



# Clinical Dataset Structure: A Universal Standard for Structuring Clinical Research Data and Metadata

Bhavesh Patel, PhD<sup>1</sup>, Sanjay Soundarajan, MS<sup>1</sup>, Aydan Gasimova, MS<sup>1</sup>, Nayoon Gim, BS<sup>2,3,4</sup>,  
Jamie Shaffer, MS<sup>2,4</sup>, Aaron Lee, MD, MSCI<sup>2,4</sup> on behalf of the AI-READI Consortium

<sup>1</sup>FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, CA, USA, <sup>2</sup>Department of Ophthalmology, University of Washington, Seattle, WA, USA, <sup>3</sup>Department of Bioengineering, University of Washington, Seattle, WA, USA, <sup>4</sup>The Roger and Angie Karalis Johnson Retina Center, Seattle, WA

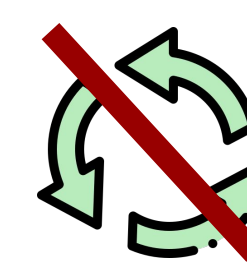
A0310

UW Medicine  
UW SCHOOL  
OF MEDICINE

## Background

During clinical research studies, different modalities of data are collected from study participants such as electrocardiogram (ECG) and eye images. There is currently no consensus on how to organize that data and include related information known as metadata. As a result, datasets from different studies are:

 **Not interoperable**, i.e. cannot be directly combined with other datasets, workflows, and tools

 **Not reusable**, i.e. cannot be easily used by researchers other than the original creators

This is preventing novel discoveries, especially through the use of artificial intelligence (AI) and machine learning (ML).

## Purpose

The purpose of this work is to develop a standard for organizing clinical research data and associated metadata.

## Methods

1. Review existing standards for organizing specific datatypes (e.g. Brain Imaging Data Structure or BIDS) and popular metadata schemas (e.g. DataCite and ClinicalTrials.gov)
2. Review AI/ML specific data documentation practices (e.g. datasheet and healthsheet)
3. Combine review findings and our understanding of clinical research data for establishing the standard
4. Use the dataset from the NIH Bridge2AI supported AI-READI project as a test case

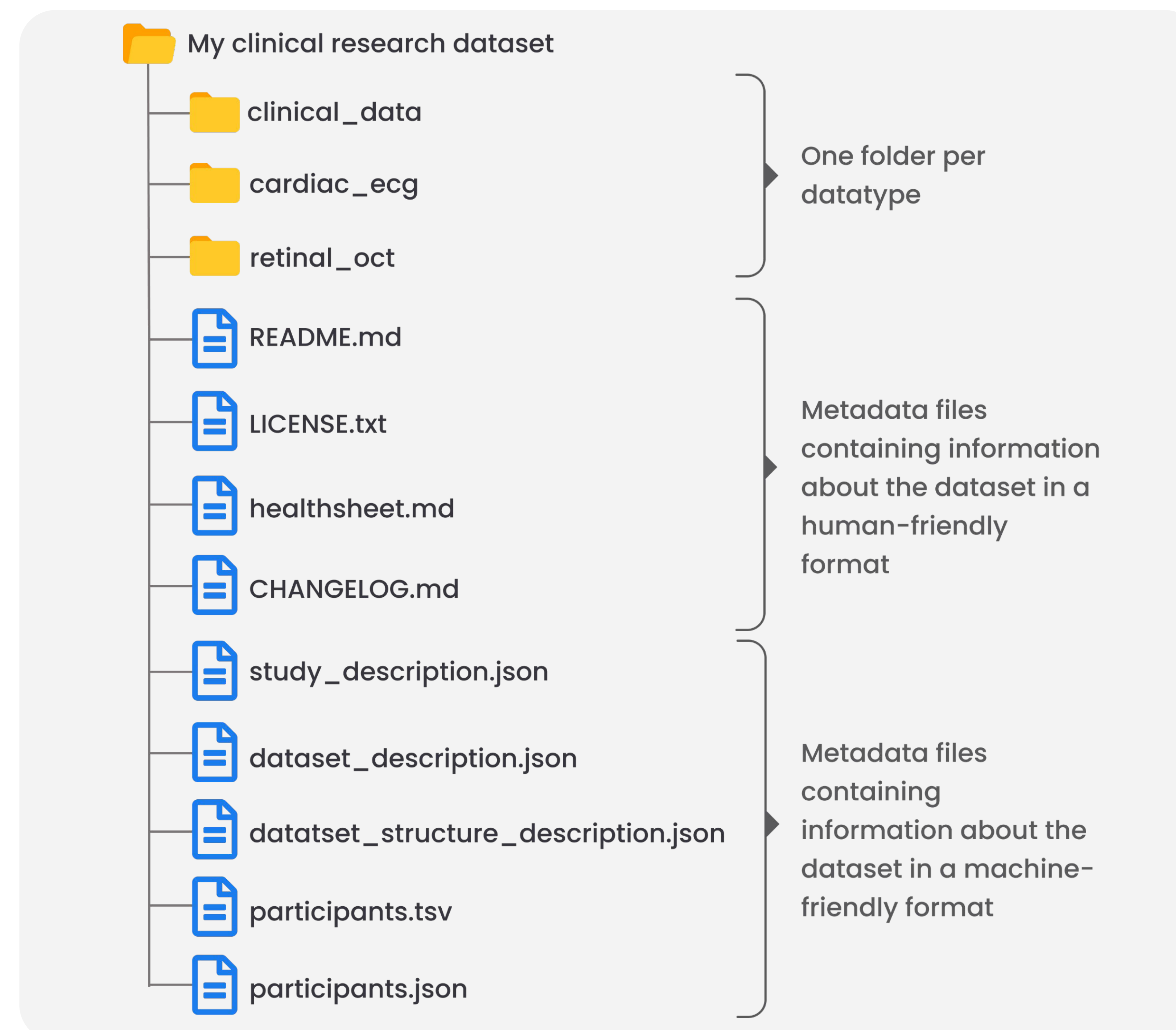


Figure 1. Illustration of a dataset organized according to the CDS.

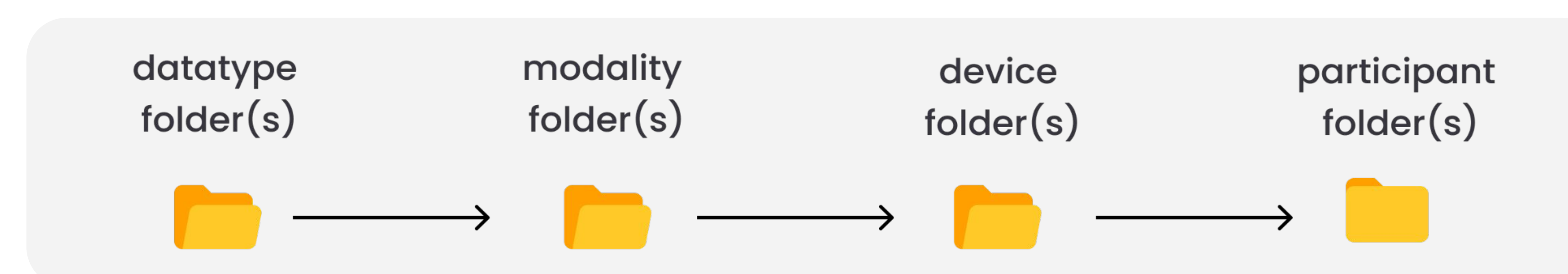
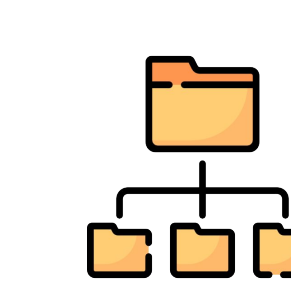


Figure 2. Illustration of the subfolder structure prescribed by the CDS. If no standard exists for structuring data pertaining to a datatype, the CDS specifies to organize data from that datatype into one folder per modality, then one per device, and finally one folder per participant.

## Results

We established the Clinical Dataset Structure (CDS) a standard for organizing clinical research data and metadata



The CDS instructs to organize data into one folder per datatype at the root level (Fig. 1). It also prescribes a specific structure within each datatype folder (Fig 2).



The CDS requires several metadata files to be included at the root-level (Fig. 1).



The CDS also provides specification for naming the different folders and files consistently.



The full specification is available at [cds-specification.readthedocs.io](https://cds-specification.readthedocs.io).



Evaluation of the AI-READI pilot dataset showed that the CDS made it easier to reuse. You can access that dataset at [fairhub.io/datasets/1](https://fairhub.io/datasets/1).

## Conclusion

The CDS provides a simple and intuitive way to organize clinical research data and metadata in line with the **FAIR (Findable, Accessible, Interoperable, Reusable) Principles**.

If everyone organizes their datasets according to the CDS, it will be easy to reuse each other's datasets and to combine datasets from different studies together, which can help increase the pace of innovations and discoveries.

We are looking for suggestions to improve the CDS! Please send your feedback and suggestions at the email below.

BRIDGE2AI

This work is funded by the NIH Common Fund's Bridge2AI program (1OT2OD032644-01)

Bhavesh Patel (bpatel@calmi2.org)  
aireadi.org  
fairdataihub.org

Find this poster  
and all related  
references here



[tinyurl.com/cdsARVO2024](https://tinyurl.com/cdsARVO2024)