The role of Data Standardization in AI-driven Clinical Data Research

Angelica Prado, Novo Nordisk
Adrian Czaban, Novo Nordisk
15-May-2025

The role of Data Standardization in AI-driven Clinical Data Research

Angelica Prado, Standards Developer, Data Standards, Novo Nordisk
Adrian Czaban, Principal Data Scientist, Biostatistics omics, Novo Nordisk

Meet the Speakers

Angelica Prado

Title: Standards Developer

Organization: Novo Nordisk

Computational Biologist with years of experience in both laboratory and computational research within the fields of Systems and Computational Biology. Her academic work integrated hands-on experimental techniques with the acquisition, visualization, and analysis of omics data, bridging wet lab and bioinformatics approaches. She has a strong interest in big data, data standardization, data visualization, and automation.

Adrian Czaban

Title: Principal Clinical Data Scientist

Organization: Novo Nordisk

Adrian Czaban is a Clinical Data Scientist with close to 10 years of experience. He has been involved in multiple submission projects in his career and is very interested in different new data types. His curiosity has led him to work with Omics data, and he is currently leading a Phuse Omics Project focused on technical aspects of handling those data in a Clinical Study setting.

Meet the Contributors

Ashwin Vishwanathan

Title: Lead Data Scientist

Organization: Novo Nordisk

Technical lead for projects, taking it from an ideation phase to production and regulatory interactions. Also involved in projects that look at leveraging AI/ML tools in the clinical development process.

Vivek Das

Title: Lead Data Scientist

Organization: Novo Nordisk

Leads a team of Clinician Data Scientists and Computational Biologist. Has been involved in design, build, deploy, & implement integrative multi-omics strategies, in-silico analytical frameworks & interpretability in Late-Stage Clinical Trial (Cardio-Renal and neuronal) settings for treatment efficacy and biomarker discovery.

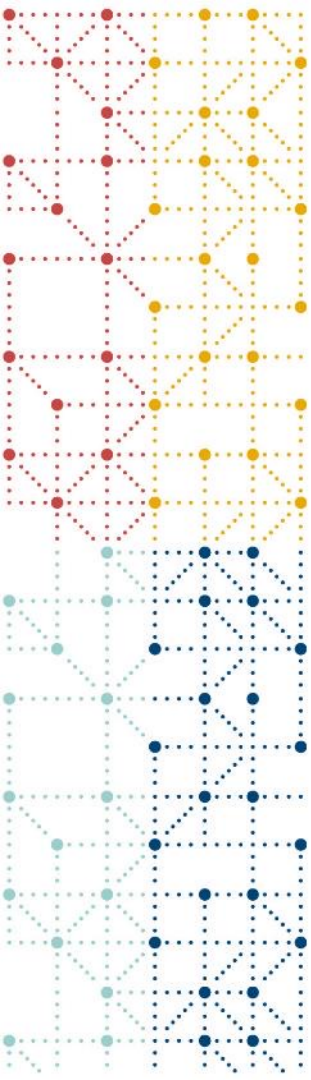
Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

- *The author(s) have no real or apparent conflicts of interest to report.*

Agenda

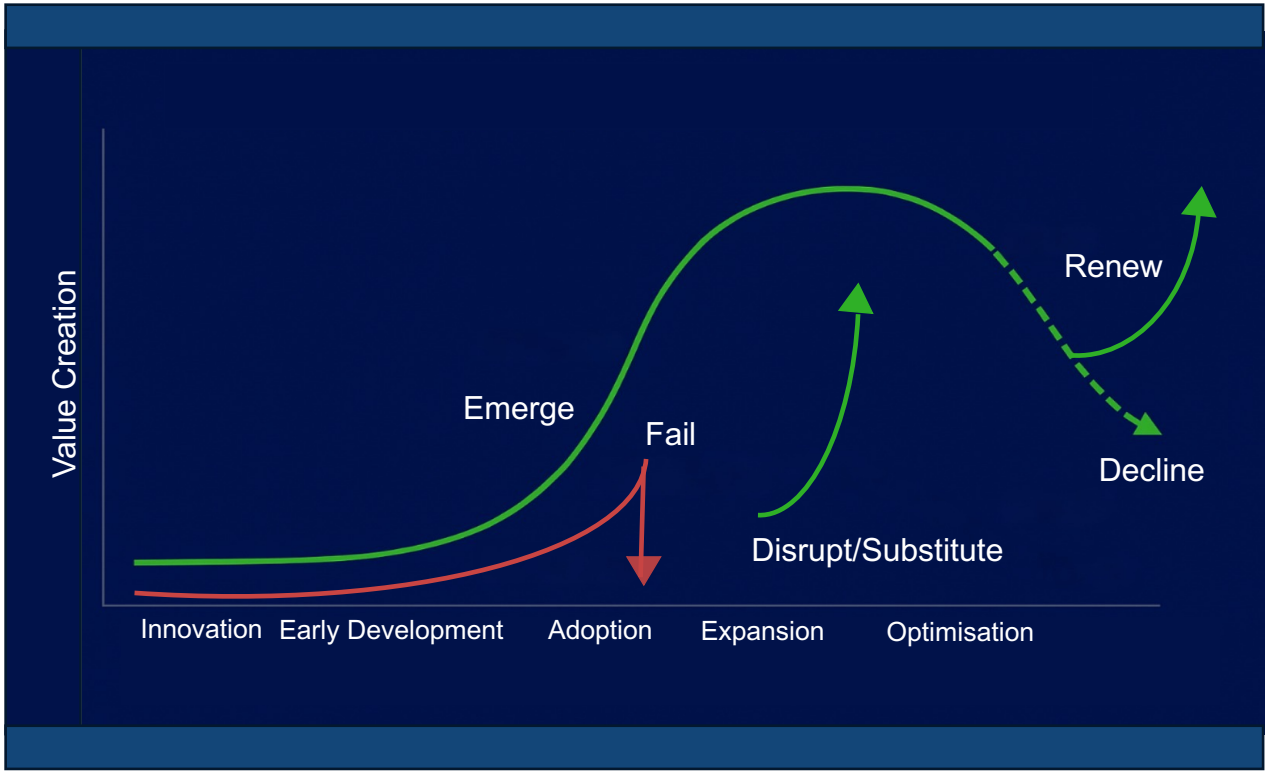
1. Introduction
2. Data standardization and AI today
3. Key challenges in data standardization for AI
4. Gaps in data standards
5. Future directions
6. Conclusions
7. Q&A



Introduction

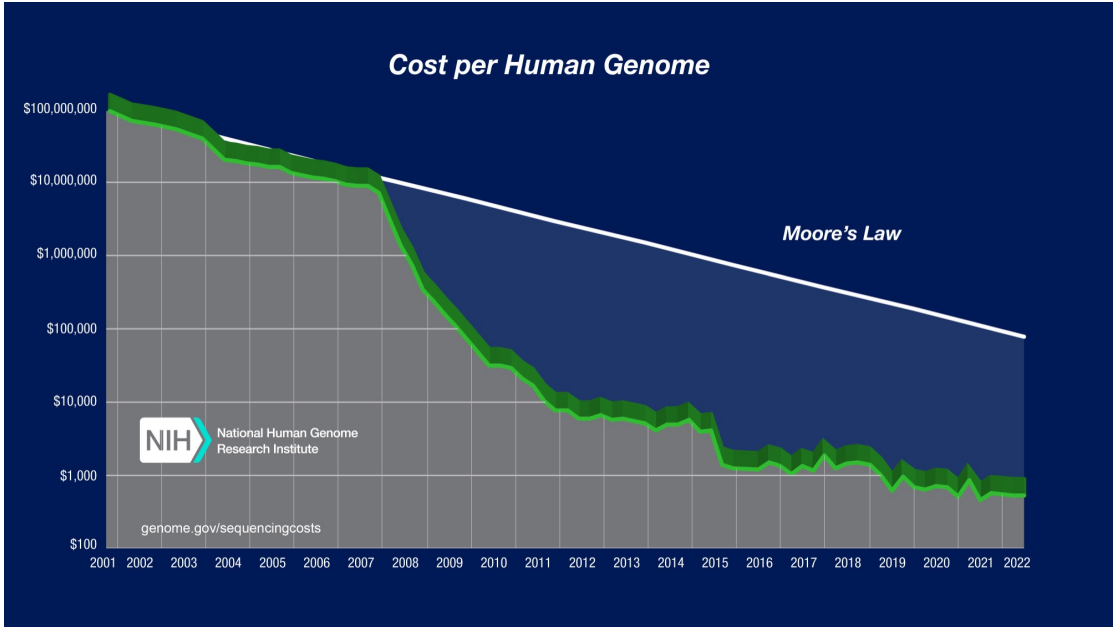
- Emerging technologies
- Repurposing data for AI and analytics

Emerging Technologies



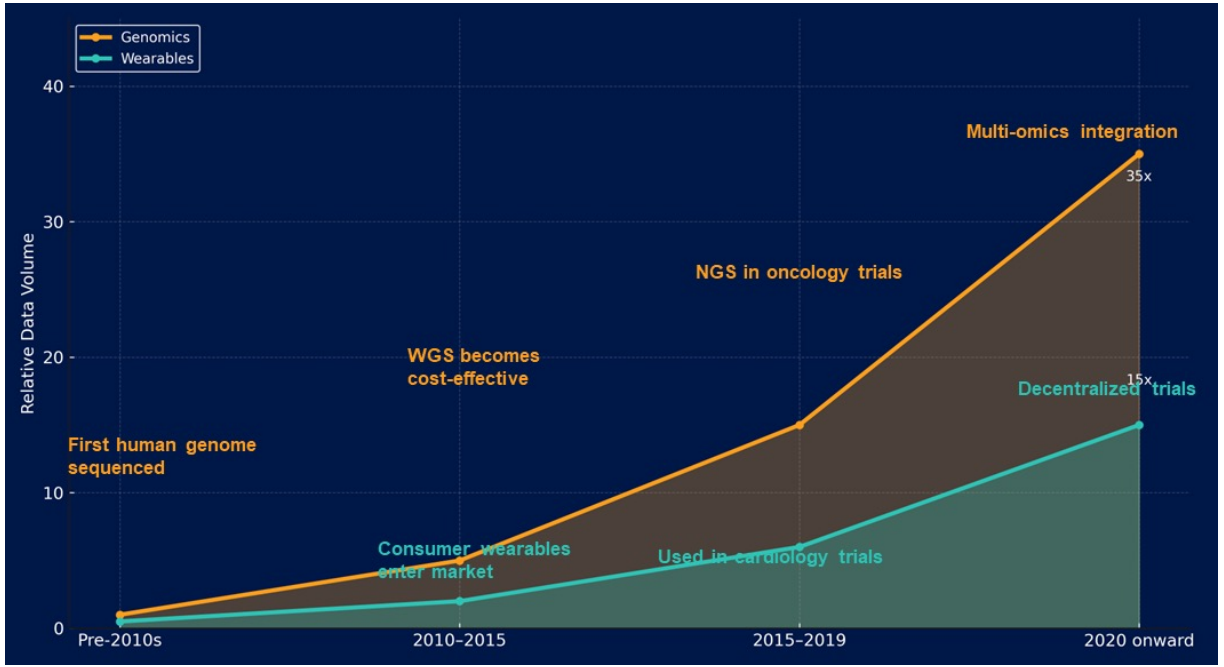
- Wearable Biosensors
- Omics
- Liquid Biopsies
- Immunotherapy Biomarkers
- CRISPR-Cas9
- AI assisted analysis

Expansion: Cost

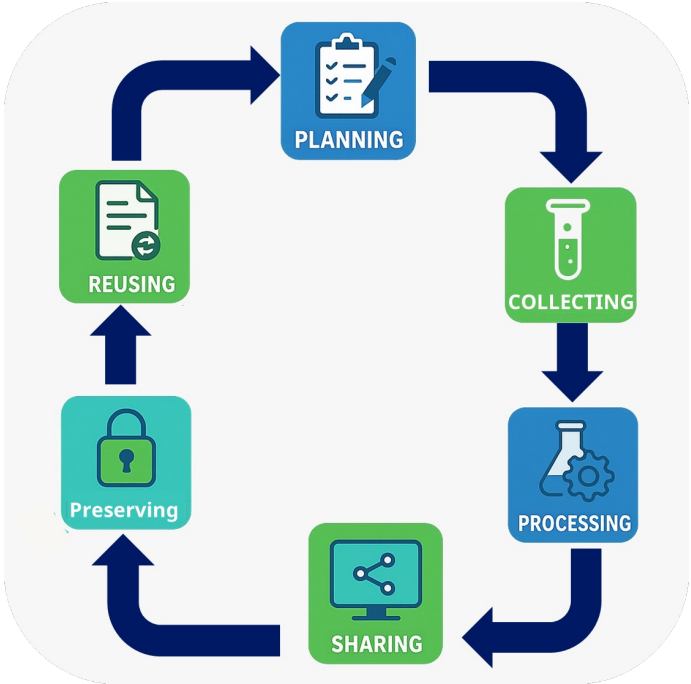


Source: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (Wetterstand, 2023)

Expansion: Volume



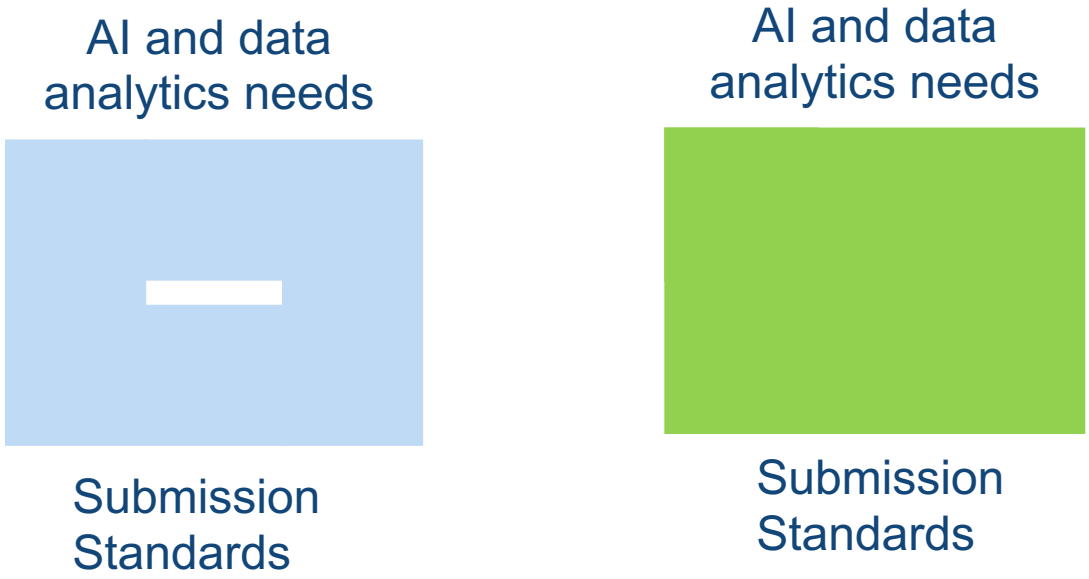
Data Repurposing for AI

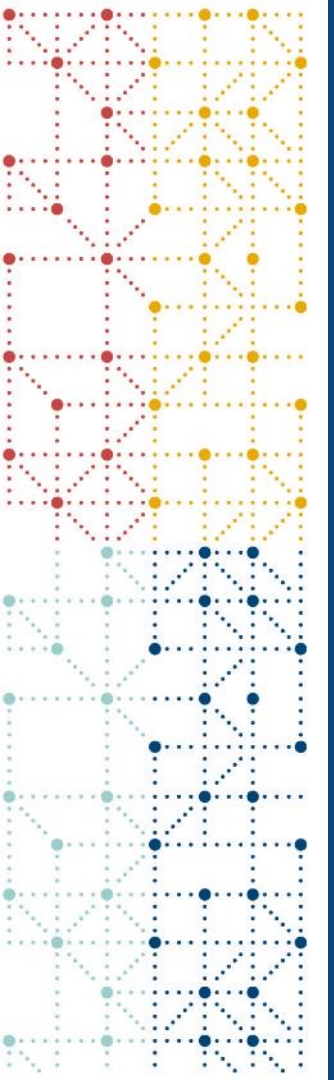


Questions

- How to streamline generation of SDTM data standards for emerging data sources?
- Are data standards only needed for submission?
- What sort of information should be covered by those standards when it comes to multidimensional data?
- Can we generate data standards for unstructured data?

How can we effectively generate data standards for new data sources?

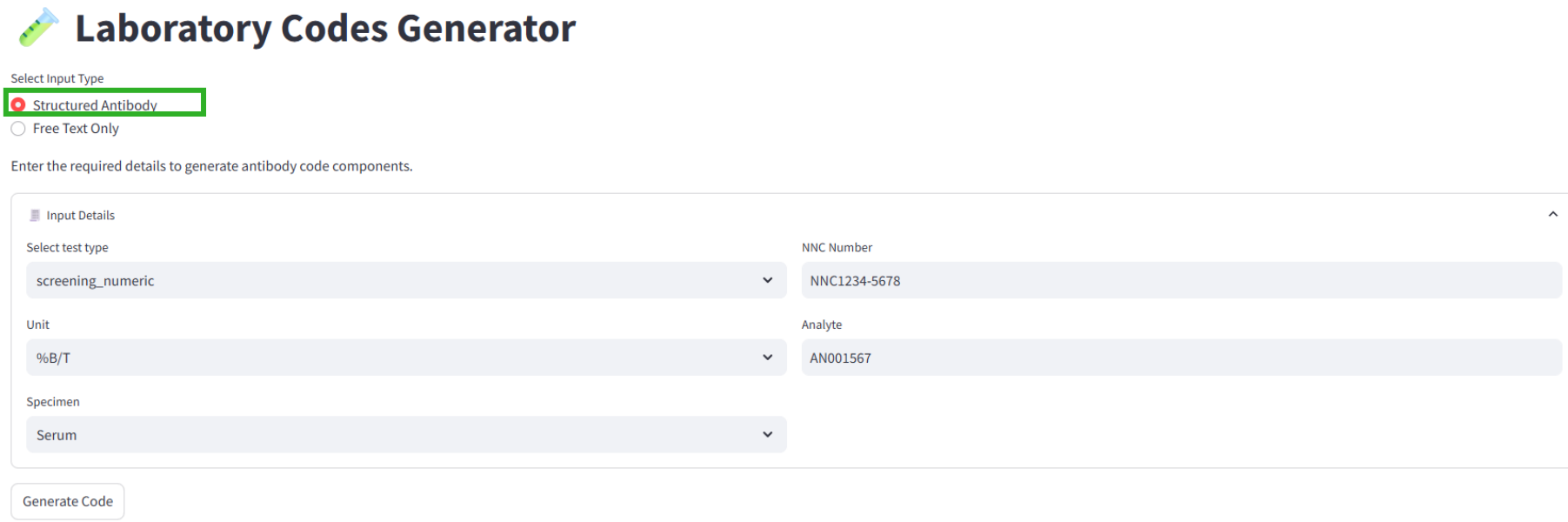




Data Standardization and AI today

- Shift on how we standardize clinical data
- Emerging regulatory requirements

- **Shift on how we standardize clinical data**



• Shift on how we standardize clinical data

Laboratory Codes Generator

Select Input Type

- (•) Structured Antibody
- ( ) Free Text Only

Enter the required details to generate antibody code components.

Input Details

Select test type: screening_numeric

NNC Number: NNC1234-5678

Unit: %B/T

Analyte: AN001567

Specimen: Serum

Generate Code

Metadata attributes

	Topic Code	CT Decode/Activity	CT Code Value	TC Label/Instance	Unit	Response List	TC Short Code	Description
0	ADA_SCR_1234_5678_SERUM	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 Antibody Screening Serum	%B/T		A1567SS	Screening of the binding anti-NNC1234-5678 antibody response in serum. [__bdagnt: NNC1
1	ADA_TIT_1234_5678_BLOOD	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 Antibody Titer Blood	no unit		A1567TB	A measurement of strength of the anti-NNC1234-5678 binding antibody response in blood. [
2	ADA_IGM_1234_5678	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 IgM Antibody	%B/T		A1567M	A measurement of the binding anti-NNC1234-5678 IgM antibody in a biological specimen. [_

Shift on how we standardize clinical data

Laboratory Codes Generator

Select Input Type

- ○ Structured Antibody
- ◉ Free Text Only

Enter the required details to generate antibody code components.

Input Details

Provide free text to auto-generate code components.

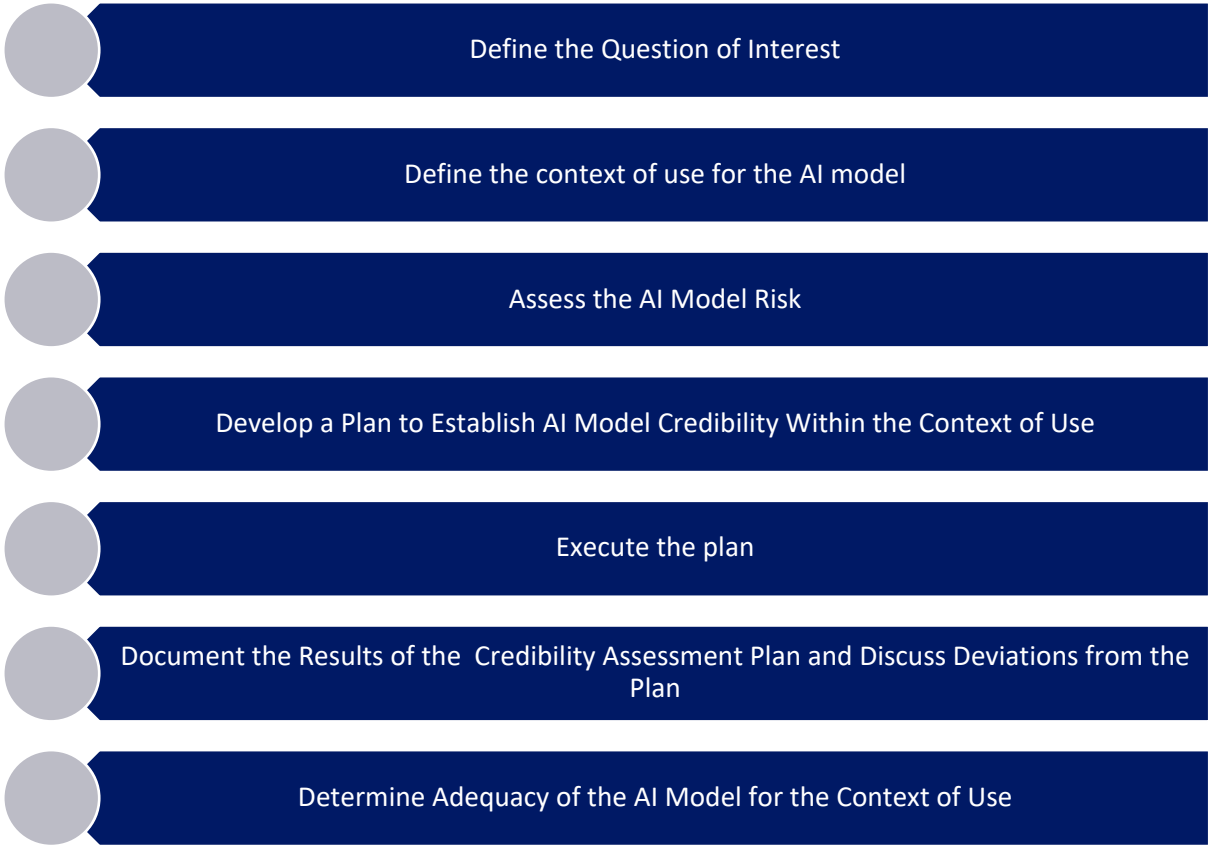
Free Text Note for Description

Relevant information about the assessment

Generate Code

	Topic Code	CT Decode/Activity	CT Code Value	TC Label/Instance	Unit	Response List	TC Short Code	Description
0	ADA_SCR_1234_5678_SERUM	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 Antibody Screening Serum	%B/T		A1567SS	Screening of the binding anti-NNC1234-5678 antibody response in serum. [__bdagnt: NNC1
1	ADA_TIT_1234_5678_BLOOD	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 Antibody Titer Blood	no unit		A1567TB	A measurement of strength of the anti-NNC1234-5678 binding antibody response in blood. [
2	ADA_IGM_1234_5678	Binding Antidrug Antibody	ADA_BAB	Anti-NNC1234-5678 IgM Antibody	%B/T		A1567M	A measurement of the binding anti-NNC1234-5678 IgM antibody in a biological specimen. [_

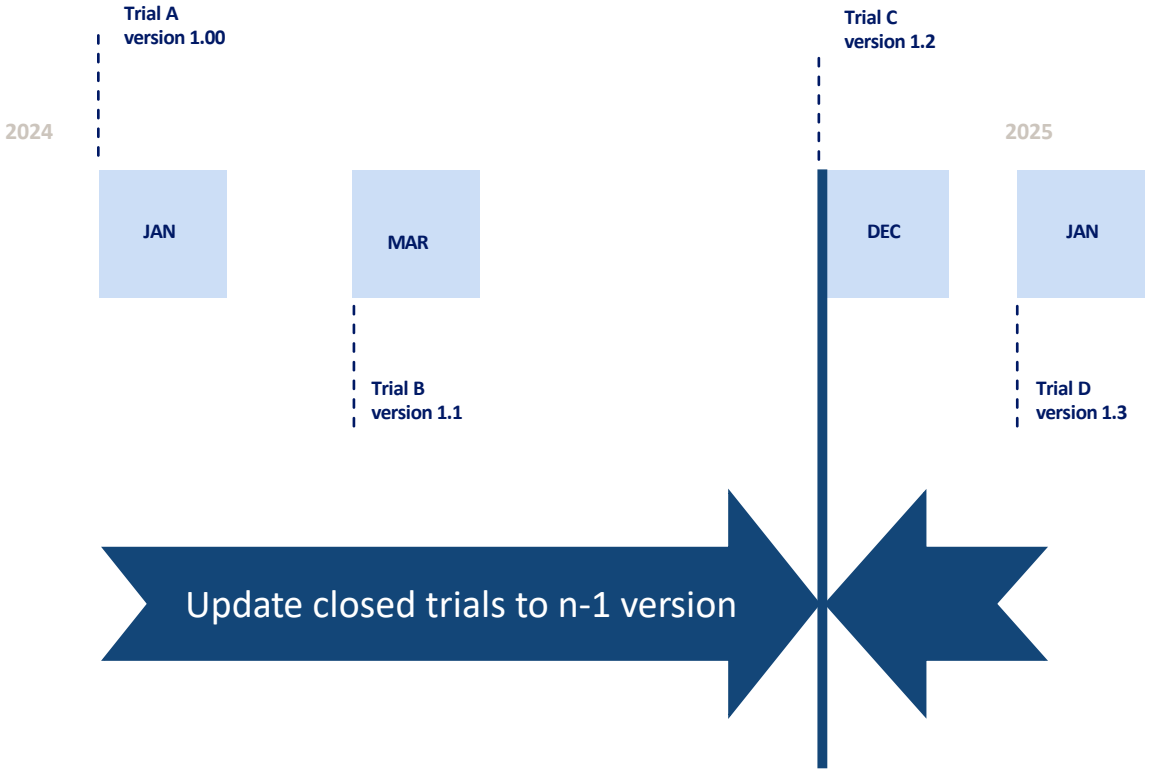
Emerging regulatory requirements for AI


Define the Question of Interest
Define the context of use for the AI model
Assess the AI Model Risk
Develop a Plan to Establish AI Model Credibility Within the Context of Use
Execute the plan
Document the Results of the Credibility Assessment Plan and Discuss Deviations from the Plan
Determine Adequacy of the AI Model for the Context of Use

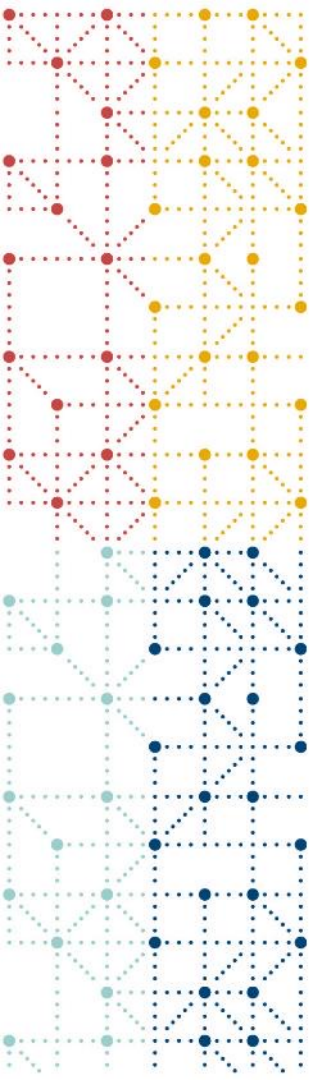
In practice it means that:

- We need to make sure the data we're using is fit for purpose.
- We need to choose when to update our models.
- We need to re-train older models if we are to compare them with new ones.
- We need to keep evaluating if what we have is still up for the job.

Illustrative example



- **AI model training:**
  - Retroactively transform to data standard version to (n-1) version.
  - Train models on n-1 version.
  - Lock model.
- **AI model inference:**
  - Transform data to n-1 version.
  - Feed to model.
- **AI model upgrade:**
  - Pick newer version of data standard.
  - Repeat training.

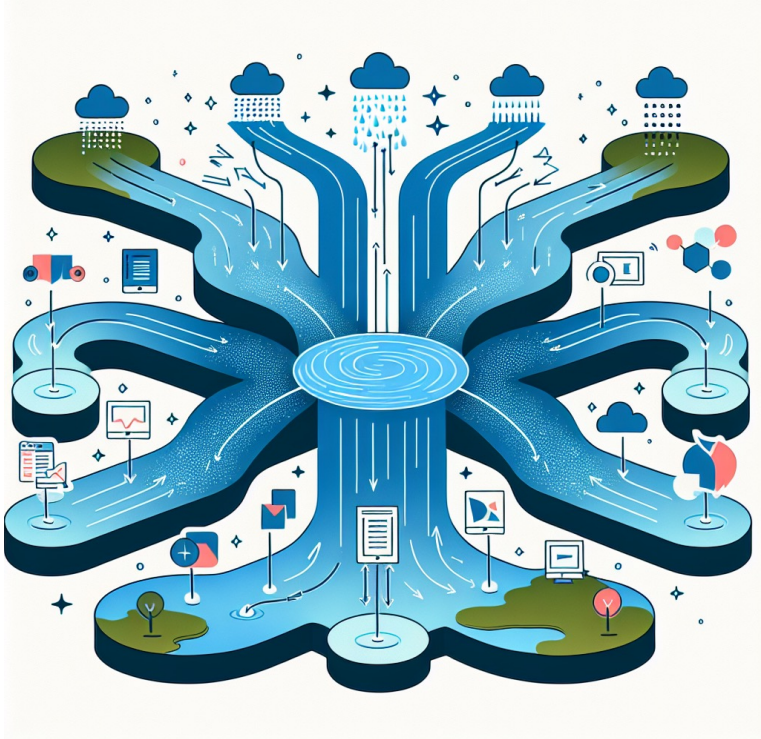


Key Challenges in data standardization for AI

- Integration of Data from multiple sources
- Keeping up with updates to Data Standards

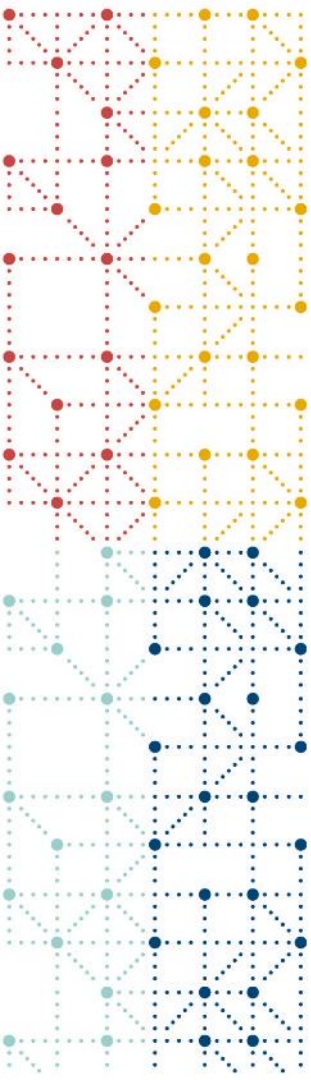
Integration of data from multiple sources

- Medical history
- Clinical studies
- Genetic databases
- Social media
- Registries
- Others



Keeping up with the updates

- MedDRA library is updated once every 6 months
- Other standards could have more dynamic/unscheduled update cycles
- I need a ‘dashboard’ of which standards are relevant for my use case and which have been updated



Gaps in SDTM data standards

- Multidimensional data
- Unstructured data

Multidimensional data sources

Wearable Biosensors

Liquid Biopsies

Immunotherapy Biomarkers

Omics

Digital Biomarkers

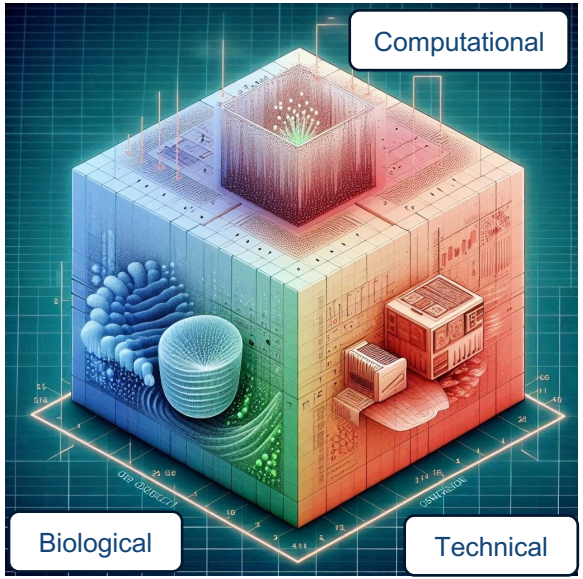
Images

AI assisted analysis

CRISPR-Cas9 Genome engineering

Multidimensional Data

- Interdependency
- Size



- Preprocessing
- Features selection
- Model
- Software versioning

- Inter-Individual differences
- Intra-Individual variation
- Tissue heterogeneity
- Cellular heterogeneity

- Method
- Sample Processing
- Assay Platform
- Operator

The Challenge

STUDYID	DOMAIN	USUBJID	LBTEST	LBTESTCD	LBSTRESN	LBSTRESC	LBORRES	LBNRIND	LBMETHOD	LBDTC
ST001	LB	ST001-001	Hemoglobin	HGB	14.2	g/dL	14.2	N	Automated	2023-06-01
ST001	LB	ST001-001	White Blood Cells	WBC	6.5	10^9/L	6.5	N	Automated	2023-06-01
ST001	LB	ST001-002	Hemoglobin	HGB	12.8	g/dL	12.8	N	Automated	2023-06-02
ST001	LB	ST001-002	White Blood Cells	WBC	7.0	10^9/L	7.0	N	Automated	2023-06-02
ST001	LB	ST001-003	Hemoglobin	HGB	15.0	g/dL	15.0	N	Automated	2023-06-03
ST001	LB	ST001-003	White Blood Cells	WBC	5.8	10^9/L	5.8	N	Automated	2023-06-03

The Challenge

Preprocessing

Calibration

Transformations

Data provenance

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

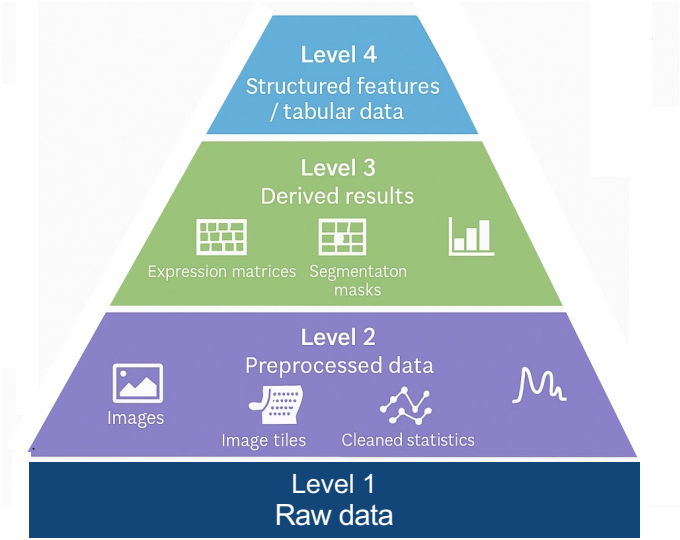
- Parallel systems:
  - Submission
  - Analytics

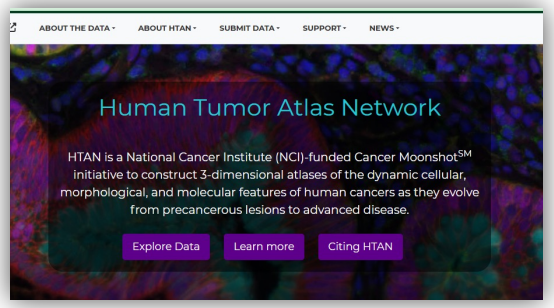
Examples of multidimensional Data Standards

Human Tumor Atlas Network (HTAN):

- Standards
- Infrastructure
- Community engagement

Human Tumor Atlas Network (HTAN) through standards, infrastructure and community engagement





- Lower Levels: Raw Data
- Higher Levels: Data processed by bioinformatics pipelines

Based on : https://www.nature.com/articles/s41592-025-02643-0?fromPaywallRec=false

What is the best approach for new multidimensional data sources?

SDTM framework, other standards and computational workflows

Current solution: BioCompute.

BioCompute: A platform for bioinformatics analysis workflow documentation

BioCompute is shorthand for the IEEE 2791-2020 standard for Bioinformatics Computational Analyses to facilitate communication between researchers, federal agencies and industires. This pipeline documentation approach has been adopted by a three FDA centers (CBER, CDER and CFSAN). Through this web portal, users can create BioCompute Objects (BCOs) in JSON format using the Builder and search for existing BCOs.

Data not governed by standards

Clinical notes (PDFs, Medical notes)

- Extracting data from PDFs using NLP models and the challenges of varying data quality.
- Need for machine-readable PDFs and the efforts within Novo to standardize data templates.

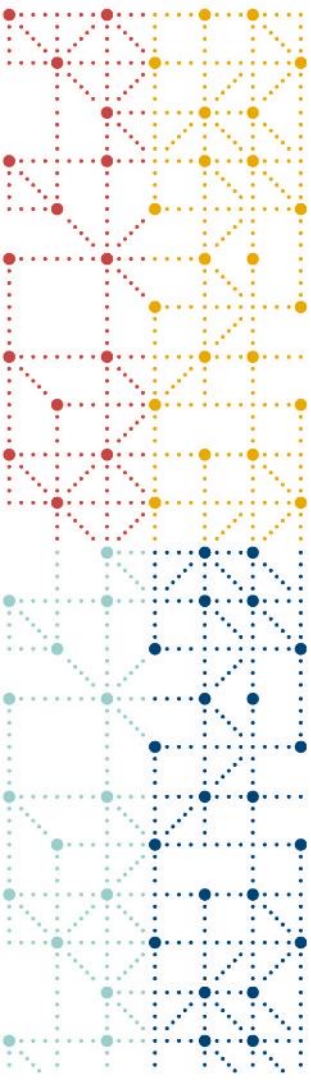
Future Directions

- Cross collaboration having in consideration emerging technologies between industry and academy consortiums.

- Rethink how new data modalities align with a tabular data standard
- Use of new technologies that could facilitate data standardisation (LLM).

- We should align into a minimal solution or expectation about metadata for new data sources.

- Integration of SDTM data standards in data analytics frameworks considering recent FDA requirements.



Q&A

Thank You!