Advancing the Use and Development of Common Data Elements in Research

Natcher Conference Center
Building 45, NIH Campus
Bethesda, MD

MARCH 6–7, 2024

Thank you Drs. Gregurick and Bertagnolli

For

Lending your voice to the importance this initiative

Setting the Stage: Making Data Interoperable

Denise Warzel, MSc

Scientific Program Analyst

Informatics Data Science Program (IDSP)

Center for Biomedical Informatics and Information Technology (CBIIT)

National Cancer Institute (NCI)

NIH NATIONAL CANCER INSTITUTE

About me…

My mother’s lung cancer diagnosis in 1998 and the mishandling of her medical records eventually led me to leave IBM to help NCI develop data standards.

- My vision was to use data mining to leverage individual patient profiles in clinical trials outcomes data to match patients to the right treatment
- Only 2% of the lung cancer data at a NCI Comprehensive Cancer Center was interoperable
- I joined NCI in 2000 and we established NCI’s Cancer Data Standards Repository and Registry (caDSR) to support something we called common data elements (CDEs).

Today: NCI CDEs are used in 2,200+ sites across US, Canada and internationally, 675,000+ subjects, 3,500+ trials

- While I’ve never had the opportunity to revisited my original idea, the problem of not enough standardized, interoperable data has been successfully addressed through the use of CDEs.

Our mission is clear:

Encourage adoption of CDEs to:

- Improve ***data quality and consistency***
- Integrate and leverage data to support ***advanced analytic methods to better support Data Scientists***

* **Achieve Better Health Outcomes**

→**Increase Life Expectancy**
→**Improved National Health Equity**

Setting the Stage

Elevate	Importance of CDEs
Appreciate	Unique value of CDEs for achieving data integration and the goals outlined by Drs. Gregurick and Bertagnolli
Leverage	CDEs help integrate data, improve quality and consistency and advance data analytics

CDEs Pivotal Role

- Generally understood → common data standards improve interoperability
- Typically →shared set questions on forms, a shared data dictionary, or a common data collection system

➡ ***Data Scientists require high quality, interoperable data that enables new discoveries***

What is a CDE?

Definition:

- **Question or field** [what] **and it’s allowable responses** [how]
- **Used *systematically across different*** sites, studies, or clinical trials
- Helps **ensure consistent data collection**

**Basic Definition Adapted from [NLM CDE Repository “Definition of CDE”]*

Interoperability

- Defined:
  - Ability to seamlessly and efficiently exchange or reuse data with clear unambiguous meaning
- Principles:
  - Standards for exchange
  - Rich metadata
  - Shared semantic alignment and mapping
  - Governance among stakeholders

Applies to CDE interoperability and Data Interoperability

What is unique about CDEs?

Basic Definition Plus Deeper Characteristics and Benefits:

1. **Standard Terminology Concepts** → unambiguous, shared, and computable meaning
2. **Standardized Structure** → machine computability
3. **Independent Semantics** → reusable across physical data models, forms, datasets and supports ***different allowable responses across the same CDE meaning (what)***
4. **Persistent Unique Identifier** → identifiable, outside specific data collection systems
5. **Supports FAIR data** → rich metadata, web accessible repository (Findable, Accessible, Interoperable, Reusable)

Standard Terminologies

Unambiguous, Computable Meaning for CDEs

Interoperability Principle: Shared semantic alignment and mapping

What Do You Mean?

- Context is important in conveying meaning
  - Words have different meanings depending on words around it.
- Some examples:
  - **Agent:** chemical compound or government employee?
  - **Alcohol:** disinfectant or a drink?
  - **Colon:** sentence punctuation or biological organ?
  - **Mole:** animal, blemish, unit of measure, or spy?
  - **Probe:** examination, investigation, or instrument?

  → **The above words are SEMANTICALLY AMBIGUOUS.**

- Words can mean different things in different contexts.

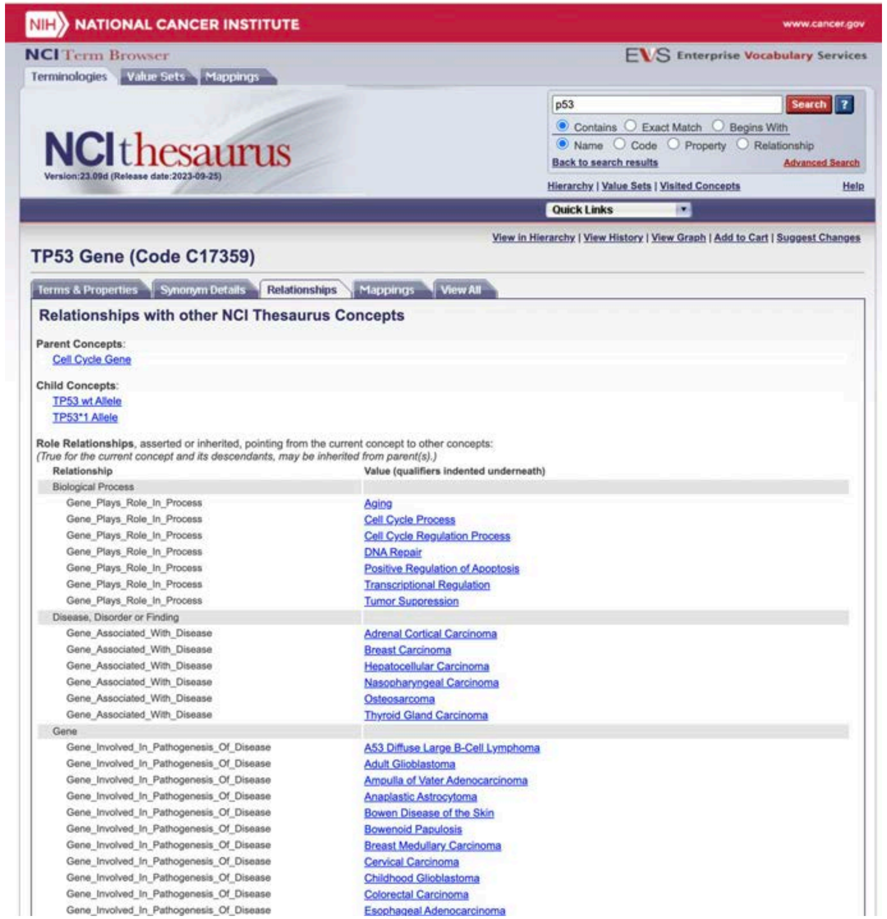
What are Standard Terminologies?

- Vocabulary for a specific field of study, domain or context
    - Terms used in a profession i.e. Healthcare and Biomedical Research
- Independent unique identifiers (concept id or code)
    - Ensures people and computers attach the same meaning
- Provide consistency, clarity
    - Unambiguous semantics
- May include text definitions, synonyms, mappings to other terminologies
    - Improves understanding and facilitates mapping of terminology unique identifier across terminologies
- Ontologies → special kind of terminology
    - Encode Knowledge through **context specific concept relationships**

A word about Ontologies

Knowledge Expansion, Access and Compare Meaning

- “TP53 Gene” Code C17359
- Concept Relationships
  - Gene_Plays_Role_In_Process
  - Gene_Associated_With_Disease
  - Gene_Involved_In_Pathogensis_Of_Disease
  - Gene_Has_Abnormality
  - Gene_Found_In_Organism


NIH NATIONAL CANCER INSTITUTE
www.cancer.gov
NCI Term Browser
EVS Enterprise Vocabulary Services
Terminologies
Value Sets
Mappings
NCI thesaurus
Version:23.09d (Release date:2023-09-25)
p53
Search
?
● Contains ○ Exact Match ○ Begins With
● Name ○ Code ○ Property ○ Relationship
Back to search results
Advanced Search
Hierarchy | Value Sets | Visited Concepts
Help
Quick Links
View in Hierarchy | View History | View Graph | Add to Cart | Suggest Changes
Terms & Properties
Synonym Details
Relationships
Mappings
View All

TP53 Gene (Code C17359)

Relationships with other NCI Thesaurus Concepts

Parent Concepts:
Cell Cycle Gene

Child Concepts:
TP53 wt Allele
TP53*1 Allele

Role Relationships, asserted or inherited, pointing from the current concept to other concepts:
(True for the current concept and its descendants, may be inherited from parent(s).)

Relationship	Value (qualifiers indented underneath)
Biological Process	
Gene_Plays_Role_In_Process	Aging
Gene_Plays_Role_In_Process	Cell Cycle Process
Gene_Plays_Role_In_Process	Cell Cycle Regulation Process
Gene_Plays_Role_In_Process	DNA Repair
Gene_Plays_Role_In_Process	Positive Regulation of Apoptosis
Gene_Plays_Role_In_Process	Transcriptional Regulation
Gene_Plays_Role_In_Process	Tumor Suppression
Disease, Disorder or Finding	
Gene_Associated_With_Disease	Adrenal Cortical Carcinoma
Gene_Associated_With_Disease	Breast Carcinoma
Gene_Associated_With_Disease	Hepatocellular Carcinoma
Gene_Associated_With_Disease	Nasopharyngeal Carcinoma
Gene_Associated_With_Disease	Osteosarcoma
Gene_Associated_With_Disease	Thyroid Gland Carcinoma
Gene	
Gene_Involved_In_Pathogenesis_Of_Disease	A53 Diffuse Large B-Cell Lymphoma
Gene_Involved_In_Pathogenesis_Of_Disease	Adult Glioblastoma
Gene_Involved_In_Pathogenesis_Of_Disease	Ampulla of Vater Adenocarcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	Anaplastic Astrocytoma
Gene_Involved_In_Pathogenesis_Of_Disease	Bowen Disease of the Skin
Gene_Involved_In_Pathogenesis_Of_Disease	Bowenoid Papulosis
Gene_Involved_In_Pathogenesis_Of_Disease	Breast Medullary Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	Cervical Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	Childhood Glioblastoma
Gene_Involved_In_Pathogenesis_Of_Disease	Colorectal Carcinoma
Gene_Involved_In_Pathogenesis_Of_Disease	Esophageal Adenocarcinoma

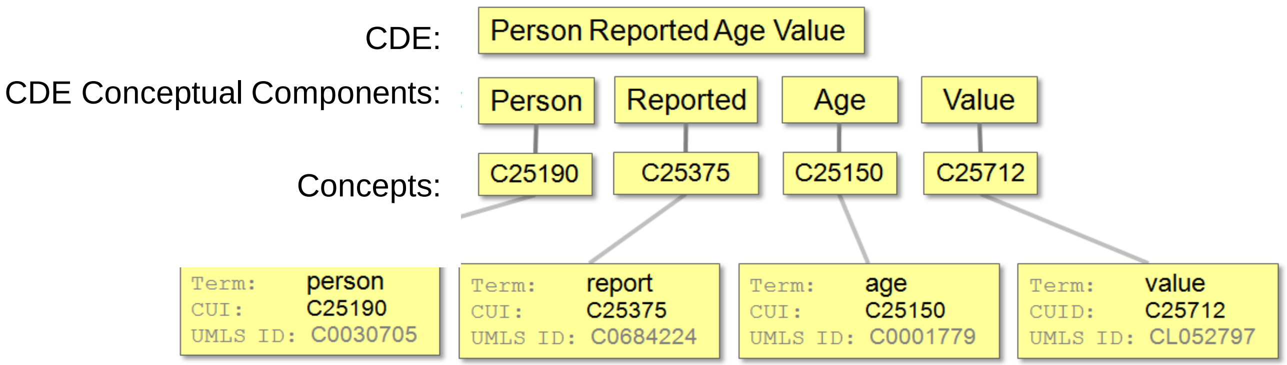
Key message

- Words can have different meanings
- Use Standard Terminology for clear, shared meaning

Standard Structure

Machine computable meaning

Standard Structure for Concepts representing CDE meaning (What)



CDE: Person Reported Age Value

CDE Conceptual Components: Person | Reported | Age | Value

Concepts: C25190 | C25375 | C25150 | C25712

Term: person CUI: C25190 UMLS ID: C0030705	Term: report CUI: C25375 UMLS ID: C0684224	Term: age CUI: C25150 UMLS ID: C0001779	Term: value CUID: C25712 UMLS ID: CL052797

Term: the word in the terminology (human readable and understanding what is meant)

CUI: in this case →NCI Thesaurus Concept Codes

UMLS CUI: the Unified Medical Language System (UMLS) Concept Unique Identifier

Benefit: Organizing terminology concepts in a standard structure enables computable and comparable meaning

Standard Structured for Concepts representing allowable responses (How)

Terminology Code Mapping

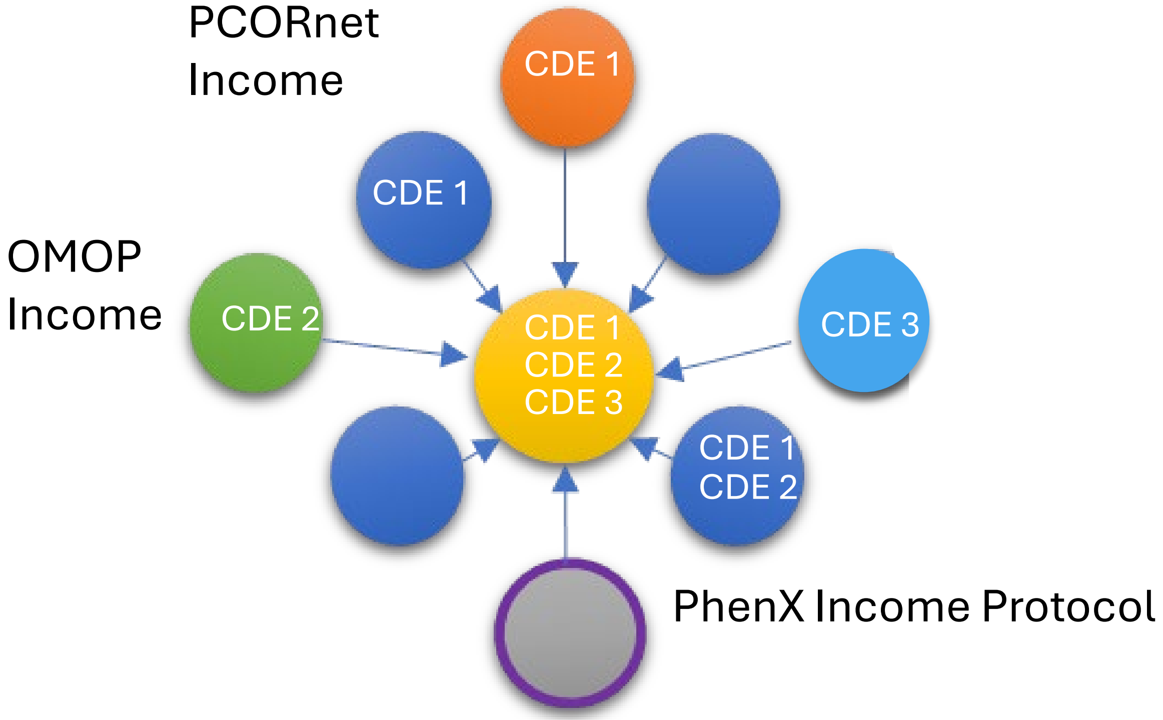
	NCI Thesaurus →	UMLS CUI →	LOINC
American Indian or Alaska Native	C41259	C0282204	LA10608-0
Asian or Asian American	C41260	C0003988	LA6156-9
Black of African American	C16352	C0085756	LA10610-6
Hispanic, Latino, or Spanish	C17459	C0086409	LA6214-6
Native Hawaiian or Other Pacific Islander	C41219	C1513907	LA10611-4
Middle Eastern or North African	C43866	C1553353	No Match
White	C41261	C0043157	LA4457-3

One last word on Independent Semantics/Meaning

Machine computable meaning

Common Semantics vs Common Data Model

- Independent - Common Semantics
  - *Provides clear, unambiguous meaning*
  - Meaning is independent of any data model
  - Mapping at CDE level instead of Model Level
- Common Data Model
  - *Describes how data is organized for data storage*
  - Driven by query and analysis requirements

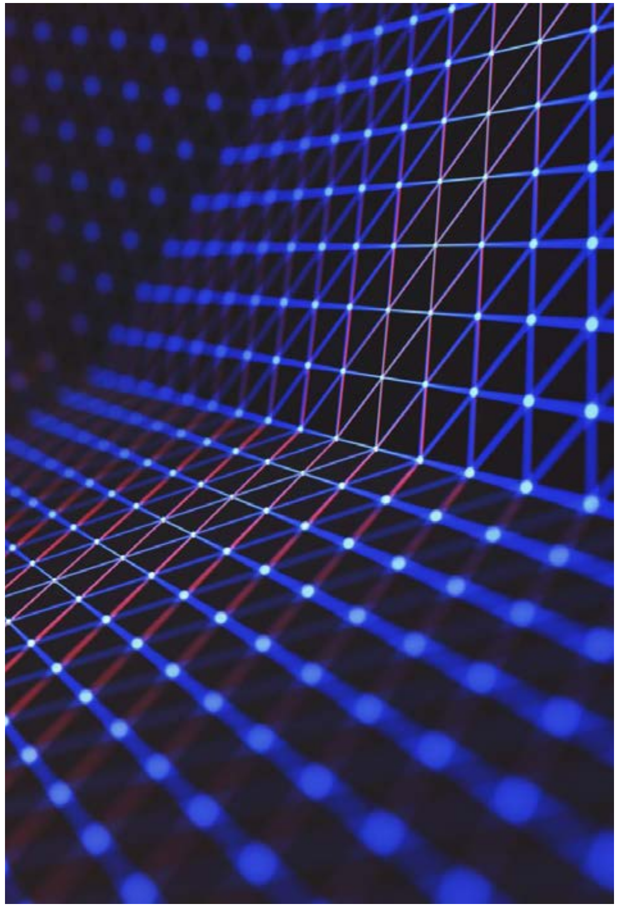


Key message

- Concepts are the foundation for machine computable meaning
  - Standard Structure for CDE Concepts
    - The field or question (“What”)
    - The allowable responses (“How”)

Adopt CDEs to Data Interoperability →Advance Research

Improve	Support	Enhance	Simplify	Reduce
Improve data quality and consistency	**Support Data harmonization**	**Enhance knowledge acquisition**	**Simplify collaboration**	**Reduce project start-up**
Unambiguous meaning people and computers	Mapping and transformation	Advanced data analytics Data Science	Healthcare and Research	Well designed and vetted



Advancing the Use and Development of **Common Data Elements** in Research

Natcher Conference Center
Building 45, NIH Campus
Bethesda, MD

MARCH 6–7, 2024

Thank you for your time and attention!

We will now break until 11:30 a.m.