Use of SNOMED CT Compositional Grammar to Annotate Human Clinical Trials Data Elements

Kin Wah Fung, Vojtech Huser, Cynthia Burke, Minh-Diep Nguyen, Liz Amos

National Library of Medicine
National Institutes of Health
Bethesda, MD, USA

Routine Healthcare and Research

- "Big data" analytics research studies are increasingly combining data from multiple sources e.g. healthcare claims data, electronic health record (EHR) data, data captured in Case Report Forms from clinical trials
  - An example: US National Institutes of Health's AllOfUs research program will use EHR Data and healthcare claims data https://allofus.nih.gov/
- Challenge:
  - how to make the disparate data sources interoperable?
  - Answer: Data standardization
- Routine healthcare – use of standard terminologies
  - SNOMED CT, LOINC, ATC, RxNorm (RxNorm Extension [outside USA drugs])
- Research – data structure standardization
  - Common Data Elements (CDEs)
  - Patient Reported Outcomes (PROs)

Standardizing data collection in research studies

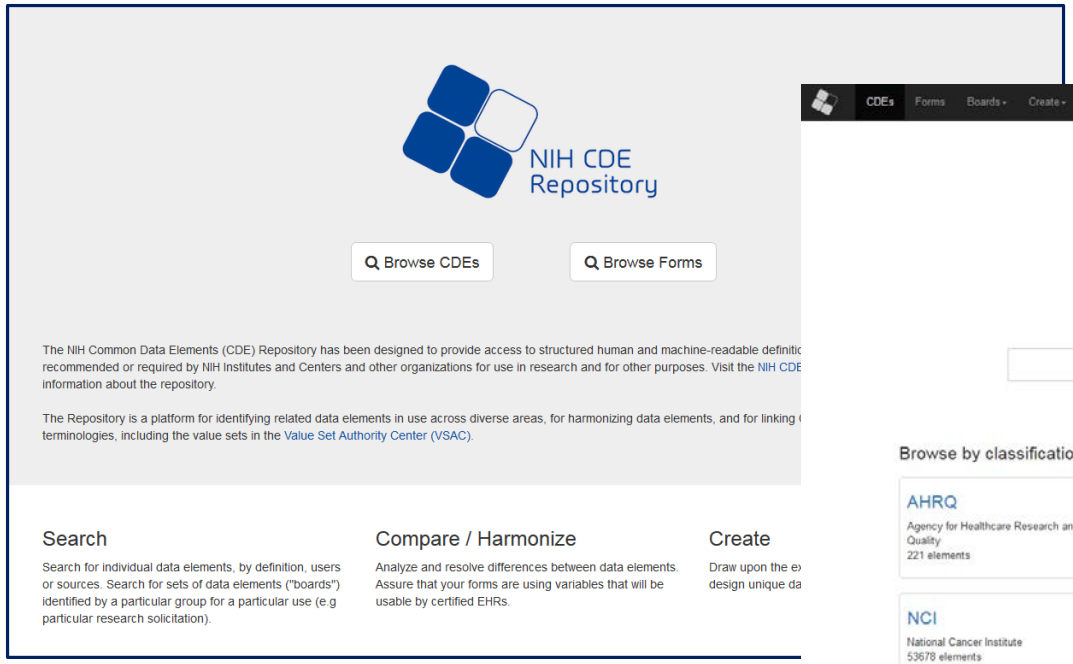
- Data standardization for submission to regulatory authorities (FDA, EMA, JPMC)
  - CDISC data structure standards and CDISC Controlled Terminology (since early 2000s)
- Since 2010s: emergence of Common Data Elements initiatives
  - Standardize data structure for any study (not just for studies undergoing regulatory submissions)

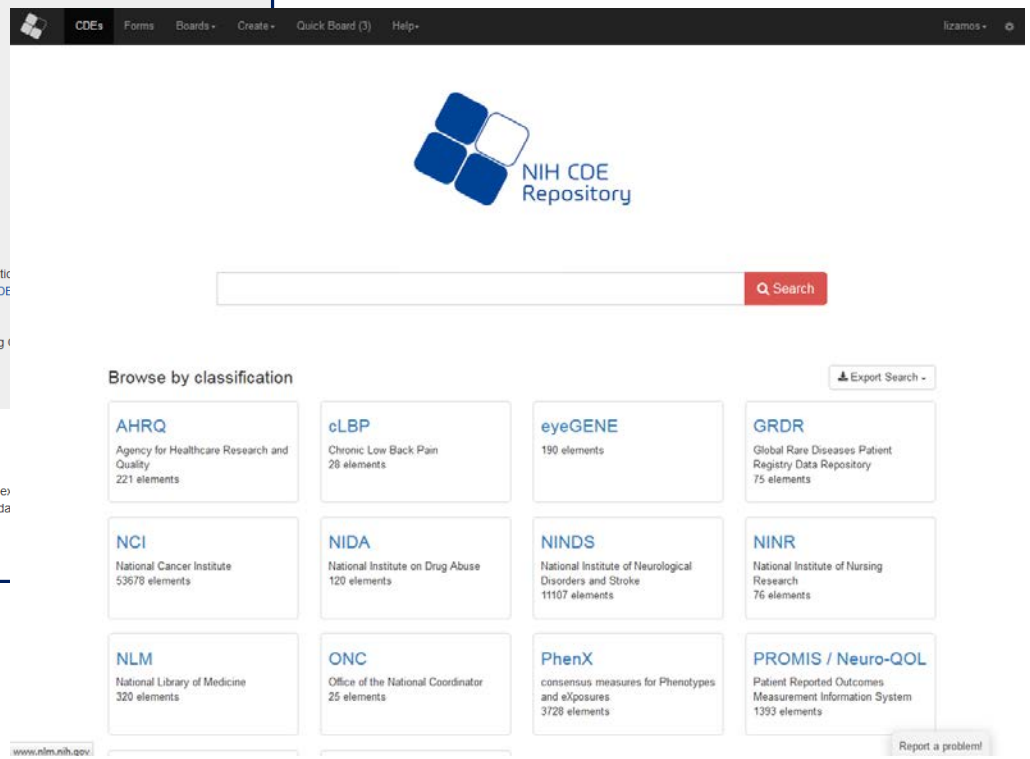
NIH Common Data Element Repository

- Common data element (CDE) is a representation of a variable (usually a question) common to multiple studies e.g., how often do you consume alcohol?
- Usually the response is a fixed list of values e.g., not at all, some days, everyday
- In the repository, CDEs are defined unambiguously in human and machine readable terms
- Sets of CDEs can be combined into more complex questionnaires, survey instruments, and case report forms

The NIH CDE Repository is a tool to search across CDE initiatives, harmonize differences and create new CDEs

https://cde.nlm.nih.gov

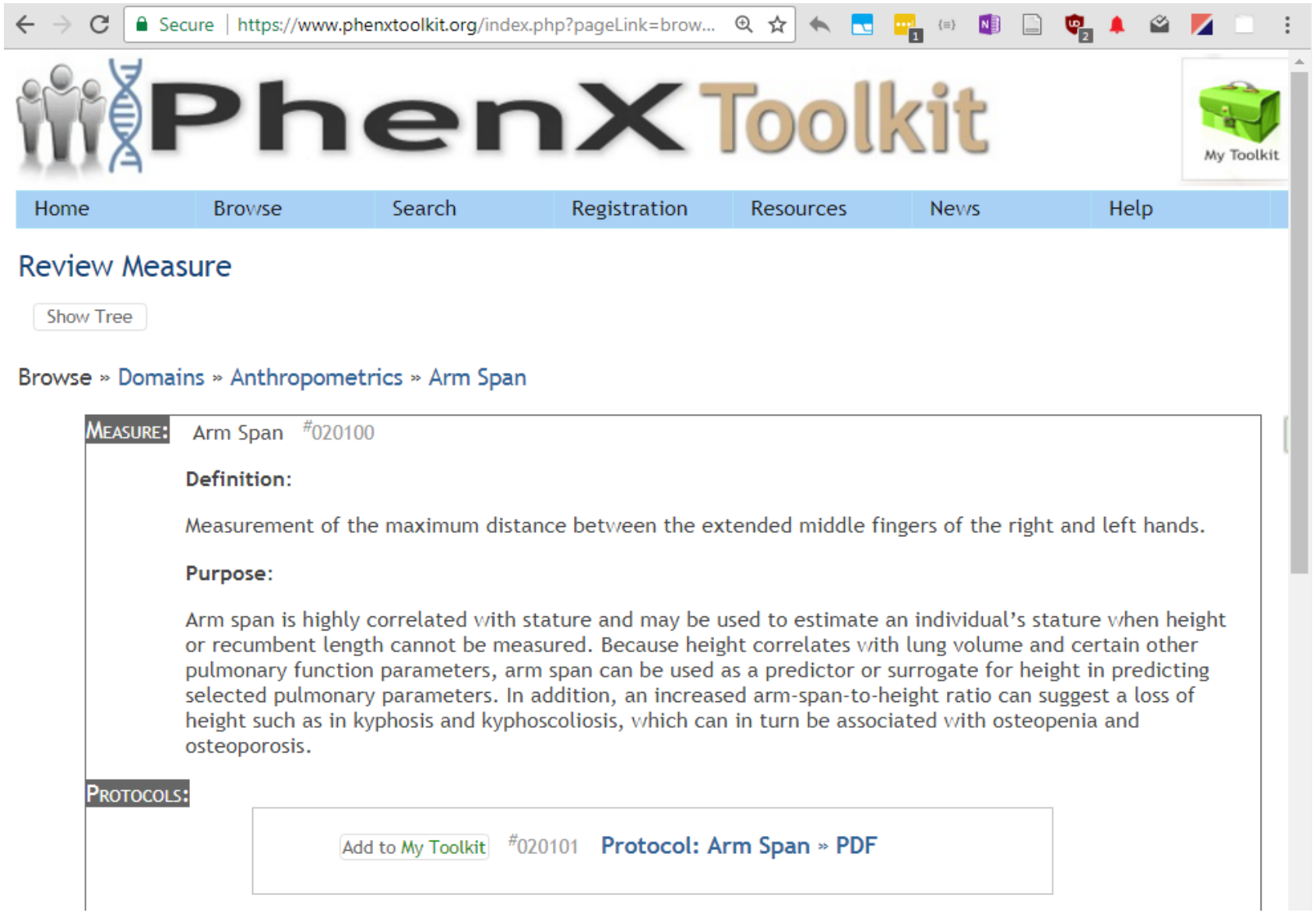




Benefits of re-using CDEs

- Consistent data collection of core set of variables from different sources (sites, projects, initiatives) to allow:
  - Aggregation of data to increase statistical power
  - Rigorous comparison of data & results
- Can be used to promote research:
  - Efficiency – off-the-shelf data elements
  - Quality – validated instruments & measures
  - Clarity – unambiguously defined data elements
  - Reproducibility – from rigorous comparison

CDE initiative example: PhenX



Challenges in developing research data elements

- Overlap of research data element initiatives
  - Each research group (e.g., cancer, behavioral researchers) may have a separate initiative

...

- Discovering existing defined research data elements is challenging for investigators (reliance of simple text search)
  - Portals: https://cde.nlm.nih.gov or https://medical-data-models.org
- Researchers may not be intimately familiar with routine healthcare terminologies (e.g., SNOMED CT)
  - overlap of research data elements with terminologies; duplication

Possible solution: Annotate CDEs using SNOMED CT

- **Discover overlapping** (or related) research **data elements**
  - Note there are three levels for annotation
    - Case Report Form
    - Case Report Form Question
    - Case Report Form Answer
- Improve discoverability of research data elements - **improve search**
- **Demonstrate utility of** routine healthcare **terminologies** (e.g., SNOMED CT) to researchers → (bring closer research and routine healthcare EHR data)
- Assess the feasibility of using **SNOMED CT Compositional Grammar** to support the annotation
  - Extend SNOMED CT Compositional Grammar
    - it has known limitations

Methods

- Source of CDEs:
  1. PhenX data elements - 564 forms and 22705 elements
     - developed by two institutes at US National Institutes of Health (NIH)
  2. CRF library published by Elli Lilly - 914 forms and 28310 elements
  3. subset of REDCap Consortium library - 10 forms and 879 elements.

- Separate annotation subtasks structured by CDE data type
  - such as boolean, date, text, number, pick-list question (radiobutton/checkbox)
- Pilot experiment/ feasibility assessment
  - source #1 (PhenX) + convenience sample of CDEs for each data type

CDE Input data (PhenX, RedCAP .CSV format)

Cancer Personal and Family History

Variable / Field Name	Form Name	Section	Field Typ	Field Label	Choices, Calculations, OR Slider Labels
instructions_070601	phenx_cancer_personal_and_family_history		descriptive	Please answer the following questions about you and your "blood" relatives. Do not include adopted, half and	
self_gender	phenx_cancer_personal_and_family_history	YOURSELF	radio	Gender?	UNDEFINED_CODE,Male|UNDEFINED_CODE_1,Female
self_alive	phenx_cancer_personal_and_family_history		radio	Alive?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
self_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
self_ever_cancer_diagnosis	phenx_cancer_personal_and_family_history		radio	Never had cancer?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
self_first_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	First cancer diagnosis?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Ovarian Canc
self_first_cancer_diagnosis_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
self_second_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	Second or other cancer diagnoses?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Ovarian Canc
mother_gender	phenx_cancer_personal_and_family_history	MOTHER	radio	Gender?	UNDEFINED_CODE,Female
mother_alive	phenx_cancer_personal_and_family_history		radio	Alive?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
mother_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
mother_ever_cancer_diagnosis	phenx_cancer_personal_and_family_history		radio	Never had cancer?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
mother_first_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	First cancer diagnosis?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Ovarian Canc
mother_first_cancer_diagnosis_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
mother_second_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	Second or other cancer diagnoses?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Ovarian Canc
father_gender	phenx_cancer_personal_and_family_history	FATHER	radio	Gender?	UNDEFINED_CODE,Male
father_alive	phenx_cancer_personal_and_family_history		radio	Alive?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
father_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
father_ever_cancer_diagnosis	phenx_cancer_personal_and_family_history		radio	Never had cancer?	UNDEFINED_CODE,Yes|UNDEFINED_CODE_1,No
father_first_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	First cancer diagnosis?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Lymphoma |
father_first_cancer_diagnosis_age	phenx_cancer_personal_and_family_history		radio	Current age or age at death?	UNDEFINED_CODE,<50|UNDEFINED_CODE_1,50-64|UNDEFINED_COD
father_second_cancer_diagnosis	phenx_cancer_personal_and_family_history		checkbox	Second or other cancer diagnoses?	UNDEFINED_CODE, Breast Cancer | UNDEFINED_CODE_1, Colon/Rectal
sibling_gender_01	phenx_cancer_personal_and_family_history	SIBLING 1	radio	Gender?	UNDEFINED_CODE,Male|UNDEFINED_CODE_1,Female

Methods

- Three annotators
  - One with SNOMED CT training (Foundation, Implementation, Content Development)
  - Two medical librarians with no prior SNOMED CT experience (internal training provided for the project)

- Feasibility pilot study – exploratory phase
  - First identify SNOMED CT concepts
    - by simply enumerating terms linked to a give CDE
  - Construct SNOMED CT expression later
  - Identify level of match (exact, partial-high, partial-low, no match)
  - Expectation setting: “lessons learned” type of results (not final quantitative numbers)

  - Cloud based spreadsheet document (with limitations and challenges)
    - Ideal system would be a web-based system with user friendly searching and entering SCT terms

Results: Annotations made (Question level)

- By data type

TYPE	COUNT
a-boolean	74
a-booleanExt	31
date_mdy	483
integer	10
number	96
radio	32
string	44
TOTAL	**770**

- Convenience sample
  - ideally we would have CDE's usage data
- Data available on project website (GitHub repository)
  - https://github.com/lhncbc/CDE/tree/master/annotation

Annotation (Entering Related Concepts) on Question level

Field Label	DONE	TYPE	A-sul	match clas	A-term1	A-term2	A-term3
Low Blood Pressure Date of Ons	1	date_mdy		exact	298059007 |Date of onset (observ	45007003 | Low blood pressure (disorder) |	
Mental Retardation Date of On	1	date_mdy		exact	298059007 |Date of onset (observ	91138005 | Mental retardation (disorder) |	
Palpitations Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	80313002 | Palpitations (finding) |	
Rheumatic Fever Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	58718002 | Rheumatic fever (disorder) |	
Rheumatism Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	396332003 | Rheumatism (disorder) |	
Seizures Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	91175000 | Seizure (finding) |	
Syphilis Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	76272004 | Syphilis (disorder) |	
Tuberculosis Date of Onset	1	date_mdy		exact	298059007 |Date of onset (observ	56717001 | Tuberculosis (disorder) |	
Measles - Date	1	date_mdy	date	exact	410672004 | Date property (qualif	14189004 | Measles (disorder) |	
Rubella - Date	1	date_mdy	date	exact	410672004 | Date property (qualif	36653000 | Rubella (disorder) |	
Pertussis/Whooping Cough - D	1	date_mdy	date	exact	410672004 | Date property (qualif	27836007 | Pertussis (disorder) |	
Pneumonia - Date	1	date_mdy	date	exact	410672004 | Date property (qualif	233604007 | Pneumonia (disorder) |	
Mumps - Date	1	date_mdy	date	exact	410672004 | Date property (qualif	36989005 | Mumps (disorder) |	
Scarlet Fever - Date	1	date_mdy	date	exact	410672004 | Date property (qualif	30242009 | Scarlet fever (disorder) |	
Arthritis Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	3723001 | Arthritis (disorder) |	
Asthma Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	195967001 | Asthma (disorder) |	
Bronchitis Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	32398004 | Bronchitis (disorder) |	
Dizziness Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	404640003 | Dizziness (finding) |	
Emphysema Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	49158009 | Emphysema (morphologic abnormality) |	
Epilepsy Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	84757009 | Epilepsy (disorder) |	
Glaucoma Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	23986001 | Glaucoma (disorder) |	
Gonorrhea Date of Onset	1	date_mdy		exact	298059007 | Date of onset (obser	15628003 | Gonorrhea (disorder) |	
Hearing Impairment Date of On	1	date_mdy		exact	298059007 | Date of onset (obser	15188001 | Hearing loss (disorder) |	

Postcoordination

- Research data elements can be very detailed and granular
  - Challenge: When is an element too complex to model using SNOMED CT?
- SNOMED CT expressions are needed
    - Example:
      - 68526006|removal of device from abdomen|:425391005|using access device|= 6174004|laparoscope|
    - https://confluence.ihtsdotools.org/display/DOCSTART/7.+SNOMED+CT+Expressions
    - Expressions must follow concept model
      - most relevant is Observable Entity; a chosen model may undergo revisions
- Alternatives to postcoordination
  - Expression repository (for research data elements)
  - Information model postcoordination
    - We are considering that for some elements or element groups
      - Similar to defining ”CDE concept model”
      - Also considering limitations of the SNOMED CT Compositional grammar (extension?)

Results: Example 1

- Simple term

- Form: phenx_personal_history_of_allergies_infectious_dis
- Question: *Gonorrhea Date of Onset*

- Annotation terms
  - 15628003 | Gonorrhea (disorder) |
  - 298059007 | Date of onset (observable entity) |

Results: Example 2

- Form: phenx_peripheral_arterial_disease
- Question: *Measure systolic blood pressure of: Left dorsalis pedis artery (mm Hg)*
- Annotation terms
  - 445731001 |Dorsalis pedis arterial pressure (observable entity)|
  - 259018001 |Millimeter of mercury (qualifier value)|
  - 7771000 |Left (qualifier value)|
  - 259018001 |Millimeter of mercury (qualifier value)|

- SNOMED CT Expression (using SNOMED CT compositional grammar)
  - 363787002 | Observable entity (observable entity) | :
    118598001 | Measurement property | = 72313002 | Systolic arterial pressure (obs. entity)|,
    704327008 | Direct site (attribute) | = (
    86547008 | Structure of dorsalis pedis artery (body structure) | :
    272741003 | Laterality (attribute) | = 7771000 | Left (qualifier value) |),
    246514001 | Units (attribute) | = 259018001 |Millimeter of mercury (qualifier value)|

List of Examples

- Using syntax Form: *Question*
- Listed in the order of increasing complexity

- phenx_spirometry: *Date*
- phenx_spirometry: *Date of birth*
- phenx_medical_history: *Hemodialysis Date of Onset*
- phenx_knee_height: *Knee Height 3*
- phenx_hip_circumference: *Hip Circumference measured in centimeter, first measurement*
- phenx_arrhythmia_atrial_and_ventricular *Year inserted permanent pacemaker.*
- phenx_smoking_quit_attempts: *Are you considering quitting smoking during the next 6 months?*
- ninds_mutation_analysis*: Are there additional genes sequenced with no mutations detected? (Yes/No)*
- phenx_body_composition_suprailiac_skinfold_thickness: *Was a caliper placed perpendicular to the skinfold, which should have been sloping downward and forward at a 45 degree angle extending toward the pubis symphysis, about 2.0 cm medial to the fingers?*

Results: Example 3

- Question mapping vs. question+answer mapping
  - On the answer level, the mapping can be more accurate

- Form: Screening NCI Standard Template:
- *Question: Does the participant meet all screening criteria? (Yes/No)*
  - *Answer: No*:
    - 444734003 |Does not meet eligibility criteria for clinical trial (finding)|
  - Answer: Yes
    - 399223003 |Patient eligible for clinical trial (finding)|

Permissible Values (Answers) Annotation

- Checklist and radiobutton research data elements enumerate list of possible answers
- Example:
  - Alcohol Use Disorders Identification Test (AUDIT) (Interview Version):
    3. How often do you have six or more drinks on one occasion?
    - Never; Less than monthly; Monthly; Weekly; Daily or almost daily (permissible values)
- Smaller set of terms to annotate (n=6336 unique answer terms in PhenX pick-list questions)
- Very common terms
  - Yes, No, Often, Rarely, True, False, Daily, 1, 2, 3, Don't Know/Refused, Never
- Scale terms
  - Extremely Important, Quite Important, Quite Unimportant, Somewhat Important, Somewhat Unimportant
  - Agree strongly, agree somewhat, disagree somewhat, disagree strongly
- Other terms
  - Bleeding from probing detected, No evidence of bleeding, Talking on phone, Tennis, Travel by bicycling, Discrete pitting of the enamel exists

Exact Matches for Permissible Values

ANSWER	FREQUENCY	MATCH	A-term1-SNOMED-CODE	A-term1-SNOMED Definition
Yes	1748	exact	373066001	yes (qualifier value)
No	1661	exact	373067005	no (qualifier value)
Refused	380	exact	443390004	Refused (qualifier value)
Never	276	exact	664011000124114	Never (qualifier value)
Other	167	exact	74964007	Other (qualifier value)
A little	127	exact	255507004	Small (qualifier value)
Rarely	122	exact	89292003	Rare (qualifier value)
FALSE	115	exact	64100000	False (qualifier value)
TRUE	113	exact	31874001	True (qualifier value)
Some	108	exact	27768009	Some (qualifier value)
UNCERTAIN	94	exact	64957009	Uncertain (qualifier value)
Never	92	exact	664011000124114	Never (qualifier value)
Often	89	exact	70232002	Frequent (qualifier value)
None	84	exact	260413007	None (qualifier value)

Partial Matches for Permissible Values

ANSWER	FREQUENCY	MATCH	A-term1-SNOMED-CC	A-term1-SNOMED Definition	A-term2-SNC	A-term2-SNOMED Definit
DON'T KNOW / REFUSED	300	partial-high	443390004	Refused (qualifier value)		
Not Asked	265	partial-high	1631000175102	Patient not asked (contextual qualifier) (qv)		
Not at all	213	partial-high	260353006	Nothing at all (qualifier value)		
Refused to answer	174	partial-high	2834984018	Refused (qualifier value)		
Bleeding from probing detected	168	partial-low	131148009	Bleeding (finding)	260373001	Detected (qualifier value)
Somewhat true	129	partial-low	31874001	True (qualifier value)		
Very close	87	partial-high	260358002	Very (quanlifier value)		
Not At All Important =1	80	partial-high	260353006	Nothing at all (qualifier value)		

Lessons Learned (1/2)

- Annotation requires significant SNOMED CT training and annotation guidelines (we are drafting those)
  - E.g., how to deal with misclassified data types; flavors of null (in pick-list questions)

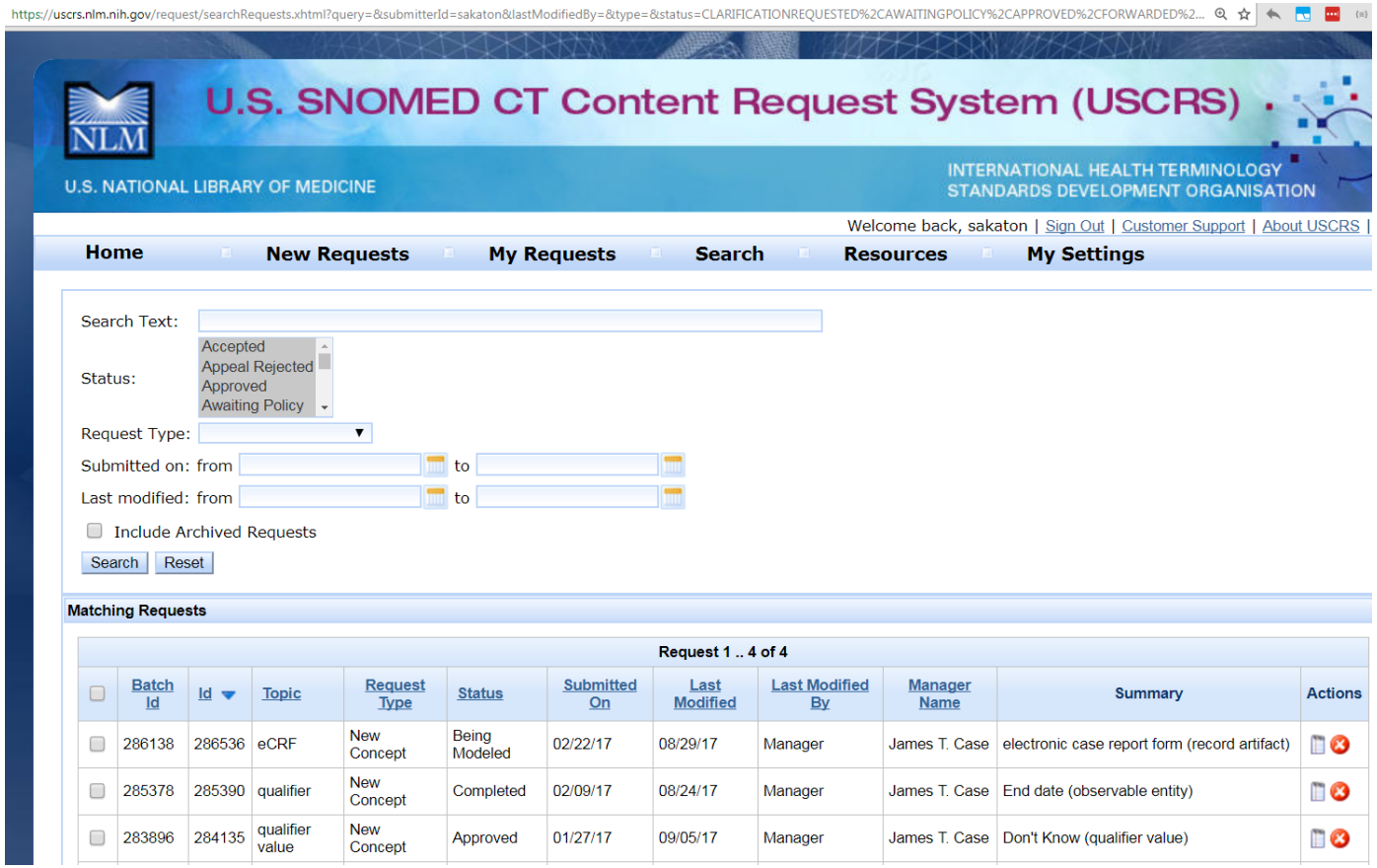
- Determining type of match is challenging
  - completed only for 191 elements out of 770 (24%);
    - preliminary results; not ready for presentation
    - 2007 study in JAMIA (Andrew et al.) found poor agreement of 3 annotators (data annotation)
  - Working categories: Exact, Partial (low/high later scrapped), No-match (+ Complex/Out of scope)
    - phenx_alcohol_30_day_quantity_and_frequency: *During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?*
      - Related concepts
        - 160573003 |Alcohol intake (observable entity)|
        - 408731000 |Temporal context (attribute)|
        - 258734002 |Counts (qualifier value)|
      - Classify as partial-high or partial-low? (what % of question can be captured by the expression?)
  - In formal maps: type of match would correspond to CorrelationIds (e.g., LOINC map)
    - 447557004 |exact match from SCT to target|
    - 447559001 |broad to narrow map from SCT to target| (CDE is narrower than SCT expression)

Lessons Learned (2/2)

- Lack of software tools for working with SNOMED CT expressions
  - Learn from LOINC map (expression expander)
    - E.g., what is 363787002: 370134009=123029007,704327008=258459007, 704318007= 118556004, 246093002=273948005,370132008=30766002, 704319004=31773000
  - Expression builder
    - Load and validate concept model rules (ongoing effort around Machine Readable Concept Model)
  - Expression checker
    - Web version no longer available
    - Only as application download + install
      - https://github.com/ldthomas/apg-js2
- Helpful advice obtained from community
  - https://confluence.ihtsdotools.org/questions

Value to SNOMED CT: New Requests

- Examples of accepted terms
  - Electronic case report form
  - Arm Span
- Examples of terms in Clarification status (International edition)
  - Don’t know (qualifier value)



Conclusion and Future work

- Pilot effort to relate research data elements with SNOMED CT using compositional grammar
- Research case report forms can be of high granularity (compared to EHR data)
    - Rather extreme use case for creating SNOMED CT Expressions
- Future work
    - Prioritize mapping of research data elements that have usage data linked to them
        - elements used in at least 2 studies shared via a data sharing platform
    - Annotations can be used to pre-populate research Case Report Forms from EHR data
        - E.g., Date permanent pacemaker was implanted

Acknowledgement / Questions?

- Thank you
  - Vojtech Huser
  - Cynthia Burke
  - Minh-Diep Nguyen
  - Liz Amos
  - NIH CDE Task Force
  - Linda Bird

- Questions
  - Vojtech Huser
    - vojtech.huser@nih.gov
  - Project repository
    - https://github.com/lhncbc/CDE