HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2017 December 01.

Published in final edited form as:

Clin Trials. 2016 December ; 13(6): 671–676. doi:10.1177/1740774516653238.

Improving the value of clinical research through the use of Common Data Elements (CDEs)

Jerry Sheehan[1], **Steven Hirschfeld**[2], **Erin Foster**[1], **Udi Ghitza**[3], **Kerry Goetz**[4], **Joanna Karpinski**[1], **Lisa Lang**[1], **Richard P. Moser**[5], **Joanne Odenkirchen**[6], **Dianne Reeves**[5], **Yaffa Rubinstein**[7], **Ellen Werner**[8], and **Michael Huerta**[1]

[1]National Library of Medicine, National Institutes of Health, USA

[2]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, USA

[3]National Institute on Drug Abuse, National Institutes of Health, USA

[4]National Eye Institute, National Institutes of Health, USA

[5]National Cancer Institute, National Institutes of Health, USA

[6]National Institute of Neurological Disorders and Stroke, National Institutes of Health, USA

[7]National Center for Advancing Translational Sciences, National Institutes of Health, USA

[8]National Heart Lung and Blood Institute, National Institutes of Health, USA

Abstract

The use of Common Data Elements (CDEs) can facilitate cross study comparisons, data aggregation and meta-analyses, simplify training and operations, improve overall efficiency, promote interoperability between different systems, and improve the quality of data collection. A CDE is a combination of a precisely defined question (variable) paired with a specified set of responses to the question that is common to multiple datasets or used across different studies. CDEs, especially when they conform to accepted standards, are identified by research communities from variable sets currently in use or are newly developed to address a designated data need. There are no formal international specifications governing the construction or use of CDEs. Consequently, CDEs tend to be made available by research communities on an empiric basis.

Some limitations of Common Data Elements are that there may still be differences across studies in the interpretation and implementation of the Common Data Elements, variable validity in different populations, and inhibition by some existing research practices and the use of legacy data systems. Current National Institutes of Health efforts to support Common Data Element use are linked to the strengthening of National Institutes of Health Data Sharing policies and the investments in data repositories. Initiatives include cross-domain and domain-specific resources, construction of a Common Data Element Portal, and establishment of trans-National Institutes of Health working groups to address technical and implementation topics. The National Institutes of

Corresponding Author: Mike Huerta, National Library of Medicine, 38 Center Drive, Rm 2S20, MSC-3811, Bethesda, MD, 20894-3811, USA mike.huerta@nih.hhs.gov.

Health is seeking to lower the barriers to Common Data Element use through greater awareness and encourage the culture change necessary for their uptake and use. As National Institutes of Health, other agencies, professional societies, patient registries, and advocacy groups continue efforts to develop and promote the responsible use of Common Data Elements, particularly if linked to accepted data standards and terminologies, continued engagement with and feedback from the research community will remain important.

Keywords

Common data elements; data collection; data sharing; interoperability; data standards

Consistency in data collection is a fundamental principle of scientific research in general and clinical trials in particular. In any given study, each opportunity for data collection is expected to meet specifications independent of time, location, or people involved. While consistency of data collection within an individual study is essential for maintaining data quality and enabling analysis, consistency of data collection across multiple studies brings additional value. As biomedical research becomes more data-intensive, and as policy and practices promote increased data-sharing, greater scientific opportunities emerge from the comparison and secondary use of biomedical research data. Data sharing to support the combination of data across data sets for strengthening inferences and performing new analyses is rapidly becoming a general expectation.

Absent a unifying framework for all biomedical information and the concurrent existence of multiple ontologies, each serving different purposes, the linking and convergence of collected data occurs in niches and pockets of activity. One empiric approach for achieving consistency in data collection within and across research studies is the use of Common Data Elements.

What are Common Data Elements?

The term “Common Data Element” was initially developed by Silva and Wittes in 1999 for Case Report Forms used in National Cancer Institute clinical trials, and has continued to evolve.[1] As used currently, a “Common Data Element” is a combination of a precisely defined question (variable) paired with a specified set of responses to the question that is common to multiple datasets or used across different studies..[2] The primary context for CDEs is in research where precision, reproducibility, and cross-study comparison are priorities. A CDE can stand alone as a single variable, or may be included in a structured collection of elements such as a multi-item scale or index or a complex case report form.[3]

One critical characteristic of CDEs is the use of a defined value set, where, for a question that is designated as a variable for data collection, the permissible responses are restricted to a fixed list. For example, if the variable is current pregnancy status, the fixed value set could be limited to Yes or No. If the variable is type of brain tumors that are gliomas of the highest grade, the fixed value set could be, based on current classifications, glioblastoma multiforme, gliosarcoma, or gliomatosis cerebri.

For some CDEs, precision in defining the method of assessment may be part of the specification. For example, if a CDE for a clinical study is defined as the result of an immunoassay, the CDE may specify the specific way in which the assay is to be conducted. For example, with the enzyme-linked immunospot (ELISPOT) assay to detect either antibody or cytokine secretion, results of several studies show ELISPOT results vary from laboratory to laboratory but can be harmonized through rigorous training, quality assurance, and quality control measures.[4, 5] Defining the specifications for the ELISPOT assay, perhaps including the need for a central laboratory, and other parameters will produce much greater value for a single study and for any collection of studies than just achieving consensus on the use of ELISPOT as a specific outcome measure. The principle of, when appropriate, defining the acceptable methods as well as the concept for a CDE, can improve the value and utility.

In practice, CDEs are identified by research communities from variable sets currently in use or are newly developed to address a designated data need. CDE development and selection is an iterative process guided by feasibility, utility, and acceptability that benefits from multiple stakeholders including clinicians, informaticists, terminologists, statisticians, patients and others. CDEs that are specified using standardized vocabularies, codesets and terminologies can ease the burden of data collection, data exchange and promote discovery and interoperability between systems, including patient registries and electronic health records.

There are no formal international specifications governing the construction or use of CDEs. Consequently, CDEs tend to be made available by research communities on an empiric basis.

What is the value of CDEs?

CDE use has some advantages within a single study if they are perceived and implemented as a standard or specification. CDEs can provide consistency and efficiency in establishing data collection infrastructure and minimize variability in training and implementation. Consequently the use of CDEs can increase the efficiency, quality, clarity, and reproducibility of the overall research process and results.

CDEs can be used to design the logic of data collection, can be embedded in case report forms, patient registries, and integrated into collected and analytic datasets. CDEs can be expressed in machine readable formats to be used in data analytic plans and structured routines and scripts to incorporate the CDE variables.

Enhanced value of CDEs is across studies to pool and combine data for meta-analyses, modeling and post hoc construction of synthetic cohorts for exploratory analyses. CDEs can also be a tool to link data sets and examine relationships even if there is not a one to one mapping across all data elements in multiple data sets. CDEs can be used to link and aggregate variables across multiple datasets by identifying the CDEs and pulling the associated values into a new hybrid analytic dataset. CDEs can also be used to map associations across datasets. Well constructed and implemented CDEs increase the precision

and can eliminate the errors that come with other methods such as ad hoc transformations, conversion and manual linking.

CDEs that are used in multiple studies are a tool to leverage the substantial investment made to collect quality data from clinical trials by increasing the consistency of data collection across studies. The use of CDEs, especially when they conform to accepted standards, can facilitate cross study comparisons, data aggregation and meta-analyses, simplify training and operations, improve overall efficiency, promote interoperability between different systems, and improve the quality of data collection.

What are challenges to CDE adoption and use?

Despite its potential benefits, adoption and use of CDEs across clinical research studies face several challenges. First, while bringing greater standardization to research data collection, there may still be differences across studies in the interpretation and implementation of the data elements. Thus pooling and merging data may appear to be feasible based on variable names and even the value sets that apply to those variables. However, unless the criteria for assigning values are consistently and uniformly applied, the validity of such an operation may be compromised and the resulting conclusions weak.

Caution must also be taken to ensure that CDEs are valid in the different populations that may be recruited for a particular study. Many CDE collections, for example, make use of specific data collection instruments that have been validated in specific populations. Using them in populations for which they have not been validated can mean that the results are not truly comparable with those derived from studies done in populations for which they have been validated. This can be particularly challenging in international studies and national studies that recruit participants with different cultural and linguistic backgrounds.

In addition, by facilitating the use of clinical research data beyond the original purpose (study) for which it was collected, use of CDEs can exacerbate concerns about privacy and confidentiality. Researchers who wish to combine participant-level data from multiple studies, for example, must ensure that the use of the data is consistent with the informed consent under which the data were collected. They must also ensure that the combination of data from multiple sources does not undermine privacy protections by facilitating reidentification of human subjects. To a large extent, these concerns are similar to those involved in any effort to combine data from multiple pre-existing studies, but use of CDEs makes such combinations easier and more reliable.

As the trend to consolidated oversight for human research protection progresses with the revision of the Common Rule, the parochialism of multiple and disparate Institutional Review Boards is being replaced by centralized or federated models.[6, 7] These newer consolidated models provide the opportunity for consistent informed consent and policy regarding data sharing. A federated model has been proposed for data sharing among health care provider information systems that relies on several principles including transparency, representation, and local benefit.[8] A system where access to patient level data is controlled

through data access boards or committees that screen requests and evaluate the relative merits and risks is a resource intensive but workable solution.

Adoption of CDEs can also be inhibited by some existing research practices and legacy data systems. Although CDEs are often designated by research communities based on expert consensus, their use may entail changes in pre-existing approaches to data collection by those individual researchers and research institutions that have collected a certain type of data in another way (e.g. have used a different instrument than the one designated for assessing mental health). It also entails changes in the way researchers design and develop case report forms, to ensure that they incorporate designated CDEs rather than developing their own specifications for data collection. Researchers need to be aware of CDEs relevant to their research and strive to incorporate them inlace with this paragraph:

What are some NIH CDE related activities?

The National Institutes of Health (NIH) is taking steps to promote the use of CDEs, taking into account their associated benefits and challenges. For more than 20 years, NIH Institutes and Centers have worked to develop and identify CDEs for use in a variety of research domains, but these efforts have intensified in recent years as clinical research has become more data-centric and opportunities for data sharing have increased.[9]

Recognizing cross-domain patterns and needs across the NIH community and beyond, NIH supports initiatives that transcend the conventional domains of individual NIH programs. Examples of these cross-cutting initiatives include patient-reported outcomes (Patient Reported Outcomes Measurement Information System or PROMIS®), phenotypic and exposure measures (Phenotypex and eXposures or PhenX), and neurological and behavioral function (NIH Toolbox).

In addition to these broadly applicable CDE efforts individual NIH components have developed CDE collections that are targeted to particular disorders or research projects or topics of interest within their respective missions. These CDEs cover domains such as cancer, neurological disorders, ophthalmic disease, and substance abuse. Beyond these research-oriented CDEs, there are CDEs that have been developed for patient registries, specifically the NIH Global Rare Disease Patient Registry Data Repository.[10] Examples of NIH-supported CDE resources are in Table 1.

NIH also supports resources that contribute to the formulation and use of CDEs. This includes terminology sets, metadata registries, and tools for collecting and selecting amongst CDE options. For example, the use of terminology-based tools that construct content are part of the cancer Data Standards Registry and Repository, a metadata registry from the National Cancer Institute. The growing maturity of the use and implementation of CDEs is evident in the development of data repositories specifically designed to capture data from studies or patient registries which use CDEs to facilitate the secondary use of the collected data. These NIH resources include the Federal Interagency Traumatic Brain Injury Research Informatics System, the National Database for Autism Research, and the Database of Genotype and Phenotype.

NIH is also improving the coordination and communication of CDE efforts across NIH and beyond. Much of this work is led by the trans-NIH Biomedical Informatics Coordinating Committee's CDE Working Group. The CDE Working Group and its members have contributed to the Office of the National Coordinator for Health Information Technology's Structured Data Capture initiative to identify standards for creation and exchange of data elements between case report forms use in clinical research and electronic health records used in clinical care. It has also coordinated NIH's participation in the Coalition for Accelerating Standards and Therapies initiative to develop standards for reporting clinical trial data in 60 high-priority therapeutic areas designated by the Food and Drug Administration, an initiative that engages the global pharmaceutical industry. CDE Working Group has also engaged with the European Union's Core Outcome Measures in Effectiveness Trials initiative, which promotes the development and application of agreed standardized sets of outcomes in all clinical trials of a specific condition. Future efforts will engage standards organizations such as Health Level 7 and the Clinical Data Interchange Standards Consortium.

What is NIH doing to encourage responsible CDE use?

NIH is taking steps to encourage and facilitate the use of CDEs across the research community. To raise the visibility of NIH-supported CDE collections, the CDE Working Group launched a NIH CDE Resource Portal (http://cde.nih.gov) in January 2013. The Portal provides a single point of entry for information about NIH CDE collections, resources, tools and related standards. The Portal is not a source for distributing CDEs, but links to websites and repositories with detailed information about each initiative. The Portal also can be used to compare the subject areas or domains addressed by CDEs in each collection and should be a first stop for those interested in NIH CDEs.

NIH also launched a prototype NIH CDE Repository in 2015 (http://cde.nlm.nih.gov). This platform offers an infrastructure for both searching for existing CDEs and for assembling new CDE collections and developing new CDEs in a manner that is both parsimonious – avoiding duplication of effort and promoting the reuse of existing CDEs – and transparent – using versioning and inclusion of provenance. The repository supports efforts to harmonize CDEs by providing tools that identify similar CDEs and consolidating them where possible. In addition the repository contains several standardized assessment instruments (from which some CDEs have been derived) and has the ability to represent case report forms.

Several NIH programs have taken steps to encourage uptake and use of CDEs through their funded research. The National Institute of Drug Abuse strongly encourages the use of the Substance Abuse and Addiction Collection (part of the PhenX Toolkit) in human subject research it supports.[11] In 2015, almost 40 active NIH Funding Opportunity Announcements explicitly call for the use of CDEs in NIH-funded research.12 Several Funding Opportunity Announcements issued by the National Human Genome Research Institute for genomewide association studies direct investigators to use PhenX measures.[12] Investigators funded under any of several National Institute of Neurologic Disease and Stroke programs are expected to use the institute’s CDEs, with those funded for work on progression of chronic traumatic encephalopathy required to use the institute’s CDEs for Traumatic Brain Injury.[13] Additional

Funding Opportunity Announcements from other programs directed at international research are currently targeted for issuance.

There is emerging evidence that these efforts are promoting use of CDEs. As of June 2015, more than 90 articles identified in Pubmed had been published that cite the use of PhenX measures, and some 448 published articles describe the use or development of PROMIS measures. The National Institute of Neurologic Disease and Stroke has funded more than 25 clinical trials that make use of the institute's CDEs, and 47 of its funded grants under Funding Opportunity Announcements that encourage CDE use have generated more than 270 publications.[12] Since the early 1990s the National Cancer Institute has used CDEs in the data collection portion of their enterprise clinical trials activities; since 2002 the common use of CDEs was formally adopted in the intramural program. While much of this use is in research settings, there is also evidence of CDE use in clinical care.[9]

Current NIH efforts to support CDE use are linked to the strengthening of NIH Data Sharing policies and the investments in data repositories. For examples see https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html. An expectation to provide a data sharing plan in funding applications, develop the plan during the funded project, and implement the data sharing plan shortly after project completion will all contribute to the acceptance and use of CDEs as part of more general initiative to disseminate scientific data that are interoperable.

Conclusion

Through the development of new CDE resources and the inclusion of recommendations or terms and conditions that encourage or require the use of CDEs, the NIH is seeking to lower the barriers to CDE use through greater awareness and encourage the culture change necessary for their uptake and use. As more clinical studies make use of CDEs and more data sets that use CDEs become available, the opportunities and risks for comparing data across studies and pooling data from multiple studies will grow, and the incentives for other researchers to use CDEs will become stronger. Additional incentives may come as researchers recognize the ability to ask new research questions that can be answered by drawing on the use of CDEs across research disciplines. As NIH, other agencies, professional societies, patient registries and advocacy groups continue efforts to develop and promote the responsible use of CDEs, particularly if linked to accepted data standards and terminologies, continued engagement with and feedback from the research community will remain important. As CDEs are used more broadly, the resources needed to deliver high quality data will become more efficient and the ability to leverage a data intensive environment will continue to improve, ultimately benefitting science and patients.

Acknowledgments

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Silva J, Wittes R. Role of clinical trials informatics in the NCI's cancer informatics infrastructure. Proc AMIA Symp. 1999:950–954. [PubMed: 10566501]

2. National Institutes of Health. [accessed 9 September 2015] What is a CDE?. 2015. http://www.nlm.nih.gov/cde/glossary.html#cdedefinition

3. National Institutes of Health. [accessed 9 September 2015] Glossary. 2015. http://www.nlm.nih.gov/cde/glossary.html

4. Cox JH, Ferrari G, Kalams SA, et al. Results of an ELISPOT proficiency panel conducted in 11 laboratories participating in international human immunodeficiency virus type 1 vaccine trials. AIDS Res Hum Retroviruses. 2005; 21:68–81. [PubMed: 15665646]

5. Janetzki S, Panageas KS, Ben-Porat L, et al. Results and harmonization guidelines from two large-scale international Elispot proficiency panels conducted by the Cancer Vaccine Consortium (CVC/SVI). Cancer Immunol Immunother. 2008; 57:303–315. [PubMed: 17721781]

6. Check DK, Weinfurt KP, Dombeck CB, et al. Use of central institutional review boards for multicenter clinical trials in the United States: a review of the literature. Clin Trials. 2013; 10:560–567. [PubMed: 23666951]

7. Slutsman J, Hirschfeld S. A federated model of IRB review for multi-site studies: a report on the national children's study federated IRB Initiative. IRB: Ethics Hum Res. 2014; 36:1–6.

8. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. Nat Biotechnol. 2015; 33:360–363. [PubMed: 25850061]

9. Long Range Planning Committee. Translating cancer research into cancer care: final report of the Long Range Planning Committee. In: Silva, JS.; Ball, MJ.; Douglas, JV., editors. Cancer informatics essential technologies for clinical trials. New York: Springer; 2002. p. 5-16.

10. Rubinstein Y, McInnes P. NIH/NCATS/GRDR Common Data Elements: A leading force for standardized data collection. Contemp Clin Trials. 2015; 20:78–80.

11. Ghitza UE, Gore-Langton RE, Lindblad R, et al. NIDA clinical trials network Common Data Elements initiative: advancing big-data addictive-disorders research. Front Psychiatry. 2015; 6:33. [PubMed: 25784882]

12. National Institutes of Health. [accessed 9 September 2015] Guidance to encourage the use of CDEs. 2015. http://www.nlm.nih.gov/cde/policyinformation.html

13. National Institutes of Health. [accessed 9 September 2015] NINDS Common Data Elements. 2015. https://commondataelements.ninds.nih.gov

Table 1
Examples of NIH Common Data Elements Resources cited in the text

Resource	Description	URL
PROMIS	Patient Reported Outcomes Measurement Information System, of highly reliable, precise measures of patient–reported health status for physical, mental, and social well–being.	http://www.nihpromis.org
PhenX	Consensus measures for Phenotypes and eXposures	https://www.phenxtoolkit.org/
NIH Toolbox	NIH Toolbox is a multidimensional set of brief measures assessing cognitive, emotional, motor and sensory function from ages 3 to 85 years	http://www.nihtoolbox.org/
GRDR	The Global Rare Diseases Patient Registry Data Repository program is to develop information from different registries for rare diseases	https://ncats.nih.gov/grdr
caDSR	cancer Data Standards Registry and Repository comprises tools and resources to develop and implement reusable metadata that describe common data elements (CDEs), information models, and case-report forms (CRFs)	https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/metadata-and-models#caDSR
NDAR	National Database for Autism Research is a data repository that aims to accelerate research through	https://ndar.nih.gov/

Resource	Description	URL
	data sharing, data harmonization, and the reporting of research results	
FITBIR	The Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System was developed to share data across the entire Traumatic Brain Injury research field	https://fitbir.nih.gov/
dbGAP	The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and study results related to the interaction of genotype and phenotype in Humans	Http://www.ncbi.nlm.nih.gov/gap