

Data Linkage for AI-Enabled Public Health

A practical guide to data linkage and better-connected health systems

This practical guide helps public health leaders and implementation teams plan, govern and implement data linkage. It outlines the decisions that shape a linkage investment, the steps to produce reliable, linked information, and the opportunities this creates for AI-enabled public health. It focuses not only on completing a linkage exercise, but also on building capabilities institutions can sustain and reuse.

Countries across the Global South are investing in the digital foundations of public health: digitizing services and records, building data infrastructure, strengthening civil registration and identification systems, and adopting standards.^{1,2} **When guided by shared standards and national digital health plans, these investments help institutions move toward more interoperable systems that can routinely share and use information.** Yet the work of connecting them is still developing. Data about the same person, service or outcome often remain spread across programs, facilities and public institutions, with each holding only part of the picture.³ Achieving interoperability requires more than a single platform, standard or technology purchase.⁴ It depends on reliable data, consistent use of standards, reliable unique identifiers, secure infrastructure, clear authority and responsibilities, and institutions that can work together across their different roles. These capabilities take time to build.^{5,6,7}

Data linkage—the process of connecting multiple datasets—offers a practical middle ground. It allows institutions to link selected records across separate datasets to answer a specific public health question

without waiting for every system to be fully interoperable. Governments can therefore derive greater value from existing data investments while broader interoperability foundations continue to develop. Linked data can answer questions that no individual source can address adequately on its own, for example, whether people complete a continuum of care, how service use relates to later outcomes, where surveillance systems are missing cases, or which populations remain underserved. Its value lies not in producing a larger dataset but in creating a more complete basis for public health analysis, decisions and action.^{8,9,10}

Data linkage is more than matching records. It often requires institutions that hold different parts of the data to work together and trust how the data will be handled, used and shared. A strong linkage effort therefore needs clear authority, usable data, secure ways to handle information, methods suited to the purpose, checks on match quality, and clear responsibility for the linked data it creates. Data linkage can support one-off analyses to answer specific questions, and the methodology can also generate insights not routinely available from a single dataset. When linkage is repeated, it can

also strengthen the wider path toward interoperability by revealing gaps in data quality, governance, identity and infrastructure, while building agreements, working relationships and methods that can be used again.¹¹ Linkage is not interoperability: It connects records across systems that remain separate, while interoperability allows systems to exchange and use information routinely. The aim is to create value in public health now while building stronger connections between systems and institutions over time, rather than creating a permanent workaround to address fragmentation.¹²

Institutions need not choose between waiting for an ideal interoperable architecture or continuing to make decisions from fragmented information. Well-governed data linkage can answer priority questions now while strengthening the capabilities required for more connected public health intelligence over time.

Data Linkage Use Cases

Data linkage can be used in different ways depending on the public health question and the data available. The examples below illustrate some common use cases and the types of records that may be linked to support them. This list is not exhaustive; the appropriate linkage will depend on the intended use, available data and local context.¹³

Use case	Records to link and what linkage can add
Find missed cases	Link disease or injury surveillance records with laboratory, hospital or death-registration records to identify potentially missed cases. Review these against the agreed case definition, population and time period to improve case capture and reveal reporting gaps.
Fill data gaps	Link health service records with household, social protection or other administrative records to add information such as household circumstances or participation in other services. This can help explain differences in health and service use.
Improve data quality	Link health records with civil registration or another verified administrative source to check missing or inconsistent fields, such as date of birth or place of residence, and investigate differences between sources.
Track care pathways	Link primary care, referral, hospital and treatment records to examine whether people reach the next service, complete treatment, experience delays or have gaps in follow-up.
Link services to outcomes	Link service or program records with later hospital admissions, follow-up records or death-registration data to examine survival, readmission and other outcomes that the original service record may not capture.
Identify underserved groups	Link birth registration or other population register records with immunization or other health service records to assess service reach among populations represented in the source records and identify possible gaps for further investigation.

Decisions That Shape a Data-Linkage Investment

Data linkage should begin with four connected decisions that determine the investment's scope, cost, risk and credibility. These decisions are interdependent and may need revisiting as institutions examine their data and implementation constraints.

1

Public Health Purpose and Intended Use

The intended use determines which datasets and institutions are needed, how up-to-date the information must be, what output to produce, and how much uncertainty can be accepted. The standard of evidence must reflect the consequences of use. Population-level analysis may allow a different level of linkage uncertainty compared to work that identifies individuals for program or clinical action.^{6,7} Make the objective specific. For example, linkage may help find cases that a main source has missed, add information it does not collect, improve poorly recorded fields, follow care pathways, or connect services with later outcomes. Agree how the team will judge whether that objective has been met. Define the population, geographic area, time and inclusion criteria, and whether the records being linked concern people, households, facilities or events. Where relevant, agree on a case definition and the population used to calculate rates. Check out differences in definitions and coverage across sources; these should not be mistaken for linkage failures, and records from outside the agreed area should not be counted as additional cases within it.

DECIDE

What will be understood, decided, or done differently because the records have been linked, and how will we know the objective has been met?

CAPABILITY

Institutional coordination and public health intelligence: Agreeing on the purpose brings participating institutions together around a shared question. Defining the population, intended use and measures of success helps teams connect their data to the decisions and actions they need to support.

2

Linkage Operating Model and Continuity

Is the linkage required once, routinely¹⁴ or continuously? A one-time linkage provides a snapshot for a defined question. Routine linkage repeats the process at agreed intervals as source data is updated. Continuous or near-real-time linkage requires ongoing data flows and a service that can process new information as available. Each model has different requirements. Moving from a one-time exercise to a recurring process requires reusable agreements, maintained methods and code, trained staff, secure infrastructure, and continuing quality checks. The most advanced model is not automatically the most appropriate. The investment should reflect how quickly and how often information is needed. Alongside the linked information, this work can strengthen institutional coordination, data quality and standards, technical expertise, governance and data stewardship, and public health intelligence. These capabilities develop across several stages, as institutions work together, solve practical problems, and retain the agreements, methods, and lessons for future use.

DECIDE

What operating model is justified by the intended use, and can the participating institutions sustain it?

CAPABILITY

Technical expertise, governance and data stewardship: Putting the operating model into practice helps institutions build the staff skills, maintenance routines and clear responsibilities needed to repeat the linkage. Retaining documented methods, code and quality checks reduces the need to start from scratch each time.

3

Governance, Institutional Roles and Data Stewardship

Governance decisions are needed throughout the linkage process.

- **Before linkage:** Who owns and manages the source data, who can authorize its use, and what conditions apply to access?
- **During linkage:** Who can handle identifiable information, how will it be processed securely, and who will oversee the methods and validate the results?
- **After linkage:** Who owns and manages the linked data, where will it be stored, who can access it, what uses are permitted, and how long will it be retained?

Formal approval is necessary, but it does not make data available by itself. The people who approve data use, manage source systems, and extract or prepare records may work in different teams or institutions. Governance must connect these groups, turn approval into clear working responsibilities, and establish how problems will be resolved. This requires coordination both within each institution and between participating institutions. Trust and collaboration also shape access. Institutions need to understand how their data will be protected and used, and what they will receive in return. Access can develop in stages: reviewing data dictionaries and collection forms, walking through the data with the staff who manage it, then requesting a limited sample before seeking the larger extract needed for the agreed purpose. Each stage should be justified by the objective and supported by the necessary permissions.

DECIDE

Who holds authority, operational responsibility and accountability at each stage, and how will the participating institutions build trust and work together?

CAPABILITY

Institutional coordination and governance; data stewardship: Working through access, responsibilities and safeguards helps institutions establish trusted working relationships and practical ways to resolve problems together. Clear roles, documented agreements and access procedures provide a basis for future collaboration, subject to the permissions required for each use.

4

Data Readiness, Linkage Design and Quality Assurance

Readiness and feasibility answer different questions. **Readiness** asks whether the records can be linked and what needs to change. **Feasibility** asks whether the linkage is worth undertaking, given its likely cost and value for the intended use.

The readiness assessment should examine the actual identifying fields, their completeness and consistency, whether definitions align, and how local naming, language, and recording practices affect matching. Distinguish problems that data preparation can address, such as standardizing formats, removing duplicates or combining identifying fields, from weaknesses in source systems that must be improved or accepted as limitations. The choice of deterministic, probabilistic or combined methods should reflect the data, dataset size, available computing capacity, required speed, and consequences of false or missed matches. Methods, infrastructure and validation must therefore be considered together. Quality cannot be judged by the match rate alone. Teams must assess linkage errors, how they affect different population groups, and whether the remaining uncertainty is acceptable for the intended use. The assessment must also examine whether the linked information would improve the analysis or decision identified in **decision 1 (Public Health Purpose and Intended Use)**. Successful matching is not enough if the information needed for that purpose is missing or unusable. For example, identifying additional injury cases will not improve an analysis of road crash locations if those records contain location information that is incomplete, unclear, incorrect, or otherwise unusable”

DECIDE

What must be prepared, improved or accepted? Who is responsible? Will the linked information be useful enough to justify the work?

CAPABILITY

Data quality and standards; technical expertise: Assessing real data helps teams recognize recording problems, agree on definitions and judge which methods are suitable. Documenting these findings builds experience in deciding what needs improvement, what limitations can be accepted, and when linkage is unlikely to deliver enough value.

Data Linkage Process

Once participating institutions agree to proceed after the feasibility assessment. The technical process of data linkage can begin. The stages are presented here in order, but teams may need to return to earlier steps as the work proceeds. Validation may require changes to data transformations, identifying fields, candidate generation, comparison rules, or thresholds.



1 Source Data

The process begins with the data dictionary, which describes what each system is intended to capture, followed by data anatomy, metadata and actual records. A field may exist in the database structure without being populated; structure alone is not usable without data. Profiling should establish the completeness, validity, consistency, duplication, stability and discriminatory value of the available fields. It should also examine how the data is produced. Information recorded during a busy clinical encounter may have different limitations from information required for registration, payment or civil registration. Understanding this context is necessary to describe the data anatomy and determine what can reasonably be linked. Create versioned extracts, preserving the original records unchanged.¹⁵



2 Data Transformation

Working copies of the data are cleaned and standardized before comparison. This may include harmonizing date and code formats, removing placeholder values, parsing compound fields, standardizing names and addresses, resolving abbreviations, correcting inconsistencies, and preparing derived fields. Transformations must reflect language, culture and system context. Phonetic or textual rules developed for Portuguese names, for example, should not be assumed to work in French, English or other languages. Naming practices, imprecise dates of birth, transliteration and the system's original purpose may all affect how records should be prepared. Original values must be preserved. Provide version and document transformation rules, reference tables and code so the process can be reproduced.



3 Identifying Fields

Any unique identifier must be tested for coverage, validity and duplication. When no sufficiently reliable identifier is available, explore the data to develop combinations of identifying fields. Name, date of birth and mother's name may form a useful composite identifier in some settings, but there is no universal combination. The appropriate fields depend on what the systems capture, how consistently those fields are populated, and how well they distinguish one person from another. This stage should produce two technical resources: a data dictionary describing the fields and their meaning, and a set of comparison rules documenting how the fields will be assessed and how similar rules have performed in previous linkages.¹⁶



4 Candidate Generation

Large datasets cannot usually be compared across every possible pair of records. Blocking or indexing is therefore used to identify records that could plausibly belong to the same person, household, facility or event. Candidate groups may be generated using different combinations or transformed portions of identifying fields. Several blocking passes may be needed because a single restrictive rule can exclude true matches before comparison begins. In some approaches, candidate generation and record matching are treated as a single step. Whether separate or combined, the approach must keep processing manageable without excluding too many genuine matches. When candidate generation is a separate step, assess its performance separately from the matching method that follows.



5 Record Matching

Method follows the data. It must also reflect the requirements of the downstream analysis or digital product: Does the process need to be fast, precise, or highly sensitive? Deterministic methods apply defined agreement rules. Probabilistic methods of weight agreement and disagreement across several fields. Hybrid methods commonly accept strong deterministic matches first and apply probabilistic scoring to unresolved records. Machine-learning methods may also classify, rank or cluster of candidate records where sufficient training data and processing capacity are available. Comparison may use exact, phonetic, textual, temporal or geographic similarity. Thresholds should classify candidate pairs as accepted matches, rejected matches or uncertain cases requiring review. Conflicting, duplicate or many-to-one matches must then be resolved so that pairwise comparisons produce coherent person-, household-, facility- or event-level groups. The method, infrastructure, and operating requirements are bound together. A large dataset, short turnaround time or limited processing environment may require a different approach from a smaller linkage intended to recover as many matches as possible.^{17, 18}



6 Validation

A complete reference standard for evaluating linkage results is rarely available. Teams can develop an approximate standard by using manually reviewed results from a prior linkage or by asking a second reviewer to assess the same record pairs without seeing the first reviewer's decisions. Manual review can also contain errors, so describe the standard as approximate. Selecting pairs entirely at random can yield mostly true non-matches, leaving too few true matches to assess performance adequately. Assess quality through several forms of evidence: review samples of proposed matches, examine uncertain cases, compare methods, and test different thresholds. Validation should estimate precision, recall, false matches, and missed matches. **The F-measure combines precision and recall into a single summary of matching performance.** Accuracy and specificity can be misleading because most possible record pairs are non-matched, so these measures may look high even when many true matches are missed.¹⁹ A high match rate is not evidence of quality, and a low match rate is not, on its own, evidence of underreporting. Missing or poorly recorded identifying fields can prevent genuine matches, so unmatched records should not automatically be counted as newly identified cases. Assess the quality of matching fields before matching, rather than inferring it from the match rate afterward. Validation should also examine whether errors are concentrated in facilities, source systems, or population groups, and whether they affect the analysis or subsequent action. The acceptable balance between false and missed matches should be agreed with the operational teams that will use the results. They understand the consequences of incorrectly including someone, missing someone, or identifying more cases than the program can manage. Where the required standard is not met, return to the relevant earlier step: data transformation, candidate generation, comparison rules, method or threshold selection.²⁰

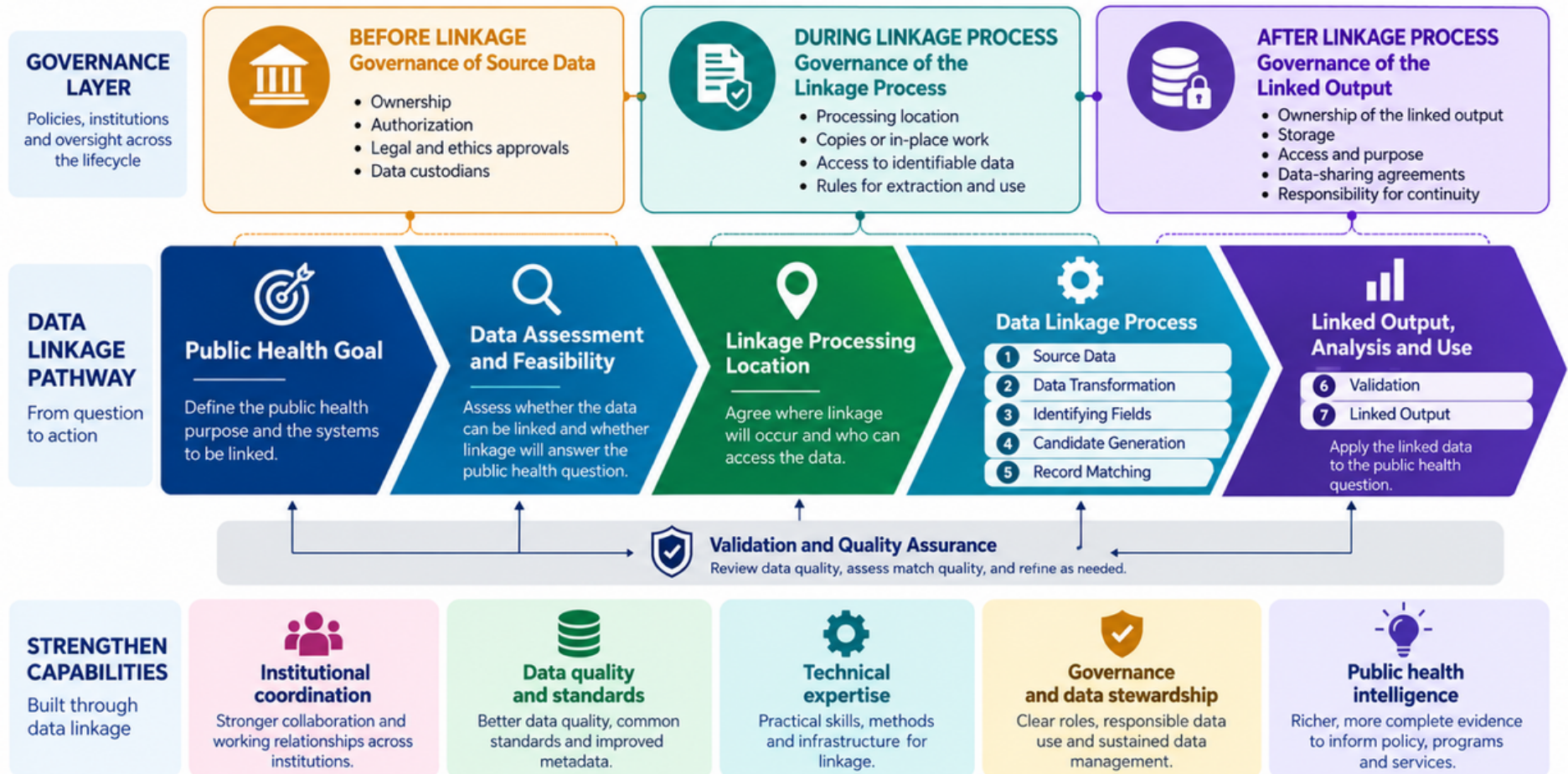


7 Linked Output

The output may be a linked dataset, analytical cohort, crosswalk, dashboard, indicator or list of records for program action. Its form should be determined by what will be done after the linkage. Each output should retain the information required to interpret and reproduce it: source provenance, linkage identifier, match status, method, score or confidence level, rule and software version, validation evidence, and known limitations. Treat the resulting output as a new information asset with defined ownership, storage, access, permitted use, retention, and stewardship. The value of the process ultimately lies in what you can do after linkage, not in the matching operation itself.²¹

Data Linkage for AI-Enabled Public Health

From trusted data to healthier populations



What Linkage Makes Possible (AI Use Case)

Data linkage and AI have two distinct relationships: Linked data can support AI analysis, and AI methods can help match records.

Data linkage as a foundation for AI analysis

Data linkage provides an important foundation for AI-enabled public health when the needed information is spread across systems. It connects information about a person’s circumstances, care and outcomes, creating data that models can learn from and teams can use to test predictions. For example, linking birth and death records connects conditions at birth with later survival. Linking emergency department records with hospital and mortality records allows teams to investigate which signs at presentation are associated with serious outcomes. The value is not simply more data, but the connection between earlier information and what happened afterward. These connections can support risk prediction, earlier identification of possible cases and analysis of care pathways. In this type of application, AI analyzes the linked information; it does not necessarily perform the matching.

What AI Does With the Linked Data

These examples show how linked records can create the longitudinal or cross-system information needed for AI analysis.

Linkage example	Records linked	Implementation maturity	What AI does with the linked data
ClarIA: Early identification of gender-based violence Recife, Brazil Municipal implementation	Primary-care electronic records linked with violence notifications in SINAN, Brazil’s Notifiable Diseases Information System, bringing together open-text health records and reported violence. ²²	In use across more than 40 health facilities, with 664 professionals trained as of August 2026. This demonstrates implementation, but impact on violence or femicide has not yet been measured.	Natural-language processing and machine learning analyze clinical narratives and patterns over time to identify possible violence risk and generate alerts in the municipal electronic health record.
Neonatal-mortality forecasting using BRNeoDeath (Brazilian neonatal mortality prediction or decision-support model) and NeMoR Data-driven neonatal mortality rate forecasting method Brazil: national data; modeling evaluated in São Paulo Research and model development.	SINASC live-birth records were linked with SIM (Brazil’s Mortality Information System.) death records using the shared live-birth declaration number. The study dataset included approximately 28.4 million records from 2006–2016. ²³	Forecasting was evaluated using São Paulo data and reported in a 2021 preprint; this is not evidence of routine clinical deployment.	Linked records were summarized by health region and combined with health-service information to develop machine-learning models forecasting neonatal mortality up to four months ahead.

Linkage example	Records linked	Implementation maturity	What AI does with the linked data
Predict severe COVID-19 outcomes South Africa: Western Cape Province Model development and external validation.	Emergency-department records were linked, using unique hospital patient numbers, with laboratory results, hospital and critical-care admissions, death records and other provincial health data. ²⁴	The study included 282,051 complete cases from seven public emergency departments. The authors recommended further validation and an impact assessment before clinical introduction.	An XGBoost model used the linked information to predict death or critical-care admission within 30 days among patients presenting with suspected COVID-19.

AI supporting record matching

AI can also contribute to the matching process itself. A supervised learning model can learn from examples of records known to belong to the same or different people, helping it assess new pairs. Some approaches learn to recognize similarities despite spelling differences, formatting changes or incomplete identifiers. Others use optimization methods to find better combinations of matching weights and penalties, rather than relying entirely on manual adjustment. These methods can work alongside deterministic and probabilistic approaches rather than replace them.

Improvement must be demonstrated, not assumed. Compare an AI-assisted method with an appropriate alternative using the same data and a credible reference standard. The relevant question is whether it reduces false or missed matches, or makes processing more manageable, without introducing unacceptable errors. The examples below show both a substantial improvement in an existing linkage method and a comparison in which overall performance was broadly similar.

AI and Statistical Learning Supporting Record Matching

These examples show computational and statistical learning methods used within the linkage process itself to help determine which records belong to the same person.

Linkage example	Records linked	Implementation maturity	How AI supported record matching
AI-assisted optimization of CIDACS-RL (Centre for Data and Knowledge Integration for Health.) Brazil: three-state evaluation using national systems. Method development evaluation	SINASC (Sistema de Informações sobre Nascidos Vivos) live-birth records linked with SIM, Brazil's Mortality Information System. Mortality records in Paraná, Santa Catarina and Rio Grande do Sul. ²⁵	Compared with the original configuration, false-positive matches fell by about 90%, while recall remained near 98%.	A genetic algorithm tested different combinations of matching weights and disagreement penalties to improve the existing probabilistic linkage method.
eTrace: Community-to-facility patient matching Malawi: Neno District-level method evaluation.	Community health-worker records linked with facility electronic medical-record systems for HIV and noncommunicable disease services. ²⁶	Both approaches achieved average precision of 0.86. Logistic regression achieved higher recall at a selected precision threshold, reaching 0.74 recall at 0.90 precision.	A supervised logistic-regression model was compared with conventional probabilistic linkage to identify records belonging to the same patient.

Whether AI matches records or analyzes the resulting data, linkage errors can affect the findings. False links can attach the wrong history or outcome to a person. Missed links can leave histories incomplete or exclude people whose identifying information is poorly recorded. The absence of a matched outcome record should therefore not automatically be treated as evidence that the outcome did not occur. These problems can affect some population groups more than others and distort the evidence used to develop or evaluate a model. Teams should retain information about the data sources, matching methods, linkage confidence, validation results and known limitations throughout model development and evaluation. They should examine how linkage errors affect model performance across population groups and whether different matching decisions change the results. For predictive models, information used to make a prediction must be available when the prediction is made; later outcomes are used to train and evaluate the model, not supplied as advance knowledge.

Endnotes

- ¹ World Bank Group. [National Health Compacts: Country-Led Health Reform Plans](#). Washington, DC: World Bank Group.
- ² Inter-American Development Bank. [Pan-American Highway for Digital Health](#). Washington, DC: IDB.
- ³ Africa Centres for Disease Control and Prevention. [African Union Health Information Exchange Guidelines and Standards](#). Preliminary edition. Addis Ababa: Africa CDC; 2023.
- ⁴ Mamuye AL, Yilma TM, Abdulwahab A, et al. Health Information Exchange Policy and Standards for Digital Health Systems in Africa: A Systematic Review. *PLOS Digital Health*. 2022;1(10):e0000118.
- ⁵ World Health Organization. [WHO and HL7 Collaborate to Support Adoption of Open Interoperability Standards](#). Geneva: WHO; 2023.
- ⁶ MEASURE Evaluation. [Health Information Systems Interoperability Maturity Toolkit](#). Chapel Hill, NC: MEASURE Evaluation; 2019.
- ⁷ World Health Organization. [Web Annex C: Record-Linkage Exercises](#). In: Consolidated Guidance on Tuberculosis Data Generation and Use. Module 1: Tuberculosis Surveillance. Geneva: WHO; 2024.
- ⁸ Gilbert R, Lafferty R, Hagger-Johnson G, et al. [GUILD: Guidance for Information about Linking Data Sets](#). *Journal of Public Health*. 2018;40(1):191–198.
- ⁹ Office for National Statistics. [Data Linkage and Matching Policy](#). Newport: ONS.
- ¹⁰ Mamuye AL, Yilma TM, Abdulwahab A, et al. Health Information Exchange Policy and Standards for Digital Health Systems in Africa: A Systematic Review. *PLOS Digital Health*. 2022;1(10):e0000118.
- ¹¹ Mutemaringa T, Heekes A, Smith M, Boule A, Tiffin N. Record Linkage for Routinely Collected Health Data in an African Health Information Exchange. *International Journal of Population Data Science*. 2023;8(1):07.
- ¹² World Health Organization. [Sharing and Reuse of Health-Related Data for Research Purposes: WHO Policy and Implementation Guidance](#). Geneva: WHO; 2022.
- ¹³ Barbosa GCG, Ali MS, Araujo B, et al. CIDACS-RL: A Novel Indexing Search and Scoring-Based Record Linkage System for Huge Datasets with High Accuracy and Scalability. *BMC Medical Informatics and Decision Making*. 2020;20:289.
- ¹⁴ World Health Organization. [Ethics and Governance of Artificial Intelligence for Health](#). Geneva: WHO; 2021.
- ¹⁵ Gilbert R, Lafferty R, Hagger-Johnson G, et al. [GUILD: Guidance for Information about Linking Data Sets](#). *Journal of Public Health*. 2018;40(1):191–198.
- ¹⁶ Mutemaringa T, Heekes A, Smith M, Boule A, Tiffin N. Record Linkage for Routinely Collected Health Data in an African Health Information Exchange. *International Journal of Population Data Science*. 2023;8(1):1771.
- ¹⁷ World Health Organization. [Web Annex C: Record-Linkage Exercises](#). In: Consolidated Guidance on Tuberculosis Data Generation and Use. Module 1: Tuberculosis Surveillance. Geneva: WHO; 2024.
- ¹⁸ Almeida D, Gorender D, Ichihara MY, et al. Examining the Quality of Record Linkage Process Using Nationwide Brazilian Administrative Databases to Build a Large Birth Cohort. *BMC Medical Informatics and Decision Making*. 2020;20:173.
- ¹⁹ Hand DJ, Christen P. A Note on Using the F-Measure for Evaluating Record Linkage Algorithms. *Statistics and Computing*. 2018;28:539–547.
- ²⁰ Hand DJ, Christen P. A Note on Using the F-Measure for Evaluating Record Linkage Algorithms. *Statistics and Computing*. 2018;28:539–547.
- ²¹ Cao C, Pillai J, Daraei S, Ghadermarzi S. [Linking Patient Records at Scale with a Hybrid Approach Combining Contrastive Learning and Deterministic Rules](#). *Biology Methods and Protocols*. 2026;11(1):bpag009.
- ²² Vital Strategies. [Pioneering Technology to Enable Early Identification of Women at Risk of Violence and Femicide Launched in Recife](#). 2026.
- ²³ Morsoleto R, Silva VA, Caliri JS, Miranda SMF, Ferreira HNM. Prediction of Infant Mortality in Brazil Using Machine Learning and Entity Matching on Brazilian Unified Health System's Data. In: Proceedings of the Symposium on Knowledge Discovery, Mining and Learning. 2025:113–120.
- ²⁴ Fuller GW, Hasan M, Hodkinson P, et al. Training and Testing of a Gradient-Boosted Machine-Learning Model to Predict Adverse Outcome in Patients Presenting to Emergency Departments with Suspected COVID-19 Infection in a Middle-Income Setting. *PLOS Digital Health*. 2023;2(9):e0000309.
- ²⁵ Grand Challenges. [Decision-Making Support Platform Based on Visual Analytics and Machine Learning to Support Public Policies Focused on Gestational Health](#). Grand Challenges Brazil; 2018.
- ²⁶ Dixon A, Thengo L, Kitsao E, et al. [Community and Facility Health Information System Integration in Malawi: A Comparison of Machine Learning and Probabilistic Record Linkage Methods](#). *ACM Journal on Computing and Sustainable Societies*. 2023;1(2):1–16.