Journal of the Society for Clinical Data Management

Jentoft K, et al. CDISC Implementation in an Academic Research Organization. *Journal of the Society for Clinical Data Management*. 2022; 2(3): 3, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.164

OPINION PAPER

CDISC Implementation in an Academic Research Organization

Katie Jentoft[*], Eric Tustison[*] and Hong Yu[†]

Introduction: The United States Food and Drug Administration (FDA) requirement for standardized data submissions led our Academic Research Organization (ARO) to use CDISC data standards in clinical trials since January 2018. Implementing CDISC data standards effectively enables standardized data collection and facilitates data submissions to the FDA.

Objectives: The objective of this paper is to illustrate the positives and negatives of our ARO's three-phased implementation of CDISC data standards, inclusive of partially automated dataset conversion, CDASH case report forms, and Pinnacle 21 data checks. Our ARO shares our experience to support other organizations in standardizing their data for FDA submissions.

Methods: Our ARO went through three phases of CDISC data standardization implementation: phase one – application of CDISC SDTM conversion to non-standardized datasets, phase two – utilization of CDASH case report forms, phase three – leveraging ongoing Pinnacle 21 data checks to identify data issues.

Results: Phase one required significant time to create a standardized dataset upon study conclusion. Phase two required additional resources for start-up activities but proportionally reduced the overall effort to produce the final dataset. Phase three required investment upon start-up and ongoing targeted data review but aims to reduce the production cost of the final standardized dataset.

Conclusion: This evolution of CDISC data standards implementation refined our standardization process to meet FDA requirements, streamlining data collection and overall efficiency of clinical trials. We support collaborations to develop open-source training materials and examples of CDISC data standards implementation to improve the standardization process for other AROs.

Keywords: Manage Clinical Research Data; Collect data; Define Data; Design Form; Process Data; Clean data

Introduction

The Data Management group at the Sean M. Healey & AMG Center for ALS (amyotrophic lateral sclerosis) and Neurological Clinical Research Institute at Massachusetts General Hospital is an academic research organization (ARO) responsible for data management of multicenter clinical trials. We strive to effectively adhere to the United States Food and Drug Administration (FDA) requirements. Particularly with Clinical Data Interchange Standards Consortium (CDISC) standards for data submissions,[1–2] our team's best practices have evolved from reactive to proactive as we developed standardization for multiple trials. We followed an iterative approach, modifying existing processes, developing new tools, and working with CDISC experts. By sharing our experiences, we hope to foster collaboration and assist other AROs facing similar challenges.

Background

The FDA requirement for standardized data submissions prompted our ARO's use of CDISC data standards in clinical trials since January 2018.[1–3] The currently supported data standards require CDISC Study Data Tabulation Model (SDTM) for clinical data, which "provides a standard for organizing and formatting data to streamline processes in collection, management, analysis and reporting."[4] This model is used for the final dataset submitted to the FDA.[3] There are multiple ways to create a compliant dataset. The first way is to convert non-standardized data to standardized study data.[3] A second way to create a compliant dataset is to use CDISC Clinical Data Acquisition Standards Harmonization (CDASH) to collect standardized data from the beginning of a study, easing the conversion to SDTM.[5] Further, the FDA's Pinnacle 21 data checks can be run on CDASH-compliant data collection to support ongoing data review.[6] Our ARO's experiences with each of these three approaches illustrate the evolution of our

[*] NCRI, Mass General Hospital

[†] Mass General Hospital

Corresponding author: Katie Jentoft (KJENTOFT@mgh.harvard.edu)

standardization procedures. This article will discuss each approach, along with lessons learned and ideas for future improvements.

Methods

Phase one: Early experience with data conversion to SDTM format

One of our early experiences with CDISC data standards was supporting a clinical trial with more than 100 participants started in 2017. We designed the electronic Case Report Forms (eCRFs) according to our internal standards following Good Clinical Data Management Practices but did not use specific CDISC recommendations.[7] As we neared study completion, we developed a plan to convert the collected data into the SDTM format. This became a multi-team project with Data Managers (DMs), Systems Analysts (SAs), and an external CDISC consultant working together to achieve the compliant format. The overall goal was to convert the existing dataset to SDTM and then to create the analysis datasets using the Analysis Data Model (ADaM) and the Clinical Study Report (CSR).[8]

The DMs reviewed the data and issued queries to the sites to resolve discrepancies, ensuring the dataset was as clean as possible prior to conversion. DMs also provided advice to the SAs on SDTM mappings and specification questions. The SAs produced SDTM tables for the data and developed a proprietary data conversion tool to partially automate the process. The consultant mentored us on conversion questions, reviewing, and troubleshooting. The consultant received our dataset tables, exported them to the Pinnacle 21 data review validator,[9] and provided feedback. The consultant also helped prepare the final ADaM dataset and CSR. As multiple individuals worked on this effort over varying lengths of time, we performed a retrospective review of the study timeline to determine estimated hours required to convert the data into the SDTM format.

Phase two: Implementation of CDASH-compliant data collection methods

In 2018, we incorporated CDISC data standards during the design phase of our next trial by developing CDASH compliant data collection to ease the conversion of the final dataset to SDTM.[4–5] Data collection directly into SDTM format would be unwieldy primarily due to its vertical data structure. The CDASH standards document provides for one-to-one conversion to SDTM for many data fields and prescribes ways to bridge to SDTM when one-to-one conversion is not available.[10] For this trial, we began by identifying the common CDASH domains provided in the CDASH standards that we planned to use for data collection.[10] Then, using the domain query text recommendations and CDASH eCRF design principles,[10] we designed our eCRFs to include the relevant questions for our trial. CDASH questions that were optional and not relevant were excluded. Additional data points desired by the trial sponsor but not in the CDASH domain were either included as additional questions in the same eCRF or as separate eCRFs. For all eCRFs, the data collected had to be coded based on the CDASH standards to enable conversion to SDTM.[10] The coding rules are also prescribed in the CDASH standards and could be used as-is for fields taken directly from a domain or customized for original fields as long as the standard structure was maintained.[10] This specified approach to eCRFs—standard design principles, query text, and coding—significantly transformed how we developed the entire eCRF package and required more up-front work than legacy trials not following CDASH.

The second half of this experience involved converting the dataset to SDTM, which was completed by an external CDISC consultant. Our data conversion tool developed for the 2017 trial was not sufficient for the 2018 trial. Therefore, our ARO decided to outsource the SDTM conversion based on the time and resources available.

Phase three: Ongoing SDTM conversion and Pinnacle 21 checks throughout trial

In 2020, the third trial we intentionally managed with CDISC in mind followed CDISC data standards more holistically than the previous two trials. In addition to creating eCRFs with CDASH-compliant fields, we also developed a process for ongoing SDTM conversion throughout the trial. We provided SAS data exports every two weeks to an external CDISC consultant, which they used to create and update the SDTMs. After receiving each updated dataset, the consultant exported the SDTM output to the Pinnacle 21 data review validator and communicated back to us any data or structural issues, which we then worked to resolve.[9]

Results

For the first trial, based on retrospective review of study timeline, it took one full-time equivalent (FTE) approximately seven months, or 1,120 hours, to convert the data into SDTM format.

For the second trial, it took one FTE approximately two months, or 320 hours, to convert the data into SDTM format. The entire conversion effort was outsourced to an external CDISC consultant.

For the third trial, which is still in progress, SDTM conversion is ongoing. Producing the final dataset is expected to require less relative effort than the previous two trials, because we no longer need to complete the entire data conversion at the end of the trial. Specialized data review and data cleaning processes evenly distribute the preparation work for data conversion throughout the trial. While producing the final SDTM dataset is expected to take less time than in the previous trials, it is important to emphasize the significant time investment during study start-up and throughout the trial to prepare for SDTM.

Discussion

Over the course of these three trials, we developed expertise in CDISC data standards by gradually incorporating CDISC principles, query text, and coding over time, learning it was better to plan for standardization as early as possible in the trial management process rather than assuming it was something best left as part of trial closeout. From not using CDISC data standards prior to 2018 to incorporating

CDISC in the earliest trial design phases in 2020, our ARO has come a long way in implementing standardization.

In our 2017 trial, it took significantly more time and effort than an average non-standardized trial to produce a final standardized dataset. As this was our first SDTM conversion, there was a lot to learn while doing the project. Fundamentally we discovered we could not convert all data points successfully as our eCRF data collection tools were not designed to facilitate SDTM. For example, free text data that had to be parsed would often be placed in the Comments (CO) domain in SDTM, which is not as easily accessible for analysis as pre-defined SDTM fields. Additionally, some of the issues identified by the Pinnacle 21 report could not be resolved and had to be explained in the Study Data Reviewer's Guide.

The limitations imposed by the initial design proved to be challenging and a key area for improvement in subsequent studies.

In our 2018 trial, we reduced the overall time and effort to produce a similar size dataset, although this approach also required more work to be done at study start-up. This experience gave us a strong understanding of how to design custom eCRFs to build robust and compliant data collection tools. Additionally, the eCRFs we developed for common CDASH domains could be reused in other trials. The SDTM conversion was smoother and more efficient than in the first trial due to the decision to outsource the entire process. Our main challenge in this experience was delayed data cleaning, because Pinnacle 21 data validation was not run while the trial was in progress. We did not fully understand the necessity of Pinnacle 21 and did not allocate resources toward it until SDTM conversion. At that point, we could not implement Pinnacle 21 proactively.

In our 2020 trial, additional effort at start-up was now expected, and ongoing targeted data cleaning and conversion efforts required more work than studies that do not have these processes in place. The overall data cleaning process for CDISC compliance has become more effective, with internal logic checks run on the raw data in real-time and frequent Pinnacle 21 data validation. Because issues can be identified soon after they occur, they can be resolved before they worsen or develop into problematic patterns. We estimate the final effort to produce a standardized dataset upon study closeout will be minimal, because most standardization will have already been completed.

One benefit of our most recent experience implementing concurrent SDTM conversion and Pinnacle 21 checks while a trial was ongoing was to identify eCRF data points that did not convert cleanly to SDTM variables. While the eCRF fields were developed according to CDISC data standards, the actual data entered did not always convert well to SDTM. For example, the general CDISC principle to avoid blank fields did not work well for recording adverse event (AE) outcome dates for ongoing AEs. The original design of our AE eCRF required an outcome date for ongoing AEs to document when the assessment was made that the AE was "recovering/resolving" or "not recovered/not resolved." However, we learned from the Pinnacle 21 output that best practice was for AE outcome date to be left blank if the outcome was ongoing, as outcome dates for ongoing events did not convert to SDTM outputs.

The results of ongoing Pinnacle 21 checks also led us to revise eCRF completion guidance to clinical research personnel, which improved the consistency and quality of data entered in the electronic data capture system. As the entire study team aligned to follow CDISC data standards from the first moment of data collection, we pivoted away from traditional wide-sweeping data cleaning methods prior to database lock. Instead, we focused our attention with laser precision on key fields and critical data flow in nearly real-time to fully support the goal of SDTM conversion. This made it a seamlessly integrated step in trial management rather than an awkward burden at the end of a trial. Overall, this iterative process led to improved data quality for the trial through real-time data cleaning that led to more accurate interim analyses and deepened our understanding of CDISC-compliant design for implementation in future trials.

In our efforts to achieve data standardization, we learned the hard way through missed opportunities. We identified areas that needed improvement too late in the process to benefit our early trials. However, these experiences proved to be invaluable for understanding how to revise our processes for subsequent trials to achieve CDISC compliance. Based on our experiences implementing CDISC data standards, we feel there is a real need for AROs to have comprehensive and continuous CDISC training. Ideally it would be broken down into bite-sized pieces, with practice material and many detailed examples. Online resources similar to W3Schools for SQL training,[11] which is highly interactive and easy to reference on the Web, would be hugely beneficial for organizations of all sizes. For example, an online module could display a sample eCRF and prompt for conformant CDASH field annotations; the module could autodetect deviations from CDASH annotation principles and display a correct alternative. It would also be less overwhelming than a day or week of formal CDISC instruction from an expert, as it takes ongoing practice to fully understand the principles and goals of these standards. While having expert-led CDISC training can be a great place to start, it would be cost-prohibitive to contract an expert on retainer to answer all the questions that inevitably arise during CDISC implementation, especially for small organizations just getting started with CDISC. Additionally, while we appreciate the extent of CDISC reference material freely available online, we wish it was easier to understand which CDISC documents are needed for which tasks. A virtual look-up tool or visual schematic would help, such as a quick start guide that provides a high-level view with guidance on where to go for more detailed information. Open-source training in both technical and design principles would be key to help all users, especially those who are learning CDISC for the first time.

Reflecting on the results of the upfront work to implement CDISC compliance, our organization saved increasing amounts of time in the preparation of the datasets for analysis during our three phases of CDISC data standardization implementation. It was progressively easier and faster to finalize the second dataset compared to the first, and the third dataset is poised to continue this trend. Because datasets can be finalized more quickly due to CDISC preparation, analysis can also begin more quickly. However, the amount of time spent on data analysis is independent of the time spent on data preparation. Therefore, the absolute analysis time is not affected positively or negatively by using CDISC.

Because data standardization leads to faster data preparation, our experience as an ARO leads us to advocate for required standards for National Institutes of Health (NIH) data sharing. CDISC is a strong contender for data standards, given its widespread use for clinical trial data submitted to the FDA. The challenge is that most of the institutions running NIH-funded studies do not necessarily have the resources to create CDISC-compliant datasets. While it seems redundant to create separate standards, perhaps another standard would be simpler or more cost-effective to implement than CDISC while still enabling NIH studies to achieve standardized data.

We believe it would benefit the research community dramatically if we converted all existing CRFs to standards. In general, the research landscape has changed significantly with the COVID-19 pandemic. Many people are using big data, artificial intelligence, or machine learning in health care research. Having data standards is critical for data aggregation and effective analysis of large datasets. With increased standardization, more knowledge can be derived more quickly than in the past, ultimately leading to new treatments for devastating diseases and improved health care for everyone.

Conclusion

Overall, the third approach we took to CDISC implementation is the experience we would recommend based on our experiences so far: starting with CDISC data standards in mind from the earliest stages of database development, using CDASH, and running SDTM conversion and Pinnacle 21 checks concurrently with active data collection. Given the resources we had for each trial, we made the best decisions we could to produce CDISC-compliant datasets. Each experience helped refine our understanding and influenced our data management processes for future trials. For an ARO to proactively implement CDISC data standards, we advocate for open-source educational resources and ongoing community discussion to enable standardization for all clinical trials.

Competing Interests

The authors have no competing interests to declare.

References

1. **U.S. Department of Health & Human Services/ U.S. Food & Drug Administration.** Study Data Standards: What You Need to Know. Published September 2017. https://www.fda.gov/media/98907/download. Accessed January 4, 2022.
2. **CDISC.** cdisc.org. https://www.cdisc.org. Accessed May 20, 2022.
3. **U.S. Department of Health & Human Services/ U.S. Food & Drug Administration.** Study Data Technical Conformance Guide: Technical Specifications Document. Published July 2020. https://www.fda.gov/media/136460/download. Accessed April 11, 2022.
4. **CDISC.** SDTM. https://www.cdisc.org/standards/foundational/sdtm. Accessed April 14, 2022.
5. **CDISC.** CDASH. https://www.cdisc.org/standards/foundational/cdash. Accessed April 14, 2022.
6. **Pinnacle 21.** Pinnacle21.com. https://www.Pinnacle21.com. Accessed April 14, 2022.
7. **GCDMP.** scdm.org. https://scdm.org/gcdmp/. Accessed April 14, 2022.
8. **ADaM.** cdisc.org. https://www.cdisc.org/standards/foundational/adam. Accessed April 14, 2022.
9. **Pinnacle 21.** Downloads. https://www.Pinnacle21.com/downloads. Accessed February 17, 2022.
10. **CDISC.** CDASH v1.1. Published January 18, 2011. https://www.cdisc.org/system/files/members/standard/foundational/cdash/cdash_std_1_1_2011_01_18.pdf. Accessed April 15, 2022.
11. **W3schools.** SQL Tutorial. https://www.W3Schools.com/sql/default.asp. Accessed February 17, 2022.

How to cite this article: Jentoft K, Tustison E, Yu H. CDISC Implementation in an Academic Research Organization. *Journal of the Society for Clinical Data Management*. 2022; 2(3): 3, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.164

Submitted: 18 February 2022 **Accepted:** 28 November 2022 **Published:** 23 December 2022

JSCDM *Journal of the Society for Clinical Data Management* is a peer-reviewed open access journal published by Society for Clinical Data Management. OPEN ACCESS