Improving Usability, Safety and Patient Outcomes with Health Information Technology
F. Lau et al. (Eds.)

doi:10.3233/978-1-61499-951-5-125

Development of Data Validation Rules for Therapeutic Area Standard Data Elements in Four Mental Health Domains to Improve the Quality of FDA Submissions

Maryam GARZA[a,1], Emel SEKER[a], Meredith ZOZUS[a]
[a] *University of Arkansas for Medical Sciences, Little Rock, Arkansas*

Abstract. Data standards are now required for many submissions to the United States Food and Drug Administration (FDA). The required standard for submission of clinical data is the Clinical Data Interchange Standards Consortium (CDISC) Submission Data Tabulation Model (SDTM). Currently, 45 business rules and 115 associated validation rules exist for SDTM data. However, such rules have not yet been developed for therapeutic area data standards developed under the last reauthorization of the Prescription Drug User Fee Act (PDUFA V). The objective of this effort was to develop data validation rules for new therapeutic area data standards in four mental health domains, assess the metadata required to associate such rules with standard data elements, and assess the level of data validation possible for therapeutic area data elements.

Keywords. data standards, therapeutic area standards, validation rules, CDISC SDTM, regulatory submissions

1. Introduction

With acceptance of risk-based approaches in clinical trials, the era of expecting zero-defect data and the correspondingly high expenditures in data cleaning is coming to a close [1-3]. As a result, data cleaning activities in clinical research are becoming more targeted towards data needed for study endpoints [4]. At the same time, advances in data standards for regulatory submission in the United States are aimed at standardizing data sufficiently for use of software to facilitate the regulatory review process.

Data standards are now required for submission of data from clinical and nonclinical studies as of December 17, 2016 [5]. In particular, the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) is required for clinical study data. To assist the regulated industry with data submission, the Food and Drug Administration (FDA) has published 45 business rules and 115 validator rules that check that the study data are conformant to the standard and will support regulatory review and analysis [6]. These business and validator rules, however do not yet exist for efficacy data, but could be developed as part of therapeutic area data standards.

[1] Maryam Garza, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, 4301 W Markham St., Little Rock, AR.

Detection and resolution of data discrepancies (called data cleaning) during a study should be informed by submission data standards. The extent of data cleaning for a study should depend on the scientific and operational needs of a study. In particular, data collection and cleaning scope should be defined by those activities necessary to ensure that the data are capable of supporting study conclusions [7]. This is operationalized through establishing acceptance criteria for data error and designing data collection, processing, and control procedures that consistently produce data meeting the criteria [4].

Thus, while we report methodology for developing data cleaning rules along with standardized data elements, these are of a different sort than FDA business and validator rules. While over time some may prove useful and be desired by reviewers, we currently recommend their use on any particular study only to the extent that they are needed to assure that data are capable of supporting research conclusions. We specifically do not include rules toward pursuit of zero defects or out of concerns from the past that a single defect would call an entire submission into question and delay the review process [7]. Rather, we offer them for use when necessary to meet the scientific aims of a study and based on standards to facilitate efficient sharing and implementation.

It has been shown from first principles that delaying detection and resolution of discrepancies foregoes opportunity to resolve discrepancies in some cases, renders some discrepancies unresolvable, and increases the cost of resolving others [7-9]. Thus, any such rules, especially those to which data must conform for regulatory submission, should be implemented as far upstream in the data collection and management process as possible [4].

2. Background

A rule is a statement of a condition against which data are evaluated. For example, *"lab values from a complete blood count can't be negative"* or *"measured physical quantities captures as percentages must range between zero and one"*. In clinical research data management, such rules are referred to as query rules, edit checks, or discrepancy checks.

The earliest reports of data processing in clinical research included accounts of rules-based data cleaning [10-18]. In the therapeutic development industry, rules-based data cleaning has occurred on most if not all studies [4]. In fact, fear that notice of an errant data value would substantially delay a regulatory submission prompted the practice of developing and running often hundreds of rules for a clinical study. On older studies, the clinical investigational site was contacted in attempts to resolve each discrepancy against the source, often the medical record [7]. The disposition and resolution of each discrepancy was tracked from origination to resolution. The discrepancies often numbered in the thousands for a small study of a few hundred patients. It was not uncommon for 10-30% of the cost of a clinical study to be spent on data cleaning and monitoring [2].

With widespread use of web-based electronic data capture (EDC) software and associated processes, efficiencies have been gained and data discrepancies are usually communicated to sites upon data entry where they can be resolved quickly and where tracking is automated. Using EDC systems, data discrepancies are usually communicated to sites upon data entry where they can be resolved quickly and where tracking is automated. The process continues to rely on development and use of rules. Similar to clinical decision support rules in medical informatics, in the absence of standards data

models and associated controlled terminology or standard data elements, rules could not be widely shared or reused, leaving untapped inefficiency.

With respect to regulatory decision making, the Center for Drug Evaluation and Research (CDER) receives more than 150,000 submissions each year, adding up to millions of data values considered in regulatory decision-making. The FDA has been supportive as the regulated industry organized to develop standards through CDISC. The FDA looked toward further data standardization to facilitate handling such large volumes of data and under the Prescription Drug User Fee Act (PDUFA) to improve the efficiency of the review process required data standardization. In 2010, CDER established the Data Standards Program with the goal of standardizing efficacy data not yet tackled under the CDISC SDTM.

As part of the CDER Data Standards Program, four therapeutic area data standards have been developed in the mental health domain: Schizophrenia, Major Depressive Disorder (MDD), Bipolar Disorder (BPD), and Generalized Anxiety Disorder (GAD). Briefly, candidate data elements were identified from data collection forms from recent marketing applications and NIMH-funded studies. The initial set of data elements was consolidated. Unique data elements were defined and then vetted by clinical experts, regulatory authorities, professional societies, and informatics experts. Each set of data elements was then represented in Unified Modeling Language (UML) use case and activity diagrams and class models. The four therapeutic area standards were balloted through Health Level Seven (www.hl7.org, HL7), an ANSI-accredited standards development organization, and after passing ballot were published as HL7 standards. Once published, the standards were provided to CDISC for use in Therapeutic Area User Guides (TAUGs) to support the evaluation of marketing applications submitted to the FDA for drug development and clinical trials (one TAUG per domain). The TAUGs were developed under the Coalition for Accelerating Standards and Therapies (CFAST) initiative.

At the time of this publication, CDISC had already developed the therapeutic area user guides (TAUGs) for two of the four domains, Schizophrenia and MDD [19-24]. The existing user guides are provisional standards that demonstrate how to represent therapeutic area specific data using the CDISC foundational standards (www.cdisc.org).

Data checking rules have not traditionally been a part of these standards but are desired by the regulators and regulated industry alike. Further, existence of standard data elements and common data models in which they are structured, enables definition and sharing of data checking rules to accompany the new standards and assist in submission and use of data submitted in the standards.

3. Methods

Standard data elements for the four therapeutic areas were mapped to the CDISC SDTM Implementation Guide (SDTMIG) version 3.2 [25] and SDTM version 1.4 [26]. Mapping began with the Schizophrenia and MDD data elements, as the TAUGs had already been developed and were able to be referenced for mapping to SDTM. Independent reviews of the data elements and their subsequent mappings to SDTM were conducted by two study members. Each reviewer manually compared the standard data elements against the respective TAUGs using a simple search, mapping those identified in the user guides. This was done by comparing the definitions to determine if they were semantically related. Data elements that did not directly map to one of the standard SDTM variables

were represented in the Supplemental Qualifiers (SUPPQUAL) domain and associated back to the parent record in one of the general, observational domains. SUPPQUAL is an extension mechanism utilized for representing and relating data values not accommodated in existing domains [25].

While the user guides did offer insight into where to best store the therapeutic area-specific data elements within the CDISC models, not all data elements were covered in the TAUGs and no direct mapping document was provided for Schizophrenia. Supplementary documentation was provided for MDD with mapping examples for several common data elements (shared across the four therapeutic areas) and a number of MDD-specific data elements to the CDASH model, which offered suggestions for mapping to SDTM; but, again, not all data elements were covered. Leveraging the methodology implemented for Schizophrenia and MDD mappings, the BPD and GAD data element mapping followed suit.

Upon completion of the mapping, data validation rules (or edit checks) were written according to the Good Clinical Data Management Practices (GCDMP) against the therapeutic area-specific data elements and the SDTM with the intent of identifying all possible relationships that could be leveraged for data validation. The edit checks aim to detect inconsistencies in the data or potential data errors, which will ultimately improve the quality of the data [4]. The complete list of therapeutic area-specific data elements was reviewed to determine those most necessitating checks. The foundational SDTM model and its existing standards' data elements were also considered when developing the rules. The rules were written using ANSI standard SQL (American National Standards Institute, Structured Query Language), the de facto standard for relational databases.

4. Results

In total, 415 data elements were mapped. Of the 415 total data elements, 215 (51.8%) mapped to general observation classes and 200 (48.2%) mapped to special-purpose domains. A total of 41 data elements were shared across all four models (Figure 1). Several additional data elements, while they may not have been common across all four models, were shared between two or three of the four. For example, among the 85 total Schizophrenia data elements, only 26 were unique to Schizophrenia (Table 1).

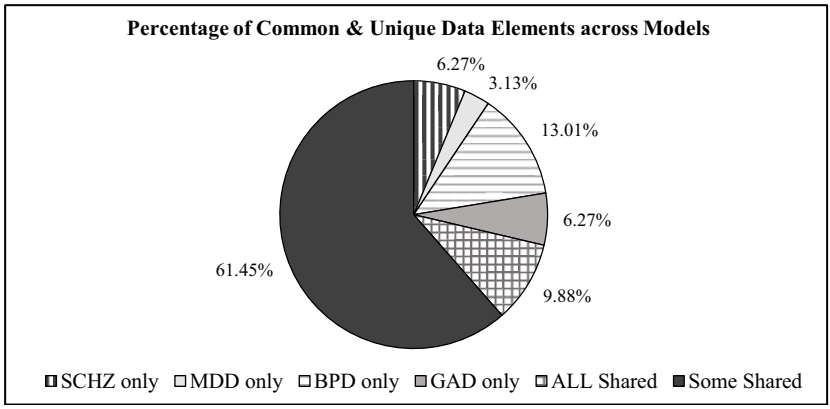

Figure 1. Percentage of common & unique data elements across MH models.

Table 1. Data element counts and percentages per therapeutic area data model

	Schizophrenia (N = 85)		MDD (N = 94)		BPD (N = 144)		GAD (N = 92)	
	n	%	n	%	n	%	n	%
Therapeutic Area Specific	26	30.6	13	13.8	54	37.5	26	28.3
Common with *any* Model	59	48.2	81	43.6	90	28.5	66	44.6
Common in *all* Models	41	69.4	41	86.2	41	62.5	41	71

Upon completion of the mapping, a total of 371 rules were developed on 191 individual data elements across the four therapeutic areas. The rules were classified into three categories: range checks (1.1%), logical inconsistencies (56.0%), and missing values (42.9%). Approximately one-third (30.4%) of the checks were written against date fields to verify consistency across other date fields or fields with data dependencies.

On average, 79.1% of the data elements had validation rules that were written against common data elements (those common across the therapeutic areas) versus 20.9% against unique, therapeutic area-specific data elements. For each therapeutic area, rules were written against a total of 49 Schizophrenia data elements, 41 for MDD, 54 for BPD, and 47 for GAD (Table 2). Less than one-third (27.1%) of the Schizophrenia data elements had rules programmed against other SDTM fields external to the Schizophrenia standard data element set. MDD, BPD, and GAD each had similar results: 27.5%, 26.5%, and 15.9%, respectively.

Table 2. Validation rules written against common data elements versus unique data elements per therapeutic area compared to full data element list. (DEs = data elements)

	Schizophrenia (N = 85)		MDD (N = 94)		BPD (N = 144)		GAD (N = 92)	
	n	%	n	%	n	%	n	%
Rules against Common DEs	33	67.4	40	97.6	44	81.5	34	72.3
Rules against Unique DEs	16	32.6	1	2.4	10	18.5	13	27.7
Total DEs with Rules	49	25.7	41	21.4	54	28.3	47	24.6

5. Discussion

It was anticipated, and confirmed, that many of the general observations would map to the Medical History (MH) and Disposition (DS) domains, given the nature of the therapeutic area data elements and the typical data collected during clinical trials. This would likely translate across other therapeutic areas, as in many cases, the data points of interest tend to align with those two domains: the MH domain captures historical data relevant to the study endpoints (i.e., prior and concomitant conditions), while the DS domain captures most of the data relevant to the study milestones [26]. It was also predicted that at least one-third would not map directly to an SDTM general observation domain and would require mapping to SUPPQUAL, based on previous experience with data element mappings to the common data models such as SDTM [27]. We predict that this would also be the case in other therapeutic areas, but realize that this could vary based on the complexity of and data points of interest for a particular therapeutic area.

As SDTM does not allow for the creation of new variables, the SUPPQUAL domain is used to capture additional data elements (or "additional Qualifiers for an observation") that do not "fit" within the current set of standard variables within the general observation

classes [26,27]. These variables are then associated back to parent records within a general observation class using a domain identifier.

With the mapping and validation rule development came a series of challenges. As previously mentioned, the Schizophrenia and MDD TAUGs were referenced in order to complete the mappings. However, not all data elements were covered in the TAUGs, nor were there direct SDTM mapping documents for either model. Approximately 20.0% of the Schizophrenia data elements were not explicitly mapped in the Schizophrenia TAUG, whereas close to 60.0% of the MDD elements were not mapped in the MDD TAUG (although some were covered by what had been mapped in Schizophrenia, as these models shared common data elements). On average, 13.8% of the data elements had not been mapped in either TAUG, which required that the SDTM and the SDTMIG be referenced.

Furthermore, a few of the common data elements shared by both Schizophrenia and MDD were not mapped consistently within the TAUGs. As several of these elements were also shared by BPD and GAD, a decision needed to be made as to which SDTM variable to map to so as to allow for consistency across all models (and to allow for standard query rules for common data elements). Challenges were also met when developing the validation rules for data elements that were mapped to the SUPPQUAL domain. Data elements requiring multivariate rules in which two or more data elements were from the SUPPQUAL domain, complicated the structure of the query due to the nature of the table generated for SUPPQUAL elements.

A limitation of this effort is that both the mappings and the validation rules have only been validated internally. However, as the mappings leveraged existing TAUGs and preliminary mappings from CDISC, we are confident in the accuracy of the mappings. Additionally, the validation rules have only been developed and written in the SQL code, but the rules have yet to be programmed and tested/validated. Currently, both the mappings and validation rules have been submitted to CDISC for review and collaboration between both teams continues in an effort to complete validation. This external review may result in changes to the mappings and/or validation rules. However, as nearly 90.0% of the Schizophrenia and MDD data elements had been previously mapped by CDISC in the TAUGs, and since the BPD and GAD elements were either common or similar in structure, it is anticipated that any changes or updates to the mappings would be minimal.

The collaboration with CDISC continues and TAUG development for the BPD and GAD models is underway; the mappings from this effort will be leveraged for their development. Additionally, continued FDA engagement is planned so that the team responsible for implementing the validation rules have also had the opportunity to review the rules and provide feedback. It is critical for all three groups be in sync as the mappings and the TAUG development will greatly affect the validation rules. The final rule set will be turned over to the FDA for implementation and dissemination to industry. It is recommended that a continuous feedback loop be maintained as the rules are implemented and executed for continuous quality improvement.

6. Conclusion

Standardized data elements and validation rules can improve the quality of data that is submitted by sponsors for regulatory decision-making. Validation rules accompanying standard data elements support sponsors in checking data consistency as early as possible

in the data collection process, a clear best practice. Existence of such rules can decrease the cost of data management and increase the quality of data submitted to the FDA.

References

[1] United States Food and Drug Administration (FDA) Guidance for Industry Oversight of Clinical Investigations —A Risk-Based Approach to Monitoring. OMB Control No. 0910-0733, U.S. FDA, August 2013.

[2] E.L. Eisenstein, P.W. Lemons, B.E. Tardiff, K.A. Schulman, M.K. Jolly, and R.M. Califf, Reducing the costs of phase III cardiovascular clinical trials, *Am. Heart J.* **149** (2005) 482–488. doi:10.1016/j.ahj.2004.04.049.

[3] E.L. Eisenstein, R. Collins, B.S. Cracknell, O. Podesta, E.D. Reid, P. Sandercock, Y. Shakhov, M.L. Terrin, M.A. Sellers, R.M. Califf, C.B. Granger, and R. Diaz, Sensible approaches for reducing clinical trial costs, *Clin Trials*. **5** (2008) 75–84. doi:10.1177/1740774507087551.

[4] Society for Clinical Data Management (SCDM), Good Clinical Data Management Practices (GCDMP). Society for Clinical Data Management (www.scdm.org). October 2013.

[5] United States Food and Drug Administration (FDA) Guidance for Industry Providing Regulatory Submissions In Electronic Format —Standardized Study Data. December 2014.

[6] United States Food and Drug Administration (FDA), Study Data Standards Resources web page available at https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm accessed October 20, 2017.

[7] Institute of Medicine (US) Roundtable on Research and Development of Drugs, Biologics, and Medical Devices; Davis JR, Nolan VP, Woodcock J, Estabrook RW, editors. Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report. Washington (DC): National Academies Press (US); 1999.

[8] M. Nahm, J. Bonner, P. L. Reed, and K. Howard, Determinants of accuracy in the context of clinical study data. Proceedings of the International Conference on Information Quality (ICIQ) November 2012. Available from http://mitiq.mit.edu/ICIQ/2012

[9] M.N. Zozus, The data book: collection and management of research data, CRC Press, Taylor & Francis Group, Boca Raton, 2017.

[10] W.H. Forrest, and J.W. Bellville, The use of computers in clinical trials, *British Journal of Anaesthesia*. **39** (1967) 311–319. doi:10.1093/bja/39.4.311.

[11] R.A. Kronmal, K. Davis, L.D. Fisher, R.A. Jones, and M.J. Gillespie, Data management for a large collaborative clinical trial (Cass: Coronary Artery Surgery Study), *Computers and Biomedical Research*. **11** (1978) 553–566. doi:10.1016/0010-4809(78)90034-4.

[12] G.L. Knatterud, Methods of quality control and of continuous audit procedures for controlled clinical trials, *Control Clin Trials*. **1** (1981) 327–332.

[13] N S.L. Norton, A.V. Buchanan, D.L. Rossmann, R. Chakraborty, and K.M. Weiss, Data entry errors in an on-line operation, *Comput. Biomed. Res.* **14** (1981) 179–198.

[14] A.E. Cato, G. Cloutier, and L. Cook, Data entry design and data quality (1985).

[15] A. Bagniewska, D. Black, K. Molvig, C. Fox, C. Ireland, J. Smith, and S. Hulley, Data quality in a distributed data processing system: the SHEP Pilot Study, *Control Clin Trials*. **7** (1986) 27–37.

[16] A.G. DuChene, D.H. Hultgren, J.D. Neaton, P.V. Grambsch, S.K. Broste, B.M. Aus, and W.L. Rasmussen, Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT), *Control Clin Trials*. **7** (1986) 34-45.

[17] I.K. Crombie, and J.M. Irving, An investigation of data entry methods with a personal computer, *Computers and Biomedical Research*. **19** (1986) 543–550. doi:10.1016/0010-4809(86)90028-5.

[18] S.P. Fortmann, W.L. Haskell, P.T. Williams, A.N. Varady, S.B. Hulley, and J.W. Farquhar, Community surveillance of cardiovascular diseases in the Stanford Five-City Project. Methods and initial experience, *Am. J. Epidemiol.* **123** (1986) 656–669.

[19] M. Zozus, M. Younes, and A. Walden, Schizophrenia Standard Data Elements. Health Level 7, Release 1. HL7 V3 DAM SCHIZ, R1_2014OCT Domain Analysis Model (DAM) Document of the HL7 Version 3 Domain Analysis Model: Schizophrenia, Release 1 - US Realm October 2014 HL7 Informative Document: HL7 V3 DAM SCHIZ, R1. A Technical Report prepared by Health Level Seven International and registered with ANSI: 10/26/2014.

[20] M. Zozus, M. Younes, C. Kluchar, J. Topping, and A. Walden, Major Depressive Disorder (MDD) Standard Data Elements. Health Level 7, Release 1. HL7 Version 3 Domain Analysis Model: Major Depressive Disorder, Release 1. HL7 V3 DAM MDD, R1_2014OCT Domain Analysis Model (DAM)

Document of the HL7 Version 3 Domain Analysis Model: Major Depressive Disorder, Release 1 - US Realm October 2014 HL7 Informative Document: HL7 V3 DAM MDD, R1. A Technical Report prepared by Health Level Seven International and registered with ANSI: 10/26/2014.

[21] Clinical Data Interchange Standards Consortium (CDISC). Clinical Data Acquisition Standards Harmonization (CDASH), Version 1.1. Prepared by the CDISC CDASH Team, January 2011. Available at https://www.cdisc.org/standards/foundational/cdash.

[22] Clinical Data Interchange Standards Consortium (CDISC). Clinical Data Acquisition Standards Harmonization (CDASH): Implementation Guide for Human Clinical Trials, Version 2.0. Prepared by the CDISC CDASH Team, September 2017. Available at https://www.cdisc.org/standards/foundational/cdash/cdash-20.

[23] Clinical Data Interchange Standards Consortium (CDISC). Therapeutic Area Data Standards User Guide (TAUG) for Schizophrenia, Version 1.0 (Provisional). Prepared by the CFAST Schizophrenia Standards Team, May 2015. Available at https://www.cdisc.org/standards/therapeutic-areas/schizophrenia/schizophrenia-therapeutic-area-user-guide-v10.

[24] Clinical Data Interchange Standards Consortium (CDISC). Therapeutic Area Data Standards User Guide (TAUG) for Major Depressive Disorder (MDD), Version 1.0 (Provisional). Prepared by the CFAST MDD Standards Team, November 2016 Available at https://www.cdisc.org/standards/therapeutic-areas/major-depressive-disorder.

[25] Clinical Data Interchange Standards Consortium (CDISC). Study Data Tabulation Model (SDTM), Version 1.4. Prepared by the CDISC Submission Data Standards Team, November 2013. Available at https://www.cdisc.org/standards/foundational/sdtm.

[26] Clinical Data Interchange Standards Consortium (CDISC). Study Data Tabulation Model Implementation Guide (SDTMIG): Human Clinical Trials, Version 3.2. Prepared by the CDISC Submission Data Standards Team, November 2013. Available at https://www.cdisc.org/standards/foundational/sdtmig.

[27] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, and M.N. Zozus, Evaluating common data models for use with a longitudinal community registry, *J Biomed Inform*. **64** (2016) 333–341. doi:10.1016/j.jbi.2016.10.016.