Building on the Rich Metadata from Decades of Health Behavior Studies: The Potential for Common Data Elements (CDEs) to Enhance the Identification of Health Data across Different Research Projects

PROJECT GOAL

Overall Strategy

Select datasets, use cases, and ontologies. Add ontology and NLM CDE terms to variable metadata. Test for improved search results. Analyze inter-rater reliability.

Our goal in this pilot project is to test an approach aimed at helping researchers easily find health-related variables among the five million variables in the ICPSR archives. We focused on two areas of vital research: drug addiction and dementia. Variables in NAHDAP and NACDA studies were searched and tagged with terms from the widely-used PROMIS ontology and National Library of Medicine CDE Repository. Our ultimate goal is to create a body of gold-standard hand-curated metadata to support machine learning models for automated metadata tagging in existing datasets.

Secondary analyses of key datasets multiply the benefits of our national investment in health science. At ICPSR we work hard to increase the use of legacy data. With funding from NIDA, NIA, and OBSSR, ICPSR is working to increase the use of extant data for health research by making health-related variables easier to identify. To pilot this work, we are adding variable-level metadata from subsets of controlled vocabularies (e.g., CDEs from the NIH CDE Repository and ontology terms from SNOMED, PROMIS, and PICO), focusing on opioid use and abuse in the National Addiction and HIV Data Archive Program (NADHAP) and dementia and cognitive function in the National Archive of Computerized Data on Aging (NACDA). The enhanced metadata allows the search to find individual variables where each question is narrowly focused (e.g., participants are asked about the use of specific types of opioids, but the term opioid was not used) and to reveal variables where search returns are overwhelming, with hundreds of studies containing potentially thousands of variables. This pilot has yielded unique insights into the strengths and limitations associated with applying CDEs and ontologies (and other controlled vocabularies) to existing studies to make data more discoverable and more usable.

PROJECT CHALLENGES

- **Difficulty finding variables:** Legacy collections without complete variable level metadata (i.e., missing question text, variable labels, and value labels) require visual searching through all variables and documentation in a dataset.
- **Information narrowness:** The amount of information that can or should be included in labels and question text should be more inclusive. If the survey instrument cannot be searched (e.g., poor quality scanned document), then the instrument needs to be manually consulted to properly evaluate a match.
- **Inconsistency in application of controlled vocabularies:** There are instances where multiple vocabulary terms exist for what appears to be the same survey question and somewhat different questions are grouped under the same vocabulary term. Likewise, the architecture of various vocabularies have not been reconciled.
- **Content knowledge determination:** Determining the level of content knowledge required for robust tagging is essential but difficult to establish.
- **Consistency and intra- and inter-rater reliability:** It is challenging to ensure a consistent approach within and across individuals, particularly in the more precise controlled vocabularies.

Figures 2 and 3 below show the information fields in the NLM CDE Repository for the Common Data Element (CDE) "Raw score for Copy Portion of Brief Visuospatial Memory Test — Revised (BVMT-R)." The comment boxes in the figure to the right note challenges for particular fields, and rules for matching to the CDE. One of the challenges highlighted with these comments is that multiple diseases may be identified with the CDE, but not exhaustively.

Inter-rater reliability (IRR)

Inter-rater reliability (IRR) for CDEs varied across studies (Table 1). CDE IRR was low for one study, but comparable to the other studies when we matched to PROMIS ontology terms. CDE IRR was also considerably higher between curators for the YAI study than between curators and our metadata expert. In-depth comparisons show that variable tagging can be done by trained data curators with supervision and extensive review of applied tags. Tagging would be more efficiently done by metadata experts with content knowledge.

NEXT STEP: Increase the size of the training dataset to develop machine learning models for automatically tagging data in the future.

Figure 1. Some of the challenges in using CDEs

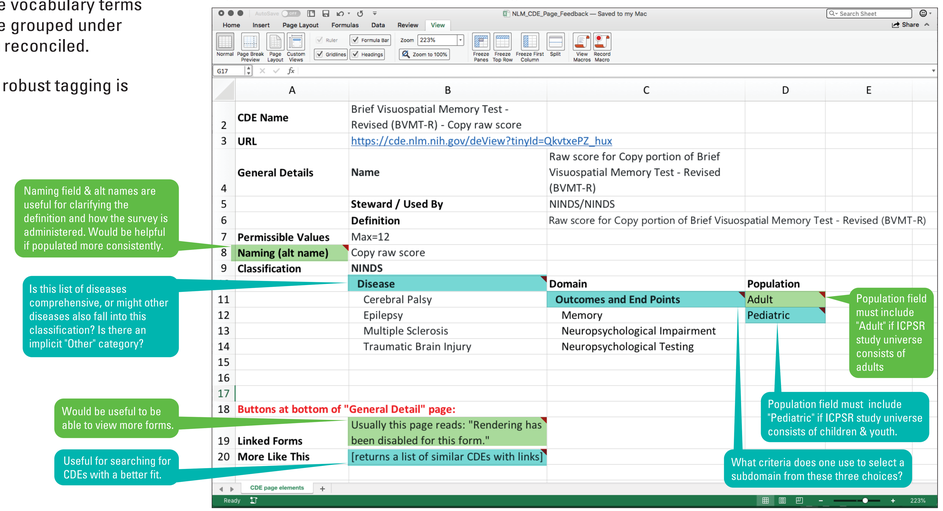


Table 1. Inter-rater reliability table

Inter-rater Reliability						
	Total Vars	Vars In-Scope	Curator 1 Agrees with Curator 2	Curator 1 Agrees with Metadata Expert	Curator 2 Agrees with Metadata Expert	All Three Agree
CDE: All In-Scope						
CFP	918	379	86.3	88.1	88.1	88.1
SAFE	377	142	96.5	97.2	98.6	96.5
YAI	533	184	89.1	13.0	10.3	8.2
Ontology: All In-Scope						
CFP	918	379	31.9	65.7	56.7	29.6
SAFE	377	142	33.1	47.2	69.7	28.9
YAI	533	184	32.6	69.0	51.1	27.7

Figure 2. NLM CDE Repository – General Detail

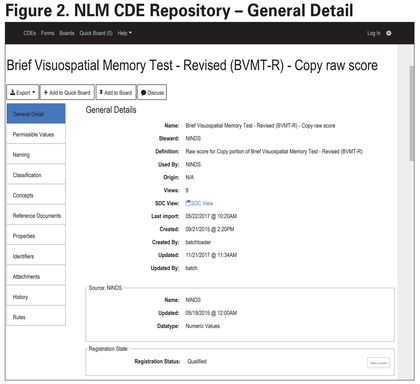
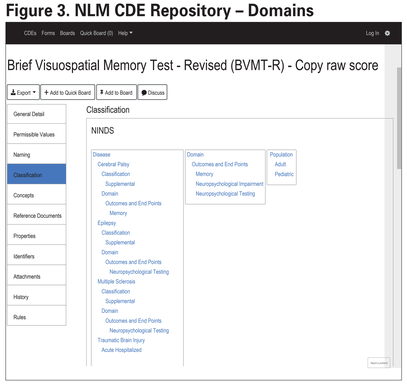


Figure 3. NLM CDE Repository – Domains



RESULTS

This project had the following results:

- Enhanced variable-level metadata for nine high-value studies with improved search results.
- Lessons regarding CDEs and ontologies:
  - Existing CDE terms are too narrowly defined to be effectively applied to legacy datasets.
  - Using more aggregate categories (NLM CDE domains and PROMIS ontology terms) made the tagging more efficient, resulted in greater consistency, and improved discoverability.
  - Aggregate categories will be easier to use in developing machine learning.
- A tagging tool built to facilitate further metadata enhancements
  - Controlled vocabulary in the tool helps consistency.
  - Domain knowledge still important to high quality enhancements, suggesting that leveraging knowledge of expert users is efficient solution.
- Supporting automation
  - ICPSR metadata that groups variables allows more efficient identification, expanding number of 'gold-standard' tags.
  - Generate a question bank from ICPSR variables; contribute these to the NLM CDE Repository and appropriate ontologies for review and adoption.

AUTHORS

Susan Hautaniemi Leonard, Vanessa Unkeless-Perez, Kaye Marz, James McNally, Amy Pienta, ICPSR, University of Michigan

Funding from the National Institute on Drug Abuse (NIDA), the National Institute on Aging (NIA), and the National Institutes of Health Office of Behavioral and Social Sciences Research (OBSSR)

NIH National Institute on Drug Abuse

NIH National Institute on Aging

NIH National Institutes of Health
Office of Behavioral and Social Sciences Research