National Institute of Diabetes and Digestive and Kidney Diseases

Preparing Data Dictionaries for AI-Research

Matthew Dancis, Booz Allen Hamilton

January 12, 2024

Announcements

- The deadline to submit your Challenge solution has been extended to **Monday, January 22, 2024, at 11:59 PM (ET).**
  - This will be the only extension and late submissions will not be accepted after this date.
  - We encourage everyone to submit in advance of this deadline to ensure we can provide any technical support as needed during submission.
- Instructions on how to submit solutions are posted on Challenge.gov under the **How to Enter** tab
- Challenge Solution Submission Form was also recently updated. Please download and complete the latest V2 of this form from the **Resources** tab on Challenge.gov

Speaker Introduction

Matthew Dancis is a Workforce Data Analyst at Booz Allen Hamilton, where he works with federal agencies to develop and deploy data-driven solutions that inform human capital strategy.

He previously supported an NIH Office of Data Science Strategy (ODSS) initiative to identify core competencies that biomedical researchers would need to prepare their data to be FAIR and AI-ready.

Prior to joining Booz Allen, Matthew worked in a knowledge management capacity, where he developed a managed metadata taxonomy for a USAID implementing partner's file system.

He holds a Master of Science in International Development Management with a concentration in data analytics.

Data Centric Challenge – Submission Requirements

1. **A single "raw" dataset** A single "Raw" dataset resulting from data aggregation and harmonization of all study data files (from TEDDY or TrialNet – as per level of participation), but has not otherwise been altered. This dataset must be represented as a single rectangular file (i.e., tabular, spreadsheet, or matrix) in .csv file format available within your Workspace.

2. **An "AI-ready" version** of the raw dataset that has been enhanced for AI-readiness. This dataset must be represented as a single rectangular file (i.e., tabular, spreadsheet, or matrix) in .csv file format available within your Workspace

3. **The code script,** in Python or R, used to generate the raw and AI-ready files submitted to the private GitHub repository created for your team.

4. **A human-readable data dictionary/codebook** documenting the AI-ready dataset (Excel format preferred, with the following information included at a minimum: variable name, variable label/description, variable type, measurement unit as applicable (e.g., pounds, kilograms), and corresponding code lists as needed (e.g., 0 = No, 1 = Yes) and submitted to Challenge.gov as an attachment.

5. **Challenge Solution Submission Form**, describing the 1) AI-ready dataset, 2) methods for preparing the AI-ready dataset, and 3) potential use cases for the prepared dataset as it relates to T1D, or other disease areas of interest to NIDDK, and submitted to Challenge.gov as an attachment.

Please follow the **Submission Instructions** posted to Challenge.gov for Phase 2: Data Enhancement

Frequently Asked Questions are also posted to Challenge.gov under the **FAQs** tab

National Institute of Diabetes and Digestive and Kidney Diseases

Agenda

- Characteristics of a human- and machine-readable data dictionary
- Importance of data dictionaries for AI-readiness
- Demonstration of preparing a data dictionary using R

What is a data dictionary?

What is a data dictionary?

A data dictionary is a document that **outlines the structure, content, and meaning of a given variable**...a data dictionary is a CSV file containing information on the variables and the structure of [a] REDCap database (NNLM).

A data dictionary is a set of information **describing the contents, format, and structure of a database** (IBM).

A data dictionary is a **metadata repository** (Wikipedia)

A codebook is a **human readable document that provides information on each data element**. (NNLM).

What is a data dictionary?

Data dictionaries typically take one of two forms: active or passive

Active data dictionaries

- Live on digitized and controlled data entry platforms
- Have data entry constraints
- Update automatically when changes are made
- Useful for mitigating human error and ensuring data integrity
- Limiting for novel research questions
- Example: Protein Data Bank

Passive data dictionaries:

- Live on Excel Workbooks or CSV files
- Are typically ad hoc in format and information
- Are stand-alone reference documents
- Useful for contextualizing data and data handling decisions for novel research questions that require disparate datasets
- Example: Rankin, et al. study on renal failure within 90 days of dialysis *(previous office hours presentation)*

Characteristics of a human-readable data dictionary/codebook

There is no cross-domain/cross-industry standard for what information should go into a data dictionary

Minimum viable product:

- Variable names
- Variable labels/descriptions
- Variable types
- Measurement units as applicable
- Corresponding code lists (e.g., 0 = No, 1 = Yes)
- Missingness representation
- A README tab to orient the reader

Good practices:

- Differentiation of raw and manipulated data
- Data that was not used

To strive for:

- Dataset documentation, i.e., a thorough write up about the data in a separate tab that describes the data's provenance (i.e., lineage and context)
- Summary statistics
- FAIR and CARE oriented data

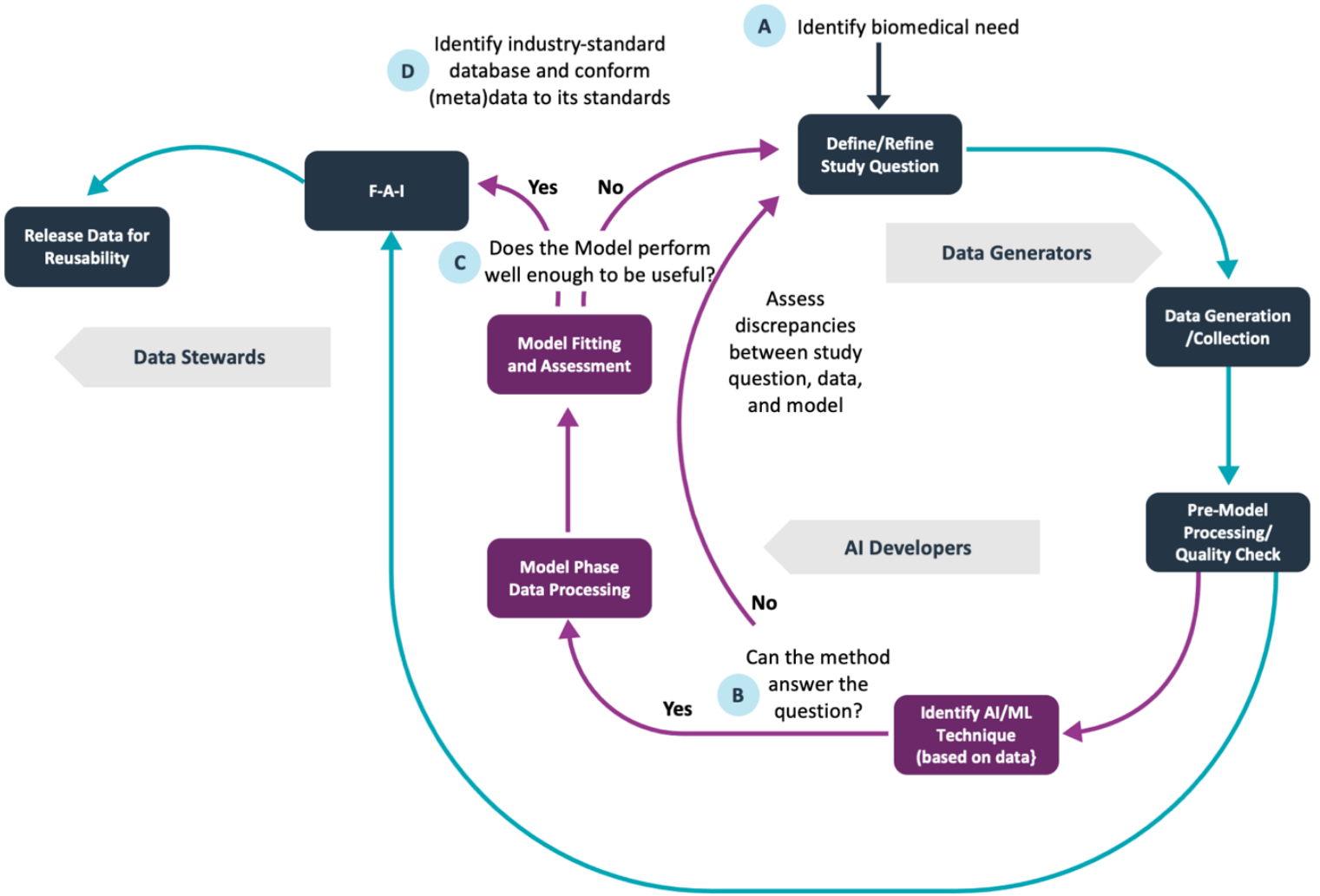
USRDS Dataset	Category	Feature Description	Variable Name	Variable Type	Operational Variable?
Patients	Demographics	Age	inc_age	Measure (Years)	No
Patients	Demographics	Race	race	Factor with 7 levels: 1=White, 2=Black/African American, 3=American Indian or Alaska Native, 4=Asian, 5=Native Hawaiian or Pacific Islander, 6=Other or Multiracial, 9=Unknown	No
Patients	Demographics	Sex	sex	Factor with 3 levels: 1=M, 2=F, 3=Unknown	No
Patients	Demographics	Ethnicity	ethn	Factor with 5 levels: 1=Hispanic-Mexican, 2=Hispanic Other, 3=Non-Hispanic, 5=Hispanic Non-Specified, 9=Unknown	No
Medical Evidence	Clinical Variables	BMI	bmi	Measure (kg/m^2)	No
Medical Evidence	Clinical Variables	Weight	weight	Measure (kg)	No
Medical Evidence	Clinical Variables	Height	height	Measure (cm)	No
Medical Evidence	Clinical Variables	Albumin	album	Measure (g/dl)	No
Medical Evidence	Clinical Variables	Serum Creatinine	sercr	Measure (mg/dl)	No
Medical Evidence	Clinical Variables	Hemoglobin	heglb	Measure (g/dl)	No
Medical Evidence	Clinical Variables	Estimated glomerular filtration rate (GFR)	gfr_epi	Measure (mL/min)	No
Medical Evidence	Comorbidities	Congestive heart	como_chf	Factor with 3 levels:	No

ReadMe | 1 - Features Direct from USRDS | 2 - Constructed Features | 2a - PDIS Recode | 3 - Non-Feature Variables

ONC Training Data for ML Implementation Guide, Rankin, et al.

Importance of data dictionaries for AI-readiness

- **AI-readiness inherently presumes a "hand-off"** of data from a "data generator" to an AI developer who can merge the data into larger relational datasets they can use to train models
- Dictionaries help facilitate seamless transfer of data from generators to developers to stewards



How Great Data Dictionaries Can Address/Reduce Bias in AI

POTENTIAL CHALLENGES	RECOMMENDED SOLUTIONS
Data diversity due to limited population representation	• Assess the limitations • Identify the strategy for mitigating a lack of diversity as part of the research design
Overreliance on machine learning solutions	• Ensure interdisciplinary approach and continuous human involvement • Conduct follow-up studies to ensure results are meaningful
Algorithms based on biased data	• Identify the target population and select training and testing sets accordingly • Build and test algorithms in socioeconomically diverse health care systems • Ensure that key variables that are related to race, gender, etc. are being captured and included in algorithms where appropriate • Test algorithms for potential discriminatory behavior throughout processing • Develop feedback loops to monitor and verify output and validity
Non-clinically meaningful algorithms	• Focus on clinically important improvements in relevant outcomes rather than strict performance measures • Impose human values in algorithms at the cost of efficiency

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#!po=15.6250

Importance of data dictionaries for AI-readiness

What are the FAIR and CARE Principles?

The Role of Metadata in Adherence to FAIR and CARE Principles

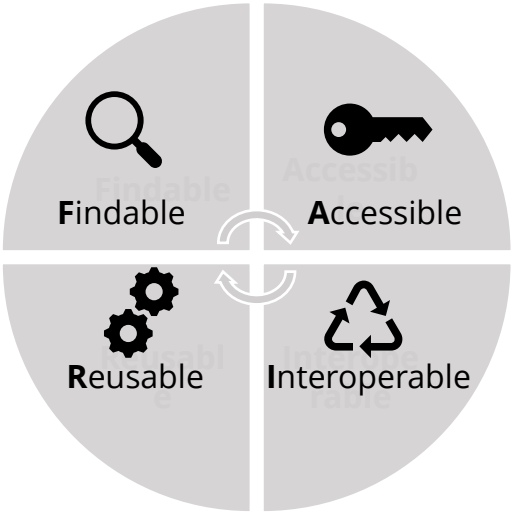
AI data scientists rely on metadata for:

- **Findability:** Metadata enable **search engines** to crawl relational databases to easily find data that are relevant to developers' research questions. The more robust the taxonomy of metadata that describes a dataset, the more findable it will be. *An example of metadata servicing findability is the Digital Object Identifier (DOI) used to find journal articles*

- **Accessibility:** Metadata are what make data retrievable when a user searches for it. Metadata can also serve as an **authentication or authorization procedure** for protected datasets. Metadata in data dictionaries can outline well-defined **license/access conditions**

- **Interoperability:** FAIR datasets should have metadata that adhere to at least one standardized metadata **vocabulary** to ensure that secondary data users can easily combine the data with other datasets through common identifiers

- **Reusability:** Metadata within data dictionaries should contain thorough and accurate descriptions of the data's attributes to ensure reusability



The Role of Metadata in Adherence to FAIR and CARE Principles

AI data scientists rely on metadata for:

- **Collective Benefit:** Metadata on marginalized communities are "designed to support Indigenous nations. Use/reuse of data for resource allocation is consistent with community values"
- **Authority to Control:** Metadata on marginalized communities are designed to preserve the indigenous communities' right to control data about them
- **Ethics:** Data documentation in data dictionaries is designed to "minimize harm, maximize benefits, promote justice, and allow for future use."
- **Responsibility:** Data generators who generate/collect data about marginalized communities assume the responsibility to engage with the communities about the data collection/generation process

Useful Resources

A FAIR and AI-ready Higgs boson decay dataset (Chen, Y., Huerta, E.A., Duarte, J. *et al.* A FAIR and AI-ready Higgs boson decay dataset. *Sci Data* **9**, 31 (2022). https://doi.org/10.1038/s41597-021-01109-0)

dataMeta: Making and appending a data dictionary to an R dataset (Dania M. Rodriguez, Michael A Johansson, Luis Mier-y-Teran-Romero, moiradillon2, eyq9, YoJimboDurant, … Daniel Mietchen. (2017). cdcepi/zika: March 31, 2017 [Data set]. Zenodo. http://doi.org/10.5281/zenodo.439543)

Operationalizing the CARE and FAIR Principles for Indigenous data futures (Carroll, S.R., Herczog, E., Hudson, M. *et al.* Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* **8**, 108 (2021). https://doi.org/10.1038/s41597-021-00892-0)

GitHub with Zika dataset

DEMO

Questions

Appendix

What is a data dictionary?

Data Dictionary Example

ONC Training Data for ML Implementation Guide, Rankin, et al.

USRDS Dataset	Category	Feature Description	Variable Name	Variable Type	Operational Variable?
Patients	Demographics	Age	inc_age	Measure (Years)	No
Patients	Demographics	Race	race	Factor with 7 levels: 1=White, 2=Black/African American, 3=American Indian or Alaska Native, 4=Asian, 5=Native Hawaiian or Pacific Islander, 6=Other or Multiracial, 9=Unknown	No
Patients	Demographics	Sex	sex	Factor with 3 levels: 1=M, 2=F, 3=Unknown	No
Patients	Demographics	Ethnicity	ethn	Factor with 5 levels: 1=Hispanic-Mexican, 2=Hispanic Other, 3=Non-Hispanic, 5=Hispanic Non-Specified, 9=Unknown	No
Medical Evidence	Clinical Variables	BMI	bmi	Measure (kg/m^2)	No
Medical Evidence	Clinical Variables	Weight	weight	Measure (kg)	No
Medical Evidence	Clinical Variables	Height	height	Measure (cm)	No
Medical Evidence	Clinical Variables	Albumin	album	Measure (g/dl)	No
Medical Evidence	Clinical Variables	Serum Creatinine	sercr	Measure (mg/dl)	No
Medical Evidence	Clinical Variables	Hemoglobin	heglb	Measure (g/dl)	No
Medical Evidence	Clinical Variables	Estimated glomerular filtration rate (GFR)	gfr_epi	Measure (mL/min)	No
Medical Evidence	Comorbidities	Congestive heart	como_chf	Factor with 3 levels:	No

ReadMe | 1 - Features Direct from USRDS | 2 - Constructed Features | 2a - PDIS Recode | 3 - Non-Feature Variables

Codebook Example

CDC Sample Child Interview, 2021

2020 NATIONAL HEALTH INTERVIEW SURVEY (NHIS)
Codebook for Sample Child file (Document Version: 16 September 2021)
PUBLIC USE

Variable:	RECTYPE
Module:	General
Section:	IDN : Identifier
File(s):	Adult, Adultinc, Child, Childinc, Paradata, Adultlong, Adultpart
Data Type:	Numeric
Length:	2
Question Text:	
Description:	Record type
Recode:	
Universe:	HHX ne ' '
Universe Description:	All households
Sources:	
Question ID:	
Keywords:	
Notes:	
Evaluation Report:	

Unweighted frequencies:

RECTYPE Record type

Code	Description	Frequency	Percent
10	Sample Adult	0	0.00
20	Sample Child	5790	100.00
30	Sample Adult Income	0	0.00
40	Sample Child Income	0	0.00
50	Paradata	0	0.00
60	Sample Adult Longitudinal	0	0.00
70	Sample Adult Partial	0	0.00

Frequency Missing: