AI-readiness Criteria for Biomedical Data

Authors: Timothy Clark*,†,1, Harry Caufield[2], Jillian A. Parker[3], Sadnan Al Manir[1], Edilberto Amorim[4], James Eddy[5], Nayoon Gim[6], Brian Gow[7], Wesley Goar[8], Jan N. Hansen[9], Nomi Harris[2], Henning Hermjakob[10], Marcin Joachimiak[2], Gianna Jordan[11], In-Hee Lee [8], Shannon K. McWeeney[12], Camille Nebeker[3], Milen Nikolov[11], Sarah J. Ratcliffe[1], Justin Reese[2], Jamie Shaffer[6], Nathan Sheffield[1], Gloria Sheynkman[1], James Stevenson[8], Jake Y. Chen[13], Chris Mungall[2], Alex Wagner[8], Sek Won Kong[14], Satrajit S. Ghosh[7], Bhavesh Patel[15], Andrew Williams∮,[16], Monica C. Munoz-Torres*,†,[17]

Institutional Affiliations: [1] University of Virginia, [2] Lawrence Berkeley National Laboratory, [3] University of California San Diego, [4] University of California San Francisco, [5] Avantiqor, [6] University of Washington, [7] Massachusetts Institute of Technology, [8] Nationwide Children's Hospital, [9] Stanford University, [10] European Molecular Biology Laboratory - European Bioinformatics Institute, [11] Sage Bionetworks, [12] Oregon Health and Science University, [13] University of Alabama at Birmingham, [14] Boston Children's Hospital, [15] California Medical Innovations Institute, [16] Tufts University, [17] University of Colorado Anschutz

*** Corresponding Authors:** Timothy Clark (twclark@virginia.edu), Monica C. Munoz-Torres (monica.munoztorres@cuanschutz.edu) † equal contribution

∮ **In memoriam** - August 22, 2025

Abstract

Biomedical research is rapidly adopting artificial intelligence (AI). Yet the inherent complexity of biomedical data preparation requires implementing actionable, robust criteria for ethical and explainable AI (XAI) at the "pre-model" stage, encompassing data acquisition, detailed transformations, and ethical governance. Simple conformance to FAIR (Findable, Accessible, Interoperable, Reusable) Principles is insufficient.

Here, we define criteria and practices for reliable AI-readiness of biomedical data, developed by the **NIH Bridge to Artificial Intelligence (Bridge2AI) Standards Working Group** across seven core dimensions of dataset AI-readiness: **FAIRness, Provenance, Characterization, Ethics, Pre-model Explainability, Sustainability, and Computability.** Conformance to these criteria provides a basis for pre-model scientific rigor and ethical integrity, mitigating downstream risks of bias and error prior to AI modeling. We apply and evaluate these standards across all four Bridge2AI flagship datasets, spanning functional genomics to clinical medicine, and encode them in machine-actionable metadata bound to the datasets.

This framework sets a benchmark for preparing ethical, reusable datasets in biomedical AI and provides standardized methods for reliable pre-model data evaluation.

Introduction

Artificial intelligence and machine learning (AI/ML) are transforming 21st-century biomedicine, as research now generates massive volumes of high-velocity, multi-modal data [1,2]. Although data is the foundation for AI/ML[3,4], adequate standardized definitions of *AI-readiness* for biomedical research remain elusive [5]. A recent report by the National Academies of Sciences, Engineering, and Medicine (NASEM) identifies this as a critical 'readiness gap' in automated research workflows [6]. Recent reviews highlight ethical data acquisition and pragmatic constraints [5,7], but often treat datasets as objective *ground truths*, ignoring transparency in data preparation as a prerequisite for the epistemic validity of data as evidence, as argued by Leonelli [10]. Without it, metadata specifications risk treating complex biological and computational derivations as fictitious 'ground truths' devoid of context.

Appropriate AI-readiness characterization must include complete lifecycle transparency before modeling begins, incorporating metadata on ethical sample and cohort acquisition; processing and instrumentation; and sustainable, FAIR (Findable, Accessible, Interoperable, Reusable) data governance [11], in addition to computational readiness. This framework incorporates lessons learned in the NIH Bridge to AI (Bridge2AI) program, which produces significant multi-domain, AI-ready datasets of laboratory, clinical, and behavioral data [12].

The criteria we present here are intended to support the current large-scale transition to biomedical AI and should be considered for wide adoption. Dataset characterizations across these criteria must be available as both human- and machine-readable metadata to ensure utility.

Approach

The AI-readiness criteria presented here were developed within the Bridge2AI Standards Working Group, comprising experts in AI/ML, data standards, ethics, and data engineering from across four Bridge2AI Grand Challenges (GCs) and the Bridge Center (BC).

We developed recommendations based on the following inputs:

- Domain Expertise: Ongoing feedback from GC data generation teams preparing data releases.
- Standards: Alignment with established ontologies and NIH metadata guidelines.
- Research Literature: We reviewed extensive literature focused on FAIRness, ethical AI, pre-model explainability, and data preparedness for AI.
- Validation: Formal ongoing analysis of released Bridge2AI datasets against the proposed criteria.

Our framework builds on the 2019 NIH Advisory Committee to the Director Working Group on AI (ACD AI WG) report [13], which identified *Provenance, Description, Accessibility, Sample Size, Multimodality, Perturbations, Longitudinality,* and *Growth as core pillars* of AI readiness. We refined and expanded on these basic ideas to address requirements of preparing large-scale multimodal datasets for AI analysis.

Practices involving Bridge2AI datasets grounded our criteria. We urge data generators to also address any study-specific or clinical-domain issues, as well as task-specific pre-model data engineering, when preparing their datasets.

Fundamental Requirements of AI-Readiness

FAIRness, deep provenance, and full characterization

FAIRness is essential for biomedical data; however, the FAIR Principles do not sufficiently account for the intricate derivation histories inherent in biomedical AI applications. We require FAIRness as a Level 0 baseline, but emphasize *Deep Provenance* as a separate requirement.

High-stakes biomedical models, particularly those driving clinical decisions, require full transparency into their processing, beyond simple entity-to-entity assertions. We specify provenance using the W3C Provenance Ontology (PROV-O) extended by a biomedical-specific profile in the Evidence Graph Ontology (EVI) [14,15]. EVI profiles treat processed scientific data as assertions of computational arguments, providing machine-interpretable representations of how these assertions were derived and of the inputs, software, models, services, computations, and outputs to which they are resolvable. Transparent derivation provides the *warrant for justified true belief* (JTB) [16] that separates rigorous knowledge from findings obtained by epistemic luck from opaque processes and resolves the *Gettier problem* of provenance opacity [17–19].

Deep provenance ensures that data presented to an AI model derive from verifiable, transparent processes.

Recent high-profile article retractions, amounting to over 10,000 in 2023 [20], can be traced to systemic failures in data integrity and provenance [21]. These appear to have undergone metastasis since Begley and Ioannidis first pointed to systemic issues of “sloppy science” in 2015 [22]. End-to-end traceability in AI, is the specific remedy for these issues. It requires that data be resolvable back to their unmodified sources (e.g., raw EHR, NGS reads, imaging, proteomics, sensors, psychometric testing, or clinical trials) through all intermediate steps, including software, models, and computations. If any statistical or computational (e.g., AI/ML) models are used in the processing pipeline, whether for imputation, or for complex inference (as in Bridge2AI Cell Maps for Artificial Intelligence) [23,24], they should be described using standard representations equivalent in rigor to Model Cards [25,26].

Data must be fully characterized in depth with descriptive metadata on structure, statistical distributions, standards conformance, ethical derivation, biological assumptions, use cases, and inherent biases that may influence model weights. Full provenance graphs in biomedical AI may be constrained to some extent by operational considerations (e.g., *black-box* outsourced laboratory protocols), but we strongly recommend them as aspirational goals.

Ethics, Regulatory Compliance, and Governance

Associated Criteria: These core factors in biomedical AI require verifiable criteria to ensure that data are well-governed and ethically and legally compliant. We define criteria in this section to ensure that (1) the intended applications align with consent and institutional agreements; (2) governance attributes including Data Use Agreements are represented in machine-readable format; (3) socio-technical context of data collection, whether directly from individuals in a cohort, or from banked samples, are represented, to make potential sources of bias transparent.

Human-mediated Oversight: Our criteria provide for human-mediated oversight. We require documentation of individuals responsible for ethics review, IRB and protocol-specific authorization, and designation of a data governance chairperson. Depending upon project scale, governance may be overseen by a Principal Investigator or by a well-organized Governance Committee. The responsible governance chair and ethics contact will ensure appropriate governance of the project, supervise compliance with Belmont, Menlo, and CARE principles of ethical guidance [27–30], and other standards relevant to protection of subjects and privacy. The governance chair is the point of contact and certifies all data releases.

Ethical Licensing: We require that explicit license terms be identified in metadata, as appropriate to regulatory requirements, using a granular Data Use Agreement (DUA) or common licenses, such as Creative Commons CC-BY and its variants [31]. **We do not recommend** the Creative Commons CC0 Public Domain Dedication, as it fails to preserve provenance and authorial responsibility and does not provide enforceable qualifications for downstream use in regulatory compliance; consequently, it is highly inappropriate for biomedical data.

Sustainability

Sustainability ensures that the compute-intensive investment in AI training remains reproducible through long-term archival in trusted, sustainable repositories [32]. For controlled-access sensitive data, additional security measures will apply to ensure sound security practices, in accordance with applicable government regulations and funding agency requirements. In the US, NIH policy on Controlled Access Data Repositories is evolving, seemingly toward requiring conformance to NIST-SP-800-17 [33] certification (see NIH NOT-OD-24-005). We require that the chosen repository supports long-term persistence of data and metadata packages with project-appropriate security levels.

Pre-model Explainability

Pre-model Explainability is the transparency and reliability summit of our criteria, summarizing all preceding criteria into a human- and machine-readable Datasheet and providing cryptographic hash integrity seals for the metadata plus data packaging. Our Datasheets substantially extend the well-known Gebru et al. (2021) Datasheets for Datasets model [34], making it both actionable and accessible through an architecture for the specification and analysis of these metadata. Developing this framework was a central challenge of our project, requiring a consensus-driven approach among domain scientists, computational biologists, clinicians, and ontologists, each of whom was previously committed to distinct modeling frameworks. Our framework uses a middle exchangeability layer to transform heterogeneous, domain-specific data into standard interoperable packages.

Within these standardized packages, a W3C PROV provenance backbone is extended by Evidence Graph Ontology (EVI) domain profiles [14,15,35] that characterize PROV Entities as Datasets, Software, Models, Instruments, Reagents, and Samples. These Entity subclasses may be further annotated with terms from standard ontologies. EVI models data outputs as results of computational arguments, supported by resolvable provenance components that serve as *epistemic credentials* [16–19] for the resulting data. This formalization transforms derivation history into a verifiable warrant and answers the *Gettier problem* within the limits of current knowledge. Formalizing the relationship between evidence and assertions strengthens the claims of data to represent provable knowledge.

Finally, the architecture includes LinkML metadata [36] along with the provenance graphs in a community-standard, lightweight RO-Crate-based Exchange Layer [37]. LinkML provides a basis for later deep semantic enrichment. We also provide software (see Dataset and Software Availability section) for the automated evaluation of criterion conformance and its visual representation in human-readable datasheets. We maintain close collaboration with the LinkML and RO-Crate centers of practice at Lawrence Berkeley National Laboratory and the University of Manchester to ensure continued harmonization and integration of these approaches.

While we mention the metadata model and provide supporting tools to facilitate its implementation, the model itself is generalizable and does not require these tools. The specific tools and formal specification of the architecture will be described in detail in forthcoming publications.

Dimensions of AI-Readiness

AI-readiness is a dynamic, context-dependent developmental property of specific data sets. We do not score it pass/fail overall, but along multiple dimensions based on readiness scores for major components, yielding a characteristic readiness profile **(Fig. 1).** Achieving it in any particular use case is a collaborative, developmental, research-driven task [38–40].

AI-Readiness, as defined here, extends beyond considerations of utility, convenience, or tractability for computer scientists and informaticians. It emphasizes the reuse of data and the generation of results that are ethical, scientifically valid, explainable, interpretable, and sustainable. Our criteria for AI-readiness directly support these goals. We consider scientific validity and research integrity as ethical practices [41,42]. Ultimately, our principal goal is data that are available, reusable, deeply characterized, standardized where possible, and that provide foundational support for ethical explainability of results. Compliance with the biomedical AI-readiness criteria specified here, and their documentation in metadata, is required for fully responsible AI practices.

AI-Readiness Criteria

Fundamental Criteria

A set of foundational criteria may define biomedical AI-readiness. While *Fundamental FAIRness* is considered a Level 0 requirement for NIH research data, AI-readiness imposes properties that extend far beyond basic FAIR compliance, requiring more rigorous specification and novel metadata extensions.

AI-readiness implies that data must be: **FAIR, Provenanced, Characterized, Pre-model Explainable, Ethical, Sustainable, and Computable**. As illustrated in the AI-Readiness Wheel (**Fig. 1**), these seven dimensions are not isolated silos; rather, they form an interdependent ecosystem in which the strength of one, such as Deep Characterization, directly enables the success of another, such as Pre-model Explainability.

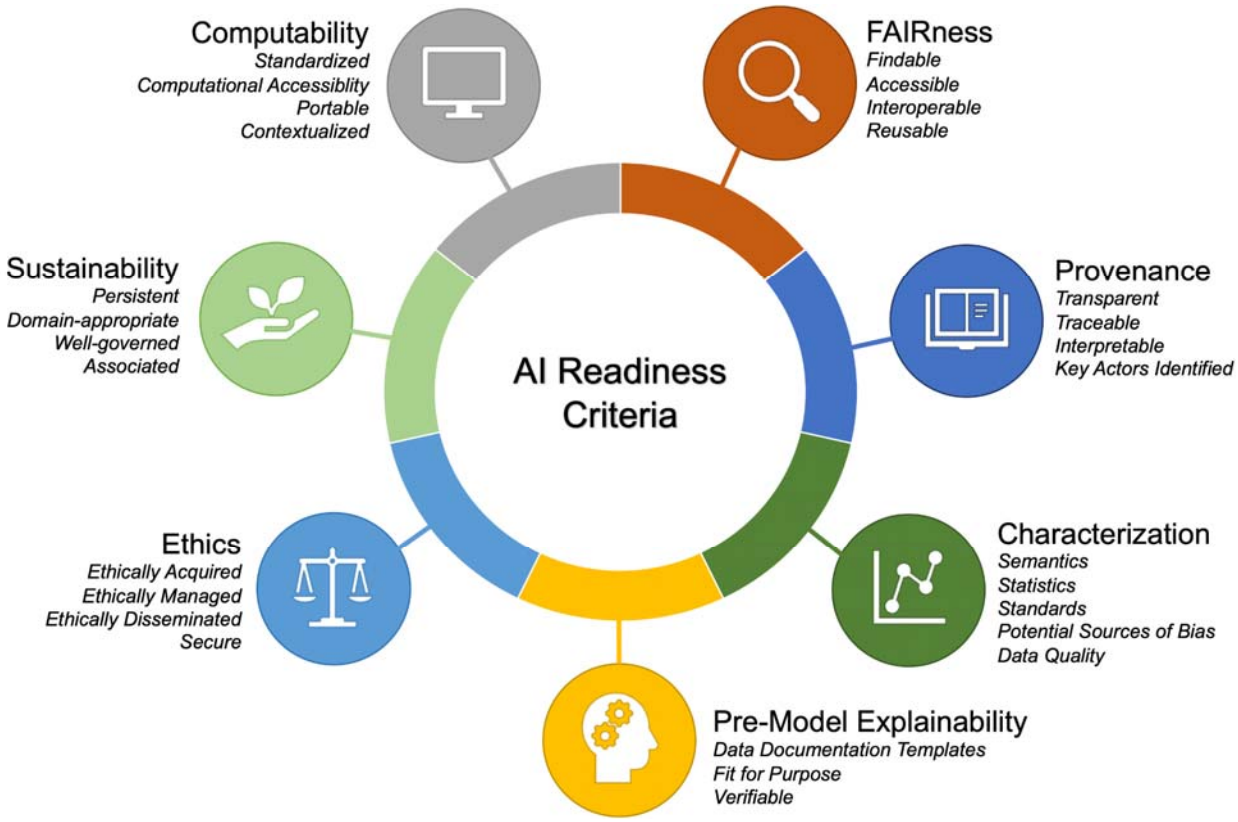


Figure 1. Seven principles of AI-readiness were developed for Bridge2AI datasets, along with their relevant subcriteria (*italics*) as detailed in Table 1.

Detailed AI-Readiness Criteria

To provide more precise implementation guidance, we developed the following detailed criteria and supporting practices. We have reviewed these practices against a related but less comprehensive effort in another domain, Earth and Space Sciences, as a consistency check [40]. **Table 1** describes AI-readiness criteria and relevant practices that support their achievement, which constitutes a novel output of this research and a major advancement in AI-readiness practices.

Table 1 – AI-readiness Criteria and Practices

ID	Criterion	Practice[†]	Suggested resources
0	FAIRness		
0.a	Findable	Deposit datasets in a searchable FAIR-compliant data repository.	FAIRsharing.org [43] NIH GREI-repositories [44]
0.b	Accessible	Descriptive metadata should always be available and accessible, even if the dataset is restricted, unavailable, or de-accessioned. Ensure metadata conforms to standards such as DCAT (Data Catalog Vocabulary) or schema.org.	DCAT [45] schema.org [46]
0.c	Interoperable	Wherever possible, provide data and metadata using formally defined specifications for digital objects.	RDF [47] JSON-LD [48]
0.d	Reusable	Attach a clear and accessible data usage license that allows the responsible use of AI/ML applications or link to a Data Use Agreement (DUA).	Creative Commons licenses (other than CC0) [31]
1	Provenance		
1.a	Transparent	Identify data sources traceable to a reasonable ground truth, e.g., clinical data from EHR at a given hospital, clinical trials, or laboratory data.	OMOP [49] RRID [50]
1.b	Traceable	Identify key data transformation steps, with links to software, at an appropriate level of detail, ideally using a machine-readable representation.	W3C PROV [37], EVI [14,15]
1.c	Interpretable	Make software for key data transformation and analysis steps available in a sustainable repository.	Zenodo [51–53] Software Heritage [54] GitHub [55]
1.d	Key Actors Identified	Identify people and organizations responsible for obtaining and processing the data, the samples, and subject groups involved in producing the data. Reference these parties along with other dataset metadata.	ORCID [56] ROR [57]
2	Characterization		
2.a	Semantics	Use comprehensive descriptive metadata for datasets, including a detailed abstract, dataset keywords, and subject-specific vocabularies (e.g., MeSH) to enable detailed search and discovery.	Datacite schema [58] Schema.org [46]
2.b	Statistics	Provide domain-appropriate statistical characterizations of key features of the dataset to inform analysis planning. Ensure missing values are encoded consistently.	
2.c	Standards	Provide a machine-readable data dictionary or schema for each dataset, linked to the dataset metadata, and referencing any relevant standards.	JSON Schema [59] Frictionless Data [60] Pydantic [61]
2.d	Potential Sources of Bias	Describe known sources of bias in the data and assumptions made in collecting, processing, or interpreting the data. Include any known explanations regarding missing values, including reasons for missingness, and the degree to which data represents a	Croissant RAI [9]

		state vs. a control (e.g., disease state vs. healthy state).	
2.e	Data Quality	Have quality control procedures been applied? If so, provide a link to a description.	
3	Pre-model Explainability		
3.a	Data Documentation Template	Machine-readable metadata and a linked human-readable document should support a domain-appropriate extension of the original Gebru Datasheets concept for pre-model explainability, as detailed in the present article. Healthsheets information may be a helpful supplement.	Datasheets [34] Healthsheets
3.b	Fit for Purpose	Identify appropriate and inappropriate use cases for the data set. Link to any previously published analyses using this data.	Croissant RAI [9]
3.c	Verifiable	Provide a mechanism for ensuring the integrity of each raw or processed dataset**,** such as a cryptographic hash.	SHA256 [62] Merkel Trees [63]
4	Ethics		
4.a	Ethically Acquired	Describe ethical data acquisition consistent with accepted principles (e.g., Belmont Report, Menlo Report, CARE Principles) sufficient to evaluate for intended use, along with a management plan.	Belmont Report [27,28] Menlo Report [29] CARE Principles [30]
4.b	Ethically Managed	Data management must align with ethical principles throughout the health AI lifecycle. Indicate privacy-protection processing, if any, sufficient to evaluate ethical status for intended use.	HL7 privacy protection metadata [64]
4.c	Ethically Disseminated	Specify a licensing agreement and/or data use agreement (DUA) on as open terms as ethical and sustainability considerations permit. Specify contact information for a data access committee.	
4.d	Secure	Specify security requirements for storing and accessing this data, e.g., “public”, “controlled access only”, etc.	HL7 privacy protection metadata [64]
5	Sustainability		
5.a	Persistent	Ensure unprocessed data is preserved in an archive that complies with privacy laws and retention guidelines, enabling future reprocessing and publishing of revised data.	TRUST Principles [32]
5.b	Domain-appropriate	Ensure single-domain raw or processed data are deposited in a FAIR-domain-appropriate specialist repository if available.	FAIRsharing.org [43] NIH GREI-participating repositories [45]
5.c	Well-governed	Implement a data governance model before data capture, including a repository that facilitates data stewardship in the future and governance that accounts for maintenance, terms, policy changes, and fairness.	FAIRsharing.org [44] NIH GREI-participating repositories [44]
5.d	Associated	Document project-level connections between data components and elements in a machine-readable manner.	RO-Crate [37]
6	Computability		

6.a	Standardized	Datasets adhere to established, documented standards, including metadata standards, and their adherence can be validated deterministically.	Croissant [8], Croissant Responsible AI [9]
6.b	Computationally Accessible	Provide a mechanism to access data through established exchange protocols or a well-documented API.	FAIRsharing.org [44] NIH GREI-participating repositories [44]
6.c	Portable	Maximize portability across computational resources where possible. If data use requires specific resources, provide machine-readable documentation defining these resources.	
6.d	Contextualized	Include any considerations regarding splits of the data, including any information withheld at any point of data collection and processing. Provide examples of data components to aid understanding of their general structure and content.	Croissant [8]
[†] Practices may impact multiple criteria; the most relevant relationship is shown for brevity			

The AI-readiness of biomedical data can be evaluated against machine-readable metadata using these criteria, yielding a multidimensional score matrix. We provide a subjective evaluation report in a spreadsheet (Supplementary Information) and tools to perform automated evaluation within the RO-Crate packages [65,66].

Fig. 2 shows a sample evaluation using a pass/fail rating for each criterion, as a percentage of the category total, and plotted in a radar plot. Current status is the purple line; the target (100%) status is blue. Results from automated evaluations of AI-readiness for each of the Bridge2AI Grand Challenges were obtained using automated tools and data evaluation forms (Supplementary Information). Machine-readable AI-readiness metadata provides both a basis for evaluation and for human-readable Datasheets, building on the concepts of Gebru et al. 2021.

Challenges and Limitations

Preparation of AI-ready biomedical datasets requires additional effort beyond simply capturing measurements or observations for statistical analysis. This effort increases when datasets are meant, as in Bridge2AI, to meet multiple use cases and be sustainable over time, rather than addressing one-off, highly focused research questions. For use-case-specific studies, computability criteria, such as detailed feature engineering, should be made available by the data providers.

AI-readiness data preparation requires a deep understanding of the data, the predictive task(s) for which the data will be used, the scientific domain of the data, statistical methods, AI technologies, biomedical data standards, and appropriate ethical practices. It also requires at least some attention, depending upon the project scope and intended longevity, to sustainability within the biomedical data ecosystem. Our approach respects the need for pre-model explainability (XAI) by clearly defining provenance using four unique, stand-alone sub-criteria (1.a-1.d; see **Table 1**).

In clinical studies, the ethical treatment of human subjects and the protection of their data are major concerns. It requires close attention to proper de-identification techniques (anonymization), privacy-preserving practices, and responsible data stewardship, further emphasizing the need for the Provenance and Ethics criteria to ensure that data-use limitations, compliance, intellectual property, and other restrictions are clearly stated and adhered to in downstream data use.

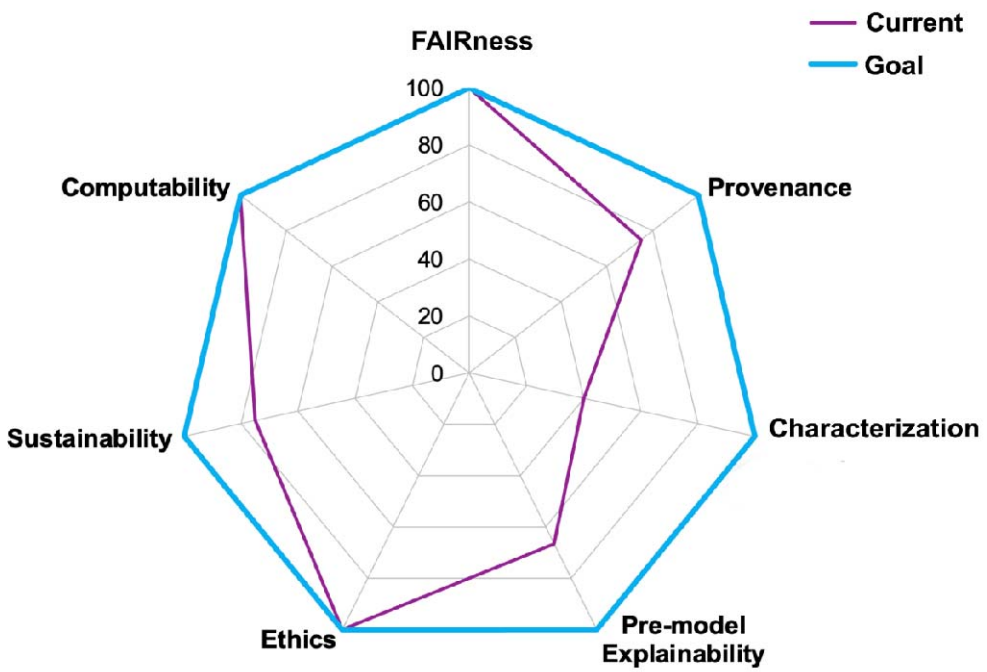


Figure 2. Example AI-readiness scored in a radar plot. Presence/absence of each criterion yields a percentage present score for each axis. Project evaluations represented using such plots are provided in **Figure S.1** in Supplementary Information.

Certain inherent limitations are implied by the time-, place-, technology-, and culture-boundedness of our efforts. AI/ML applications and capabilities are a rapidly progressing, revolutionary scientific and societal development. Our understanding of data ethics and society's ability to democratically control and adapt AI technologies for the widest possible social benefit must surely evolve. Cultural, ethnic, and gender definitions used in these datasets today may seem archaic in ten or twenty years, and what we do not conceive of as biases today may seem biased tomorrow. Thus, best practices must continue to evolve alongside the field of biomedical AI/ML.

Such limitations and challenges require teamwork and demand a Team Science approach, which is particularly important as project ambition and scope increase [67].

Translating AI-readiness Criteria into Exchangeable Metadata

The AI-readiness criteria described in this article help support data reuse and evaluation across heterogeneous biomedical and computational contexts. To enable practical reuse, these criteria must be operationalized as structured, exchangeable metadata describing the datasets along the major criteria axes. Within Bridge2AI, we translate the criteria into machine-actionable metadata wrapped in standard, lightweight RO-Crate version 1.2 packages [68]. This approach employs widely adopted serializations and vocabularies for interoperability across domains. Bridge2AI RO-Crate packages provide metadata using: JSON-LD graphs and the schema.org vocabulary for semantic content [46,48]; JSON Schema dataset structure descriptions [59]; and Frictionless Data schema evaluation [60]. Provenance information is represented using the W3C PROV data model as a structural backbone, supplemented by EVI profiles [14,15,35]. EVI is a domain-specific extension of PROV-O [69] that provides biomedical specializations of the core PROV classes (Entity, Activity, and Agent). Dual use of PROV-O and EVI preserves

compatibility with existing provenance-aware tools and infrastructures while enabling more semantically precise biomedical domain descriptions and epistemic justification. Bridge2AI FAIRSCAPE tools [70,71] generate RO-Crate packages conforming to this model and can optionally translate them to LinkML semantic modeling language for deeper semantic integration across projects when required.

RO-Crate packages may be generated through many types of workflows and validated at varying levels of rigor. While tooling such as Bridge2AI's FAIRSCAPE provides one concrete implementation of this approach, the translation of AI-readiness criteria into exchangeable metadata is not dependent on any specific software system and is intended to be broadly applicable across biomedical data ecosystems.

Conclusions and Future Directions

AI-ready biomedical data preparation and evaluation require practices to establish data and software **FAIRness**; **Deep Provenance**; **Statistical, Semantic, and Ethical-Regulatory Governance Characterization**; with support for **Pre-model Explainability**; **Sustainability**; and **Computability**. These should be reflected in the human- and machine-readable metadata associated with an AI-ready dataset. We have outlined criteria reflecting our recommended practices and multiple practices to evaluate adherence to these criteria. The criteria proposed here are currently used in NIH's Bridge2AI program and have evolved across data releases through intensive discussions among multi-domain experts. We believe these datasets and their associated deep metadata and technologies will enable novel, significant, and transformational discoveries. Developing such data resources has enabled participating investigators to deeply examine requirements for AI-readiness. It prompted passionate methodological discussions amongst experts and ultimately motivated the criteria and evaluation methods described in this article.

If widely adopted, the standards proposed here will significantly benefit the biomedical AI/ML community at large and users of AI-enhanced biomedical research, including clinicians and patients. Ensuring AI-readiness enables downstream data users to apply emerging AI capabilities ethically and reliably. Such capabilities will improve our understanding of many diseases and aid the development of novel, effective treatments and technologies.

Our major contributions outlined in this article include:

- Defined practices and criteria for AI-readiness of biomedical data;
- A formal evaluation approach against these criteria;
- Detailed evaluation of the Bridge2AI Grand Challenge datasets.

Some open-source tools supporting AI-readiness developed in Bridge2AI include:

- RO-Crate structures for multi-modal data with Deep Provenance;
- Extended Evidence Graph Ontology (EVI) PROV profiles;
- JSON Schema, Pydantic, and LinkML models for AI-readiness Datasheets; and
- The FAIRSCAPE AI-readiness framework.

These tools, though not required to implement the AI-readiness criteria, will continue to be developed and fully extended across all Bridge2AI program components as the project enters its fourth year.

We invite feedback on this work and collaboration with the broader biomedical AI/ML community, including users of Bridge2AI datasets. The perspectives presented here reflect the

experience of teams producing datasets intended to serve as flagship examples of best practices. As these data are used to develop effective AI methods, their application will help identify strengths of the proposed approach and guide future improvement. We encourage the use of these metrics to improve AI-readiness across data products and to assess the suitability of reused datasets.

Data and Software Availability Statement

Dataset Evaluation

An AI-Readiness self-evaluation worksheet is available here

- Parker, J.,A., et al. 2024 - AI-Readiness Self-Evaluation Worksheet https://doi.org/10.5281/zenodo.13961091 [72]

Datasets

The four Bridge2AI Grand Challenge datasets cited below were evaluated against the AI-readiness criteria established here (Supplementary Figure S.1 and Table S.2). A companion paper [Clark et al., doi:10.1101/2024.12.23.629818v4] describes the FAIRSCAPE framework that produced and packaged datasets and AI-readiness metadata for these datasets and presents them as production examples of that framework.

- Clark, T; Parker, J; Al Manir, S; et al. 2025, "Cell Maps for Artificial Intelligence - October 2025 Data Release (Beta)", University of Virginia Dataverse, V2. https://doi.org/10.18130/V3/K7TGEM
- Rosenthal, Eric S.; Kamaleswaran, Rishikesan; Strekalova, Yulia Levites; et al. 2026, "Data Manifest for Collaborative Hospital Repository Uniting Standards (CHoRUS) April 2026", University of Virginia Dataverse, V1. https://doi.org/10.18130/V3/XNBOPG
- Bensoussan, Y., Sigaras, A., Rameau, A., et al. (2025). Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information (version 3.0.0). [Data set] PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/k81f-qr68
- AI-READI Consortium. (2025). Flagship Dataset of Type 2 Diabetes from the AI-READI Project (3.0.0) [Data set]. FAIRhub. https://doi.org/10.60775/fairhub.3

Software and Tooling

Bridge2AI-funded assistive tools for AI-readiness metadata packages are available here:

- **FAIRSCAPE AI-readiness Framework:**
  - Levinson, M.A., et al. 2026 - FAIRSCAPE CLI (v1.1.21). Zenodo. https://zenodo.org/records/18234493 [70]
  - Levinson, M.A., et al. 2026 - FAIRSCAPE Server (v1.0.3). Zenodo. https://zenodo.org/records/18714417 [71]
- **RO-Crate Validation Classes:**
  - Niestroy, J., et al. 2026 - FAIRSCAPE Pydantic Models (v1.0.1). Zenodo. https://doi.org/10.5281/zenodo.18234523 [65]
  - Leo, S., et al. 2025 - ro-crate-py (0.14.2). Zenodo. https://doi.org/10.5281/zenodo.17342107 [66]
- **LinkML-based Semantic Modeling and Translation:**

- Moxon, S; et al. 2025. LinkML: An Open Data Modeling Framework. GigaScience, giaf152, https://doi.org/10.1093/gigascience/giaf152 [36]

Acknowledgements

This work was funded by National Institutes of Health Bridge2AI program awards OT2OD032742, OT2OD032644, OT2OD032720, OT2OD032701, U54HG012510, 5U54HG012517, and 5U54HG012513; and by the Frederick Thomas Fund of the University of Virginia. JNH has been supported with an EMBO Postdoctoral Fellowship (ALTF 556-2022).

We thank Carole Goble, Trey Ideker, Maryann Martone, and Randall Moorman for very helpful discussions, which improved this article.

Use of Generative AI in the Writing Process

The authors used Gemini 3 Flash (released December 2025) during copyediting to refine structural flow, improve argument clarity, and ensure compliance with Scientific Data editorial policies. Following the use of this tool, the authors independently edited the content as needed. The authors take full responsibility for the content of the publication.

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

TC and MCMT conceived the study, oversaw the collaboration, standards alignment, and discussions, led development of the metadata architecture, and prepared the initial manuscript. TC, SAM, JN, and MAL developed the FAIRSCAPE RO-Crate packaging tools. TC and SAM developed the EVI profiles. JHC, MJ, JR, and MCMT developed the LinkML-based schema and AI-assist tools. TC and JAP managed CM4AI metadata preparation and release testing. SSG, BP, and AW managed metadata packaging for Bridge2AI clinical programs. TC prepared the final manuscript with review and edits by MCMT, JAP, and SJR. All authors contributed significantly to key discussions that informed the preparation, content, editing, and final review of the manuscript. All authors reviewed and approved the manuscript.

Competing interests:

The authors declare no competing interests.

REFERENCES

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56 (2019).
2. Clissa, L., Lassnig, M. & Rinaldi, L. How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry. *Front. Big Data* **6**, 1271639 (2023).
3. Strickland, E. Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big. *IEEE Spectr.* **59**, 22–50 (2022).
4. Walsh, I. *et al.* DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* **18**, 1122–1127 (2021).
5. Hiniduma, K., Byna, S. & Bez, J. L. Data Readiness for AI: A 360-Degree Survey. *ACM Comput. Surv.* **57**, 1–39 (2025).

6. Committee on Realizing Opportunities for Advanced and Automated Workflows in Scientific Research *et al. Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. 26532 (National Academies Press, Washington, D.C., 2022). doi:10.17226/26532.

7. Ng, M. Y. *et al.* Perceptions of Data Set Experts on Important Characteristics of Health Data Sets Ready for Machine Learning: A Qualitative Study. *JAMA Netw Open* **6**, e2345892 (2023).

8. Akhtar, M. *et al.* Croissant: A Metadata Format for ML-Ready Datasets. in *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning* 1–6 (2024). doi:10.1145/3650203.3663326.

9. Jain, N. *et al.* A Standardized Machine-readable Dataset Documentation Format for Responsible AI. Preprint at https://doi.org/10.48550/arXiv.2407.16883 (2024).

10. Leonelli, S. Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science. *Harvard Data Science Review* https://doi.org/10.1162/99608f92.17405bb6 (2019) doi:10.1162/99608f92.17405bb6.

11. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).

12. National Institutes of Health. Bridge to Artificial Intelligence (Bridge2AI).

13. ACD AI WG. Report of the Advisory Committee to the Director Working Group on AI. (2019).

14. Al Manir, S., Niestroy, J., Levinson, M. A. & Clark, T. Evidence Graphs: Supporting Transparent and FAIR Computation, with Defeasible Reasoning on Data, Methods, and Results. in *Provenance and Annotation of Data and Processes* (eds Glavic, B., Braganholo, V. & Koop, D.) vol. 12839 39–50 (Springer International Publishing, Cham, 2021).

15. Al Manir, S., Niestroy, J., Levinson, M. & Clark, T. EVI: The Evidence Graph Ontology, OWL 2 Vocabulary. Zenodo (2021).

16. Plantinga, A. *Warrant: The Current Debate*. (Oxford University PressNew York, 1993). doi:10.1093/0195078624.001.0001.

17. Gettier, E. L. Is Justified True Belief Knowledge? *Analysis* **23**, 121–123 (1963).

18. Goldman, A. I. A Causal Theory of Knowing. *The Journal of Philosophy* **64**, 357 (1967).

19. Harman, G. *Thought:* (Princeton University Press, 2015). doi:10.1515/9781400868995.

20. Van Noorden, R. More than 10,000 research papers were retracted in 2023 — a new record. *Nature* **624**, 479–481 (2023).

21. Ioannidis, J. P. A., Pezzullo, A. M., Cristiano, A., Boccia, S. & Baas, J. Linking citation and retraction data reveals the demographics of scientific retractions among highly cited authors. *PLoS Biol* **23**, e3002999 (2025).

22. Begley, C. G. & Ioannidis, J. P. A. Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circulation Research* **116**, 116–126 (2015).

23. Clark T *et al.* Cell Maps for Artificial Intelligence: AI-Ready Maps of Human Cell Architecture from Disease-Relevant Cell Lines - Data Release (0.6 Beta). University of Virginia Dataverse https://doi.org/10.18130/V3/B35XWX (2025).

24. Lenkiewicz, J. *et al.* Cell Mapping Toolkit: an end-to-end pipeline for mapping subcellular organization. *Bioinformatics* **41**, btaf205 (2025).

25. Mitchell, M. *et al.* Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (2019) doi:10.1145/3287560.3287596.

26. Crisan, A., Drouhard, M., Vig, J. & Rajani, N. Interactive Model Cards: A Human-Centered Approach to Model Documentation. in *2022 ACM Conference on Fairness, Accountability, and Transparency* 427–439 (2022). doi:10.1145/3531146.3533108.

27. The National Commission for the Protection of Human Subjects of & Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. (1979).

28. Adashi, E. Y., Walters, L. B. & Menikoff, J. A. The Belmont Report at 40: Reckoning With Time. *Am J Public Health* **108**, 1345–1348 (2018).

29. Bailey, M., Dittrich, D., Kenneally, E. & Maughan, D. The Menlo Report. *IEEE Secur. Privacy Mag.* **10**, 71–75 (2012).

30. Carroll, S. R. *et al.* The CARE Principles for Indigenous Data Governance. *Data Science Journal* **19**, 43 (2020).

31. Carroll, M. W. Creative Commons and the New Intermediaries. *Mich. St. L. Rev* **45**, (2006).

32. Lin, D. *et al.* The TRUST Principles for digital repositories. *Sci Data* **7**, 144 (2020).

33. Ross, R., Pillitteri, V., Dempsey, K., Riddle, M. & Guissanie, G. *Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations*. NIST SP 800-171r2 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-171r2.pdf (2020) doi:10.6028/NIST.SP.800-171r2.

34. Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).

35. Gil, Y. *et al.* PROV Model Primer: W3C Working Group Note 30 April 2013. (2013).

36. Moxon, S. A. T. *et al.* LinkML: An Open Data Modeling Framework. *GigaScience* giaf152 (2025) doi:10.1093/gigascience/giaf152.

37. RO-Crate Community. Research Object Crate (RO-Crate). (2023).

38. Kidwai-Khan, F. *et al.* A Roadmap to Artificial Intelligence (AI): Methods for Designing and Building AI ready Data for Women's Health Studies. *medRxiv* 2023.05.25.23290399 (2023) doi:10.1101/2023.05.25.23290399.

39. Thomas, D. M. *et al.* Transforming Big Data into AI-ready data for nutrition and obesity research. *Obesity* **32**, 857–870 (2024).

40. Poduval, B. *et al.* AI-ready data in space science and solar physics: problems, mitigation and action plan. *Front. Astron. Space Sci.* **10**, 1203598 (2023).

41. Lindemann, T. & Häberlein, L. Contours of a research ethics and integrity perspective on open science. *Front. Res. Metr. Anal.* **8**, 1052353 (2023).

42. Committee on Responsible Science, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. *Fostering Integrity in Research*. 21896 (National Academies Press, Washington, D.C., 2017). doi:10.17226/21896.

43. The FAIRsharing Team. FAIRsharing.org. (2024).

44. National Institutes of Health, O. of the D. Generalist Repository Ecosystem Initiative. (2023).

45. Riccardo Albertoni *et al.* Data Catalog Vocabulary (DCAT) - Version 3. (2024).

46. Guha, R. V., Brickley, D. & Macbeth, S. Schema.org: evolution of structured data on the web. *Communications of the ACM* **59**, 44–51 (2016).

47. Brickley, D. & Guha, R. V. *RDF Vocabulary Description Language 1.0: RDF Schema*. http://www.w3.org/TR/rdf-schema/ (2004).

48. Sporny, M. *et al.* JSON-LD 1.1: A JSON-based Serialization for Linked Data. *W3C Recommendation 16 July 2020* https://www.w3.org/TR/json-ld/ (2020).

49. Observational Medical Outcomes Partnership (OMOP). Standardized Data: The OMOP Common Data Model. (2024).

50. Bandrowski, A. E. & Martone, M. E. RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. *Neuron* **90**, 434–436 (2016).

51. European Organization For Nuclear Research & OpenAIRE. Zenodo. (2013) doi:10.25495/7GXK-RD71.

52. Troupin, C., Muñoz, C., Fernández, J. G. & Rújula, M. A. Scientific results traceability: software citation using GitHub & Zenodo. in *Proceedings of the International Conference on Marine Data and Information Systems* 4 (Barcelona (Spain), 2018).

53. Potter, M. & Smith, T. Making code citable with Zenodo and GitHub. (2015).

54. Abramatic, J.-F., Di Cosmo, R. & Zacchiroli, S. Building the universal archive of source code. *Commun. ACM* **61**, 29–31 (2018).

55. Blischak, J. D., Davenport, E. R. & Wilson, G. A Quick Introduction to Version Control with Git and GitHub. *PLoS Comput Biol* **12**, e1004668 (2016).

56. Akers, K. G., Sarkozy, A., Wu, W. & Slyman, A. ORCID Author Identifiers: A Primer for Librarians. *Medical Reference Services Quarterly* **35**, 135–144 (2016).

57. Lammey, R. Solutions for identification problems: a look at the Research Organization Registry. *Sci Ed* **7**, 65–69 (2020).

58. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.6. https://doi.org/10.14454/MZV1-5B55 (2024) doi:10.14454/MZV1-5B55.

59. Wright, A., Andrews, A., Hutton, B., & Dennis, G. JSON Schema: A Media Type for Describing JSON Documents. (2022).

60. Fowler, D., Barratt, J. & Walsh, P. Frictionless Data: Making Research Data Quality Visible. *IJDC* **12**, 274–285 (2018).

61. Pydantic. Pydantic: Documentation for version: v2.12.5.

62. National Institute of Standards and Technology (US). *Secure Hash Standard*. NIST FIPS 180-4 https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf (2015) doi:10.6028/NIST.FIPS.180-4.

63. Kelsey, J., Change, S. & Perlner, R. *SHA-3 Derived Functions: cSHAKE, KMAC, TupleHash and ParallelHash*. NIST SP 800-185 https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-185.pdf (2016) doi:10.6028/NIST.SP.800-185.

64. HL7. HL7 Terminology 6.1.0 - Confidentiality. (2024).

65. Justin Niestroy, Max Levinson, Sadnan Al Manir & Clark, T. Fairscape Pydantic Models. Zenodo https://doi.org/10.5281/ZENODO.18234523 (2026).

66. Simone Leo *et al.* ro-crate-py. Zenodo https://doi.org/10.5281/ZENODO.17342107 (2025).

67. National Academies. Enhancing the Effectiveness of Team Science. (2015).

68. Sefton, P. *et al.* RO-Crate Metadata Specification 1.2.0. Preprint at https://doi.org/10.5281/ZENODO.13751027 (2025).

69. Lebo, T. *et al.* PROV-O: The PROV Ontology W3C Recommendation 30 April 2013. http://www.w3.org/TR/prov-o/ (2013).

70. Levinson, M. A., Niestroy, Justin & Al Manir, S. FAIRSCAPE CLI. Zenodo https://doi.org/10.5281/ZENODO.18234493 (2026).

71. Levinson, M. A., Al Manir, S., Niestroy, J. & Clark, T. FAIRSCAPE Server. Zenodo https://doi.org/10.5281/ZENODO.18244417 (2026).

72. Parker, J.A., et al. AI-Readiness Self-Evaluation Worksheet. Preprint at https://doi.org/10.5281/ZENODO.13961091 (2024).

73. Clark, T; Parker, J; Al Manir, S; et al. 2025, "Cell Maps for Artificial Intelligence - October 2025 Data Release (Beta)", University of Virginia Dataverse, V2. https://doi.org/10.18130/V3/K7TGEM

74. Rosenthal, Eric S.; Kamaleswaran, Rishikesan; Strekalova, Yulia Levites; et al. 2026, "Data Manifest for Collaborative Hospital Repository Uniting Standards (CHoRUS) April 2026", University of Virginia Dataverse, V1. https://doi.org/10.18130/V3/XNBOPG

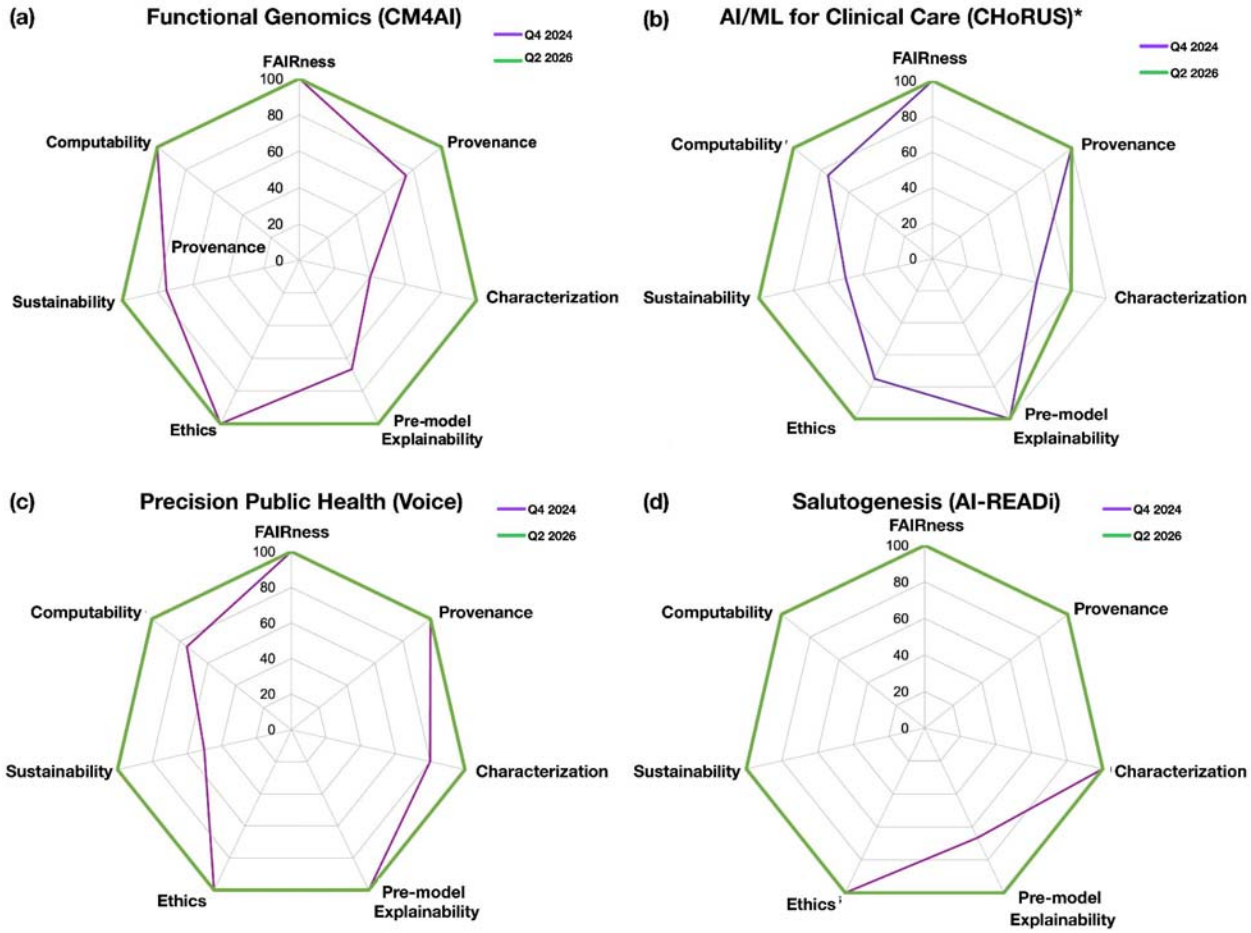
75. Bensoussan, Y., Sigaras, A., Rameau, A., et al. (2025). Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information (version 3.0.0). [Data set] PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/k81f-qr68

76. AI-READI Consortium. (2025). Flagship Dataset of Type 2 Diabetes from the AI-READI Project (3.0.0) [Data set]. FAIRhub. https://doi.org/10.60775/fairhub.3

Supplementary Information

Supplementary Figure S-1: AI-readiness evaluation

We evaluated AI-readiness for each Bridge2AI Grand Challenge (GC) dataset across the major dimensions defined in Table 1 by treating each criterion as an axis in a radar plot (**Supplementary Fig. S-1**). If a sub-criterion is addressed satisfactorily, it was given a score of "1"; a score of "0" was assigned if the GC did not address it. We then computed the overall criterion score, on a scale of 0-100% satisfaction, by totaling the number of sub-criterion scored as "1" and dividing by the total number of sub-criteria. For example, scoring three out of four total sub-criteria as "1" would produce an overall score of 75% satisfaction for that criterion. shows radar plot evaluations for each of the four NIH Bridge2AI GCs in its 2024 state, and its 2026 Q1 state. AI-readiness scores of 100% on all criteria are planned for the end of the project.



Supplementary Figure S.1: AI-readiness radar plots for Bridge2AI Grand Challenges (GCs): (a) Functional Genomics (CM4AI); (b) AI/ML for Clinical Care (CHoRUS); (c) Precision Public Health (Voice); (d) Salutogenesis (AI-READI). Purple lines represent the AI-readiness evaluated in Q4 of Year 2. Green lines indicate the extent to which each GC's data and metadata practices currently meet the seven AI-readiness criteria for data collected as of Q2 in Year 4 of the program. All GCs are on track to achieve full compliance by Q4 Year 4.

Supplementary Table S.2: Supporting Data for Figure S.1

Functional Genomics (CM4AI) - 2024 Q4			
Criterion		Criterion met? (Y=1; N=0)	Total Score for Criterion (%)
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	0	75
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	40
	Statistics (2.b)	0	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	0	
	Data Quality (2.e)	0	
Pre-model explainability (3)	Data documentation templates (3.a)	0	67
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	0	50
	Domain-appropriate (5.b)	0	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	100
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

Functional Genomics (CM4AI) - 2026 Q1			
Criterion		Criterion met? (Y=1; N=0)	Total Score for Criterion (%)
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	

	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	100
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	100
	Statistics (2.b)	1	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	1	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	100
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	100
	Domain-appropriate (5.b)	1	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	100
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

Salutogenesis (AI-READI) - YEAR 2 Q4			
Criterion		**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	100
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	100
	Statistics (2.b)	1	

	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	1	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	67
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	0	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	100
	Domain-appropriate (5.b)	1	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	100
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

Salutogenesis (AI-READI) - YEAR 3 Q4			
Criterion		**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	100
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	100
	Statistics (2.b)	1	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	1	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	67
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	0	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	

	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	100
	Domain-appropriate (5.b)	1	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	100
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

AI/ML for Clinical Care (CHoRUS) - 2024 Q4			
Criterion		**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	100
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	0	60
	Statistics (2.b)	1	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	0	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	100
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	75
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	0	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	50
	Domain-appropriate (5.b)	0	
	Well-governed (5.c)	1	
	Associated (5.d)	0	
Computability (6)	Standardized (6.a)	1	75

	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	0	

AI/ML for Clinical Care (CHoRUS) - 2026 Q1			
Criterion		**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**
FAIRness (0)	Findable (0.a)	1	0
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	0
	Traceable (1.b)	1	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	0
	Statistics (2.b)	0	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	1	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	0
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	0
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	0
	Domain-appropriate (5.b)	1	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	0
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

Precision Public Health (Voice) - 2024 Q4		
Criterion	**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**

FAIRness (0)	Findable (0.a)	0	50
	Accessible (0.b)	0	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	75
	Traceable (1.b)	0	
	Interpretable (1.c)	1	
	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	40
	Statistics (2.b)	0	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	0	
	Data Quality (2.e)	0	
Pre-model explainability (3)	Data documentation templates (3.a)	1	67
	Fit for purpose (3.c)	0	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	50
	Domain-appropriate (5.b)	0	
	Well-governed (5.c)	1	
	Associated (5.d)	0	
Computability (6)	Standardized (6.a)	1	75
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	0	

Precision Public Health (Voice) - 2026 Q1			
Criterion		**Criterion met? (Y=1; N=0)**	**Total Score for Criterion (%)**
FAIRness (0)	Findable (0.a)	1	100
	Accessible (0.b)	1	
	Interoperable (0.c)	1	
	Reusable (0.d)	1	
Provenance (1)	Transparent (1.a)	1	100
	Traceable (1.b)	1	
	Interpretable (1.c)	1	

	Key actors identified (1.d)	1	
Characterization (2)	Semantics (2.a)	1	100
	Statistics (2.b)	1	
	Standards (2.c)	1	
	Potential Sources of Bias (2.d)	1	
	Data Quality (2.e)	1	
Pre-model explainability (3)	Data documentation templates (3.a)	1	100
	Fit for purpose (3.c)	1	
	Verifiable (3.d)	1	
Ethics (4)	Ethically acquired (4.a)	1	100
	Ethically managed (4.b)	1	
	Ethically disseminated (4.c)	1	
	Secure (4.d)	1	
Sustainability (5)	Persistent (5.a)	1	100
	Domain-appropriate (5.b)	1	
	Well-governed (5.c)	1	
	Associated (5.d)	1	
Computability (6)	Standardized (6.a)	1	100
	Computational Accessibility (6.b)	1	
	Portable (6.c)	1	
	Contextualized (6.d)	1	

Supplementary Note S.3: Glossary

Artificial Intelligence (AI): Artificial Intelligence is the ability of a computer to perform tasks commonly associated with intelligent beings. AI is an umbrella term encompassing many rapidly evolving interdisciplinary subfields, including knowledge graphs, expert systems, and machine learning, and has many applications, such as speech recognition and natural language processing, image processing, robotics, and intelligent agents.

AI-ready Data: AI-ready Data is data that has been prepared such that it can be considered ethically acquired and optimally used for training, classification, prediction, text/image generation, or simulation, and has explainable results based upon it, using appropriate AI and/or machine learning methods in biomedical and clinical settings. The degree and nature of such preparation and requirements placed upon it depend upon the specific type of data, its ownership and derivation, and the set of use cases to which it will be applied. The realm of biomedical research constitutes one such set of use cases.

AI System: An AI system is a machine-based system that, for explicit or implicit objectives, infers from its inputs how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Biomedical Data: Biomedical data is laboratory, clinical, omics, environmental, or behavioral data obtained to study and/or intervene in the biology, psychology, ecosystems, health, clinical care, and other characteristics of biological systems.

Data: (a) Information in a specific representation, as a sequence of meaningful symbols; (b) (Computing) The quantities, characters, or symbols on which operations are performed by a computer, being stored on various media and transmitted in the form of electrical signals.

Data Element: A basic unit of information that has a unique meaning; an attribute, field, feature, or property in a dataset.

Dataset: A collection of data and metadata, or a set of datasets, constituting a body of structured information describing some topic(s) of interest.

Explainable Artificial Intelligence (XAI): In AI systems and applications, XAI is the availability of sufficient information, tools, and methods to enable an AI/ML classification, simulation, or prediction to be explained based on inputs to the model and model methods. XAI aims to explain the information that grounds an AI model's decisions or predictions.

FAIR, FAIRness: A set of defined characteristics of data, tools, and infrastructures that aid discovery and reuse by third parties. Compliance with the FAIR (Findable, Accessible, Interoperable, Reusable) Principles is a prerequisite for proper data management and data stewardship and is strongly recommended by NIH.

Machine Learning: Machine learning is a set of techniques that generate models automatically from training data, enabling the identification of patterns and regularities rather than through explicit instructions from a human.

Metadata: Data that describes and gives information about other data.

Provenance: Provenance is a record of the history, authorship, ownership, and transformations of a physical entity or information object, such as data or software. Provenance provides an essential basis for evaluating the validity of information and the nature of physical entities.