Interoperability between value sets for clinical research and healthcare

Mapping value sets between CDISC & MU

Raja Cholan
Oregon Health & Science University, MS Student Clinical Informatics
NLM Summer Fellowship Research Project
Mentor: Olivier Bodenreider
8/11/2017

Motivation

- Historically, clinical research data distinct from clinical care data
- Now, shift towards pragmatic trials: notion of clinical research on patients directly from EHR data
- Therefore, we assessed interoperability between clinical research & EHR data

Definitions

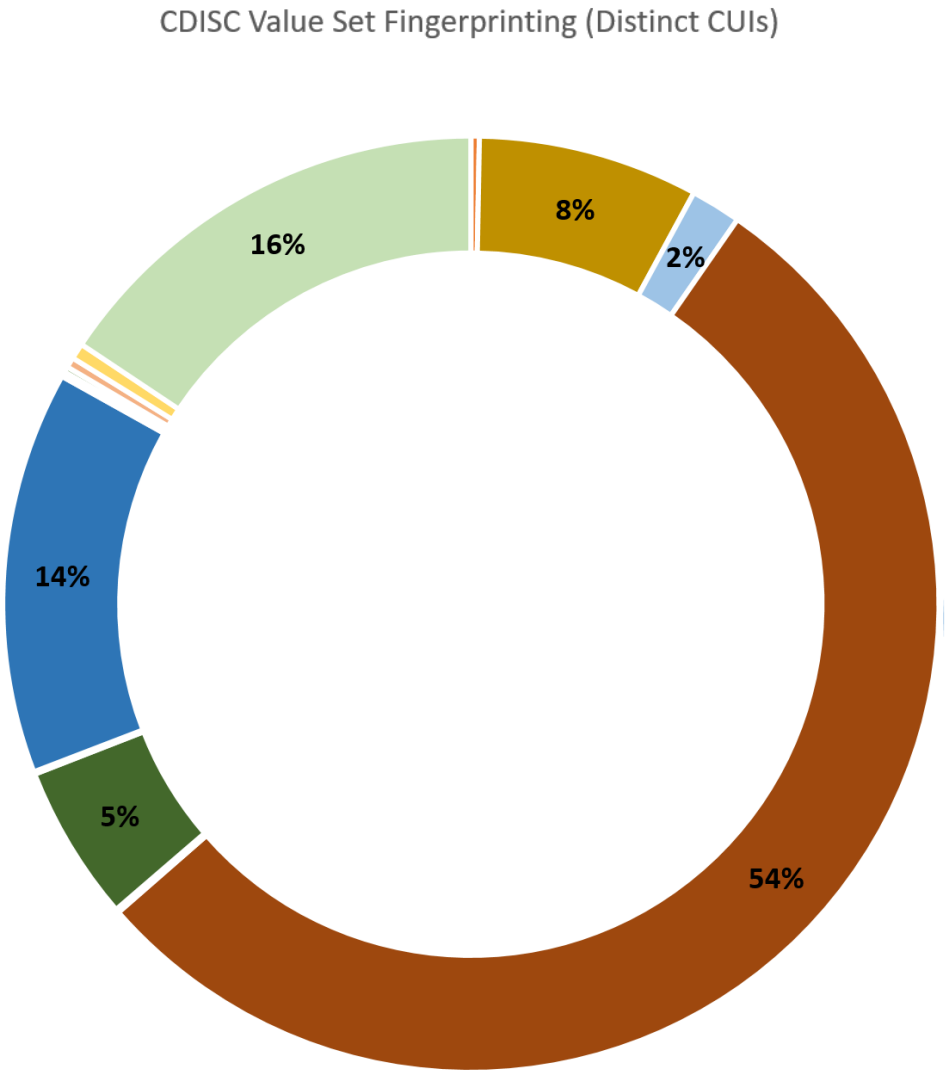
- Value set: "lists of specific values (terms, codes) that define clinical concepts derived from standard vocabularies"-*VSAC FAQ*
- Clinical Data Interchange Standards Consortium (CDISC)
  - "[CDISC] defines platform-independent standards that support the electronic acquisition, exchange, submission and archiving of study data and metadata for pharmaceutical companies and the Food and Drug Administration."–*CDISC website*
  - Codes, terms from NCI Thesaurus
- Value Set Authority Center (VSAC)
  - "Value sets designed for many purposes and programs, including… CMS eCQMs, [and Meaningful Use]" – *VSAC FAQ*
  - Codes, terms from standard clinical terminologies (e.g., SNOMED CT, RxNorm, LOINC)
- Unified Medical Language System (UMLS) Metathesaurus
  - Contains terms and codes from over 150 source vocabularies organized by concept, relationship, attribute, and meaning.
  - We used the UMLS Metathesaurus to connect codes from NCI to codes from VSAC
  - Concepts have Concept Unique Identifiers (CUIs)

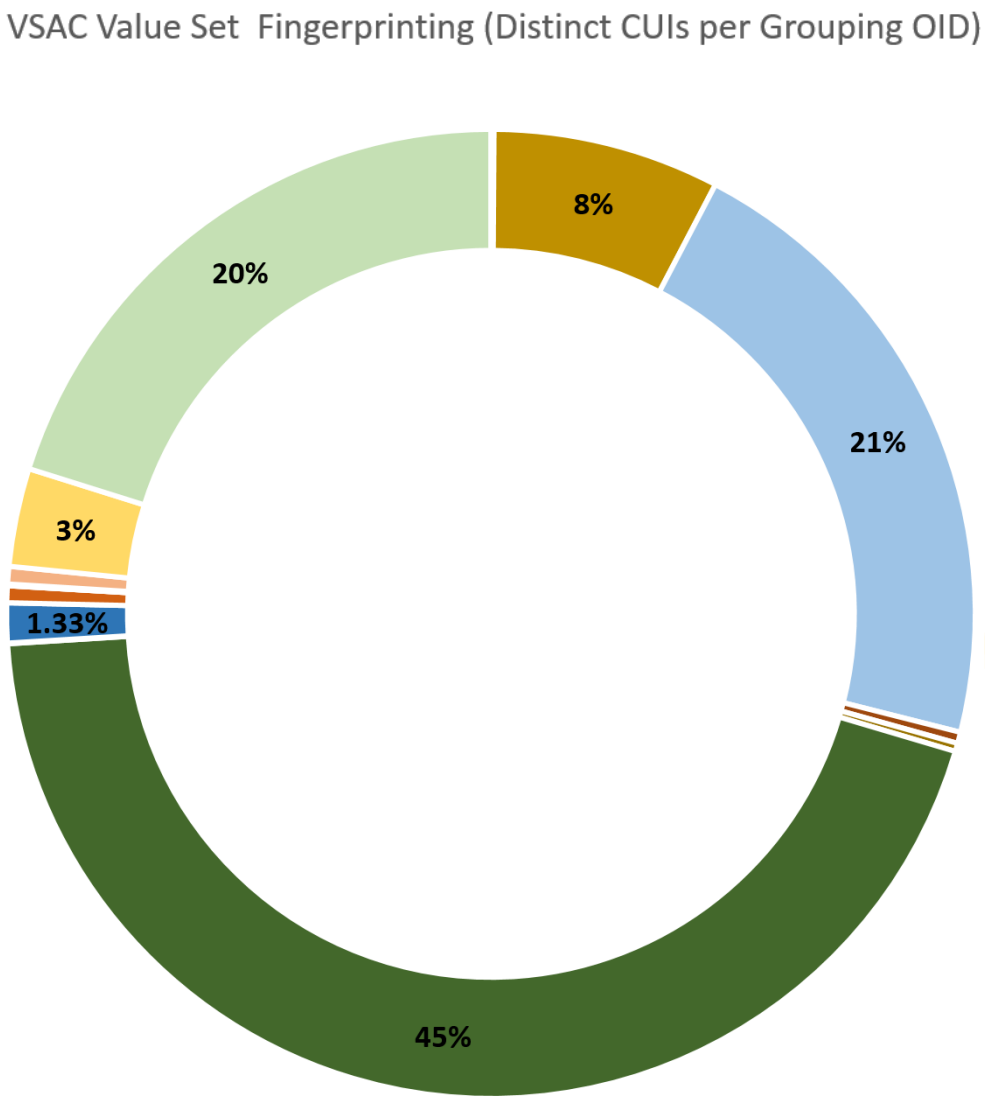
Research Questions

1. What are the semantic characteristics of the concepts used in CDISC and VSAC value sets?

2. To what extent do existing value sets in the VSAC represent value sets in CDISC?

1. Value Set Semantic Profiles (i.e. Fingerprinting)





Activities & Behaviors
Anatomy
Chemicals & Drugs
Concepts & Ideas
Devices
Disorders
Genes & Molecular Sequences
Geographic Areas
Living Beings
Objects
Occupations
Organizations
Phenomena
Physiology
Procedures

Example

Research

CDISC

- Value Set Name (i.e. Codelist Name): PROCEDURE
- Value Set ID (Codelist Code): C101858
  - -60 NCI Codes → 60 UMLS CUIs

Healthcare

VSAC

-Value Set Name: Coronary Artery Bypass Graft

-Object Identifier, OID (Value Set ID): 2.16.840.1.113883.3.464.1003.104.11.1004

-52 SNOMED CT Codes → 51 UMLS CUIs

1 NCI Code
Code: C51998
Name: CORONARY ARTERY BYPASS GRAFT

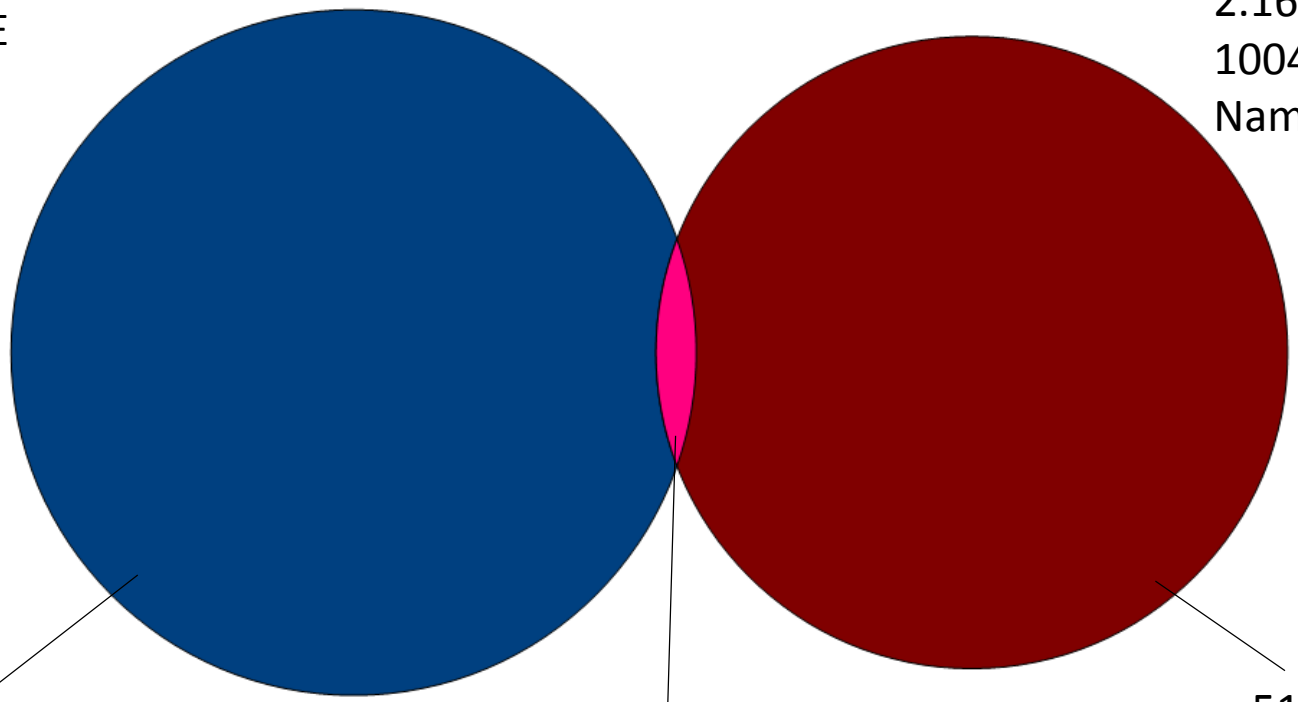
2 SNOMED CT Codes
Code: 232717009 Code: 67166004
Name: Coronary Artery Bypass Surgery

UMLS CUI: C0010055

Example, continued

CDISC Value Set ID: C101858
Name: PROCEDURE

VSAC OID:
2.16.840.1.113883.3.464.1003.104.11.1004
Name: Coronary Artery Bypass Graft



59 UMLS CUIs distinct to CDISC
i.e.
C3272249, Pericardial Stripping
C0348007, Laser ablation

1 UMLS CUI **C0010055** in both

51 UMLS CUIs distinct to VSAC
i.e.
C0190233, Coronary artery bypass with autogenous graft, three grafts

Overview of Methods

- **Establishing lists of value sets for clinical research (CDISC) and healthcare (MU)**
  - CDISC provided SDTM value set.
    - 573 value sets; 20,132 total codes; 12,891 distinct codes
  - VSAC: Retrieved full value set expansions for the 05/05/2017 release of CMS eCQM Value Sets using VSAC API.
    - 3,606 extensional value sets; 605,522 total codes; 389,539 distinct codes
- **Mapping codes (i.e. SCUIs) in CDISC and VSAC value sets to UMLS**
  - Use UMLS API to map CDISC SCUIs to UMLS CUIs
  - Use UMLS API to map VSAC SCUIs to UMLS CUIs
- **Characterizing semantics of concepts in CDISC and MU value set**
  - Use UMLS API to map each CUI to a semantic type. Then, mapped each CUI to one of fifteen Semantic Groups using the “Semantic Group File” provided by MetaMap
- **Comparing value sets between CDISC and VSAC**
  - Use R, SQL to calculate Jaccard similarity scores (intersection/union) and inclusion scores
- **Evaluating gaps and similarities**
  - Qualitatively evaluate discrepancies above/below certain thresholds

Results: UMLS Mappings

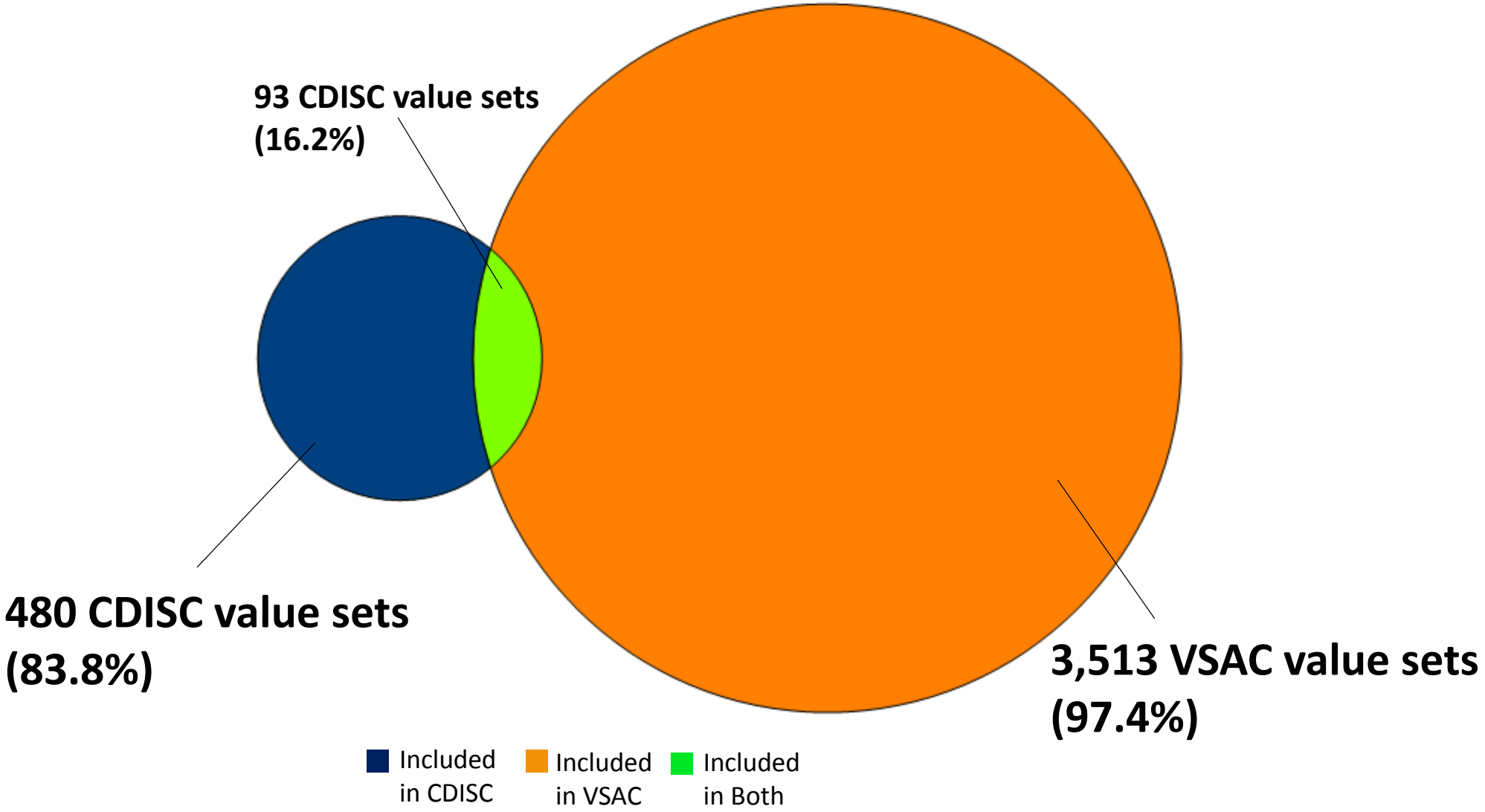
CDISC

- Only ~92% of the CDISC SCUIs could be mapped to a UMLS CUI
  - Sampling the 8% not mapped shows provisional codes added to NCI thesaurus in 12/2016.
  - NCI terminology was not updated in the 2017AA (May) release of UMLS.
  - NCI codes are scheduled to be updated in the forthcoming November 2017AB version.

VSAC

- 99.8% of VSAC SCUIs were mapped to a UMLS CUI
  - Of the 590 VSAC SCUIs not mapped to a UMLS CUI, ~93% were RxNorm codes, and ~7% were NCI codes. 1 code belonged to the CVX terminology.
  - 1.25% of VSAC SCUIs mapped to more than 1 UMLS CUIs

2. Coverage of CDISC by VSAC



- 93/573 (16.2%) CDISC value sets share at least 1 UMLS CUI with a VSAC value set.

2. Coverage of CDISC by VSAC, Jaccard

CDISC_VS_ID	CDISC_VS_Name	VSAC_OID	VSAC VS Name	VSAC_CUI_Count	CDISC_CUI_Count	Intersect	Union	Jaccard	Jaccard Mod	IS
C66726	Pharmaceutical Dosage Form	2.16.840.1.113883.3.88.12.3221.8.11	Medication Product Form	164	170	151	183	0.82513661	0.82294792	0.1726372
C66729	Route of Administration Response	2.16.840.1.113883.3.88.12.3221.8.7	Structured Product Labeling Drug Route of Administration	125	130	114	141	0.80851064	0.80566876	0.16888889
C74457	Race	2.16.840.1.113883.3.2074.1.1.3	Race Category Excluding Nulls	6	6	5	7	0.71428571	0.65465367	0
C99074	Directionality	2.16.840.1.113883.3.2074.1.1.24	Anatomical Site Modifier	25	30	22	33	0.66666667	0.65443321	0.4
C74457	Race	2.16.840.1.114222.4.11.836	Race	7	6	5	8	0.625	0.57282196	0.27777778
C66790	Ethnic Group	2.16.840.1.114222.4.11.837	Ethnicity	2	5	2	5	0.4	0.30983867	1
C66729	Route of Administration Response	2.16.840.1.113762.1.4.1018.98	Route of Administration	148	130	77	201	0.38308458	0.38108933	0.08598015
C66732	Sex of Participants Response	2.16.840.1.113883.3.2074.1.1.14	Biological Sex	4	4	2	6	0.33333333	0.25819889	0
C66790	Ethnic Group	2.16.840.1.114222.4.11.877	Detailed Ethnicity	3	5	2	6	0.33333333	0.25819889	0.33333333
C66731	Sex	2.16.840.1.113883.3.2074.1.1.14	Biological Sex	4	5	2	7	0.28571429	0.22131333	0.1
C124307	Treatment Intent	2.16.840.1.113762.1.4.1116.232	Palliative Intent	1	4	1	4	0.25	0.1118034	1
C124304	Subject Status Response	2.16.840.1.113762.1.4.1116.294	Unknown Result	1	4	1	4	0.25	0.1118034	1
C78738	Never/Current/Former Classification	2.16.840.1.113883.3.464.1003.124.11.1038	Confirmed as Current	1	4	1	4	0.25	0.1118034	1
C101848	Risk Assessment	2.16.840.1.113762.1.4.1116.294	Unknown Result	1	5	1	5	0.2	0.08944272	1
C66731	Sex	2.16.840.1.113762.1.4.1116.294	Unknown Result	1	5	1	5	0.2	0.08944272	1
C66742	No Yes Response	2.16.840.1.113762.1.4.1116.294	Unknown Result	1	5	1	5	0.2	0.08944272	1
C66790	Ethnic Group	2.16.840.1.113762.1.4.1116.294	Unknown Result	1	5	1	5	0.2	0.08944272	1

- 573 CDISC value sets * 3,606 VSAC value sets = 2,066,238 comparisons
- Of these, 461 comparisons had 1 or more CUIs in common between value sets
- 17 comparisons had Jaccard > 0.2!

	Jaccard	Jaccard.Mod	IS
median	0.00395257	0.00234143	0.096995
mean	0.03541273	0.02342295	0.2333919
SE.mean	0.00430013	0.00377937	0.0142513
CI.mean.0.95	0.00845034	0.00742696	0.0280058
var	0.00852441	0.00658475	0.0936292
std.dev	0.09232774	0.08114644	0.3059889
coef.var	2.60719082	3.46439934	1.3110522

Revisit Research Questions

(1) What are the semantic characteristics of the concepts used in CDISC and VSAC value sets? *CDISC and VSAC value sets have different semantic profiles.*

(2) To what extent do existing value sets in the VSAC represent value sets in CDISC? *Barely*.

New Question:

(3) Can we create a surrogate source of value sets that wouldn't already exist in the VSAC – by using the UMLS to represent CDISC value sets from standard terminologies (e.g., SNOMED CT or RxNorm)?

Example, Surrogate Coverage

CDISC Value Set ID: C74456
(VS Name: Anatomical Location)

CDISC Code: C33301
Name: PERINEUM

Mapping via UMLS API

UMLS CUI: C0031066

Retrieve 'Source Concept' via UTS API

SNOMEDCT_US	38864007
SNOMEDCT_US	261157007
SNOMEDCT_US	38864007

3. Surrogate Value Set and CUI Coverage

- # of distinct Source CUIs from CDISC: 12,890
- # of distinct CUIs from UMLS: 12,014

	# of CDISC Value Sets Covered (%)	# of UMLS CUIs covered in CDISC (%)
SNOMED CT	133 (23.2%)	3491 (29.1%)
LOINC	130 (22.7%)	1066 (8.9%)
ICD-9-CM	14 (2.4%)	227 (1.9%)
ICD-10-CM	10 (1.7%)	28 (0.2%)
CPT	10 (1.7%)	227 (1.9%)
RxNorm	7 (1.2%)	126 (1.0%)
HCPCS	2 (0.3%)	5 (0.04%)
ICD10PCS	1 (0.2%)	1 (0.01%)

CDISC vs. Surrogate, Coverage by SNOMED

CDISC_VS_ID	VS_Name	CDISC_Code_Count	CDISC_CUI_Count	Count_CUI_Null	Intersect.cnt.SNOMED	Coverage % SNOMED
C111108	Employment Status	4	4	0	4	1
C116111	SDTM Species	118	118	0	116	0.983050847
C66770	Units for Vital Signs Results	15	15	0	14	0.933333333
C85491	Microorganism	1495	1487	1	1380	0.927419355
C99074	Directionality	33	30	0	27	0.9
C99073	Laterality	8	8	0	7	0.875
C66781	Age Unit	6	6	0	5	0.833333333
C66731	Sex	5	5	0	4	0.8
C66728	Relation to Reference Period	8	8	0	6	0.75
C119013	Ophthalmic Focus of Study Specific Interest	4	4	0	3	0.75
C66733	Size Response	4	4	0	3	0.75
C66769	Severity/Intensity Scale for Adverse Events	4	4	0	3	0.75
C74561	Skin Type Response	4	4	0	3	0.75
C78738	Never/Current/Former Classification	4	4	0	3	0.75
C74456	Anatomical Location	833	824	5	609	0.734620024
C95120	Physical Properties Test Name	11	11	0	8	0.727272727
C95121	Physical Properties Test Code	11	11	0	8	0.727272727
C78734	Specimen Type	85	85	0	61	0.717647059
C101852	Sudden Death Syndrome Type	7	7	0	5	0.714285714
C99078	Intervention Type Response	10	10	0	7	0.7
C66741	Vital Signs Test Code	31	31	0	21	0.677419355
C67153	Vital Signs Test Name	31	31	0	21	0.677419355
C101834	Normal Abnormal Response	6	6	0	4	0.666666667
C74457	Race	6	6	0	4	0.666666667
C125923	BRIDG Activity Mood	3	3	0	2	0.666666667
C78737	Relationship Type	3	3	0	2	0.666666667
C71113	Frequency	73	73	0	48	0.657534247
C124299	Biospecimen Characteristics Test Name	14	14	0	9	0.642857143
C124300	Biospecimen Characteristics Test Code	14	14	0	9	0.642857143
C71148	Position	16	16	0	10	0.625
C99075	Portion/Totality	8	8	0	5	0.625
C66729	Route of Administration Response	130	130	0	80	0.615384615
C66742	No Yes Response	5	5	0	3	0.6
C90013	ECG Lead	21	21	0	12	0.571428571
C127262	Environmental Setting	16	16	0	9	0.5625
C101858	Procedure	60	60	0	33	0.55
C66726	Pharmaceutical Dosage Form	170	170	0	90	0.529411765
C124309	Tumor or Lesion Properties Test Result	12	12	0	6	0.5
C76348	Marital Status Response	10	10	0	5	0.5
C101843	Coronary Artery Disease Presentation	6	6	0	3	0.5
C101865	Acute Coronary Syndrome Presentation Cate	6	6	0	3	0.5
C124304	Subject Status Response	4	4	0	2	0.5
C124307	Treatment Intent	4	4	0	2	0.5
C66732	Sex of Participants Response	4	4	0	2	0.5
C101815	Eastern Cooperative Oncology Group Perforn	2	2	0	1	0.5
C101816	Eastern Cooperative Oncology Group Perforn	2	2	0	1	0.5
C116105	Pharmacogenomics Findings Test Name	2	2	0	1	0.5
C116106	Pharmacogenomics Findings Test Code	2	2	0	1	0.5
C66789	Not Done	2	2	0	1	0.5

	Coverage % SNOMED
median	0.28571429
mean	0.37670022
SE.mean	0.02216883
CI.mean.0.95	0.04385214
var	0.0653638
std.dev	0.25566345
coef.var	0.67869206

CDISC vs Surrogate, Coverage by LOINC

CDISC_VS_ID	VS_Name	CDISC_Code_Count	CDISC_CUI_Count	Count_CUI_Null	Intersect.cnt.LNC	Coverage % LNC
C111108	Employment Status	4	4	0	4	1
C102580	Laboratory Test Standard Character Result	6	6	0	5	0.833333333
C66742	No Yes Response	5	5	0	4	0.8
C76348	Marital Status Response	10	10	0	8	0.8
C101848	Risk Assessment	5	5	0	4	0.8
C119013	Ophthalmic Focus of Study Specific Interest	4	4	0	3	0.75
C66733	Size Response	4	4	0	3	0.75
C124304	Subject Status Response	4	4	0	3	0.75
C66732	Sex of Participants Response	4	4	0	3	0.75
C116107	Death Details Test Name	4	4	0	3	0.75
C116108	Death Details Test Code	4	4	0	3	0.75
C95120	Physical Properties Test Name	11	11	0	8	0.727272727
C95121	Physical Properties Test Code	11	11	0	8	0.727272727
C101834	Normal Abnormal Response	6	6	0	4	0.666666667
C74457	Race	6	6	0	4	0.666666667
C125923	BRIDG Activity Mood	3	3	0	2	0.666666667
C78737	Relationship Type	3	3	0	2	0.666666667
C124299	Biospecimen Characteristics Test Name	14	14	0	9	0.642857143
C124300	Biospecimen Characteristics Test Code	14	14	0	9	0.642857143
C99073	Laterality	8	8	0	5	0.625
C66731	Sex	5	5	0	3	0.6
C99078	Intervention Type Response	10	10	0	6	0.6
C66790	Ethnic Group	5	5	0	3	0.6
C114119	Pharmacogenomics Biomarker Medical Statem	5	5	0	3	0.6
C90013	ECG Lead	21	21	0	12	0.571428571
C128690	Ethnicity As Collected	14	14	0	8	0.571428571
C99074	Directionality	33	30	0	17	0.566666667
C78734	Specimen Type	85	85	0	46	0.541176471
C66781	Age Unit	6	6	0	3	0.5
C78738	Never/Current/Former Classification	4	4	0	2	0.5
C99075	Portion/Totality	8	8	0	4	0.5
C66789	Not Done	2	2	0	1	0.5
C120990	Model for End Stage Liver Disease Clinical Clas	2	2	0	1	0.5
C120991	Model for End Stage Liver Disease Clinical Clas	2	2	0	1	0.5
C124305	Subject Status Test Code	2	2	0	1	0.5
C124306	Subject Status Test Name	2	2	0	1	0.5

	Jaccard.LNC
median	0.25
mean	0.3110918
SE.mean	0.02117406
CI.mean.0.95	0.0418934
var	0.05828431
std.dev	0.24142143
coef.var	0.77604565

Research Question 3

- Can we create a surrogate source of value sets that wouldn't already exist in the VSAC – by using the UMLS to represent CDISC value sets from standard terminologies?

  Better than VSAC; but, still not great.

Conclusions & Implications

- VSAC/MU value sets mainly cover clinical concepts of interest such as diagnoses, drugs, procedures, and not many administrative concepts.
- CDISC value sets essentially cover administrative concepts, and a small subset of disorders and procedures.
- Interestingly, there are a number of value sets for questionnaires, functional assessments, experience scales etc. in CDISC with little or no coverage by LOINC nor SNOMED CT. One suggestion is for LOINC to look into these and include them in future versions, if appropriate.
- Currently, there are no code samples provided for the VSAC API. One outcome of this study, is that we can provide Perl code sample.

- One limitation of this study is that the UMLS annual update process can lead to significant discrepancies between source terminologies and those stored in the UMLS.