Data Dictionary/Codebook

Best Practice

General Description

Category:

Data Management Planning

Purpose of this best practice:

A data dictionary consists of definitions of every data item (variable) that is being collected for a study. It is an essential part of successful data management and should be updated whenever a variable is modified, added, or deleted. Such changes include all aspects of the data field, including field name, data type, menu definition (when appropriate), changes to the way the question is asked, validation requirements, etc.

Categories

Best Practice
Data Management
Data Management Planning
Data Sharing

Recommendations and Considerations

① The data dictionary:

1.1. Serves as a communication tool within the study team as data collection requirements are being developed.

1.2. Documents, in full detail, a description of the data elements captured in the study, allowing new study team members and future users of the data to understand exactly what they are working with.

1.3. Serves as the foundation for accurate report generation.

1.4. Allows statisticians to accurately interpret the data for analysis.

1.5. Is a living document that should be maintained throughout the project life-cycle.

1.6. Serves as a primary reference tool when data are shared with external researchers.

② Field names should be unique and unambiguous. Field names should provide an indication as to the content of the field. The field names may not be

intuitive to "everyone", but should be helpful to the experts within that discipline who will have access to the data.

2.1. e.g. 'platelets_k_cu_mm'
(In this case abbreviations are included since they are commonly used and widely understood in many disciplines.)

2.1.1. Poor Field Names:
Age (age of what?)
Height (measured in?)
Today_Date (meaningless!)
Q1

Better Field Names:
Age_Dx (age at diagnosis... very specific)
Height_cm
PHQ_CompletionDate (date the PHQ form was completed)
PHQ_Q1 (not "great", but at least we know what form Q1 is associated with)

③ Avoid abbreviations whenever possible (e.g. 'sodium_serum', not 'na_serum').

④ Include units of measure in the variable name, if appropriate.

4.1. Examples:
height_cm
weight_kg

⑤ The data dictionary should describe what type of data can be stored in each variable

5.1. The titles and definitions of variable types are usually very similar across data management systems. Commonly used types include:
a. Date
b. Integer
c. Float - decimal
d. String - alphanumeric
e. Text
f. Select one option
g. Select all options that apply
h. Calculated

⑥ Provide a descriptive label / definition

6.1. May include the 'Question' or text that appears on a Case Report Form with the variable. It clearly instructs users what information should be entered in that variable.

⑦ Indicate Length and Format

7.1. Record how long the variable is - how many characters or numbers may be entered - how the data should be displayed and stored. Examples:
   a. Date - MM/DD/YYYY
   b. Decimal - 6 characters, ###.##
   c. String - 15 characters
   d. Option - select response from a dropdown menu

⑧ Validation Rules

8.1. The criteria a response must meet to be considered a valid response.
Examples:
Must be >10
Admission date must be between date A and date B

⑨ Branching logic rules, where appropriate.

9.1. The conditions under which data should not be collected (skipped) for this variable.
Example:
Rule: If subject is male, the pregnancy test result field should be disabled.
Suggestion: Include the code/logic (e.g. Sex=1) that indicates when a field is to be collected.

⑩ Version Information

10.1. Changes in variable attributes should documented over time and a version number/date changed should be recorded for each iteration. These updates should be 'versioned' (archived) as changes are made.

⑪ Variable Codes

11.1. Example: If responses are selected from a list of options, what code for each option should be stored in the database. Examples:

Option = 'Yes' Code = 1
Option = 'No' Code = 0

Resources and Examples

Useful Links:

DDI - Document, Discover and Interoperate

Topic Experts:

Johns Hopkins School of Public Health Biostatistics Center
ICTR
Oncology Clinical Research Office

Responsibilities

Position:	Responsibility:
Principal Investigator	The Principal Investigator (PI) should participate in development of the database. At a minimum, the PI should review the final data dictionary prior to going live.
Data Manager, Study Coordinator	The Data Manager and Study Coordinator are responsible for generating the data dictionary and updating the data dictionary as needed.
Data Analyst / Biostatistician	Because the Data Dictionary will be a primary reference during data analysis, the Data Analyst / Biostatistician should review the Data Dictionary to ensure variables are being created appropriately for analysis.