Development of the EDDA Study Design Terminology: Enhancing Retrieval of Clinical and Bibliographic Records in Dispersed Repositories

Ashleigh Faith, MLS, CRM, Eugene Tseytlin, MS, Tanja Bekhuis PhD, MS, MLS, AHIP

1 School of Library and Information Sciences & 2 Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, US

BACKGROUND

Medical terminology varies across disciplines and reflects linguistic differences in communities of clinicians, researchers, and indexers. Inconsistency of terms for the same concepts and lack of machine-readable metadata impede discovery of information artifacts, such as records of clinical reports and scientific articles that reside in various repositories. To facilitate discovery, retrieval, and data sharing, the medical community maintains an assortment of terminologies, thesauri, and ontologies. Valuable resources include the US National Library of Medicine Medical Subject Headings (MeSH), Elsevier Life Science thesaurus (Emtree), and the National Cancer Institute Thesaurus (NCIT). It is increasingly important to identify medical investigations by their design features, as these have implications for evidence regarding research questions.

OBJECTIVE

Bekhuis et al (2013) found that coverage of study designs was poor in MeSH and Emtree. Further investigation found that the EDDA, Emtree, and NCIT covered fewer than 30% of the 199 classes currently identified in the EDDA Study Design Terminology.

The EDDA Group at the University of Pittsburgh is further developing a terminology of study designs. In addition to randomized controlled trials, it includes terms for observational or nonrandomized designs.

METHODS

Inconsistent entry points, semantic labels, synonyms, and definitions are common.

The EDDA Group is adding hierarchy to the terminology based on lexical patterns and domain expertise.

The EDDA Study Design Terminology is freely available in the NCBO BioPortal. In the portal, users may browse the terminology or download the latest version of our terminology.

Annotiation Results. N=2,500 term annotations were recorded, including 53 MeSH, 45 NCIT, and 27 Emtree access points to EDDA Study Design terms. Both MeSH and NCIT access points enabled extraction of information. However, information from 12 of 27 (44%) Emtree access points could not be extracted. Because NCIT cross-references other controlled vocabularies, 44 Unified Medical Language System (UMLS) resources also contributed to term annotations. N=126 definitions, 5 instances of incomparability, 21 legacy terms, 102 semantic type annotations, and 1340 term variants were recorded.