainability: data protection and ethical safeguards (particularly for individual-level and novel data sources), institutional ownership, and local analytic capacity to maintain predictive systems over time.

Therefore, this review maps the evidence across these components. Specifically, it covers: (i) the data sources used, grouped as routine health systems data (e.g., HMIS/DHIS2 aggregate data, electronic medical records, HIV testing services, ANC sentinel and PMTCT data, viral load and laboratory data, and case-based surveillance), population-based surveys (DHS, PHIA, AIS), and novel data sources (e.g., geospatial and remote sensing, mobile phone, social media and internet search, and phylogenetic or recency-testing data); (ii) the surveillance targets and prediction tasks addressed; (iii) the analytical methods applied; (iv) validation practices and reported predictive performance; (v) the geographic and population coverage of the evidence; and (vi) whether predictions were implemented or used in programme decision-making. As this is a scoping review, these components structure the data charting framework (Table 1) rather than serving as criteria for formal quality appraisal.

Objectives

The overarching aim of this scoping review is to map the current landscape of routine and novel data sources used or evaluated for predictive HIV surveillance in Sub-Saharan Africa. The specific objectives are:

1.To identify and categorise all data source types that have been applied to HIV surveillance prediction in SSA, distinguishing between routine health systems data and novel/non-traditional data sources.

2.To characterise the geographic scope, temporal coverage, and methodological approaches (statistical, machine learning, mechanistic modelling) used in conjunction with these data sources.

3.To assess the range of surveillance targets addressed (HIV incidence, prevalence, transmission dynamics, service demand, hotspot identification) and the prediction horizons reported.

4.To identify the SSA countries and sub-populations most represented in the literature, and those underrepresented.

https://dx.doi.org/10.4314/mmj.v38i4.

5. To map evidence gaps and formulate a prioritised agenda for future research and investment in predictive HIV surveillance infrastructure.

Methods

This protocol follows the scoping review methodology outlined by Arksey and O'Malley (2005)[12] and the subsequent methodological refinements proposed by Levac, Colquhoun, and O'Brien (2010)[16]. Reporting will adhere to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist[13]. The protocol will be registered in Open Science Framework[14].

Research Question

The primary research question guiding this review is:

"What routine and novel data sources have been used or proposed for predictive HIV surveillance in Sub-Saharan Africa, and what methodological approaches and prediction targets have been employed?"

Eligibility Criteria

Inclusion Criteria

Studies will be included if they meet all of the following criteria:

• Conducted in, or primarily focused on, one or more countries in Sub-Saharan Africa as defined by the United Nations geoscheme.

• Describe the use, validation, or comparative assessment of any data source routine or novel for the purpose of predicting, nowcasting, or forecasting HIV incidence, prevalence, transmission dynamics, spatial hotspots, or demand for HIV-related health services. Prediction is understood as defined in Section 1.2, encompassing forecasting, nowcasting, spatial prediction, and population-level risk stratification.

• Any real-world or simulated epidemic context within SSA, including both generalised and concentrated epidemics.

• Any empirical study design (including observational, modelling, simulation, and feasibility studies), systematic reviews addressing related sub-questions, technical/methodological papers evaluating surveillance tools, and grey literature reports from credible organisations (WHO, UNAIDS, national health ministries, PEPFAR, Global Fund, major academic institutions).

• No language restriction. Non-English studies will be translated as needed.

• No date restriction. All records from database inception to the search date will be considered.

Exclusion Criteria

• Studies focused exclusively on HIV treatment outcomes (e.g., ART adherence, viral load suppression) without a surveillance prediction component.

• Studies developing individual-level clinical risk scores solely for patient management, without a population-level surveillance or prioritisation application.

• Studies set entirely outside SSA (e.g., conducted solely in North Africa, Europe, or Asia) unless they include a comparative SSA component.

• Case reports and clinical case series not contributing to population-level surveillance inference.

• Opinion pieces, letters, or editorials without original data or synthesis.

• Conference abstracts for which full-text retrieval is not possible and no associated published study exists.

Search Strategy

A comprehensive search strategy will be developed in consultation with a health sciences librarian. The strategy will incorporate Medical Subject Headings (MeSH) and free-text terms across three conceptual domains: (1) HIV/AIDS, (2) surveillance and prediction, and (3) data sources and Sub-Saharan Africa. Keywords and related terms will be used to search the relevant literature. Boolean Operators AND as well as OR will be used to develop a comprehensive search strategy. The truncations (*) will further be used to enhance the search. Initial draft of the search string for PubMed/MEDLINE is presented below:

("HIV"[MeSH] OR "HIV Infections"[MeSH] OR "Acquired Immunodeficiency Syndrome"[MeSH] OR "HIV" OR "AIDS") AND ("Disease Notification"[MeSH] OR "Sentinel Surveillance"[MeSH] OR "Epidemiological Monitoring"[MeSH] OR "predictive model*" OR "nowcast*" OR "forecast*" OR "epidemic intelligence" OR "early warning" OR "risk prediction" OR "spatiotemporal" OR "hotspot") AND ("Africa South of the Sahara"[MeSH] OR "Sub-Saharan Africa" OR "Eastern Africa" OR "Western Africa" OR "Southern Africa" OR "Nigeria" OR "Kenya" OR "South Africa" OR "Uganda" OR "Tanzania" OR "Mozambique" OR "Zimbabwe" OR "Zambia" OR "Ethiopia" OR "Malawi" OR "Botswana" OR "Eswatini" OR "Rwanda" OR "Ghana")

The search will be adapted for each database using appropriate controlled vocabulary and Boolean operators. The databases to be searched include:

• PubMed (via NCBI)

• EMBASE (via Elsevier)

• Web of Science Core Collection (Clarivate)

• Scopus (Elsevier)

• WHO Global Index Medicus (including AFRO regional index)

• WHO IRIS institutional repository

• OpenGrey and ProQuest Dissertations & Theses (grey literature)

• PEPFAR.gov, UNAIDS.org, and Global Fund document repositories

In addition, reference lists of all included studies and relevant systematic reviews will be hand-searched (backward citation tracking). Forward citation tracking will be conducted using Google Scholar and Scopus for highly cited included studies.

Study Selection

Search results will be imported into Covidence systematic review software (or an equivalent, such as Rayyan) and deduplicated. The blind screening and selection process will be initiated. Study selection will proceed in two stages:

1. Title and abstract screening: Two independent reviewers will screen all titles and abstracts against the eligibility criteria. Disagreements will be resolved by consensus or referral to a third reviewer.

2. Full-text review: All potentially eligible studies will be retrieved in full text and assessed against the full eligibility criteria. Reasons for exclusion at the full-text stage will be documented. Under this stage, the conflicts regarding decisions will be resolved by consensus

https://dx.doi.org/10.4314/mmj.v38i4.

A PRISMAflow diagram will be produced documenting the number of records at each stage[15].

Data Charting

A standardised data charting form will be developed and piloted on ten randomly selected included studies before formal data extraction. Two reviewers will independently chart data from each included study, with discrepancies resolved by discussion. The following variables will be extracted:

Data Synthesis

Findings will be synthesised narratively and thematically. The synthesis will include:

• A descriptive numerical summary of included studies by year, country, study design, and data source category.

• A taxonomy of data source types, structured around the distinction between routine health systems data and novel/non-traditional data sources, with subcategories mapped to specific surveillance applications.

• A thematic synthesis of methodological approaches organised by prediction target and analytical framework.

• An evidence gap map visualising the intersection of data source types, SSA countries/sub-regions, prediction targets, and study density.

• A narrative summary of key cross-cutting themes, including data access and governance challenges, scalability, equity implications, and integration opportunities.

Quality Considerations

Although formal quality appraisal is not a component of scoping reviews, we will document study design and data source characteristics to allow readers to contextualise the evidence base. For modelling studies, we will note whether validation against independent data was performed. For studies involving novel data sources (e.g., mobile data), we will record whether ethical approval and data anonymisation processes were described. These characteristics will be charted against the components of predictive surveillance described in Section 1.2 (Table 1) in order to describe, rather than appraise, the evidence base.

Ethical Considerations

This scoping review involves the analysis of previously published data and does not require primary ethical approval. All included studies will be cited appropriately. We will, however, apply a lens of research ethics in our data charting, specifically noting whether included studies involving individual-level or sensitive data (e.g., mobile phone records, genomic sequences) reported ethical oversight, consent frameworks, and data protection compliance.

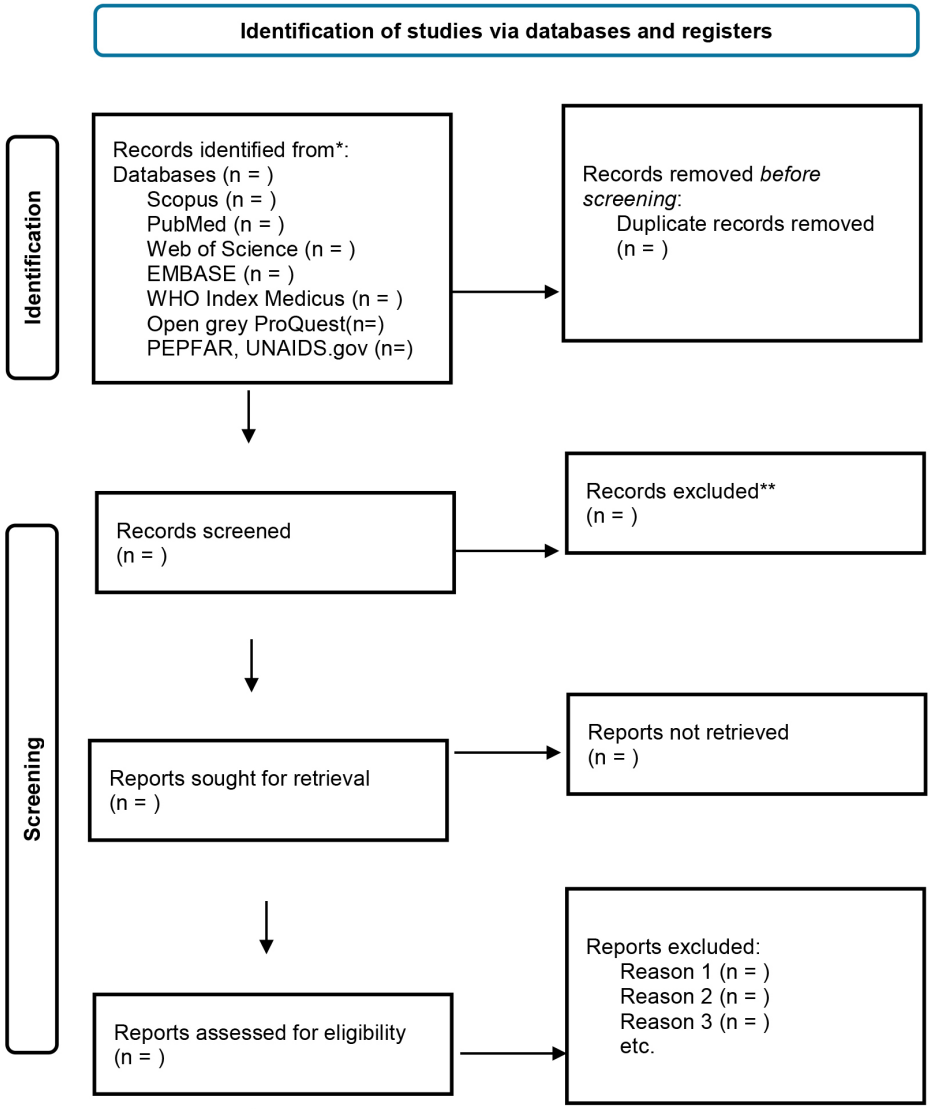


Figure 1: Prisma Flow Chart

Table 1. Data charting variables

Domain	Variables Extracted
Bibliographic	Author(s), year, title, journal/source, country of affiliation
Geographic scope	Country/countries, urban/rural setting, subnational level (region, district, community)
Study design	Observational, modelling, simulation, systematic review, mixed-methods, grey literature
Data source type	Routine health systems (ANC, HMIS, VCT, ART, PHIA/DHS/AIS) vs. novel (mobile/CDR, genomic, geospatial/RS, social media, mortality, pharmacy, etc.)
Data source details	Source name/platform, resolution (spatial/temporal), access modality (public/restricted), data custodian, reported timeliness and completeness
Prediction target	HIV incidence, prevalence, transmission network, hotspot identification, healthcare demand, cascade gaps
Analytical approach	Statistical model, mechanistic model, machine learning/AI, GIS/spatial analysis, phylogenetic analysis, mixed-methods
Prediction horizon	Nowcast, short-term (< 1 year), medium-term (1–5 years), long-term (> 5 years)
Performance metrics	Accuracy measures reported (AUC, sensitivity, specificity, RMSE, R^2, calibration metrics) if applicable
Validation and use	Validation approach (none, internal, external/out of sample), whether predictions were implemented or used in programme planning or decision making.
Population focus	General population, key populations (MSM, PWID, sex workers), pregnant women, adolescents/youth, specific age/sex strata
Key findings	Main conclusions on data source utility, limitations noted by authors
Funding & conflicts	Funding source, author conflicts of interest

This is particularly important given the stigmatised nature of HIV and the potential for re-identification in genomic and mobility datasets.

Planned Dissemination

Findings from this scoping review will be disseminated through multiple channels to reach both academic and policy audiences:

• High-impact open-access journal such as PLOS Medicine, The Lancet HIV, BMC Public Health, Systematic Reviews, or the Journal of the International AIDS Society).

• Presentation at a relevant international conference (e.g., International AIDS Conference, AIDS Impact, or the International Epidemiological Association World Congress).

• A concise policy brief targeted at national health ministries, PEPFAR/USAID country teams, Global Fund principal recipients, and UNAIDS regional offices in SSA.

• Open-access data sharing of the evidence gap map and charted dataset via an appropriate repository (e.g., OSF, Figshare, or Zenodo).

References

1. UNAIDS. Global HIV & AIDS statistics — Fact sheet 2024. https://www.unaids.org/en/resources/fact-sheet.

2. Stover J, Bollinger L, Izazola JA, Loures L, DeLay P, Ghys PD. What is required to end the AIDS epidemic as a public health threat by 2030? The cost and impact of the fast-track approach. PLoS ONE 2016;11:e0154893. https://doi.org/10.1371/journal.pone.0154893.

3. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Computational Biology 2012;8:e1002616. https://doi.org/10.1371/journal.pcbi.1002616.

4. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. Science 2012;338:267–70. https://doi.org/10.1126/science.1223467.

5. Tatem AJ, Qiu Y, Smith DL, Sabot O, Ali AS, Moonen B. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malaria Journal 2009;8:287. https://doi.org/10.1186/1475-2875-8-287.

6. Cuadros DF, Li J, Branscum AJ, Akullian A, Jia P, Mziray EN, et al. Mapping the spatial variability of HIV infection in Sub-Saharan Africa: effective information for localized HIV prevention and control. Scientific Reports 2017;7:9093. https://doi.org/10.1038/s41598-017-09464-y.

7. Grabowski MK, Lessler J, Redd AD. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Medicine 2014;11:e1001610. https://doi.org/10.1371/journal.pmed.1001610.

8. Brownstein JS, Rader B, Astley CM, Tian H. Advances in artificial intelligence for infectious-disease surveillance. New England Journal of Medicine 2023;388:1597–607. https://doi.org/10.1056/NEJMra2119215.

9. Granich R, Gilks CF, Dye C, De Cock KM, Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. The Lancet 2009;373:48–57. https://doi.org/10.1016/S0140-6736(08)61697-9.

10. Eaton JW, Johnson LF, Salomon JA. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. PLoS Medicine 2012;9:e1001245. https://doi.org/10.1371/journal.pmed.1001245.

11. Bershteyn A, Gerardin J, Bridenbecker D. Implementation and

applications of EMOD, an individual-based multi-disease modeling platform. Pathogens and Disease 2018;76:fty059. https://doi.org/10.1093/femspd/fty059.

12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005;8:19–32. https://doi.org/10.1080/1364557032000119616.

13. Tricco AC, Lillie E, Zarin W. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of Internal Medicine 2018;169:467–73. https://doi.org/10.7326/M18-0850.

14. Foster ED, Deardorff A. Open Science Framework (OSF). Jmla 2017;105. https://doi.org/10.5195/jmla.2017.88.

15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021:n71. https://doi.org/10.1136/bmj.n71.

16. Levac et al.: Scoping studies: advancing the methodology. Implementation Science 2010 5:69 doi:10.1186/1748-5908-5-69

17. Collins G S, Moons K G M, Dhiman P, Riley R D, Beam A L, Van Calster B et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods BMJ 2024; 385 :e078378 doi:10.1136/bmj-2023-078378