AI-readiness for Biomedical Data: Bridge2AI Recommendations

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], Melissa Haendel[9], Jan N. Hansen[10], Nomi Harris[2], Henning Hermjakob[11], Marcin Joachimiak[2], Gianna Jordan[12], In-Hee Lee[13], Shannon McWeeney[14], Camille Nebeker[3], Milen Nikolov[12], Jamie Shaffer[6], Nathan Sheffield[1], Gloria Sheynkman[1], James Stevenson[8], Jake Y. Chen[15], Chris Mungall[2], Alex Wagner[8], Sek Won Kong[12], Satrajit S. Ghosh[7], Bhavesh Patel[16], Andrew Williams[17], Monica C. Munoz-Torres*[,18]

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) University of North Carolina at Chapel Hill, (10) Stanford University, (11) European Molecular Biology Laboratory - European Bioinformatics Institute, (12) Sage Bionetworks, (13) Boston Children's Hospital, (14) Oregon Health and Science University, (15) University of Alabama at Birmingham,(16) California Medical Innovations Institute, (17) Tufts University, (18) University of Colorado Anschutz Medical Campus.

*** Corresponding Author:** Monica C Munoz-Torres (monica.munoz-torres@cuanschutz.edu)

On behalf of the Bridge to Artificial Intelligence (Bridge2AI) Consortium members.

Abstract

Biomedical research and clinical practice are in the midst of a transition toward significantly increased use of artificial intelligence (AI) and machine learning (ML) methods. These advances promise to enable qualitatively deeper insight into complex challenges formerly beyond the reach of analytic methods and human intuition while placing increased demands on ethical and explainable artificial intelligence (XAI), given the opaque nature of many deep learning methods.

The U.S. National Institutes of Health (NIH) has initiated a significant research and development program, Bridge2AI, aimed at producing new "flagship" datasets designed to support AI/ML analysis of complex biomedical challenges, elucidate best practices, develop tools and standards in AI/ML data science, and disseminate these datasets, tools, and methods broadly to the biomedical community.

An essential set of concepts to be developed and disseminated in this program along with the data and tools produced are criteria for AI-readiness of data, including critical considerations for XAI and ethical, legal, and social implications (ELSI) of AI technologies. NIH Bridge to Artificial Intelligence (Bridge2AI) Standards Working Group members prepared the present article to present methods for assessing the AI-readiness of biomedical data and the data standards perspectives and criteria we have developed throughout this program. While the field is rapidly evolving, these criteria are foundational for scientific rigor and the ethical design and application of biomedical AI methods.

1. Introduction

Artificial intelligence (AI) may constitute one of the most impactful advances of the early 21st century. Its innovations arrive at a crucial moment for biomedicine [1]. Scientific research produces more data than ever: a single project may generate petabytes or even exabytes of data annually in a dizzying array of types, formats, and scales [2]. The contents of electronic health records (EHRs), to cite one example, though increasingly computable [3,4] and widely adopted [5], continue to pose challenges due to their scale, complexity, heterogeneity, and missingness [6–8]. The increase of electronic health data is often complemented by diverse data types (e.g.,

‘omics, survey data, voice, video, geolocation, actigraphy) collected from varied wearables, smartphones, tablets, and instruments that capture human behavior and physiology at multiple temporal scales. The types and scale of laboratory datasets available on cells and subcellular components are constantly increasing. How to best apply newly emerging AI technologies to biomedical data is a question with evolving answers.

Applicable definitions of what constitutes AI-readiness for biomedical data have been elusive, as Hiniduma *et al.* [9] pointed out in a recent review. Ng *et al.*[10] found that ethical acquisition and societal impact with transparency and ethical reflection against pragmatic constraints were critical criteria not described in prior frameworks, which needed to fully integrate healthcare-specific AI-readiness criteria. Hiniduma *et al.* describe data as the critical fuel for AI models. The value of AI system outputs is *strongly associated with data readiness*, a crucial point in AI systems performance, fairness, and reliability. However, neither of these reviews contemplate preparation for reuse as a primary goal of data generation and tend to assume that data, as presented, are ground truths.

Availability for reuse is an essential component of the FAIR (Findable, Accessible, Interoperable, Reusable) principles for scientific data.[11] Data transformation is most often a significant feature in biomedical AI pre-model pipelines that occurs before any study-specific feature selection and engineering. Therefore, we assume a potential range of use cases and emphasize the provision of comprehensive descriptive metadata to enable the assessment of dataset fitness for particular use cases based on complete transparency.

We provide here a set of criteria for biomedical data’s AI-readiness and an evaluation method to assess dataset compliance. Our work incorporates prior scientific literature results and significant lessons learned in the Bridge to Artificial Intelligence (Bridge2AI), a flagship $130 million program of the U.S. National Institutes of Health (NIH). Bridge2AI’s goal is to produce AI-ready datasets comprised of curated cross-domain laboratory, clinical, and behavioral data to enable the advancement of AI and its use in tackling complex biomedical challenges[12]. Bridge2AI datasets must be ethically acquired, FAIR, fully reliable, robustly defined, and computationally accessible to promote use in the broader biomedical informatics AI/ML community. Along with such datasets, Bridge2AI is developing associated standards, software, tools, resources, and training materials to accelerate biomedical AI research. This program has provided a unique opportunity for AI-readiness criteria to be explored, derived, and analyzed against our own large-scale biomedical datasets for multiple types of analysis in a broad set of use cases. We define and explore these criteria here.

We also included a self-evaluation of Bridge2AI Grand Challenges projects in their current state of progress. These should not be interpreted as measures of relative excellence in achieving AI readiness. The whole focus of our effort is to provide useful points to aid in the standardization and metadata strategies for data generation efforts intended to support AI algorithm development.

AI-readiness of biomedical data is defined herein as a set of characteristics of a dataset and its associated metadata that permit reliable, ethical analysis by AI methods within defined use cases and operational limits, with sufficient metadata to support reliable, appropriate post-model explainability analysis. Further:

- *Reliability of datasets* in AI requires clearly defined, robust, transparent data acquisition and preparation methods, with adequate definition of dataset characteristics sufficient to support statistical robustness and analytic repeatability.
- *Quality of datasets* in AI demands precise, comprehensive, and unbiased data collection, with transparent metadata documentation of data origins and consistent protocols for labeling and preprocessing while minimizing bias and errors in model performance.
- *Ethical constraints* on biomedical AI datasets concern the scientific integrity of pre-model data acquisition and processing, adherence to best practices in human and animal subject protection, and proper licensing and distribution with barriers against misuse.
- *Explainability* means the ability to show how data and results were obtained with verifiable transparency, sufficient richness, and enough clarity to inspire confidence in the intended use.

Bridge2AI consists of a coordinating Bridge Center (BC) and four Grand Challenges (GCs) in Functional Genomics (Cell Maps for AI, CM4AI), AI/ML for Clinical Care (Collaborative Hospital Repository Uniting Standards, CHoRUS), Precision Public Health (Voice as a Biomarker of Health, Bridge2AI-Voice), and Salutogenesis (Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights, AI-READI), with participation from over 40 U.S., Canadian, and European institutions and more than 400 researchers. The GCs are diverse, each producing unique multimodal data. The Bridge2AI Standards Working Group (SWG) supports the Grand Challenges in establishing common approaches for defining and achieving AI-readiness and promoting data and software interoperability. The value of AI-readiness criteria and evaluation methods for datasets produced beyond the Bridge2AI program should be readily apparent. We believe the criteria specified here will broadly support the ongoing large-scale transition to the development and use of AI methods in biomedical research. This transition has already created big impacts in education [13], finance [14], and drug discovery [15], which have experienced widespread adoption of predictive computational systems. AlphaFold [16] and other revolutionary AI products are indicative of more to come. To the greatest extent feasible, dataset characterizations across these criteria should be available as machine-readable metadata.

In the remainder of this article, we outline practices and criteria contributing to the AI-readiness of any dataset for future AI/ML applications. The NIH has provided general guidance regarding data sharing and dissemination requirements and strategies to develop and publish criteria for ML-friendly datasets [17]. The 2019 Report of the Advisory Committee to the Director Working Group on AI (ACD AI WG) suggested that AI-readiness criteria should concern several categories, including *Provenance, Description, Accessibility, Sample Size, Multimodality, Perturbations, Longitudinality,* and *Growth* [17]. Our final list of AI-readiness criteria (see *Fundamental Requirements of AI-Readiness* below) reflects the integration, reorganization, and clarification of these foundational ideas in light of initial experiences of the Bridge2AI program and published recommendations of Grand Challenge researchers[18]. We also clarify preliminary definitions (Box 1) and discuss the relationship between FAIRness, AI explainability, and ethical, legal, and social Implications (ELSI) to AI readiness.

A set of preliminary definitions follows (Box 1), with references.

Box 1 – Definitions
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[19–22]**.**
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 having 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 depends 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 is one set of such use cases[23].
AI System: An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, 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.[24]

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[25].
Data: (a) Information in a specific representation, as a sequence of meaningful symbols[26]; (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[27].
Data Element: A basic unit of information that has a unique meaning; an attribute, field, feature, or property in a dataset[28].
Dataset: A collection of data and metadata, or set of datasets, constituting a body of structured information describing some topic(s) of interest[29].
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 grounding the AI model's decisions or predictions[30].
FAIR, FAIRness: A set of defined characteristics of data, tools, and infrastructures that aid discovery and reuse by third parties[8]**.** 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[11,31].
Machine Learning: Machine learning is a set of techniques that generate models in an automated manner through exposure to training data, which can help identify patterns and regularities, rather than through explicit instructions from a human[24].
Metadata: Data that describes and gives information about other data[26,27].
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[32].

2. Approach

The recommendations in this article resulted from an extensive collaborative process conducted within the Bridge2AI Standards Working Group (WG), which is composed of domain experts in AI/ML, relevant data and software standards, ethics, and data generation and preparation tasks, across the four Bridge2AI Grand Challenges (GC) and the coordinating Bridge Center. Recommendations were developed from (a) analysis, approaches, and conclusions from each GC including extensive problem-focused domain expertise; (b) special expertise from Bridge Center and GC participants in development of ontologies and data standards; (c) review and synthesis of relevant recommendations from the technical literature, with special emphasis on FAIRness, Ethical AI, and AI Explainability; (d) lessons learned from prior AI-Readiness framework development by GC participants[33–38], and (e) systematic analysis of the GC's datasets against the resultant AI-readiness criteria.

While we do not include a detailed discussion on items like "sample size" and "multimodality," which are study-dependent, we urge users to ensure these issues are considered in preparing their own project-specific datasets. Likewise, study-specific techniques of pre-modeling data engineering (e.g. feature engineering, feature extraction) are omitted, but must be documented by the data users for their specific study.

3. Fundamental Requirements of AI-Readiness

Datasets produced without consideration for emerging computational applications may pose technical barriers at best and create ethical challenges and threats to research integrity, privacy, or security concerns at worst.

We assert that it is not enough for data to be usable in creating some prediction; its provenance must also be well-documented to ensure that usability, reliability, privacy, attribution, data quality, and trust are maintained. Attributes of what is termed "pre-model explainability"[39] must also be satisfied, including consideration of biases and assumptions inherent to the data. We have integrated these and other requirements of AI-ready data based on the data's properties, the practices used to achieve those properties, and examples of the practices. AI-ready data may not be completely devoid of biases, skewness, and assumptions, but must be accompanied by documentation and metadata that describe these characteristics for downstream reuse.

We stress the importance of documentation and metadata in enhancing the capabilities of both, automated processes and human researchers when interpreting, evaluating, and validating a dataset. Datasets must have clearly defined labels, provenance, and characterization to be minimally AI-ready, i.e., the meaning and derivation of each value in the data, and the datasets as a whole, must be interpretable to both human researchers and computational processes.

3.1 FAIRness and AI-Readiness

The FAIR principles were presented in 2016 as a set of defined characteristics of data, tools, and infrastructures that aid discovery and reuse by third-parties[23,40,11]. They are increasingly recognized as essential to digital scholarship, long-term sustainability, and reuse of datasets. The zeroth-order reuse case for scientific data is their assessment for validity. As a collection of datasets designed for long-term reuse by biomedical researchers, all Bridge2AI datasets must be FAIR. In particular, they must be assigned persistent IDs, resolvable to rich descriptive metadata. The metadata must be searchable in a public resource (subject to any applicable sensitivity restrictions), and sustained beyond the life of the data itself. Since its publication, it has become clear that there will be levels of FAIRness, a spectrum ranging from simple "Findability and Accessibility" characteristics with essentially bibliographic-style machine-readable metadata, to very rich dataset and provenance descriptions in machine and human readable presentation[40]. In developing AI-readiness criteria and working to make Bridge2AI datasets compliant, we find that a number of criteria originally defined as FAIR require a great deal of additional depth and specialization, particularly along the dimensions of Provenance, Characterization, and Sustainability. Simple FAIRness is not enough. However, it does provide a useful starting point and a reliable framework to integrate datasets, software, and other research objects with the literature through direct citation[41–45], bind metadata to data, and separate full accessibility from simple dataset characterization.

Deep provenance and data characterization requirements, which we elaborate on independently, flow from the FAIR Principles. However they take on such an important role in AI-Readiness that– as will be seen later –we treat them as criteria in their own right. A recent investigation by *Science* magazine into purported fabrication of neuroscience data in over one hundred publications by a prominent Alzheimer Disease researcher[46], in which the publications in question promoted what appear to be false directions in pharmaceutical development and failed clinical trials, points sharply to the importance of these requirements. Data used for AI model training and analysis should always be capable of being traced back to the original, unmodified version from experiments, clinical trials, electronic health records, surveys, or other sources, for verification.

3.2 Pre-model Explainability and AI-readiness

AI/ML applications require structured and well-described data to ensure the conclusions drawn from analyses are understandable and interpretable. This includes explaining the data acquisition and preparation processes prior to model training and use.

Explainability refers to the ability to understand and interpret the decisions and behaviors of AI systems. It is key to supporting research integrity, ensuring compliance with regulations, facilitating debugging and improvement of AI/ML models, and informing assessments of the trustworthiness of model predictions/outputs, especially for clinical decisions or therapeutic recommendations involving patients[30,47–50]. Importantly, the explanations must be tailored to the purposes and comprehension of a particular intended audience (e.g., clinicians vs. computer scientists vs. patients). Ultimately, AI explainability enhances model reliability and acceptability for ethical use in critical domains like healthcare.

Fundamentally, any analytic prediction or classification is the assertion of a computational argument[51]. The grounds for this assertion must be shown to compel sufficient belief in the conclusion's reliability so that important actions (those with a cost or impact) are justified at the time they are made given the information available, even if ultimately the assertion turns out to be only partly correct or disproven[52]. An assertion without adequate grounds is not epistemically justified[53]. That is, it is difficult to know if belief in the assertion is justified and likely to be true. If the assertion cannot stand up to counterarguments supported by adequate evidence, it cannot be scientifically convincing[52,54,55]. This is as true for an analytic prediction as it is for a textual argument.

Bridge2AI recognizes the need for multiple AI/ML applications to operate on data generated by its four Grand Challenges, and that complete explainability for AI is an end-to-end property dependent on more than just data description. However, AI/ML applications and systems are founded on data[56], and therefore any data explainability issues propagate through the system. The five stages of AI explainability can be broadly summarized as follows, extending from Khaleghi 2019[39]:

- **Pre-Modeling Stage**: This stage involves ensuring that the data and design choices made before model training are transparent and understandable. It includes data sourcing and production transparency (provenance), feature engineering, and model selection. Clear documentation about data sources, collection methods, preprocessing steps, and the rationale behind feature selection and model choices are essential for transparency[57].
- **In-Model Stage**: This stage focuses on making the inner workings of the model more transparent and understandable. It involves providing a clear description of the model architecture, documenting the training process, and interpreting intermediate outputs. Understanding the structure of neural networks or decision trees, hyperparameter choices, and techniques like regularization and data augmentation are key components.
- **Post-Modeling Stage**: This stage involves interpreting the outputs and decisions made by the model after it has been trained. Techniques such as feature importance, local explanations (e.g., LIME[58], SHAP[59], saliency-based approaches[60]), and global interpretability methods (e.g., summary plots, rule extraction) help explain individual predictions and provide a holistic understanding of the model's behavior across different inputs.
- **Post-Deployment Stage**: This stage involves monitoring and interpreting the model's performance and decisions in a real-world setting. Continuous monitoring of performance metrics, maintaining audit trails, and collecting user feedback are crucial for detecting drifts, facilitating audits, and understanding user perceptions, i.e., Quality Control.
- **Continuous Improvement Stage**: This stage focuses on using insights gained from the explainability processes to improve the model. It includes refining the model based on feature importance, error analysis, and user feedback, and regularly updating documentation and explainability tools to reflect changes and improvements.

Given that Bridge2AI is a data generation program, pre-model XAI is the predominant stage discussed herein. The other four stages of XAI mentioned above, while important, are outside of the scope of this discussion.

Our fundamental objectives are to support Bridge2AI datasets that are as comprehensively FAIR, explainable, ethical, sustainable, and computable as possible. To this end, we define several dimensions of AI-readiness (Box 2) which may be used to guide data generation efforts and to classify interim and final results. The dimensions are composed of criteria for evaluation, and specific practices we recommend to satisfy the criteria. Compliance with these criteria and supporting practices is a goal for AI-readiness of datasets in Bridge2AI.

While our projects focus primarily on the machine learning subfield of AI, it is our expectation that other forms of AI may also leverage these datasets. With this in mind, we are seeking to build the best possible foundation for pre-model explainability across multiple studies over a long period of time, and to support long-term viability and evolution of the data generation processes we endorse.

3.3 Ethical Practices and Sustainability

Preparing biomedical data to train AI/ML models requires careful consideration of the associated ethical, legal, and social implications (ELSI)[61–63]. The impact, evaluation, and treatment of ELSI may vary by use case. Data acquisition and governance conditions must therefore be documented in metadata to allow ELSI considerations to be assessed by prospective data users and given proper weight.

Accepted ethical principles that guide much of biomedical research in the US are described in the Belmont[64,65] and Menlo Reports [66]. The latter builds off of the Belmont principles of respect for persons, beneficence and justice applied in the context of information and communication technologies. If there is a desire to include Indigenous Peoples' data, the CARE principles[67] should be used to guide ethical practices. Key ethical issues specific to acquisition, management, and use of AI in biomedical research include identification and management of biases, practices for obtaining or waiving informed consent, privacy considerations, and practices that promote trust and trustworthiness. Identification and management of ELSI across the biomedical AI lifespan requires the adoption and implementation of an appropriate governance framework[68,69]. Specifically:

- *Ethical practices* should be outlined and their implementation managed by a governing body with representation appropriate to the nature of the projects. This could include varying selections of scientists, clinicians, ethicists, patients and, in some cases, the public. To promote transparency of governance practices, access to governance documentation should be included in data release metadata. Features of governance include decisions that map to: 1) data acquisition, 2) data management, including the collection, curation, storage, access and use, and 3) sustainability of the dataset for ongoing use in advancing knowledge of human health.
- *Data acquisition and handling metadata* should include sufficient details that indicate where, from whom (i.e., what participant/patient group(s)), and how samples or subject data were obtained and processed. Research Resource Identifiers (RRIDs) should be used to document the reagent or strain IDs for data sourced from banked cell lines or model organisms, linking to the provider's data sheets. Anonymized subject IDs for data from human subjects, should link to the non-identifying subject characteristics relevant for analysis. Provenance graphs should provide detailed data acquisition and processing information.
- *Governance metadata* should indicate data licensing and/or data use, including privacy requirements, conditions and specify contact information if negotiated data use agreements (DUAs) are required. *Data license metadata*, if any, should reference commonly understood licenses such as Creative Commons licenses [70]. Do not use the CC0 public domain disclaimer (this is not a license), which disempowers any further controls over data use. Other data reuse conditions must be clearly specified if they exist, or contact information for negotiated data use agreements indicated.

- *Sustainability considerations* must be addressed for long term benefit. To be consistent with FAIR principles, data must be deposited in sustainable archives for reuse. Longer term support may require a diverse portfolio of funding to include corporate sponsors, foundations and philanthropic partnerships. Where possible, ongoing feedback from data users should be enabled. Sustainability planning should commence as early in the project as feasible.

3.4. AI-ready Data Quality

Data quality assessment critically influences the performance and reliability of AI/ML models in biomedical applications. Poor data quality, characterized by inaccuracies, incompleteness, and inconsistencies, can lead to incorrect results and negatively impact clinical implementation and decision support [71,72]. Incorporating data quality information into metadata supports the reliability and reproducibility of AI/ML models in biomedical research. It enhances transparency, facilitates data sharing, and enables researchers to assess the suitability of datasets for specific AI/ML applications Detailed metadata annotations, including data provenance and quality indicators, are important for accurately interpreting AI/ML model outputs. Large-scale data sets for which quality control/quality assurance has been conducted support training robust AI systems capable of effective generalization. Kahn and colleagues proposed a harmonized data quality assessment terminology and framework specifically designed for EHR data, enabling systematic evaluation of data quality dimensions[73]. Adhering to standards like ISO's Data Quality Management (ISO 8000-61) [74]— which offers a structured methodology for ensuring data reliability and integrity — ensures that data quality information is systematically recorded and universally understood, critical for collaborative AI/ML research. Recording data quality within metadata is crucial for the effective application of AI/ML in biomedical research. It promotes transparency, enhances dataset utility, and contributes to the development of reliable and trustworthy AI models that can improve patient outcomes. Meticulous documentation of data quality would maximize the value of data collection efforts and better enable future data reusability by providing context and interpretability. As the predictive task for which the data will be used is often unknown for resources like Bridge2AI, detailed metadata rather than filtering out data is critical to ensure maximal usage.

3.5 Dimensions of AI-Readiness

AI-readiness is a dynamic property of specific data sets. It is context-dependent and developmental. We do not score it pass-fail as a whole, but along multiple dimensions based on readiness scores for major components. Achieving it in any particular use case is a collaborative, developmental, research-driven task[23,75,76] **.**

Our vision is to answer the question: What does it mean for a biomedical dataset to be AI-ready? We hope these criteria will be helpful to others.

Ultimately, what we are seeking in AI-Readiness extends beyond simple utility, convenience, or tractability for computer scientists and informaticians. We seek to enable data that are reusable and results that are ethical, scientifically valid, explainable, interpretable, and sustainable. Our criteria for AI-readiness are in direct service of these goals. We consider scientific validity as part of ethics, having to do with research integrity. Ultimately, our principal goal is data that are available, deeply characterized, standardized where possible, and which provide foundational support for ethical explainability of results.

4. AI-Readiness Criteria

4.1 Fundamental Criteria

We outline the following criteria for biomedical AI-readiness:

Biomedical data must be FAIR. Fundamental FAIRness is a “level 0” NIH requirement for this program. While the original 2016 FAIR Principles defined a general framework for many properties more fully elaborated on here, AI-readiness of data also imposes further properties beyond basic FAIRness compliance, requiring more complete specification of some general FAIRness criteria, and extending beyond FAIRness.

AI-readiness therefore implies that data must be FAIR, Provenanced as fully as feasible, Characterized in depth, Pre-model Explainable, Ethical, Sustainable, and Computable (**Box 2** and **Figure 1**).

Box 2 – Basic Criteria for AI-Readiness in Bridge2AI

FAIRness: Digital objects are Findable, Accessible, Interoperable, and Reusable at a basic level.
Provenance: Origins and transformational history of digital objects are richly documented.
Characterization: Content semantics, statistics, and standardization properties of digital objects are well-described for datasets, and software, used to prepare the data, including any quality or bias issues.
Pre-Model Explainability: Supports explainability of predictions and classifications based on the data with regard to metadata, fit for purpose, and data integrity.
Ethics: Ethical data acquisition, management, and dissemination, are documented and maintained.
Sustainability: Digital objects and their metadata stored in FAIR, long-term, stable archives.
Computability: Standardized, computationally accessible, portable, and contextualized.

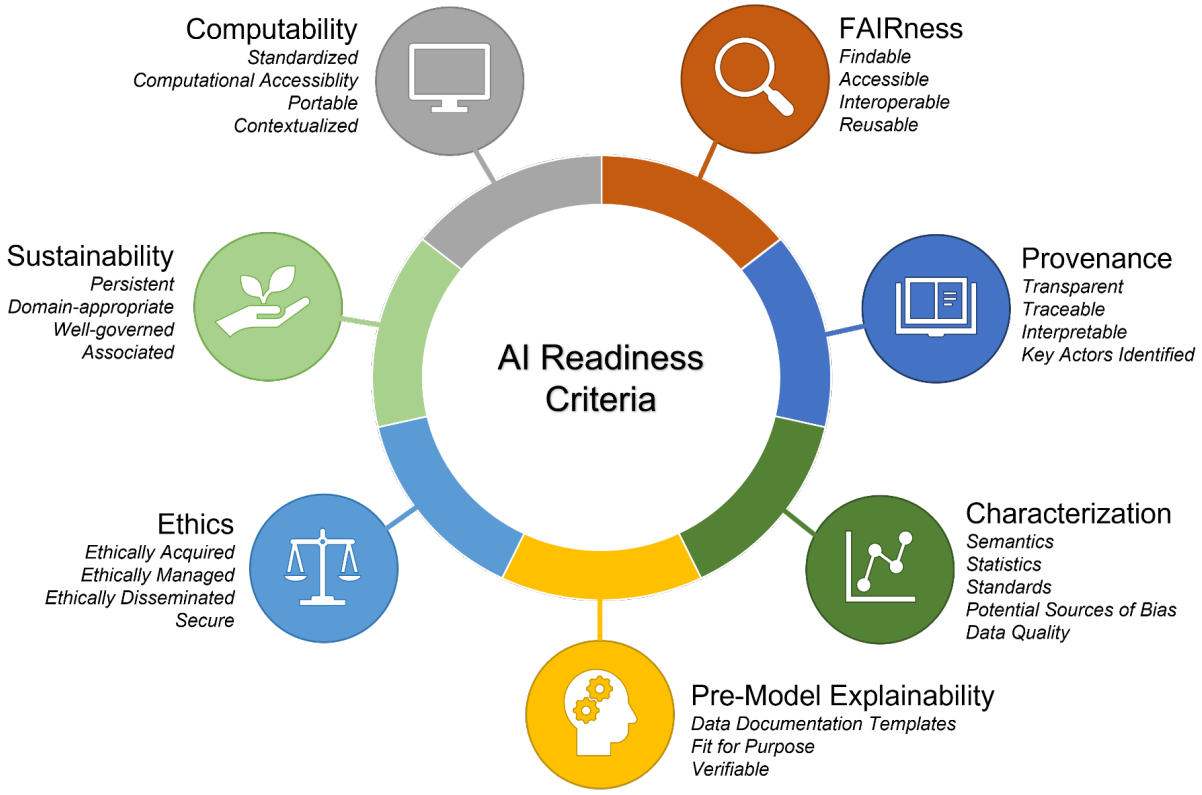

Figure 1. Seven overarching AI-readiness criteria developed for Bridge2AI datasets, along with their relevant subcriteria (*italics*) as detailed in Table 1.

4.2 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 [77]. **Table 1** describes the criteria and relevant practices that aid in achieving the criteria.

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 providing globally unique persistent identifiers (Datacite DOIs, N2T ARKs, CNRI HDLs) resolvable to searchable, machine-readable, richly-descriptive metadata, including a link to the dataset if available[78–80]. Datasets may be subject to access restrictions.	FAIRsharing.org [81] NIH GREI-participating repositories [82]
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 like DCAT (Data Catalog Vocabulary) or schema.org.	
0.c	Interoperable	Wherever possible, provide data and metadata using formally defined specifications for digital objects.	RDF, JSON LD
0.d	Reusable	Attach a clear and accessible data usage license that allows the responsible use of AI/ML applications. Alternatively, for restricted datasets, define a Data Use Agreement (DUA) and provide a means for automated acknowledgment and tracking of this agreement by data users.	Creative Commons licenses [70] (other than CC0)
1	Provenance		
1.a	Transparent	Identify sources of data back to reasonable ground-truth, e.g. clinical data from EHR at a given hospital, clinical trials, or laboratory data.	OMOP [83], RRID [84–86]
1.b	Traceable	Identify important data transformation steps, with links to software, at an appropriate level of detail, ideally using a machine-readable representation, such as W3C PROV-O or EVI.	W3C PROV [87], EVI [88]
1.c	Interpretable	Make software for key data transformation and analysis steps available in a sustainable repository [89] .	Zenodo [90], Software Heritage [91], Github
1.d	Key Actors Identified	Identify the people and organizations responsible for obtaining and processing the data, along with the samples and subject groups involved in producing the data. Reference these parties along with other dataset metadata.	ORCID [92], ROR
2	Characterization		
2.a	Semantics	Use full descriptive metadata for datasets, including a detailed abstract, dataset keywords, and subject-specific vocabularies (e.g., MeSH for biomedical data) to enable detailed search and discovery.	Datacite schema [93], Schema.org [94]
2.b	Statistics	Provide appropriate statistical characterizations of key features of the dataset (e.g. demographics) where appropriate, to assist in planning analyses. 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 important applicable standards.	

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 methodological reasons for missingness, as well as the degree to which the data represents a state of interest 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/or a linked human-readable document should support a domain-appropriate subset of the information in Datasheets[95] or Healthsheets[96]. Reference the specific information items supplied for this dataset.	Datasheets, Healthsheets
3.b	Fit for Purpose	Identify appropriate and inappropriate use cases for a given data set in AI applications. Link to any previously published analyses using this data.	
3.c	Verifiable	Provide a mechanism for ensuring the integrity of each raw or processed dataset, such as a checksum.	
4	Ethics		
4.a	Ethically Acquired	Describe ethical data acquisition consistent with accepted principles (i.e. Belmont ReportPrinciples [64,65], sufficient for its proper evaluation in context of intended use, along with a management plan.	Belmont principles, Menlo principles [66], CARE principles [67]
4.b	Ethically Managed	Data management, including processing, storage and access and use are expected to align with ethical principles throughout the health AI lifecycle. Indicate privacy-protection processing, if any, sufficient to evaluate ethical status for intended use, e.g. "anonymized" vs. "limited data set" vs. "non-PHI dataset".	
4.c	Ethically Disseminated	Specify a licensing agreement and/or data use agreement (DUA), or contact information to establish a DUA, on as open terms as ethical and sustainability considerations permit. Specify contact information for a data access committee, if needed to review requests for controlled data.	
4.d	Secure	Specify security requirements for storing and accessing this data, e.g. “public”, “controlled access only”, etc.	HL7 privacy protection metadata [97]
5	Sustainability		
5.a	Persistent	Ensure that unprocessed data is preserved in an archive adhering to privacy laws and retention guidelines, enabling future reprocessing and updated publishing of revised data.	
5.b	Domain-appropriate	Ensure single domain raw or processed data (as appropriate) is deposited in a FAIR domain-appropriate specialist repository if available.	
5.c	Well-governed	Select a repository that facilitates how data will be stewarded in the future and governance that accounts for maintenance, terms and policy changes, and fairness.	FAIRsharing.org [81] NIH GREI-participating repositories [98]
5.d	Associated	Document project-level connections between data components and elements in a machine-readable manner.	RO-Crate [99–101]
6	Computability		
6.a	Standardized	Datasets follow established, documented standards and their adherence to standards may be validated deterministically.	
6.b	Computationally Accessible	Provide a mechanism to access data either through established exchange protocols or a well-documented API.	FAIRsharing.org [81] NIH GREI-participating repositories [98]
6.c	Portable	Maximize portability across computational resources where possible. If working with the data 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. If possible, provide examples of data components to facilitate understanding of their general structure and content.	
† Practices may impact multiple criteria; the most relevant relationship is shown for brevity			

5. AI-Readiness Evaluation

We evaluated AI-readiness for each Bridge2AI Grand Challenge dataset - both released and pre-release data - along the major dimensions established in Table 1 by treating each criterion as an axis in a radar plot. If a sub-criterion is addressed in satisfactory form it was given a score of "1"; a score of "0" was assigned if the sub-criterion was not addressed by the DGP. 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-criterion. For example, scoring three out of four total sub-criteria as "1" would produce an overall score of 75% satisfaction for that criterion. **Figure 2** shows radar plot evaluations for each Bridge2AI GC in its current state, as well as the target ("goal") AI-readiness scores that each GC will strive to attain by the end of the project.

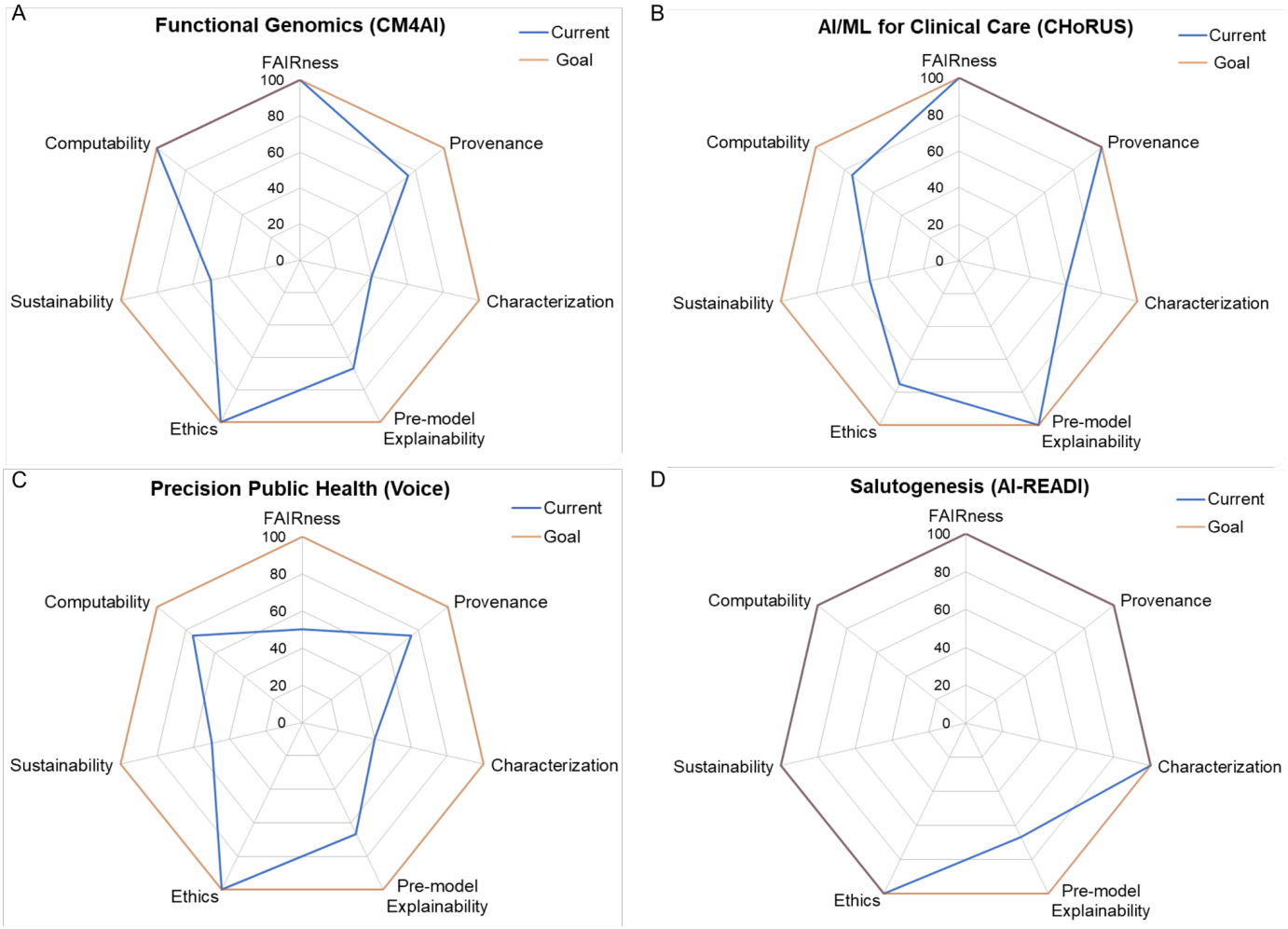

Figure 2: AI-readiness radar plots for Bridge2AI Grand Challenges: (A) Functional Genomics (CM4AI); (B) AI/ML for Clinical Care (CHoRUS); (C) Precision Public Health (Voice); (D) Salutogenesis (AI-READI). The blue lines indicate how well each GC's data and metadata practices currently meet the seven AI-

readiness criteria, for the data collected as of the end of Year Two of the program, with the orange line representing the AI-readiness goals across all criteria that each GC will try to reach by the end of the project.

Figure 2 indicates that each GC has unique opportunities and challenges to address in order to ensure that their data are AI-ready by the end of the Bridge2AI program. These radar plots help visualize AI-readiness features that would benefit the most from further development, improvement, discussion, and implementation. Completed data collection forms with detailed ratings for each criteria are available in *Supplemental Data*.

6. Challenges and Limitations

Preparation of valid AI-ready biomedical datasets requires additional effort beyond simply capturing measurements or observations for statistical analysis. This effort is increased when datasets are intended, as in Bridge2AI, to meet multiple use cases and be sustainable over time, rather than addressing one-off highly focused research questions. AI-readiness data preparation requires significant understanding of the data itself, the predictive task 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, ethical treatment of human subjects data can be a significant concern, requiring attention to proper de-identification techniques (anonymization), privacy preservation practices, and responsible data stewardship. This further emphasizes the need for the Provenance and Ethics criteria to ensure that data use limitations, compliance, intellectual property and other restrictions are clearly stated and followed in downstream use of the data.

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

Limitations and challenges like these require teamwork and demand a Team Science approach, the more so as the project ambition and scope increases [102].

7. Conclusions and Future Directions

AI-ready data preparation and evaluation requires a set of practices focused on establishing data and software FAIRness, detailed provenance, statistical characterization, support for pre-model explainability, ethical characterization, sustainability, and computability. These practices should be reflected in the metadata associated with an AI-ready dataset, and of course in the data itself. In this article, we have outlined a set of criteria reflecting our recommended practices, with methods for evaluating adherence. The criteria we propose here are currently in use in NIH's Bridge2AI program. We believe these datasets and their associated deep metadata and technologies will enable many novel, significant, and transformational discoveries. Developing these data resources has enabled and required the participating investigators to look deeply and comprehensively into the requirements for AI-readiness, and sparked the need to develop the criteria and evaluation methods described herein.

Beyond the datasets themselves, we believe the standards defined and evaluated herein will benefit the biomedical AI/ML community at large. Particularly, ensuring that data are AI-ready sets the stage for downstream users to apply the rapidly emerging capabilities of AI toward vastly improving our understanding of disease and the development of new treatments and technologies.

Our contributions 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;

Additional tools supporting AI-readiness developed in Bridge2AI include the LinkML translators; formal schemas in LinkML for Datasheets; and the FAIRSCAPE AI-readiness framework. These tools will continue to be developed along directions indicated in this article.

It should be noted that the criteria we established require significant additional metadata beyond what is required, e.g., for a Datacite DOI registration. We believe this effort will vary with the use case envisioned for a particular dataset, and may be significantly reduced by using tools we provide or are currently developing.

We welcome comments on this article and collaborations with other biomedical AI/ML researchers, including both users of the Bridge2AI datasets and those wishing to collaborate on similar projects. Our team would be grateful to users of our datasets who communicate their experiences to us, and who cite this article in their work. The ideas presented here reflect the perspectives of people embedded in the work of producing datasets that are intended to be flagship examples that embody best practices. The Bridge2AI datasets are intended to be broadly used. As users employ the data to develop impactful AI algorithms, we will learn where the ideas in this article succeed and areas for future improvement. We welcome these metrics being used by other data generators to improve AI-readiness of their products, and by those who re-use datasets produced by others, to assess their suitability.

8. Data and Software Availability Statement

The Ai-Readiness evaluation data are available in Zenodo as

AII-readiness Evaluation Data for Bridge2AI Grand Challenges

- Parker JA. 2024 - Bridge2AI Grand Challenge AI-Readiness Evaluations at Year 2 of 4. Zenodo. https//doi.org/10.5281/zenodo.13931522

Ai-readiness Evaluation Worksheet

- Parker JA, et al. 2024 - AI-Readiness Self-Evaluation Worksheet https://doi.org/10.5281/zenodo.13961091

Bridge2AI-funded assistive tools are available here:

LinkML Datasheets for Datasets Schema:

- Joachimiak MP, Caufield JH, Mungall CJ. 2024 - Datasheets for Datasets Schema (v0.1.0). Zenodo. https://doi.org/10.5281/zenodo.13964135

LinkML Translators:

- Moxon S; et al. 2024. LinkML (v1.8.4). Zenodo. https://doi.org/10.5281/zenodo.13871320

FAIRSCAPE AI-readiness Framework:

- Niestroy J, et al. 2024 - FAIRSCAPE GUI Client (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.13951906
- Levinson MA, et al. 2024 - FAIRSCAPE-CLI: A utility for packaging objects and validating metadata for FAIRSCAPE. (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.14014297
- Levinson MA, et al. 2024 - FAIRSCAPE Server, version 0.7.0. Zenodo. https://doi.org/10.5281/zenodo.13971502

9. Acknowledgements

This work was funded by the National Institutes of Health under awards OT2OD032742 [Bridge2AI: Cell Maps for AI (CM4AI) Data Generation Project], OT2OD032644 [Bridge2AI: Salutogenesis Data Generation Project], OT2OD032720 [Bridge2AI: Voice as a Biomarker of Health], OT2OD032701 [Bridge2AI: Patient-Focused Collaborative Hospital Repository Uniting Standards (CHoRUS) for Equitable AI], U54HG012510 [Bridge2AI: a FAIR AI BRIDGE Center (FABRIC)], 5U54HG012517 [Building BRIDGEs: Coordinating Standards, Diversity, and Ethics to Advance Biomedical AI], and 5U54HG012513 [Integration, Dissemination and Evaluation(BRIDGE) Center for the NIH Bridge to Artificial Intelligence (BRIDGE2AI) Program], and by the Frederick Thomas Fund of the University of Virginia. JNH has been supported with an EMBO Postdoctoral Fellowship (ALTF 556-2022).

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med [Internet]. 2019 Jan [cited 2024 Oct 3];25(1):44–56. Available from: https://www.nature.com/articles/s41591-018-0300-7
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 [Internet]. 2023 Oct 19 [cited 2024 Jun 15];6:1271639. Available from: https://www.frontiersin.org/articles/10.3389/fdata.2023.1271639/full
3. Lin AY, Arabandi S, Beale T, Duncan WD, Hicks A, Hogan WR, Jensen M, Koppel R, Martínez-Costa C, Nytrø Ø, Obeid JS, de Oliveira JP, Ruttenberg A, Seppälä S, Smith B, Soergel D, Zheng J, Schulz S. Improving the Quality and Utility of Electronic Health Record Data through Ontologies. Standards (Basel). 2023 Sep;3(3):316–340. PMCID: PMC10591519
4. Pacheco JA, Rasmussen LV, Wiley K, Person TN, Cronkite DJ, Sohn S, Murphy S, Gundelach JH, Gainer V, Castro VM, Liu C, Mentch F, Lingren T, Sundaresan AS, Eickelberg G, Willis V, Furmanchuk A, Patel R, Carrell DS, Deng Y, Walton N, Satterfield BA, Kullo IJ, Dikilitas O, Smith JC, Peterson JF, Shang N, Kiryluk K, Ni Y, Li Y, Nadkarni GN, Rosenthal EA, Walunas TL, Williams MS, Karlson EW, Linder JE, Luo Y, Weng C, Wei W. Evaluation of the portability of computable phenotypes with natural language processing in the eMERGE network. Sci Rep [Internet]. 2023 Feb 3 [cited 2024 Jun 15];13(1):1971. Available from: https://www.nature.com/articles/s41598-023-27481-y
5. Jiang J (Xuefeng), Qi K, Bai G, Schulman K. Pre-pandemic assessment: a decade of progress in electronic health record adoption among U.S. hospitals. Health Affairs Scholar [Internet]. 2023 Nov 3 [cited 2024 Jun 15];1(5):qxad056. Available from: https://academic.oup.com/healthaffairsscholar/article/doi/10.1093/haschl/qxad056/7326049
6. Abbasizanjani H, Torabi F, Bedston S, Bolton T, Davies G, Denaxas S, Griffiths R, Herbert L, Hollings S, Keene S, Khunti K, Lowthian E, Lyons J, Mizani MA, Nolan J, Sudlow C, Walker V, Whiteley W, Wood A, Akbari A, CVD-COVID-UK/COVID-IMPACT Consortium. Harmonising electronic health records for reproducible research: challenges, solutions and recommendations from a UK-wide COVID-19 research collaboration. BMC Med Inform Decis Mak [Internet]. 2023 Jan 16 [cited 2024 Jun 15];23(1):8. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-022-02093-0
7. Kim MK, Rouphael C, McMichael J, Welch N, Dasarathy S. Challenges in and Opportunities for Electronic

Health Record-Based Data Analysis and Interpretation. Gut and Liver [Internet]. 2024 Mar 15 [cited 2024 Jun 15];18(2):201–208. Available from: http://gutnliver.org/journal/view.html?doi=10.5009/gnl230272
8. Jackson N, Woods J, Watkinson P, Brent A, Peto TEA, Walker AS, Eyre DW. The quality of vital signs measurements and value preferences in electronic medical records varies by hospital, specialty, and patient demographics. Sci Rep [Internet]. 2023 Mar 8 [cited 2024 Jun 15];13(1):3858. Available from: https://www.nature.com/articles/s41598-023-30691-z
9. Hiniduma K, Byna S, Bez JL. Data Readiness for AI: A 360-Degree Survey [Internet]. arXiv; 2024 [cited 2024 Jun 19]. Available from: http://arxiv.org/abs/2404.05779
10. Ng MY, Youssef A, Miner AS, Sarellano D, Long J, Larson DB, Hernandez-Boussard T, Langlotz CP. Perceptions of Data Set Experts on Important Characteristics of Health Data Sets Ready for Machine Learning: A Qualitative Study. JAMA Netw Open. 2023 Dec 1;6(12):e2345892. PMCID: PMC10692863
11. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data [Internet]. 2016;3:160018. Available from: https://doi.org/10.1038/sdata.2016.18 PMCID: PMC4792175
12. National Institutes of Health. Bridge to Artificial Intelligence (Bridge2AI) [Internet]. National Institutes of Health Common Fund; 2023 [cited 2023 Feb 9]. Available from: https://commonfund.nih.gov/bridge2ai
13. Bahroun Z, Anane C, Ahmed V, Zacca A. Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis. Sustainability [Internet]. 2023 Aug 29 [cited 2024 Jun 25];15(17):12983. Available from: https://www.mdpi.com/2071-1050/15/17/12983
14. Cao L. AI in Finance: Challenges, Techniques, and Opportunities. ACM Comput Surv [Internet]. 2023 Mar 31 [cited 2024 Jun 25];55(3):1–38. Available from: https://dl.acm.org/doi/10.1145/3502289
15. Deng J, Yang Z, Ojima I, Samaras D, Wang F. Artificial intelligence in drug discovery: applications and techniques. Briefings in Bioinformatics [Internet]. 2022 Jan 17 [cited 2024 Jun 25];23(1):bbab430. Available from: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab430/6420092
16. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research [Internet]. 2022 Jan 7 [cited 2024 Apr 12];50(D1):D439–D444. Available from: https://academic.oup.com/nar/article/50/D1/D439/6430488
17. ACD AI WG. Report of the Advisory Committee to the Director Working Group on AI [Internet]. National Institutes of Health; 2019. Available from: https://www.acd.od.nih.gov/documents/presentations/12132019AI_FinalReport.pdf
18. Clark T, Schaffer LV, Obernier K, Al Manir S, Churas C, Dailamy A, Doctor Y, Forget A, Hansen JN, Hu M, Levinson MA, Marquez C, Nourreddine S, Niestroy JC, Pratt D, Qian G, Thaker S, Bélisle-Pipon JC, Brandt CA, Chen JY, Ding Y, Fodeh S, Krogan NJ, Lundberg E, Musmade P, Payne-Foster P, Ratcliffe S, Ravitsky V, Sali A, Schulz W, Ideker T. Cell Maps for Artificial Intelligence: AI-Ready Maps of Human Cell Architecture from Disease-Relevant Cell Lines. BioRXiv.org (submitted); 2024.
19. Chakir A, Andry JF, Ullah A, Bansal R, Ghazouani M, editors. Engineering Applications of Artificial Intelligence [Internet]. Cham: Springer Nature Switzerland; 2024 [cited 2024 Jul 4]. Available from: https://link.springer.com/10.1007/978-3-031-50300-9
20. Russell SJ, Norvig P. Artificial intelligence: a modern approach. Fourth edition. Hoboken: Pearson; 2021.
21. McCarthy J. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Magazine. 2006;24(4):12–14.
22. Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, Ektefaie Y, Kondic J, Zitnik M. Empowering Biomedical Discovery with AI Agents [Internet]. arXiv; 2024 [cited 2024 Aug 27]. Available from: http://arxiv.org/abs/2404.02831
23. Kidwai-Khan F, Wang R, Skanderson M, Brandt CA, Fodeh S, Womack JA. A roadmap to artificial

intelligence (AI): Methods for designing and building AI ready data to promote fairness. Journal of Biomedical Informatics [Internet]. 2024 Jun [cited 2024 May 23];154:104654. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1532046424000728
24. Grobelnik M, Perset K, Russell S. What is AI? Can you make a clear distinction between AI and non-AI systems? [Internet]. OECD.AI Policy Observatory; 2024. Available from: https://oecd.ai/en/wonk/definition
25. Altman RB, Levitt M. What is Biomedical Data Science and Do We Need an Annual Review of It? Annu Rev Biomed Data Sci [Internet]. 2018 Jul 20 [cited 2024 Jun 11];1(1):i–iii. Available from: https://www.annualreviews.org/doi/10.1146/annurev-bd-01-041718-100001
26. Shirey R. Internet Security Glossary, Version 2 [Internet]. Internet Engineering Task Force; 2013. Available from: https://datatracker.ietf.org/doc/rfc4949/
27. Stevenson A, editor. Shorter Oxford English dictionary on historical principles. 1: A - M / [ed.: Angus Stevenson]. 6. ed. Oxford: Oxford University Press; 2007.
28. NIST. NIST Computer Security Resource Center - Glossary [Internet]. National Institute of Standards and Technology; 2024. Available from: https://csrc.nist.gov/glossary/term/data_element
29. W3C Schema.org Community Group. Schema.org: Dataset. Schema.org; 2024.
30. Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. Sensors [Internet]. 2023 Jan 5 [cited 2024 Jun 11];23(2):634. Available from: https://www.mdpi.com/1424-8220/23/2/634
31. Juty N, Wimalaratne SM, Soiland-Reyes S, Kunze J, Goble CA, Clark T. Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR. Data Intelligence [Internet]. 2020 Jan [cited 2020 Jun 3];2(1–2):30–39. Available from: https://www.mitpressjournals.org/doi/abs/10.1162/dint_a_00025
32. Gil Y, Miles S, Belhajjame K, Deus H, Garijo D, Klyne G, Missier P, Soiland-Reyes S, Zednik S. PROV Model Primer: W3C Working Group Note 30 April 2013 [Internet]. World Wide Web Consortium (W3C); 2013. Available from: https://www.w3.org/TR/prov-primer/
33. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods [Internet]. 2018 Apr 1 [cited 2024 Mar 25];15(4):290–298. Available from: https://www.nature.com/articles/nmeth.4627
34. Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. Visible Machine Learning for Biomedicine. Cell [Internet]. 2018 Jun [cited 2024 Mar 25];173(7):1562–1565. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867418307190
35. Zheng F, Kelly MR, Ramms DJ, Heintschel ML, Tao K, Tutuncuoglu B, Lee JJ, Ono K, Foussard H, Chen M, Herrington KA, Silva E, Liu SN, Chen J, Churas C, Wilson N, Kratz A, Pillich RT, Patel DN, Park J, Kuenzi B, Yu MK, Licon K, Pratt D, Kreisberg JF, Kim M, Swaney DL, Nan X, Fraley SI, Gutkind JS, Krogan NJ, Ideker T. Interpretation of cancer mutations using a multiscale map of protein systems. Science [Internet]. 2021 Oct [cited 2022 Mar 23];374(6563):eabf3067. Available from: https://www.science.org/doi/10.1126/science.abf3067
36. Qin Y, Huttlin EL, Winsnes CF, Gosztyla ML, Wacheul L, Kelly MR, Blue SM, Zheng F, Chen M, Schaffer LV, Licon K, Bäckström A, Vaites LP, Lee JJ, Ouyang W, Liu SN, Zhang T, Silva E, Park J, Pitea A, Kreisberg JF, Gygi SP, Ma J, Harper JW, Yeo GW, Lafontaine DLJ, Lundberg E, Ideker T. A multi-scale map of cell structure fusing protein images and interactions. Nature [Internet]. 2021 Dec 16 [cited 2022 Sep 14];600(7889):536–542. Available from: https://www.nature.com/articles/s41586-021-04115-9
37. Niestroy JC, Moorman JR, Levinson MA, Manir SA, Clark TW, Fairchild KD, Lake DE. Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis. npj Digit Med [Internet]. 2022 Dec [cited 2022 Jan 27];5(1):6. Available from: https://www.nature.com/articles/s41746-021-00551-z
38. Low DM, Rao V, Randolph G, Song PC, Ghosh SS. Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings [Internet]. 2020 [cited 2024 Aug 13]. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.11.23.20235945
39. Bahador Khaleghi. The How of Explainable AI: Pre-modelling Explainability [Internet]. Towards Data Science; 2019. Available from: https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4
40. Huerta EA, Blaiszik B, Brinson LC, Bouchard KE, Diaz D, Doglioni C, Duarte JM, Emani M, Foster I, Fox G, Harris P, Heinrich L, Jha S, Katz DS, Kindratenko V, Kirkpatrick CR, Lassila-Perini K, Madduri RK, Neubauer MS, Psomopoulos FE, Roy A, Rübel O, Zhao Z, Zhu R. FAIR for AI: An interdisciplinary and international community building perspective. Sci Data. 2023 Jul 26;10(1):487. PMCID: PMC10372139

41. Stall S, Bilder G, Cannon M, Hong NC, Edmunds S, Erdmann CC, Evans M, Farmer R, Feeney P, Friedman M, Giampoala M, Hanson RB, Harrison M, Karaiskos D, Katz DS, Letizia V, Lizzi V, MacCallum C, Muench A, Perry K, Ratner H, Schindler U, Sedora B, Stockhause M, Townsend R, Yeston J, Clark T. Journal Production Guidance for Software and Data Citations [Internet]. Preprints; 2022 Dec. Available from: https://essopenarchive.org/users/536571/articles/616035-journal-production-guidance-for-software-and-data-citations?commit=637aefc4958f77e4eca3b2476f36f77fbd2daccc
42. Katz D, Chue Hong N, Clark T, Muench A, Stall S, Bouquin D, Cannon M, Edmunds S, Faez T, Feeney P, Fenner M, Friedman M, Grenier G, Harrison M, Heber J, Leary A, MacCallum C, Murray H, Pastrana E, Perry K, Schuster D, Stockhause M, Yeston J. Recognizing the value of software: a software citation guide [version 2; peer review: 2 approved]. F1000Research. 2021;9(1257).
43. Groth P, Cousijn H, Clark T, Goble C. FAIR Data Reuse – the Path through Data Citation. Data Intelligence [Internet]. 2020 Jan [cited 2020 Jun 3];2(1–2):78–86. Available from: https://www.mitpressjournals.org/doi/abs/10.1162/dint_a_00030
44. Cousijn H, Kenall A, Ganley E, Harrison M, Kernohan D, Lemberger T, Murphy F, Polischuk P, Taylor S, Martone M, Clark T. A data citation roadmap for scientific publishers. Sci Data [Internet]. 2018 Dec [cited 2020 Aug 27];5(1):180259. Available from: http://www.nature.com/articles/sdata2018259 PMCID: PMC6244190
45. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science [Internet]. 2015 May 27 [cited 2019 May 9];1:e1. Available from: https://peerj.com/articles/cs-1
46. Piller C. Picture Imperfect. Science [Internet]. 2024 Sep 26;385(6716). Available from: https://doi.org/10.1126/science.z2o7c3k
47. Kundu S. AI in medicine must be explainable. Nat Med [Internet]. 2021 Aug [cited 2023 Apr 24];27(8):1328–1328. Available from: https://www.nature.com/articles/s41591-021-01461-z
48. Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. J Healthc Inform Res [Internet]. 2022 Jun [cited 2024 Jun 22];6(2):228–239. Available from: https://link.springer.com/10.1007/s41666-022-00114-1
49. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. 2020 Nov 30;20(1):310. PMCID: PMC7706019
50. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy [Internet]. 2020 Dec 25 [cited 2021 Jul 13];23(1):18. Available from: https://www.mdpi.com/1099-4300/23/1/18
51. Al Manir S, Niestroy J, Levinson MA, Clark T. Evidence Graphs: Supporting Transparent and FAIR Computation, with Defeasible Reasoning on Data, Methods, and Results. In: Glavic B, Braganholo V, Koop D, editors. Provenance and Annotation of Data and Processes [Internet]. Cham: Springer International Publishing; 2021 [cited 2022 Mar 16]. p. 39–50. Available from: https://link.springer.com/10.1007/978-3-030-80960-7_3
52. Tohmé F, Delrieux C, Bueno O. Defeasible Reasoning + Partial Models: A Formal Framework for the Methodology of Research Programs. Found Sci [Internet]. 2011 Feb [cited 2024 Jun 12];16(1):47–65. Available from: http://link.springer.com/10.1007/s10699-010-9200-0
53. Foley R. Justification, epistemic. Routledge Encyclopedia of Philosophy [Internet]. 1st ed. London: Routledge; 2016 [cited 2024 Jun 19]. Available from: https://www.rep.routledge.com/articles/thematic/justification-epistemic/v-1
54. Bench-Capon TJM, Dunne PE. Argumentation in artificial intelligence. Artificial Intelligence [Internet]. 2007 [cited 2007 Oct 1];171(10–15):619–641. Available from: http://www.sciencedirect.com/science/article/pii/S0004370207000793
55. Carrera Á, Iglesias CA. A systematic review of argumentation techniques for multi-agent systems research. Artif Intell Rev [Internet]. 2015 Dec [cited 2020 Sep 19];44(4):509–535. Available from: http://link.springer.com/10.1007/s10462-015-9435-9
56. Huyen C. Designing machine learning systems: an iterative process for production-ready applications. First edition. Sebastopol, CA: O'Reilly Media, Inc; 2022.
57. Kamath U, Liu,John. Pre-model Interpretability and Explainability. Explainable Artificial Intelligence: An

Introduction to Interpretable Machine Learning [Internet]. Springer International Publishing; 2021. Available from: https://doi.org/10.1007/978-3-030-83356-5_2
58. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:160204938 [cs, stat] [Internet]. 2016 Aug 9 [cited 2022 Feb 4]; Available from: http://arxiv.org/abs/1602.04938
59. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems [Internet]. Long Beach, CA,USA; 2017. p. 10. Available from: https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
60. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, Nensa F. Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches. European Journal of Radiology [Internet]. 2023 May [cited 2024 Sep 26];162:110787. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0720048X23001018
61. Čartolovni A, Tomičić A, Lazić Mosler E. Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. International Journal of Medical Informatics [Internet]. 2022 May [cited 2024 Sep 11];161:104738. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1386505622000521
62. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genet Med. 2017;19(7):743–750. PMID: 27929525
63. Sen SK, Green ED, Hutter CM, Craven M, Ideker T, Di Francesco V. Opportunities for basic, clinical, and bioethics research at the intersection of machine learning and genomics. Cell Genomics [Internet]. Elsevier BV; 2024 Jan [cited 2024 Sep 17];4(1):100466. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666979X23003105
64. 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. U.S. Department of Health, Education and Welfare; 1979.
65. Adashi EY, Walters LB, Menikoff JA. The Belmont Report at 40: Reckoning With Time. Am J Public Health. 2018 Oct;108(10):1345–1348. PMID: 30138058
66. Bailey M, Dittrich D, Kenneally E, Maughan D. The Menlo Report. IEEE Secur Privacy Mag [Internet]. 2012 Mar [cited 2024 Jul 21];10(2):71–75. Available from: http://ieeexplore.ieee.org/document/6173001/
67. Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker JD, Anderson J, Hudson M. The CARE Principles for Indigenous Data Governance. Data Science Journal [Internet]. 2020 Nov 4 [cited 2024 Jul 21];19:43. Available from: http://datascience.codata.org/articles/10.5334/dsj-2020-043/
68. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. Journal of the American Medical Informatics Association [Internet]. 2020 Mar 1 [cited 2024 Sep 11];27(3):491–497. Available from: https://academic.oup.com/jamia/article/27/3/491/5612169
69. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon [Internet]. 2024 Feb [cited 2024 Sep 11];10(4):e26297. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405844024023284
70. Carroll MW. Creative Commons and the New Intermediaries. Mich St L Rev [Internet]. 2006;45. Available from: http://works.bepress.com/michael_carroll/1
71. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–1219. PMCID: PMC5070532
72. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaney-Israni S, Goldenberg A. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019 Sep;25(9):1337–1340. PMID: 31427808
73. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244. PMCID: PMC5051581
74. ISO/IEC. Information technology — Open distributed processing — Reference model: Architecture (RM/ODP) [Internet]. Geneva CH: ISO/IEC; 2009 Dec. Report No.: ISO/IEC 10746-3:2009(E). Available from: http://www.joaquin.net/ODP/Part3/0.html
75. Thomas DM, Knight R, Gilbert JA, Cornelis MC, Gantz MG, Burdekin K, Cummiskey K, Sumner SCJ, Pathmasiri W, Sazonov E, Gabriel KP, Dooley EE, Green MA, Pfluger A, Kleinberg S. Transforming Big

Data into AI-ready data for nutrition and obesity research. Obesity [Internet]. 2024 May [cited 2024 Oct 9];32(5):857–870. Available from: https://onlinelibrary.wiley.com/doi/10.1002/oby.23989
76. Poduval B, McPherron RL, Walker R, Himes MD, Pitman KM, Azari AR, Shneider C, Tiwari AK, Kapali S, Bruno G, Georgoulis MK, Verkhoglyadova O, Borovsky JE, Lapenta G, Liu J, Alberti T, Wintoft P, Wing S. AI-ready data in space science and solar physics: problems, mitigation and action plan. Front Astron Space Sci [Internet]. 2023 Jul 13 [cited 2024 Oct 9];10:1203598. Available from: https://www.frontiersin.org/articles/10.3389/fspas.2023.1203598/full
77. ESIP Data Readiness Cluster. Checklist to Examine AI-readiness for Open Environmental Datasets [Internet]. ESIP; 2022 [cited 2024 Sep 27]. p. 179221 Bytes. Available from: https://esip.figshare.com/articles/online_resource/Checklist_to_Examine_AI-readiness_for_Open_Environmental_Datasets/19983722/1
78. Cousijn H, Braukmann R, Fenner M, Ferguson C, van Horik R, Lammey R, Meadows A, Lambert S. Connected Research: The Potential of the PID Graph. Patterns [Internet]. 2021 Jan [cited 2021 Jan 24];2(1):100180. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666389920302440
79. Ferguson C, McEntrye J, Bunakov V, Lambert S, Sandt SVD, Kotarski R, Stewart S, MacEwan A, Fenner M, Cruse P, Horik RV, Dohna T, Koop-Jacobsen K, Schindler U, McCafferty S. D3.1Survey of Current PID Services Landscape - Revised. Zenodo; 2019 Oct 18 [cited 2024 Oct 7]; Available from: https://zenodo.org/record/3554255
80. Madden F, van Horik R, van de Sandt S, Lavasa A, Cousijn H. Guides to Choosing Persistent Identifiers - Version 2 [Internet]. Zenodo; 2020 [cited 2024 Oct 7]. Available from: https://zenodo.org/record/3956569
81. The FAIRsharing Team. FAIRsharing.org [Internet]. University of Oxford; 2024 [cited 2024 Oct 30]. Available from: https://fairsharing.org/
82. Barbosa S, Curtin L, Cousijn H. Generalist Repository Ecosystem Initiative Introductory Brochure [Internet]. Zenodo; 2023 [cited 2024 Oct 7]. Available from: https://zenodo.org/record/8350509
83. Observational Medical Outcomes Partnership (OMOP). Standardized Data: The OMOP Common Data Model [Internet]. Observational Health Data Sciences and Informatics; 2024. Available from: https://www.ohdsi.org/data-standardization/
84. Bandrowski AE, Martone ME. RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. Neuron [Internet]. 2016 May [cited 2019 Mar 16];90(3):434–436. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0896627316301179
85. Bandrowski, Anita, Martone Maryann, Vasilevsky Nicole, Brush Matt, Haendel Melissa. Identifying research resources in biomedical literature should be easy. Front Neuroinform [Internet]. 2014 [cited 2020 Jul 17];8. Available from: http://www.frontiersin.org/Community/AbstractDetails.aspx?ABS_DOI=10.3389/conf.fninf.2014.18.00080
86. Prager EM, Chambers KE, Plotkin JL, McArthur DL, Bandrowski AE, Bansal N, Martone ME, Bergstrom HC, Bespalov A, Graf C. Improving transparency and scientific rigor in academic publishing. Brain and Behavior [Internet]. 2018 Dec 2 [cited 2019 Jan 7];e01141. Available from: http://doi.wiley.com/10.1002/brb3.1141
87. Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, Garijo D, Soiland-Reyes S, Zednik S, Zhao J. PROV-O: The PROV Ontology W3C Recommendation 30 April 2013. 2013; Available from: http://www.w3.org/TR/prov-o/
88. Al Manir S, Niestroy J, Levinson M, Clark T. EVI: The Evidence Graph Ontology, OWL 2 Vocabulary [Internet]. Zenodo; 2021. Available from: https://doi.org/10.5281/zenodo.4630931
89. Patel B, Soundarajan S, Ménager H, Hu Z. Making Biomedical Research Software FAIR: Actionable Step-by-step Guidelines with a User-support Tool. Sci Data [Internet]. 2023 Aug 23 [cited 2024 Oct 14];10(1):557. Available from: https://www.nature.com/articles/s41597-023-02463-x
90. European Organization For Nuclear Research, OpenAIRE. Zenodo [Internet]. CERN; 2013. Available from: https://www.zenodo.org/
91. Software Heritage Foundation. SoftWare Heritage persistent IDentifiers (SWHIDs), version 1.5 [Internet]. Software Heritage Foundation; 2020 [cited 2021 Feb 5]. Available from: https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html#overview
92. Akers KG, Sarkozy A, Wu W, Slyman A. ORCID Author Identifiers: A Primer for Librarians. Medical Reference Services Quarterly [Internet]. 2016 Apr 2 [cited 2024 Oct 9];35(2):135–144. Available from: http://www.tandfonline.com/doi/full/10.1080/02763869.2016.1152139
93. DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and

Citation of Research Data and Other Research Outputs v4.5. DataCite; 2024 [cited 2024 Jul 3]; Available from: https://datacite-metadata-schema.readthedocs.io/en/4.5/
94. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Communications of the ACM [Internet]. 2016 Jan 25 [cited 2019 Jan 9];59(2):44–51. Available from: http://dl.acm.org/citation.cfm?doid=2886013.2844544
95. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K. Datasheets for Datasets [Internet]. arXiv; 2021 [cited 2023 Nov 2]. Available from: http://arxiv.org/abs/1803.09010
96. Rostamzadeh N, Mincu D, Roy S, Smart A, Wilcox L, Pushkarna M, Schrouff J, Amironesei R, Moorosi N, Heller K. Healthsheet: Development of a Transparency Artifact for Health Datasets. 2022 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Seoul Republic of Korea: ACM; 2022 [cited 2024 Jun 26]. p. 1943–1961. Available from: https://dl.acm.org/doi/10.1145/3531146.3533239
97. Health Level Seven. HL7CodeSystem: Confidentiality Version: 3.0.0 [Internet]. Health Level Seven International; 2023. Available from: http://terminology.hl7.org/CodeSystem/v3-Confidentiality
98. Natioal Institutes of Health O of the D. Generalist Repository Ecosystem Initiative [Internet]. National Institutes of Health; 2023. Available from: https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative
99. RO-Crate Community. Research Object Crate (RO-Crate) [Internet]. University of Technology Sydney and The University of Manchester UK; 2023. Available from: https://www.researchobject.org/ro-crate/
100. Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM, Garijo D, Grüning B, La Rosa M, Leo S, Carragáin EÓ, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C. Packaging research artefacts with RO-Crate. Zenodo; 2021 Aug 13 [cited 2021 Aug 21]; Available from: https://zenodo.org/record/5146227
101. Carragáin EÓ, Goble C, Sefton P, Soiland-Reyes S. A lightweight approach to research object data packaging. Zenodo; 2019 Jun 20 [cited 2021 May 23]; Available from: https://zenodo.org/record/3250687
102. National Academies. Enhancing the Effectiveness of Team Science [Internet]. National Academies Press; 2015. Available from: https://www.nationalacademies.org/our-work/the-science-of-team-science