# Interoperability between Value Sets for Clinical Research and Healthcare: Mapping Value Sets between the Clinical Data Interchange Standards Consortium (CDISC) and Meaningful Use

**Raja A. Cholan, BS[1], Olivier Bodenreider, MD, PhD[2]**
**[1]Oregon Health & Science University, Portland, OR; [2]National Library of Medicine, Bethesda, MD**

**Introduction:** Historically, there have been substantial – but, independent – efforts to standardize medical information for clinical research data and those for healthcare data. For instance, in the biomedical research domain, the Clinical Data Interchange Standards Consortium (CDISC) was developed for data exchange in clinical trials used by the Food and Drug Administration and pharmaceutical companies. CDISC uses codes and terms from clinical terminologies from the National Cancer Institute (NCI) thesaurus. The Value Set Authority Center (VSAC) maintains value sets for clinical quality measures used in value-based programs like Meaningful Use (MU). The VSAC value sets use codes and terms from standard terminologies (e.g., SNOMED CT, RxNorm, LOINC) found in most Electronic Health Record (EHR) systems.

With EHR adoption becoming nearly ubiquitous, researchers want to leverage EHRs to conduct pragmatic trials in which they can counteract recruitment challenges, burdensome data collection, and uncertain generalizability of results. Therefore, there is a need for interoperability between healthcare data (VSAC) and clinical research data (CDISC). In previous work, we and others have studied value sets found in the VSAC. Within biomedical research, meanwhile, there is a lack of methodological literature about value set engineering, particularly in CDISC controlled terminologies. In this work, our objective is to assess semantic interoperability between CDISC value sets and MU value sets. Our study questions were (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?;(3) Can we create a surrogate source of value sets that wouldn't already exist in the VSAC – by using the Unified Medical Language System (UMLS) Metathesaurus to represent CDISC value sets from standard terminologies?

**Methods:** We sought to identify gaps and similarities between clinical research value sets and healthcare quality value sets. First, we gathered the lists of value sets from CDISC (using CDISC's FTP website) and those from VSAC (using the VSAC API). Then, we mapped the codes from each source to a concept unique identifier (CUI) from the UMLS (using the UMLS API). Then, for the CUIs representing the CDISC value sets, we created surrogate value sets using associated atoms from standard terminologies in the UMLS (using the UMLS API). Next, we used the Jaccard index (intersection/union) to compare the overlap between the terminologies. Finally, we evaluated the alignments obtained by the two sources to determine what value sets and concepts in CDISC were not covered by value sets and concepts in standard terminologies.
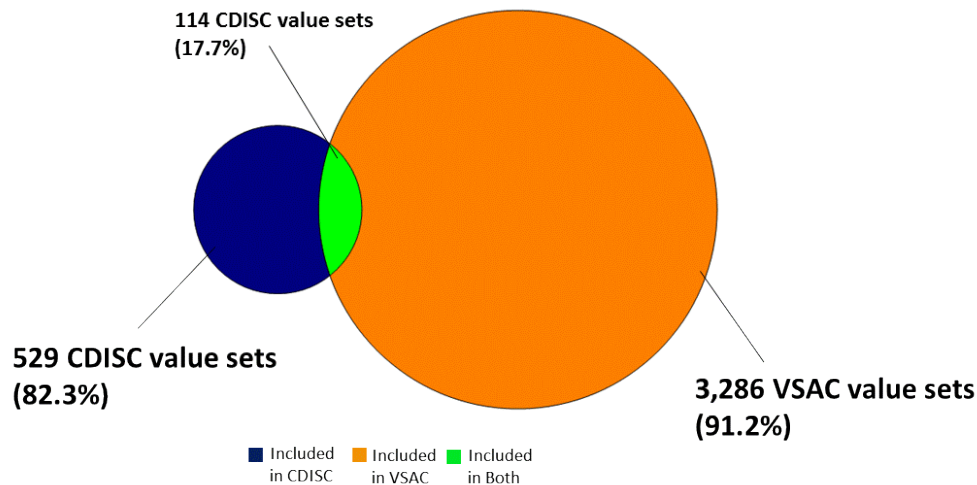
**Results:**

*1. Value Set Semantic Profiles.* The distribution of semantic groups found in CDISC concepts is significantly different from that in VSAC concepts. The CDISC value sets mostly represent administrative concepts, and a small subset of procedures and disorders. The VSAC value sets mostly represent clinical concepts, such as disorders, drugs, procedures, and very few administrative concepts. Figure 1 displays the value set semantic profiles in CDISC and VSAC.

**Figure 1. CDISC and VSAC S**emantic Profiles
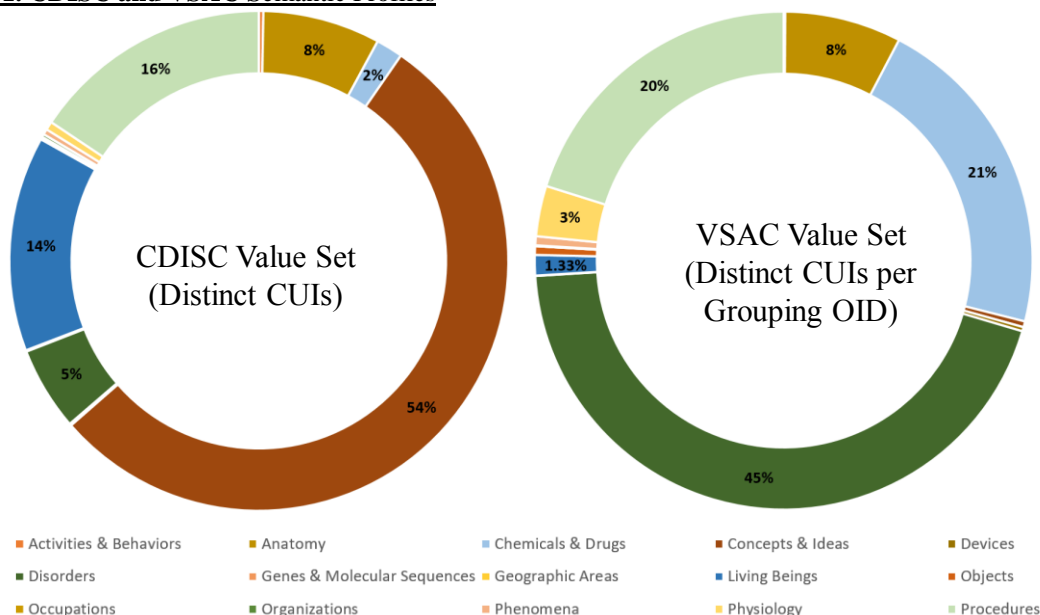
*2. Coverage of CDISC by VSAC.* ~92% of CDISC codes could be mapped to a UMLS CUI. Sampling the 8% not mapped shows provisional codes added to NCI thesaurus in 12/2016. In VSAC, meanwhile, 99.8% of the source codes were mapped to a UMLS CUI. We computed the coverage of CDISC by VSAC by analyzing the count of CDISC value sets that shared at least 1 UMLS CUI with a corresponding VSAC value set. 114/643 (17.7%) CDISC value sets share at least 1 UMLS CUI with a VSAC value set with a mean Jaccard index of 0.028.

**Figure 2. Coverage of CDISC by VSAC**



114 CDISC value sets (17.7%)

529 CDISC value sets (82.3%)

3,286 VSAC value sets (91.2%)

■ Included in CDISC    ■ Included in VSAC    ■ Included in Both

*3. Surrogate Value Set Coverage.* We were able to create surrogate value sets to cover CDISC value sets from SNOMED CT, LOINC, ICD-10-CM, ICD-9-CM, CPT, RxNorm, HCPCS, and ICD10PCS terminologies. The surrogate value sets comprised of SNOMED CT codes covered 178 (28%) CDISC value sets with a mean Jaccard index of 0.37, and LOINC surrogate value sets covered 130 (26%) CDISC value sets with a mean Jaccard index of 0.29. See Table 1 for a full list of the number of CDISC value sets covered by each standard terminology, and the number of UMLS CUIs from CDISC value sets covered by each surrogate value set. The counts of surrogate value sets represent value sets which covered an existing CDISC value set for at least 1 UMLS CUI.

**Table 1. Counts of CDISC value sets covered by surrogate value sets from standard terminologies**

|  | # of CDISC Value Sets Covered (%) | # of UMLS CUIs covered in CDISC (%) |
|---|---|---|
| **SNOMED CT** | 178 (27.7%) | 4315 (26.4%) |
| **LOINC** | 130 (25.7%) | 1271 (7.8%) |
| **ICD-10-CM** | 15 (2.3%) | 119 (0.7%) |
| **ICD-9-CM** | 14 (2.6%) | 84 (0.5%) |
| **CPT** | 10 (1.6%) | 227 (1.4%) |
| **RxNorm** | 7 (1.1%) | 126 (0.8%) |
| **HCPCS** | 2 (0.3%) | 5 (0.03%) |
| **ICD10PCS** | 1 (0.2%) | 1 (0.01%) |

**Conclusions:** VSAC/MU value sets mainly cover clinical concepts of interest such as diagnoses, drugs, procedures, and not many administrative concepts. CDISC value sets, meanwhile, essentially cover administrative concepts, and a small subset of disorders and procedures. Only about 18% of CDISC value sets can be represented by a VSAC value set with a low mean Jaccard index. Surrogate value sets created from SNOMED CT could only represent 28% of CDISC value sets, while LOINC surrogate value sets represented 26% of value sets; both SNOMED CT and LOINC surrogate value sets had higher (but less than ideal) Jaccard scores than those found in the coverage by VSAC value sets.

The overall interoperability between CDISC and MU was limited at best. Interestingly, there are a number of value sets for questionnaires, functional assessments, experience scales etc. in CDISC with little or no coverage by LOINC or SNOMED CT. One suggestion is for LOINC and SNOMED CT to look into these and include them in future versions, if appropriate.