OFFICE OF THE SECRETARY
PATIENT-CENTERED OUTCOMES
RESEARCH TRUST FUND

PROJECT REPORT

FINAL REPORT

Governance Considerations for a Public Repository of Artificial Intelligence Training Datasets

Prepared for
The Office of the Assistant Secretary for Planning and Evaluation (ASPE)
at the U.S. Department of Health and Human Services

by
NORC at the University of Chicago

July 2026

Office Of The Assistant Secretary For Planning And Evaluation

The Assistant Secretary for Planning and Evaluation (ASPE) advises the Secretary of the U.S. Department of Health and Human Services (HHS) on policy development in health, disability, human services, data, and science; and provides advice and analysis on economic policy. ASPE leads special initiatives; coordinates the Department's evaluation, research, and demonstration activities; and manages cross-Department planning activities such as strategic planning, legislative planning, and review of regulations. Integral to this role, ASPE conducts research and evaluation studies; develops policy analyses; and estimates the cost and benefits of policy alternatives under consideration by the Department or Congress.

The Office Of Health Policy

The Office of Health Policy (HP) provides a cross-cutting policy perspective that bridges Departmental programs, public and private sector activities, and the research community, in order to develop, analyze, coordinate and provide leadership on health policy issues for the Secretary. HP carries out this mission by conducting policy, economic and budget analyses, assisting in the development and review of regulations, assisting in the development and formulation of budgets and legislation, and assisting in survey design efforts, as well as conducting and coordinating research, evaluation, and information dissemination on issues relating to health policy.

Office Of The Secretary – Patient-Centered Outcomes Research Trust Fund

The Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) is charged by statute to build data capacity for patient-centered outcomes research. Coordinated by ASPE on behalf of the Department, OS-PCORTF has funded a rich portfolio of projects to meet emerging U.S. Department of Health and Human Services priorities and fill gaps in data infrastructure to enhance capabilities to collect, link, and analyze data for patient-centered outcomes research. For more information, visit https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf

This report was funded by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under Contract Number HHSP233201500048II of the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE). The work was carried out by NORC at the University of Chicago and ASPE staff. The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of HHS, ASPE, or NORC. Readers should not interpret any statement in this product as an official position of ASPE or of HHS.

Suggested Citation: Ryan S, Leaphart D, Desai P, Couture S, Dullabh P. Governance Considerations for a Public Repository of Artificial Intelligence Training Datasets. Washington, D.C.: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. July 2026.

Contributing Authors

Sofia Ryan, MSPH, Senior Research Associate II, NORC
Desirae Leaphart, PhD, Research Scientist, NORC
Priyanka Desai, PhD, Principal Research Scientist, NORC
Sara Couture, MPH, Social Science Analyst, ASPE
Prashila Dullabh, MD, Vice President and Senior Fellow, NORC

Project Officers and Project Leadership

Sara Couture (ASPE)
Jesse Anderson (ASPE)
Jennifer Wiltz (ASPE)
Marcos Trevino (ASPE)
Priyanka Desai NORC)
Prashila Dullabh (NORC)

Acknowledgements

Abel Kho, MD, Professor, Northwestern Medicine

Table of Contents

Table of Exhibits

Executive Summary

Background. Artificial intelligence (AI) technologies, particularly those that leverage natural language processing (NLP) to analyze large amounts of patient data, have potential to transform the healthcare landscape and promote a more personalized medicine model. For AI-supported tools to be safe and reliable within healthcare applications, they must be trained on high-quality datasets that reflect real-world conditions, such as those using electronic health record (EHR) data. A comprehensive repository of publicly available AI training datasets would facilitate the management and organization of de-identified EHR data and synthetic data used for training AI models. In order to establish a repository of AI training datasets, a data governance model is needed. This report synthesizes the governance requirements for a sustainable, publicly accessible repository of AI training datasets.

Methods. We conducted a literature review to identify peer-reviewed and grey literature on technical and governance requirements for AI training datasets. We also reviewed governance documents from data enclaves used in the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) projects and current publicly accessible AI training datasets to determine whether accompanying governance frameworks were available. Our findings draw from 34 sources.

Results. We identified five relevant governance domains for developing the repository. Key considerations within each domain guide the governance of a sustainable, publicly accessible repository of AI training datasets.

Governance Structure. Defines the formal system of authority, roles, and decision-making processes that ensure the ethical, legal, and effective management of the repository.

- **Accountability and Oversight.** Data governance should include clearly delineated roles that prioritize (1) preventing the misuse or harmful application of datasets and (2) supporting transparency.
- **Trust through Communication and Transparency.** Key processes for ensuring transparency include: 1) clear, public-facing documentation, 2) published governance policies, and 3) visibility into data use.
- **Public Engagement.** Continuous feedback mechanisms from the public and dataset users should be established to inform privacy, dataset use, and repository improvements.

Data Access. Defines the policies, controls, and processes that determine who can access the datasets, under what conditions, and for what purposes.

- **Access Tiers.** The repository may require two distinct tiers: Open Access (Tier 1) and Controlled or Credentialed Access (Tier 2). Tier 1 includes de-identified, low-risk data suitable for broad dissemination and Tier 2 includes de-identified but sensitive datasets.
- **Data Use Agreements (DUAs).** Comprehensive DUAs with end users should specify purpose of use, redistribution restrictions, liability and reporting, retention and time limits, and shared expectations.
- **Transparent Access.** To foster trust and usability, the repository should implement clear, user-friendly access procedures, regardless of level of access.

- **Permitted and Prohibited Uses.** Explicit documentation on prohibitive use cases for the repository's data should be provided.

Technical and Data Infrastructure. Defines the foundational systems and services that support the storage and management of datasets housed in the repository.

- **Data Preparation and Documentation.** Datasets must meet minimum preparation and documentation requirements to ensure data integrity.
- **Generation and Validation Protocols.** Synthetic datasets housed in the repository should provide clear documentation of how the data were created, their intended use, and known limitations and biases introduced during the data creation and synthesis process.
- **Interoperability and Standardized Exchange.** To be included in the repository, datasets should leverage common data models, standard data vocabularies, common data elements, and machine-readable metadata.
- **Security and Privacy Controls.** Datasets must provide information on how they follow established criteria and best practices for de-identification and privacy.

Legal and Ethical Frameworks. Defines the principles, policies, and mechanisms that ensure the lawful, responsible, and rights-respecting use of datasets in the repository.

- **Regulatory Compliance.** Data owners must provide documentation of the relevant federal, state, or jurisdictional privacy policies that govern the access and use of data in the dataset. They must also adhere to applicable privacy and ethical frameworks.
- **Patient Consent and Authorization.** Institutions must provide detailed information about how patient consent and authorization were obtained for the development of their dataset.
- **Ownership and Stewardship.** Institutions who contribute datasets to the repository will retain full ownership of the data. Different data governance practices should apply to mitigate risks applicable to vulnerable groups and to respect practices of certain communities.
- **Ethical Synthetic Data Use**. Governance of synthetic data should adhere to the same ethical principles as real-world data.

Sustainability. Describes approaches to ensure the long-term viability, adaptability, and trustworthiness of the repository

- **Governance Continuity.** The organization hosting the repository should leverage existing governance infrastructure to reduce redundancy and ensure long-term maintainability.
- **Continuous Review and Adaptation.** At least annually, governance audits and policy updates should be conducted.
- **Resource Awareness.** The organization hosting the repository should invest in long-term governance and staff capacity to sustainably maintain the repository over time.
- **Open-Source Tools.** Owners of the repository should establish a centralized policy library and community-maintained code repositories to promote reproducibility.
- **Feedback and Continuous Improvement.** Repository owners should solicit input from researchers, data users, developers, and patients to identify barriers, inform adjustments to policies and procedures, and assess the effectiveness of existing governance practices.

Conclusion. Developing a publicly accessible repository of AI training datasets requires a governance framework that integrates technical, legal, ethical, and sustainability considerations to ensure safe, interoperable, and responsible data use. As the regulatory and technological landscape continues to evolve, governance frameworks must remain adaptable to uphold ethical standards and support the long-term viability of data repositories that enable innovation in AI-driven healthcare.

1. Introduction

Artificial intelligence (AI) technologies, particularly those that leverage natural language processing (NLP) to analyze large amounts of patient data, have potential to transform the healthcare landscape and promote a more personalized medicine model.[1] For AI-supported tools to be safe and reliable within healthcare applications—such as enabling prediction of health outcomes[2]—they must be trained on high-quality datasets that reflect real-world conditions.[3] One such source of real-world data is electronic health record (EHR) data, which contains unstructured, patient-centered information that provides additional insights not available in other structured data (e.g., non-medical factors, patient-reported outcomes). EHR data represents an important input that can train AI models to reach their full potential in healthcare settings.[4]

A comprehensive repository of publicly available EHR datasets would streamline data management and organization, enabling more efficient preparation of high-quality inputs for training AI models. This repository would streamline access to critical data that would enable researchers, hospital systems, and AI tool developers to develop AI models that improve healthcare delivery. Additionally, it would support transparency and reproducibility in AI research, which would have downstream impacts on patient care.[5]

To establish a repository of AI training datasets, a data governance model is needed. Appropriate data governance, particularly of datasets derived from real-world patient data, is critical for protecting patient privacy and for building public trust in the quality and integrity of data stored within public repositories.[7] It also helps to ensure data accuracy, which is especially important when data are being used to train AI models that will help guide decision making in healthcare.[8] However, significant gaps in data governance to ensure responsible use for AI in healthcare exist,[9] and best practices for navigating governance constraints for publicly accessible repositories of AI training datasets are underexplored.[10]

> **Data governance** is defined as "the overall administration, through clearly defined procedures and plans, that assures the availability, integrity, security, and usability of structured and unstructured data available to an organization."[6]

The goal of this report is to outline technical and governance requirements for a publicly accessible repository that could house AI training datasets. Datasets housed in the repository may contain de-identified EHR data or synthetic data.

1.1 Objectives and Structure of the Report

This report synthesizes the governance requirements for a sustainable, publicly accessible repository of de-identified AI training datasets that use EHR data, including technical, infrastructure, and legal and ethical considerations. This report is guided by the following research questions:

1. What are the strengths and limitations of commonly used data access structures (e.g., open access, data use agreements (DUAs), tiered access, etc.)?
2. What are the key legal and ethical frameworks that must be considered to build a publicly accessible repository that could house NLP training datasets (and potentially other types of datasets)?
3. What are the best practices in metadata standards and documentation?
4. What are key infrastructure and interoperability considerations?

5. What standards should exist for synthetic data interoperability (e.g., Fast Healthcare Interoperability Resources [FHIR], Health Level Seven International® [HL7]), and how can we ensure these standards are consistently applied?

This report is intended for a patient-centered outcomes researcher audience, including governmental and non-governmental researchers, who are interested in using EHR data to train AI models. Groups interested in building accessible AI training datasets using EHR data may also find this report useful.

The subsequent sections of the report are organized as follows: the ***Methods*** section provides an overview of the data collection methods; the ***Findings*** section describes the identified domains guiding a governance framework for the repository, applicable considerations within each domain, and findings specific to synthetic datasets housed in the repository; and the report ends with a short conclusion section. The report also contains supplementary appendices, described in the sections that follow.

2. Methods

We conducted a literature review to identify peer-reviewed and grey literature on technical and governance requirements for AI training datasets. We also reviewed governance documents from data enclaves used in the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) projects and current publicly accessible AI training datasets to determine whether accompanying governance frameworks were available. We performed these targeted searches using PubMed, Google Scholar, and Google. The inclusion and exclusion criteria for the environmental scan are shown in Appendix A, Exhibit A-1.

The search focused on articles published over the past five years to ensure the information abstracted reflects the most current regulatory landscape, ethical frameworks, and interoperability standards. We used search strings relevant to capturing information on technical and governance requirements for AI training datasets which are shown in Appendix A, Exhibit A-2. We also reviewed references of included literature for additional resources to include in the scan.

Articles identified through the literature review were screened in two stages: first by title and abstract, then through a full-text eligibility review. The initial search yielded 38 peer-reviewed articles, supplemented by targeted searches, which added four more articles for screening. In total, we included 24 peer-reviewed articles after full-text abstraction. Additionally, we screened 10 grey literature sources and included all of these for full-text abstraction.

Information from the literature was captured using a standardized data abstraction form (Appendix A, Exhibit A-3). We then performed a thematic content analysis of the abstracted findings to develop the report.[11]

3. Findings

Based on the literature review, we identified five relevant governance domains for developing the repository. In this section, we describe these findings by detailing each of these domains and their considerations, including 1) governance structure, 2) data access, 3) technical and data infrastructure, 4) legal and ethical frameworks, and 5) sustainability. The governance domains also include synthetic data considerations.

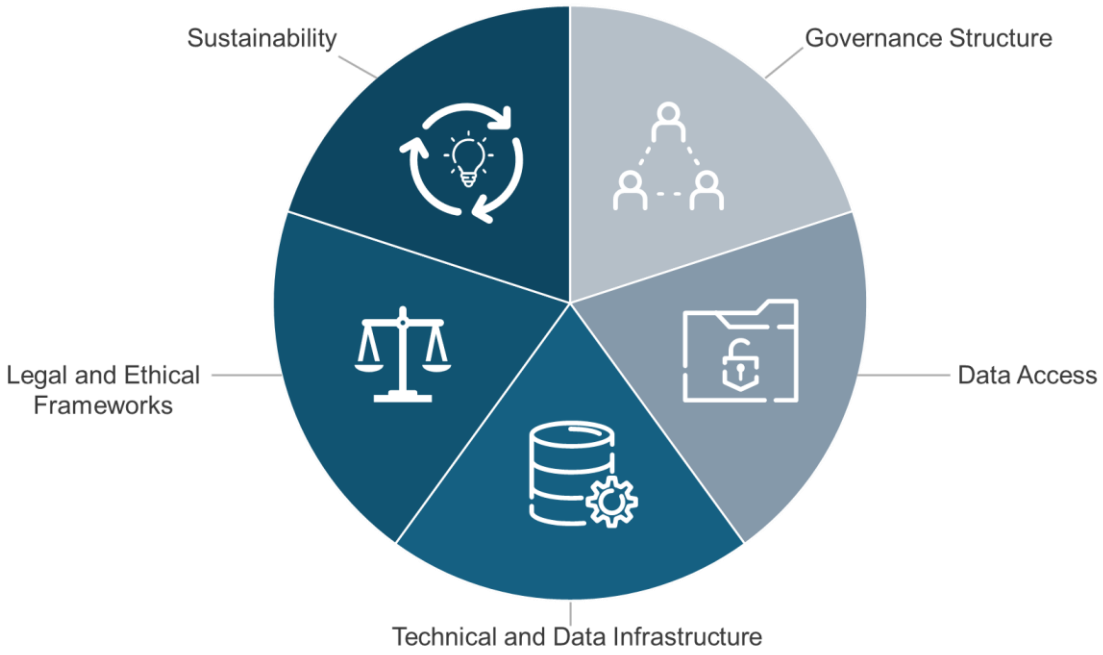
3.1 Governance Framework Domains

The purpose of the governance framework is to ensure that public repositories of de-identified AI training datasets are used ethically, legally, and consistently according to guidelines. The governance framework:

- Defines roles, responsibilities, and oversight mechanisms for managing the storage of AI training datasets;
- Recommends data stewardship and licensing practices that promote responsible sharing and reuse of stored datasets;
- Establishes technical and interoperability standards for datasets housed in the repository to support consistency and transparency;
- Outlines ethical and legal safeguards to uphold privacy and accountability; and
- Provides guidance to ensure the framework evolves alongside advances in technology, regulation, and research ethics.

The identified domains of the framework include governance structure, data access, technical and data infrastructure, legal and ethical frameworks, and sustainability. Key considerations within each domain guide the governance of a sustainable, publicly accessible repository of AI training datasets (Exhibit 1).

Exhibit 1: Governance Domains



Below, we introduce each domain and its definition, followed by sections that describe important considerations from each domain for the repository.

- **Governance Structure.** The governance structure domain defines the formal system of authority, roles, and decision-making processes that ensure the ethical, legal, and effective management of the repository.[12,13]

- **Dataset Access.** The dataset access domain defines the policies, controls, and processes that determine who can access the datasets, under what conditions, and for what purposes.[14]
- **Technical and Data Infrastructure.** The technical and data infrastructure domain defines the foundational systems and services that support the storage and management of datasets housed in the repository.[15]
- **Legal and Ethical Frameworks.** The legal and ethical domain defines the principles, policies, and mechanisms that ensure the lawful, responsible, and rights-respecting use of AI training datasets derived from EHRs.[16,17]
- **Sustainability.** The sustainability domain describes approaches to ensure the long-term viability, adaptability, and trustworthiness of the repository.

3.2 Governance Structure Considerations

Effective data governance requires a robust framework that clearly delineates roles and responsibilities to prevent misuse or harmful application of datasets, while promoting transparency and trust. This framework should encompass multiple layers of oversight and operational support.[18] We identified three considerations for the repository's governance structure.

3.2.1. Accountability and Oversight Data governance should include clearly delineated roles that prioritize (1) preventing the misuse or harmful application of datasets and (2) supporting transparency. Key roles include the governance committee, data stewards, data custodians, and support staff (Exhibit 2).

Exhibit 2: Key Governance Roles and Responsibilities

Repository Role	Responsibilities

Governance Committee	• Provides oversight of repository policies, data access rules, and governance principles.
	• May review requests for access to higher-risk data tiers (i.e., restricted data that if disclosed would result in grave or severe damage, such as biometric data).
	• Publishes governance decisions, access procedures, review criteria, and updates on approved data linkages to ensure transparency.[19]
	• Includes a range of perspectives:[20]
	  ○ **Data experts:** Researchers, clinicians, technical specialists, or others who understand data complexity and research implications.
	  ○ **Data contributors**: Organizations or individuals who contribute data (e.g., patients, advocacy groups, academic institutions) – their representation on the governance committee can ensure that participant interests are respected.
	  ○ **Ethics and compliance specialists**: Experts who can assess risks and ensure adherence to ethical standards.
	  ○ **Patients and community voices**: Patient and caregiver advocates aligned with the types of individuals whose information is being managed or who are represented in the datasets.

Repository Role	Responsibilities
Data Stewards	• Operationalizes governance policies and responsible data stewardship. • Ensures datasets meet quality, documentation, privacy, and metadata standards. • Conducts initial and ongoing compliance reviews to (1) ensure users adhere to governance requirements and (2) safeguard against misuse.[21]
Data Custodians	• Manages the technical, security, and infrastructure aspects of the repository. • Supports and monitors controlled access tiers and secure analysis environments.[22] • Ensures compliance with privacy and cybersecurity standards, including technical aspects like authentication and encryption protocols. • Plays a critical role in preventing unauthorized access.[23]
Support Staff	• Provides onboarding and help desk support for repository users.[24] • Facilitates communication between users and governance bodies. • Assists (if needed) with mandatory user training.

3.2.2 Trust through Communication and Transparency. Ensuring trust and transparency among users and data contributors of the repository must be a priority given that (1) repositories often serve as shared resources for research and innovation, and (2) stakeholders need to willingly contribute to and use the repository. Trust can be established by ensuring transparency in how datasets in the repository are collected, managed, and used, putting protections in place to safeguard privacy and security, and delineating how governance decisions are made and enforced.[13,19,25] Key processes for ensuring transparency include: (1) clear, public-facing documentation,[13] (2) published governance policies,[26] and (3) visibility into data use.[26]

Documentation should be accessible to both technical and non-technical audiences and written in plain language. According to best practices in the literature, such documentation typically includes:[21,27,28]

- The nature and scope of the datasets, including what types of data are included, their sources, and any limitations.
- Data linkage processes, including how datasets are linked and safeguards to prevent re-identification.
- Privacy protections, including methods for de-identification, anonymization, and compliance with relevant regulations (e.g., the Health Insurance Portability and Accountability Act of 1996 [HIPAA][29]).
- Intended users and use cases, including who can access the data and for what purposes.
- Data quality and provenance, such as information on data validation and completeness.
- Potential risks of misuse, ethical considerations, and mitigation strategies.
- Access procedures, including criteria for data access and tiered risk levels.
- Review processes, including how requests are evaluated and by whom.
- Update logs that record changes to datasets, policies, or infrastructure.

Additional efforts to maintain transparency and trust could include reporting dashboards that share summaries of approved projects to use the repository, data use statistics, and compliance reviews. For example, the All of Us Research Program publishes regular updates and dashboards showing participant enrollment, data use, and research outputs, fostering trust and accountability.[30]

3.2.3 Public Engagement. There should be established mechanisms for continuous feedback from the public and dataset users to inform privacy, dataset use, and repository improvements.[25] These mechanisms, such as open comment periods, public forums, Delphi processes, and listening sessions, can serve as channels for gathering various inputs that can inform continuous improvements to the repository and its policies.[9,31] Mechanisms should be put in place to facilitate patients' ability to provide feedback on the repository, as patients have historically had limited voice in the governance of research data.[9] The principle of integrating public and patient perspectives into both the ongoing governance structure and the continuous feedback process reflects models implemented in major research initiatives, which recognize that a National Trusted Research Environment must be founded on principles of social license (i.e., public approval to operate) and inclusive stewardship.[25]

3.3 Data Access Considerations

Considering the unique requirements of the repository, a multi-tiered data access structure balances the imperative of open science and research acceleration with the need to protect data privacy and maintain the trust of data contributors.[32] We provide four considerations regarding data access.

3.3.1 Access Tiers. Data access tiers function to match the sensitivity and re-identification risk of the data to the appropriate level of access control. The repository may require two distinct tiers: Open Access (Tier 1) and Controlled or Credentialed Access (Tier 2).

- **Tier 1: Open Access.** This tier includes de-identified, low-risk data suitable for broad dissemination. Examples include aggregated statistics or datasets stripped of identifiers and sensitive attributes. The primary strength of Tier 1 is that it promotes open science principles, accelerating discovery and minimizing barriers for researchers.[33]
- **Tier 2: Controlled or Credentialed Access.** This tier applies to de-identified but sensitive datasets, such as those involving rare diseases, geographic or facility-specific details, or vulnerable populations. Access to Tier 2 data should require a formal application and review, as well as execution of a DUA to ensure compliance with ethical and legal standards.[34,35]

Designing an effective data repository requires carefully weighing the benefits of open access against the need for rigorous privacy protection. Exhibit 3 summarizes strengths and limitations regarding each of the data access structures. Additionally, repository owners must determine the conditions under which data can leave the repository, as opposed to requiring users to conduct their analyses within a trusted research environment. Repository owners should clearly define qualifications for users to download data, subsequent requirements for DUAs (see section 3.3.2), as well as approaches for enforcing limitations for downloaded data specified in the DUAs.[9]

Exhibit 3: Strengths and Limitations of Data Access Structures

Type of Data Access Structure	Strengths	Limitations
Open Access	• Promotes transparency, reproducibility, and broad reuse of de-identified datasets for public benefit	• Increased risk of data misuse or re-identification • Lack of direct monitoring or tracking of who uses the data and for what specific purpose
Controlled or Credentialed Access	• Enables proportional protection by matching data sensitivity to access level • Provides legal and ethical safeguards by defining purpose limitation, data handling responsibilities, and breach reporting requirements within a DUA	• Can create inequities or barriers for smaller institutions lacking resources or credentials for higher-tier access • Can introduce complexity in implementation and management (e.g., administrative burden and variable enforcement) • Relies on the trustworthiness and self-reporting of the recipient

Regardless of access level, to help ensure responsible use of the repository, users should be required to complete baseline ethical and responsible data use training (e.g., "Responsible Conduct of Research" through the CITI program, documented privacy and security training, ethics, and data stewardship training).[34]

3.3.2 Data Use Agreements. DUAs are a critical mechanism for safeguarding privacy and clarifying responsibilities. The DUA transforms a policy requirement into a legally binding contract, ensuring that sensitive, de-identified data are handled in accordance with strict ethical and legal standards. A comprehensive DUA must delineate several specific terms to manage risk and promote accountability:[19,25,35,36,37]

- **Purpose of Use:** The agreement must explicitly define the research objectives and the user's intended applications of the data. This helps ensure that the data are used only for the project approved during the application review process.
- **Redistribution Restrictions:** A core function of the DUA is to strictly prohibit the unauthorized sharing, secondary use, or redistribution of the data. A clause describing redistribution restrictions is intended to prevent data from moving into less controlled environments, thereby maintaining the repository's chain of custody and control over sensitive data.
- **Liability and Reporting:** The DUA must clearly outline the researcher's obligations for reporting privacy breaches or misuse. Should an incident occur, the agreement specifies the timely steps the user must take, including notifying the appropriate points of contact.
- **Retention and Time Limits:** To mitigate long-term risk, the DUA should specify how long the data can be stored and used by the researcher. This timebound access ensures that data are disposed of securely or returned upon the completion of the stated research purpose, adhering to principles of data minimization.
- **Shared Expectations:** Ultimately, the DUA serves to formalize the shared expectations that both parties agree to compliance and accountability measures.

Several federal agencies have adopted core DUAs—which utilize common provisions related to confidentiality, security, data maintenance, routine uses, and reporting across all core data sources—to unify and enhance national data exchange.[38,39] The repository should aim to align with these agencies' core DUAs in order to reduce administrative burden while creating consistency and maintaining transparency. Additionally, a master DUA may be put in place for frequent data users with longstanding collaborations with the repository owner. Amendments to the master DUA need to be put in place for each additional dataset accessed or used.[40]

3.3.3 Transparent Access. To foster trust and usability, the repository should implement clear, user-friendly access procedures regardless of access level, including:[22,26]

- Published step-by-step instructions and expected turnaround times for access requests.
- Onboarding support, especially for new or resource-limited users, to understand the access permitted to them.
- Dashboards summarizing approved projects, access statistics, and compliance reviews to reinforce transparency.

3.3.4 Permitted and Prohibited Uses. A critical component of responsible data governance, particularly for AI training datasets, is providing explicit documentation on prohibitive use cases for the repository's data. This measure mitigates the risks associated with the inappropriate application of datasets that may contain subtle biases or insufficiently handle specific uses, such as to predict or diagnose certain conditions.[41] Specifically, the documentation must explicitly prohibit the use of the datasets in clinical or regulatory decision making unless the resulting AI models have undergone rigorous, independent validation demonstrating their safety, efficacy, and generalizability. The primary concern is that models trained on unvalidated datasets may perpetuate biases or produce unreliable outputs, leading to poor patient outcomes or flawed regulatory policies.[20,22,28]

3.4 Technical and Data Infrastructure Considerations

A standardized and secure technical infrastructure is needed for the repository to provide a unified, reliable foundation for effectively managing and leveraging data from multiple sources.[42] We identified three key considerations to guide technical governance of the repository.

3.4.1 Data Preparation and Documentation. Proper data preparation ensures the accuracy, consistency, and reliability of datasets and allows for more efficient analyses to be conducted.[26] AI training datasets housed in the repository must meet minimum preparation and documentation requirements to ensure data integrity:

- First, datasets must be prepared in accordance with established ethical frameworks for transparency, accountability, and purpose specification, such as those described in section 3.5.
- Second, dataset owners should implement quality assurance and validation processes, such as checks for tokenization accuracy, annotation consistency, and other relevant quality metrics.[43] For example, existing EHR databases such as the Medical Information Mart for Intensive Care (MIMIC) conduct integrity checks to ensure reliability of the database structure and consistency checks to confirm that information within the data is consistent with the known truth.[14]

- Third, to reduce the risk of incorrect use of data,[44] dataset owners should provide a comprehensive set of documentation including data dictionaries, codebooks, dataset version histories, data provenance, licensing, and/or methodology reports to promote transparency and mitigate errors.[25,26]

Synthetic Data Consideration

Generation and Validation Protocols. Currently, there is no standard approach to synthetic data governance or infrastructure requirements, which can lead to inconsistencies in synthetic data access and quality.[45] Similar to real-world datasets, synthetic datasets housed in the repository should provide clear documentation of how the data were created, including the methods, model architectures, and key parameters used.[45] Synthetic datasets should include labeling and provenance metadata that identify them as synthetic, and guidance on how they are intended to be used and interpreted.[46] They should also include documentation of known limitations and biases introduced during the data creation and synthesis process to support responsible interpretation and downstream use.[46] Finally, it is critical to assess the quality of synthetic data and its ability to mimic real data.[47] Synthetic datasets included in the repository must have been validated against real data distributions and subject to bias assessments.[22]

3.4.2 Interoperability and Standardized Exchange. Interoperability is defined as the ability of a computer or software system to exchange and use electronic health information across systems.[48] Exhibit 4 describes four key interoperability requirements that datasets should meet in order to be housed in the repository.

Exhibit 4: Interoperability Considerations for Datasets Housed in the Repository

Consideration	Description
Common Data Models	Provides a shared data language and standardized schemas to simplify sharing and management of data from different sources.[49] To be included in the repository, datasets should also leverage existing standards for health care data exchange, such as Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR).[50] Example common data models: • Observational Medical Outcomes Partnership (OMOP)[51] • Sentinel[52] • PCORnet[53] • Integrating Biology & the Bedside (i2b2)[54]
Standard Data Vocabularies	Provides a common language for describing clinical data to facilitate exchange of electronic health information.[42] Pre-specifying common vocabularies improves the setup process for databases, simplifies data exchange across datasets, and accelerates analyses using these data.[55] Examples of standard vocabularies: • Logical Observation Identifiers Names and Codes (LOINC)[56] • SNOMED Clinical Terms (SNOMED CT)[57]

Consideration	Description
	• International Classification of Diseases (ICD)/Current Procedural Terminology (CPT)[58] • RxNorm[59]
Common Data Elements	Standardized, defined questions and allowable responses that are used systematically across sites and settings.[60] The United States Core Data for Interoperability (USCDI) has a list of common data elements to help create consistency across EHRs. Example common data elements for datasets housed in the repository include: • Sex • Race/Ethnicity • Health concerns • Lab values/results • Procedures
Machine-Readable Metadata	Structured information used to describe data for identification and classification purposes (e.g., data provenance, data owner, data infrastructure) that can be automatically read and processed by computers.[61] Metadata provide a clear account of the source, transformation, and translation of data,[18] which reduces bias and improves reproducibility.[62] Example principles to guide metadata reporting: • FAIR principles (see section 3.5) • ALCOA+ principles (attributable, legible, contemporaneous, original, accurate, complete, available, consistent, and enduring)[63]

3.4.3 Security and Privacy Controls. To be housed in the repository, datasets must follow established criteria and best practices for de-identification. Dataset owners must provide clear documentation of the method of de-identification (e.g., Safe Harbor versus Expert Determination)[64] as well as formal statistical evaluation of re-identification risk.[65]

Datasets should adhere the American Statistical Association Ethical Guidelines for Statistical practice to minimize risk of data reidentification.[66] Specifically, dataset owners must provide information on how they meet these criteria:

- Pseudonymization and/or de-identification techniques via a trusted third party[22,67]
- Use of privacy-preserving record linkage (PPRL) for multi-source data integration.[68]
- Proper encryption of datasets (in transit and at rest)[13]
- Secure and time-bound data destruction protocols[21]

It is important to note that while dataset owners employ de-dentification techniques to protect patient privacy, there remains risk of re-dentification when dealing with large, multi-faceted clinical datasets.[69]

Repository and dataset owners must be careful to strike a balance between leveraging existing data for AI model development and safeguarding patient rights and privacy.[19]

3.5 Legal and Ethical Considerations

Important privacy and security issues arise when using EHR data for secondary use (e.g., AI model training) that must be addressed in order to ensure patient safety.[70] The repository should uphold all applicable privacy, consent, and ethical standards governing EHR data reuse and NLP applications. We identified three key legal and ethical considerations for the repository.

3.5.1 Regulatory Compliance. Access to and use of health information for secondary use is highly regulated at the federal level (Exhibit 5). Additionally, data privacy policies may vary widely by state, jurisdiction, type of data, or data use case, which can create difficulties when research is conducted or data are exchanged across multiple states.[67,71] To be housed in the repository, data owners must provide documentation of the relevant state or jurisdictional privacy policies that govern the access and use of data in the dataset. Data owners must adhere to these policies for the duration of time the dataset is housed in the repository and remain apprised of any applicable changes to state privacy laws.

Exhibit 5: Legal Regulations to Guide Governance of the Repository

Legal Regulation	Description
HIPAA[29]	Federal policy that establishes national standards for protecting sensitive patient health information. Sets standards for (1) who can access and use patient health information, (2) safeguarding protected health information (PHI), (3) individuals' rights to access and amend their health records, and 4) transactions of electronic health data.
Common Rule[72]	Federal policy for the protection of human subjects in research. Establishes requirements for institutions to use institutional review boards (IRBs) to oversee research, obtain informed consent from participants, and ensure ethical treatment of research participants.
21st Century Cures Act[73]	Federal policy aimed at accelerating medical research and innovation and promoting data interoperability. Establishes standards for EHRs to facilitate data sharing and analysis and delineates patient rights to access and export their EHR data. Additionally prohibits the use of Freedom of Information Act (FOIA)[74] requests to gain access to an individual's biomedical information to avoid risk of identification.
Privacy Act of 1974[75]	Federal policy that protects an individual's privacy by governing how federal agencies collect, use, and disclose personal information. This act protects against government misuse of personal data in computerized databases, requiring that an agency's collection, use, and maintenance of records must be compatible with the purposes for which it was created. It is designed to foster trust between individuals and federal data holders.
Family Educational Rights and Privacy Act (FERPA)[76]	Federal policy that protects education records and personally identifiable information contained in those records, such as health information collected for enrollment or at school-based clinics.

In addition to federal laws and policies, there are applicable privacy and ethical frameworks that provide guidance for storing, using, and sharing patient data stored in public databases, such as:

- **Fair Information Practice Principles (FIPPs)[77]:** Set of eight internationally accepted privacy principles that govern how organizations collect, use, and protect personal data. The principles focus on collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability.
- **FAIR Principles[78]:** Set of guidelines for scientific data management. The principles state that data should be findable, accessible, interoperable and reusable.
- **CARE Principles[79]:** Person-centered guidelines for the handling of data from Indigenous peoples. The CARE principles focus on collective benefit, authority to control, responsibility, and ethics.
- **Five Safes Framework[80]:** Set of principles that enable safe access to data for research. The five "safes" are (1) safe data, (2) safe projects, (3) safe people, (4) safe settings, and (5) safe outputs.

3.5.2 Patient Consent and Authorization. Consent refers to activities researchers or data users undertake to seek permission to collect, use, or share data about an individual.[67] The extent of consent necessary, including time bounds for consent and requirements for re-consent of individuals after certain periods of time, will depend on the content of the data, the purpose for which data is collected, and the location in which the data is collected.[67] Gathering consent and authorization from patients for their data to be used for secondary purposes, such as AI model training, is critical to foster trust and encourage individuals to willingly agree for their data to remain in the repository.[19] In order to be housed in the repository, institutions must provide detailed information about how patient consent and authorization were obtained for the development of their dataset, including how patients were informed that their EHR data would be used for secondary purposes.[19] If applicable, institutions should also provide information on how patient consent was gathered for broad data sharing plans.[9]

3.5.3 Ownership and Stewardship. Ownership of datasets housed in the repository must be properly attributed. It is common practice that institutions who contribute datasets to the repository will retain full ownership of the data.[68] Different data governance practices should apply to mitigate risks applicable to vulnerable groups and to respect practices of certain communities (e.g., incarcerated individuals, tribal communities, children, patients receiving care at federally qualified health centers).[35] In order to be housed in the repository, dataset owners must provide information on how data governance policies have been adjusted to protect the privacy and rights of vulnerable populations, such as specifying allowed and disallowed uses for data or additional considerations for data management.[81,82] Examples of additional considerations include:

- Institutions submitting datasets that contain American Indian/Alaska Native (AI/AN) data to the repository must include information about how tribal leadership was engaged regarding the handling of said data.[68]
- Researchers collaborating with incarcerated individuals or pediatric participants should consult stakeholders from these communities to ensure relevant legal norms and community standards are met when collecting, using, and sharing their data within AI training datasets.[35]
- Specific state laws regarding re-consent for minors once they reach a certain age may apply for pediatric populations.[67]

- Medical records relating to treatments or services for drug or alcohol abuse, HIV, and/or sickle cell anemia provided to veterans by the U.S. Department of Veterans Affairs must be kept confidential.[67]

Synthetic Data Consideration

Ethical Synthetic Data Use. Governance of synthetic data should adhere to the same ethical principles as real-world data (e.g., FAIR and CARE principles).[46] Governance of synthetic Indigenous data requires partnership with Indigenous peoples, as data available for this population historically tends to focus mainly on negative outcomes, which can bias the generation of synthetic datasets.[46]

3.6 Sustainability Considerations

It is critical for the repository to have long-term accessibility for researchers, and to be able to adapt to changing operational, financial, and technical standards. We identified five considerations for ensuring sustainability of the repository.

3.6.1 Governance Continuity. Leveraging existing governance frameworks can minimize duplication of effort. Whenever possible, the organization hosting the repository should build on established governance structures to reduce redundancy, promote efficiency, and ensure sustainable long-term management.[45] Adopting existing, vetted governance procedures for the repository can promote transparency and standardization, optimizing efficiency and sustainability.

In terms of continuity, the repository should adhere to applicable state- or jurisdiction-level policies regarding the length of time datasets can be used without requiring re-consent from patients. These policies may vary widely by jurisdiction. As noted in sections 3.3 and 3.5, DUAs with contributing institutions will include clear time bounds dictating how long datasets can be housed in the repository.

3.6.2 Continuous Review and Adaptation. At least annually, governance audits and policy updates should be conducted to ensure that the data governance policies applied to the repository can evolve and adapt to new use cases and changing legislation. The process should include:

1. Comparing governance policies with the most up-to-date federal, state, and international regulations (e.g., HIPAA[29], AI governance policies, privacy legislation) to confirm regulatory and policy alignment.[34]
2. Capturing metrics to evaluate the effectiveness of current access policies and user compliance by reviewing access logs, data use violations, data request appeals, and user feedback for areas for improvement.[43,83] Establishing these metrics at the inception of the repository will enable proactive monitoring and continuous process optimization.[20]
3. Updating repository documentation and user guidance based on the results of the annual review to ensure lasting data utility.[22]
4. Publishing a public report describing the outcomes of the governance audits and subsequent policy updates. Sharing key updates and decisions encourages transparency, promotes continuous quality improvement, and will help maintain accountability for proper governance of the repository.[20,21]

In addition to governance audits, repository owners should coordinate with dataset owners to update versions of the datasets in the repository as new releases are made available. Versioning of datasets is important to reflect expansions of existing datasets to include new patient data and variables, and changes in how existing variables are categorized or organized in the dataset.[34]

3.6.3 Resource Awareness. There are significant resource costs associated with long-term governance of the repository.[83] The organization hosting the repository should invest in long-term governance and staff capacity to sustainably maintain the repository over time.[21]

3.6.4 Open-Source Tools. Lack of reproducibility is a barrier to the generation of robust datasets to support research and AI tool development.[84] Open-source tools can be useful to improve transparency and harmonization of policies and practices for data sharing and governance.[14,21,34] Therefore, owners of the repository should establish a centralized policy library that houses governance documents, templates, and best practices across datasets to promote reproducibility. Additionally, community-maintained code repositories that contain analytical syntax, code lists, derived variables, and project documentation for the datasets housed in the repository should be open source, such as on GitHub,[85] to be discoverable and accessible to future users.[25] Existing databases of publicly available EHR data follow this model; for example, the MIMIC database developed the MIMIC Code Repository, which provides open-source code to enable reproducible studies using the freely available EHR data.[86]

3.6.5 Feedback and Continuous Improvement. Public support is essential to build legitimacy and trust in the repository.[27] Continuous dialogue with stakeholders is needed to encourage transparency and ensure improvements to the repository meet end user needs.[23,33] Repository owners should solicit input from researchers, data users, developers, and patients to identify barriers, inform adjustments to policies and procedures, and assess the effectiveness of existing governance practices. Repository owners can gather stakeholder input through surveys, open comment periods, focus groups, or appeals mechanisms for denied data access requests.[9,27] Additionally, mechanisms that introduce reciprocal value for data owners and the repository are useful for encouraging more widespread data sharing.[26] To further promote continuous community involvement in improving the repository, there should be incentives for dataset contributors, such as offering co-authorship and other collaboration opportunities to dataset owners.

4. Conclusion

Developing a publicly accessible repository of de-identified AI training datasets requires a governance framework that integrates technical, legal, ethical, and sustainability considerations to ensure safe, interoperable, and responsible data use. The five domains identified in this framework provide a foundation for mitigating risks, promoting transparency, and fostering public trust. While some governance considerations will remain stable across datasets, others may vary significantly depending on the state or jurisdiction the data were collected in. It is the responsibility of the repository owner to remain aware of these variations and ensure that all datasets included in the repository meet the minimum privacy, legal, and ethical standards required for their specific jurisdiction or data type. As the regulatory and technological landscape continues to evolve, governance frameworks must remain adaptable to uphold ethical standards and support the long-term viability of data repositories that enable innovation in AI-driven healthcare.

Appendix A. Environmental Scan Details

Exhibit A-1 describes the inclusion and exclusion criteria that guided the literature review.

Exhibit A-1. Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Publication Year	Peer-reviewed journal articles: 2020-present Grey literature: 2020-present	Peer-reviewed journal articles: Before 2020 Grey literature: Before 2020
Document Type	Peer-reviewed journal articles: Primary and secondary data analyses, scoping review, meta-analyses/systematic reviews Grey literature: Reports, working papers, evaluation studies, white papers, conference proceedings, presentations, case studies, fact sheets, issue briefs, and government documents	Peer-reviewed journal articles: Conference abstracts/proceedings Grey literature: Opinion pieces
Language	English	Non-English
Source	Academic, expert, evaluator	News outlet
Focus	Discusses technical, legal/ethical, and/or governance requirements for publicly accessible NLP-ready datasets using EHR data	No discussion of technical, legal/ethical, or governance requirements

Exhibit A-2 outlines the search terms used to identify literature via PubMed and Google.

Exhibit A-2. Literature Review Search Terms

Search Term Category	Example Terms
AI/NLP	"artificial intelligence"; Artificial Intelligence[MeSH]; "machine learning"; Machine Learning[MeSH]; "natural language processing"; Natural Language Processing[MeSH]
Data	Health Record, Electronic[MeSH]; "electronic health record"; "electronic medical record"; "publicly accessible data"; "training dataset"; "synthetic data"; "secondary use"
Governance	**General:** "data governance"; "governance framework"; "data stewardship" **Technical/Infrastructure:** "metadata standard"; "metadata documentation"; "data representation standards"; "interoperability standards"; "terminology standards";

	“interoperability”; “FHIR”; “fast healthcare interoperability resources”; “licensing structure” **Legal/Ethical: “**legal framework”; “ethical framework”; “data privacy”; “data security”; “data protection”; “data protection framework” **Access:** “data access”; “data sharing”; “DUA”; “data use agreement”; “consent”; “data storage”; “data ownership”
Reviews	"review"[Publication Type]; "systematic review"[Publication Type]; ("systematic review"[Title]; "systematic literature review"[Title]; "systematic scoping review"[Title]; "meta-analysis"[Title]; "scoping review"[Title]

Exhibit A-3 describes the key information that was abstracted from each included source from the literature review.

Exhibit A-3. Key Fields for Data Abstraction

Field Type/Category	Specific Fields
Resource Metadata	Reference details [author_pubyear_pubsource] Article name Full citation with hyperlink to the DOI if available Presents governance framework? (Y/N)
Data Access	Does article discuss data access? (Y/N) Description of data access structure used Description of strengths and limitations of data access structures Description of data use and sharing agreements Description of reporting requirements Other data access considerations
Technical/Infrastructure Considerations	Does article discuss technical/infrastructure considerations? (Y/N) Description of metadata standards and documentation, including best practices Description of licensing structure Interoperability approaches used Discuss use of synthetic data? (Y/N) If yes, describe best practices for using synthetic data

Field Type/Category	Specific Fields
	Other technical/infrastructure considerations
Legal/Ethical Considerations	Does article discuss legal/ethical considerations? (Y/N) Does article utilize existing legal or ethical frameworks? (Y/N) If yes, list framework(s) Description of data privacy and security standards Description of data ownership considerations Description of data stewardship practices Other legal/ethical considerations
Sustainability Considerations	Does article discuss sustainability considerations? (Y/N) Description of sustainability approaches used Description of best practices for sustainability Other sustainability considerations

References

[1] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98. doi:10.7861/futurehosp.6-2-94

[2] Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23(1):689. Published 2023 Sep 22. doi:10.1186/s12909-023-04698-z

[3] Polevikov S. Advancing AI in healthcare: A comprehensive review of best practices. *Clin Chim Acta*. 2023;548:117519. doi:10.1016/j.cca.2023.117519

[4] The White House. America's AI Action Plan. Washington, DC: The White House; July 2025. Accessed November 17, 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[5] Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586(7829):E14-E16. doi:10.1038/s41586-020-2766-y

[6] American Health Informatics Management Association. Healthcare Data Governance. January 2022. Accessed October 20, 2025. https://www.ahima.org/media/pmcb0fr5/healthcare-data-governance-practice-brief-final.pdf

[7] Brous P, Janssen M, Krans R. Data Governance as Success Factor for Data Science. Responsible Design, Implementation and Use of Information and Communication Technology. 2020;12066:431-442. Published 2020 Mar 6. doi:10.1007/978-3-030-44999-5_36

[8] Abraham R, Schneider J, Vom Brocke J. Data governance: A conceptual framework, structured review, and research agenda. International journal of information management. 2019 Dec 1;49:424-38.

[9] Clayton EW, Rose S, Nebecker C, et al. Biomedical data repositories require governance for artificial intelligence/machine learning applications at every step. *JAMIA Open*. 2025;8(6):ooaf134. Published 2025 Dec 1. doi:10.1093/jamiaopen/ooaf134

[10] Garg R, Gupta A. A Systematic Review of NLP Applications in Clinical Healthcare: Advancement and Challenges. In: Advances in Data-Driven Computing and Intelligent Systems. Springer; 2024. doi:10.1007/978-981-99-9521-9_3

[11] Mak S, Thomas A. Steps for Conducting a Scoping Review. J Grad Med Educ. 2022;14(5):565-567. doi:10.4300/JGME-D-22-00621.1

[12] Donsa K, Mangesius P, Lauschensky A, Baumgartner M, Tanjga N, Beyer S, Schreier G, Kreiner K. FOX BOXes as Fundamental Enablers for EHR-Standardised Data Sharing - Building the Austrian Health Data Donation Space. Stud Health Technol Inform. 2025 Apr 24;324:198-204. doi: 10.3233/SHTI250187.

[13] Jones MC, Stone T, Mason SM, Eames A, Franklin M. Navigating data governance associated with real-world data for public benefit: an overview in the UK and future considerations. BMJ Open. 2023 Oct 4;13(10):e069925. doi: 10.1136/bmjopen-2022-069925.

[14] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data.* 2023;10(1):1–9. doi:10.1038/s41597-022-01899-x

[15] Secoda. What is Data Infrastructure? September 16, 2024. Accessed October 17, 2025. https://www.secoda.co/glossary/what-is-data-infrastructure

[16] Prism Sustainability Directory. Ethical Data Governance. September 3, 2025. Accessed October 17, 2025. https://prism.sustainability-directory.com/term/ethical-data-governance/

[17] General Services Administration. Federal Data Strategy: Data Ethics Framework. Accessed October 17, 2025. https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf

[18] Heavner SF, Kumar VK, Anderson W, Al-Hakim T, Dasher P, Armaignac DL, Clermont G, Cobb JP, Manion S, Remy KE, Reuter-Rice K, Haendel M; Society of Critical Care Medicine (SCCM) Discovery Panel on Data Sharing and Harmonization. Critical Data for Critical Care: A Primer on Leveraging Electronic Health Record Data for Research From Society of Critical Care Medicine's Panel on Data Sharing and Harmonization. Crit Care Explor. 2024 Nov 15;6(11):e1179. doi: 10.1097/CCE.0000000000001179.

[19] Ho CH. Secondary Use of Health Data for Medical AI: A Cross-Regional Examination of Taiwan and the EU. Asian Bioeth Rev. 2024 Apr 6;16(3):407-422. doi: 10.1007/s41649-024-00279-4.

[20] Davis HA, Kerkman D, Hoberg AA, Countryman M, Beaver W, Bybee K, Blum JM, Knosp BM. Establishing data governance for sharing and access to real-world data: a case study. JAMIA Open. 2025 Jun 23;8(3):ooaf041. doi: 10.1093/jamiaopen/ooaf041.

[21] Heins, S., Burns, R., Hero, J., Fateh, B., Walsh, S., McCourt, A., Dubenitz, J., & Schoenbaum, M. (2024). Mortality Data Linkages for Research. Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services.

[22] Office of the National Coordinator for Health Information Technology. (2024). Training Data for Machine Learning and Patient-Centered Outcomes Research (PCOR): Final Report. U.S. Department of Health and Human Services.

[23] Munung NS, Staunton C, Mazibuko O, Wall PJ, Wonkam A. Data protection legislation in Africa and pathways for enhancing compliance in big data health research. *Health Res Policy Syst*. 2024;22(1):145. Published 2024 Oct 15. doi:10.1186/s12961-024-01230-7

[24] Smith AM. Data Governance Framework: Key Elements and Examples. October 15, 2025. Accessed November 30, 2025. https://www.dataversity.net/articles/data-governance-frameworks

[25] Boyd A, Evans KM, Turner EL, Flaig R, Oakley J, Campbell KC, et al. UK Longitudinal Linkage Collaboration (UK LLC): The National Trusted Research Environment for Longitudinal Research. Int J Popul Data Sci. 2025 Feb 17;10(1):2468. doi: 10.23889/ijpds.v10i1.2468.

[26] Tarkowski A. Data Governance in Open Source AI: Enabling Responsible and Systemic Access. January 22, 2025. Accessed November 30, 2025. https://openfuture.eu/wp-content/uploads/2025/01/250123_OSI-Data-Governance-OSAI.pdf

[27] Wieland-Jorna Y, Verheij RA, Francke AL, Tomassen M, Houtzager M, Joling KJ, Oosterveld-Vlug MG Setting up a Governance Framework for Secondary Use of Routine Health Data in Nursing Homes: Development Study Using Qualitative Interviews J Med Internet Res 2023;25:e38929

[28] Gao C, McGilchrist M, Mumtaz S, et al. A National Network of Safe Havens: Scottish Perspective. J Med Internet Res. 2022;24(3):e31684. Published 2022 Mar 9. doi:10.2196/31684

[29] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191 (1996), https://www.govinfo.gov/app/details/PLAW-104publ191.

[30] All of Us Research Hub. Data Snapshots. Accessed November 30, 2025. https://www.researchallofus.org/data-tools/data-snapshots/

[31] Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. *Health Serv Res*. 2010;45(5 Pt 2):1442-1455. doi:10.1111/j.1475-6773.2010.01140.x

[32] Marcotte JE, Rush S, Ogden-Schuette K. TIERED ACCESS TO RESEARCH DATA FOR SECONDARY ANALYSIS. J Priv Confid. 2023;13(2):10.29012/jpc.825. doi:10.29012/jpc.825

[33] Moulaei K, Akhlaghpour S, Fatehi F. Patient consent for the secondary use of health data in artificial intelligence (AI) models: A scoping review. Int J Med Inform. 2025;198:105872. doi:10.1016/j.ijmedinf.2025.105872

[34] Stanford HAI. Advancing Responsible Healthcare AI with Longitudinal EHR Datasets. February 10, 2025. Accessed November 30, 2025. https://hai.stanford.edu/news/advancing-responsible-healthcare-ai-longitudinal-ehr-datasets

[35] Bernier A, Knoppers BM. Longitudinal Health Studies: Secondary Uses Serving the Future. Biopreserv Biobank. 2021 Oct;19(5):404-413. doi: 10.1089/bio.2020.0171.

[36] Christiansen CF, Doupi P, Schutte N, Ivanković D. Piloting an infrastructure for the secondary use of health data: learnings from the HealthData@EU Pilot. Eur J Public Health. 2025 Sep 1;35(Supplement_3):iii3-iii4. doi: 10.1093/eurpub/ckaf073.

[37] Harvard University. Sample Data Usage Agreement. Accessed November 30, 2025. https://support.dataverse.harvard.edu/sample-data-usage-agreement

[38] US Centers for Disease Control and Prevention. CDC's Core Data Use Agreement Initiative. May 31, 2024. Accessed December 8, 2025. https://www.cdc.gov/data-interoperability/php/use-agreement/index.html

[39] National Institutes of Health. NIH Data Use Certification Agreement. Accessed December 18, 2025. https://grants.nih.gov/policy-and-compliance/policy-topics/sharing-policies/accessing-data/certification-agreement

[40] O'Hara A. Model data use agreements: A practical guide. In: Cole S, et al, eds. Handbook on Using Administrative Data for Research and Evidence-Based Policy. Abdul Latif Jameel Poverty Action Lab; 2020.

[41] Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370.

[42] Gliklich RE, Leavy MB. Data Standards. In: Gliklich RE, Leavy MB, Dreyer NA, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2019 Oct. Chapter 3. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551886/

[43] Alvarez-Romero C, Martínez-García A, Bernabeu-Wittel M, Parra-Calderón CL. Health data hubs: an analysis of existing data governance features for research. Health Res Policy Syst. 2023 Jul 10;21(1):70. doi: 10.1186/s12961-023-01026-1.

[44] Gierend K, Waltemath D, Ganslandt T, Siegel F. Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study. JMIR Form Res. 2023 Dec 7;7:e50027. doi: 10.2196/50027.

[45] O'Sullivan K, Caldwell J, Cole C, Harrison K, Mayor C, Chaliou A, Dymiter J, Dawson S, Brown D, Wilde K. Federating governance, access and infrastructure to support researcher use of synthetic data. International Journal of Population Data Science. 2024 Sep 10;9(5):119.

[46] Tsao SF, Sharma K, Noor H, Forster A, Chen H. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. *Stud Health Technol Inform*. 2023;302:53-57. doi:10.3233/SHTI230063

[47] Figueira A, Vaz B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*. 2022; 10(15):2733. https://doi.org/10.3390/math10152733

[48] Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap. Office of the National Coordinator for Health Information Technology. 2014. Accessed November 21, 2025. https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf

[49] Finster M, Wenzel M, Taghizadeh E. Common data models and data standards for tabular health data: a systematic review. *BMC Med Inform Decis Mak*. 2025;25(1):422. Published 2025 Nov 13. doi:10.1186/s12911-025-03267-2

[50] Heath Level 7. Welcome to FHIR. Accessed November 20, 2025. https://www.hl7.org/fhir/

[51] OHDSI CDM Working Group. OMOP Common Data Model. Accessed November 20, 2025. https://ohdsi.github.io/CommonDataModel/

[52] Sentinel Common Data Model. Accessed November 20, 2025. https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

[53] PCORnet. PCORnet Common Data Model. Accessed November 20, 2025. https://pcornet.org/data/common-data-model/

[54] i2b2 Community Wiki. i2b2 Common Data Model Documentation. Accessed November 20, 2025. https://community.i2b2.org/wiki/display/BUN/i2b2+Common+Data+Model+Documentation

[55] Bruland P, McGilchrist M, Zapletal E, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol. 2016;16(1):159. Published 2016 Nov 22. doi:10.1186/s12874-016-0259-3

[56] Bakken S, Cimino JJ, Haskell R, Kukafka R, Matsumoto C, Chan GK, Huff SM. Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. J Am Med Inform Assoc. 2000 NovDec;7(6):529-38.

[57] SNOMED International. SNOMED CT Concept Model. Accessed November 22, 2025. Retrieved from: https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model

[58] World Health Organization. International statistical classification of diseases and related health problems. 11th ed. World Health Organization; 2019. Accessed November 21, 2025. https://icd.who.int/

[59] Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc*. 2011;18(4):441-448. doi:10.1136/amiajnl-2011-000116

[60] National Library of Medicine. Common Data Elements (CDEs). Accessed November 20, 2025. https://www.nlm.nih.gov/oet/ed/cde/tutorial/03-100.html

[61] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, et al. Understanding the Nature of Metadata: Systematic Review. *J Med Internet Res*. 2022;24(1):e25440. Published 2022 Jan 11. doi:10.2196/25440

[62] Ferguson S, Tille PM. Public Datasets: A Foundation to Artificial Intelligence in Health Care. American Society for Clinical Laboratory Science. 2024 Dec 5.

[63] Bongiovanni S, Purdue R, Kornienko O, Bernard R. Quality in Non-GxP Research Environment. Handb Exp Pharmacol. 2020;257:1-17. doi:10.1007/164_2019_274

[64] Kayaalp M. Modes of De-identification. *AMIA Annu Symp Proc*. 2018;2017:1044-1050. Published 2018 Apr 16.

[65] Rodriguez A, Williams LJ, Lewis SC, et al. Evaluating re-identification risks scores in publicly available clinical trial datasets: Insights and implications. *Clin Trials*. Published online August 22, 2025. doi:10.1177/17407745251356423

[66] American Statistical Association. Ethical Guidelines for Statistical Practice. February 1, 2022. Accessed November 10, 2025. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice

[67] Office of the National Coordinator for Health Information Technology. Legal and Ethical Architecture for PCOR Data. September 28, 2017. Accessed October 2, 2025. https://www.healthit.gov/sites/default/files/page/2018-06/PCOR%20Architecture%20%28MERGE%29%20updated%20Appendix%20B.pdf

[68] NIH Center for Advancing Translational Sciences. N3C Enclave Data Overview. Accessed September 30 2025. https://ncats.nih.gov/research/research-activities/n3c/data-overview#access-requirements

[69] Knoppers BM. Framework for responsible sharing of genomic and health-related data. *Hugo J*. 2014;8(1):3. doi:10.1186/s11568-014-0003-1

[70] Basil NN, Ambe S, Ekhator C, Fonkem E. Health Records Database and Inherent Security Concerns: A Review of the Literature. *Cureus*. 2022;14(10):e30168. Published 2022 Oct 11. doi:10.7759/cureus.30168

[71] National Conference of State Legislatures. State laws related to digital privacy. Updated October 10, 2025. Accessed November 13, 2025. https://www.ncsl.org/technology-and-communication/state-laws-related-to-digital-privacy

[72] Federal Policy for the Protection of Human Subjects, 45 CFR Part 46, § 46.102, 2018 ed.` Accessed November 20, 2025. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

[73] 21st Century Cures Act. H.R. 34, 114th Congress. 2016. Accessed November 20, 2025. https://www.gpo.gov/fdsys/pkg/BILLS-114hr34enr/pdf/BILLS-114hr34enr.pdf.

[74] The Freedom of Information Act (FOIA), Pub. L. No. 89-487, 80 Stat. 250 (updated as amended by Pub. l. No. 114-185, 130 Stat. 538) (amending 5 U.S.C. § 552) (2016).

[75] Privacy Act of 1974, 5 U.S.C. § 552a (1974).

[76] 20 U.S.C. § 1232g(b)(1).

[77] Baker DB, Kaye J, Terry SF. Governance Through Privacy, Fairness, and Respect for Individuals. *EGEMS (Wash DC)*. 2016;4(2):1207. Published 2016 Mar 31. doi:10.13063/2327-9214.1207

[78] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[79] Carroll SR, Herczog E, Hudson M, Russell K, Stall S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data. 2021 Apr 16;8(1):108. doi: 10.1038/s41597-021-00892-0.

[80] Green, E. and Ritchie, F. The present and future of the Five Safes framework. *Journal of Privacy and Confidentiality*, 13(2). doi: 10.29012/jpc.831.

[81] McDonald N, Forte A. Privacy and Vulnerable Populations. In: Modern Socio-Technical Perspectives on Privacy. Springer; 2022:337-363. doi:10.1007/978-3-030-82786-1_15

[82] National Institutes of Health. Vulnerable and Other Populations Requiring Additional Protections. Accessed November 13, 2025. https://grants.nih.gov/policy-and-compliance/policy-topics/human-subjects/policies-and-regulations/vulnerable-populations

[83] Stanford Law School. AI Data Stewardship Framework. March 9, 2023. Accessed November 30, 2025. https://law.stanford.edu/2023/03/09/a-data-stewardship-framework-for-generative-ai/

[84] Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*. 2020;323(4):305-306. doi:10.1001/jama.2019.20866

[85] GitHub, Inc. GitHub. Accessed December 19, 2025. https://github.com/

[86] Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc*. 2018;25(1):32-39. doi:10.1093/jamia/ocx084