Paper DS07

Custom Domains … the Gaps in Standardized Data

Michael Beers, Pinnacle 21, Blue Bell, PA, USA

ABSTRACT

Now that six versions of SDTM Implementation Guides have been released over a period of more than 15 years, and more than 40 Therapeutic Area User Guides have been published, the industry should have a clear, standard way to map all types of clinical tabulation data, right? Why then do many studies still contain custom SDTM domains? This paper will discuss what types of data are still being submitted in custom domains, what the impact is, and what can be done about it.

INTRODUCTION

Custom domains are mostly a relic from the earlier days of CDISC standards. The early teams at CDISC started with the most typical domains to all/most trials, then expanded the guidance to new domains as much as could be done in each subsequent version. But now with 6 versions of SDTM implementation guides having been published over more than 15 years, as well as implementation guides for medical devices and associated persons, and numerous therapeutic area user guides, why are custom domains still around? Shouldn't they have gone extinct by now?
First I'll briefly discuss when and how you are supposed to use custom domains.

REGULATORY GUIDANCE

The FDA's Technical Conformance Guide, October 2022, section 4.1.1.3, states: *Prior to creating a custom domain, sponsors should confirm that the data do not fit into an existing domain. If it is necessary to create custom domains, sponsors should follow the recommendations in the SDTMIG. In addition, sponsors should present their implementation approach in the cSDRG. To provide study data that do not fit into an existing SDTM domain or draft SDTM domain, consider creating a custom dataset aligned with the Study Data Tabulation Model (SDTM). Questions about custom domains should be addressed in pre-submission meetings and documented in the SDSP.*
The PMDA's Technical Conformance Guide on Electronic Study Data Submissions, April 2022, section 4.1.1.2, states: *Depending on the characteristics of the collected data, it may not fit into an existing domain of SDTM. In such a case, it is acceptable for the applicant to create a custom domain. To perform this, the applicant must confirm that the data does not fit into existing domains, then create a custom domain according to the SDTM IG, and store the data under this domain. Explanation of the custom domain, together with the reason why it was necessary, should be described in the reviewer's guide.*

CDISC GUIDANCE

The Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.4 (Final), section 2.6, states

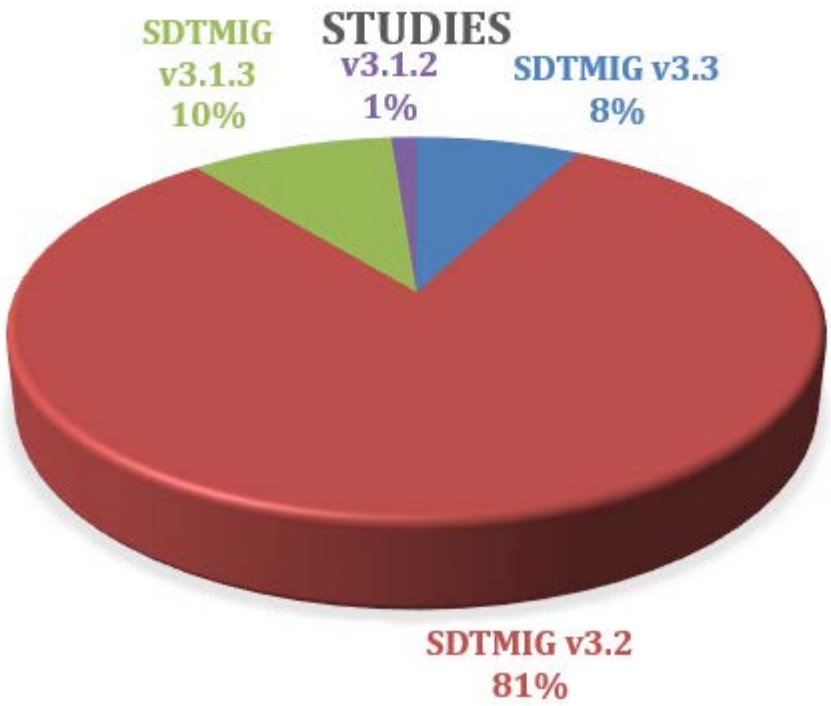
1. *Confirm that none of the existing published domains will fit the need…*
    - *Establish a domain of a common topic…*
    - *Do not create separate domains based on time…*
    - *How collected data are used (e.g., to support analyses and/or efficacy endpoints) must not result in the creation of a custom domain…*
    - *Data that were collected on separate CRF modules or pages may fit into an existing domain…*
    - *If it is necessary to represent relationships between data that are hierarchical in nature (e.g., a parent record must be observed before child records), then establish a domain pair…*
2. *Check the SDTM Draft Domains area of the CDISC wiki (https://wiki.cdisc.org/display/SDD/SDTM+Draft+Domains+Home) for proposed domains developed since the last published version of the SDTMIG. These proposed domains may be used as custom domains in a submission.*
3. *Look for an existing, relevant domain model to serve as a prototype. If no existing model seems appropriate, choose the general observation class (Interventions, Events, or Findings) that best fits the data by considering the topic of the observation…*

Also, section 4.1.6 of the same implementation guide states: *In some cases, sponsors may need to define new custom domains and may be concerned that CDISC domain codes defined in the future will conflict with those they choose to use. To eliminate any risk of a sponsor using a name that CDISC later determines to have a different meaning, domain codes beginning with the letters X, Y, and Z have been reserved for the creation of custom domains. Any letter or number may be used in the second position. Note the use of codes beginning with X, Y, or Z is optional, and not required for custom domains.*

METHODOLOGY

The vast majority of the industry uses Pinnacle 21's software, either Enterprise or Community, to validate their nonclinical and clinical (both tabulation and analysis) data against FDA Business and Validator rules, as well as CDISC Conformance rules. Pinnacle 21 does not store any subject-level data, however some metadata is available, such as dataset names and labels, and some value-level metadata. For this paper, I looked at 100 projects (typically corresponds to a drug or application or product) where studies for those projects were validated recently (from a period of roughly October 2022 to January 2023). This resulted in a sample of 555 studies validated during this date range for those 100 projects. I used this metadata to see which versions of the standards are currently being used, how many of those studies use custom domains, and what those custom domains contain. Note that my results admittedly may not be perfect, since as I stated Pinnacle 21 DOES NOT store subject-level data, and therefore further investigation into what the custom domains contain, beyond what is obtainable through metadata, is not possible and was not my goal.

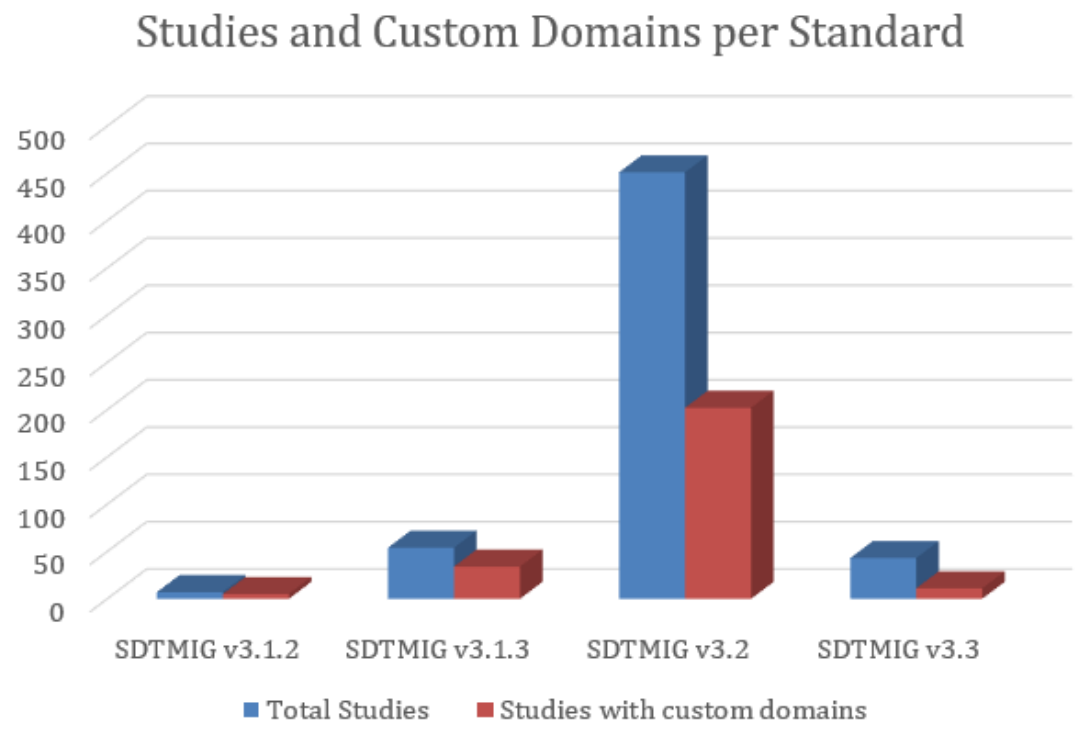
VERSIONS OF SDTMIG CURRENTLY IN USE



As we see in the pie chart, the majority of studies today are using SDTMIG v3.2. The surprising part may be that more studies are using SDTMIG v3.1.3 than are using SDTMIG v3.3. SDTMIG v3.1.3 was released all the way back in 2012, so it's very old at this point. This just shows a reluctance of the industry I suppose to move to new versions of the standards, as there are costs involved to do so, and perhaps the decision to up version relies of requirements to do so. There were a few studies still using the very old v3.1.2 still.

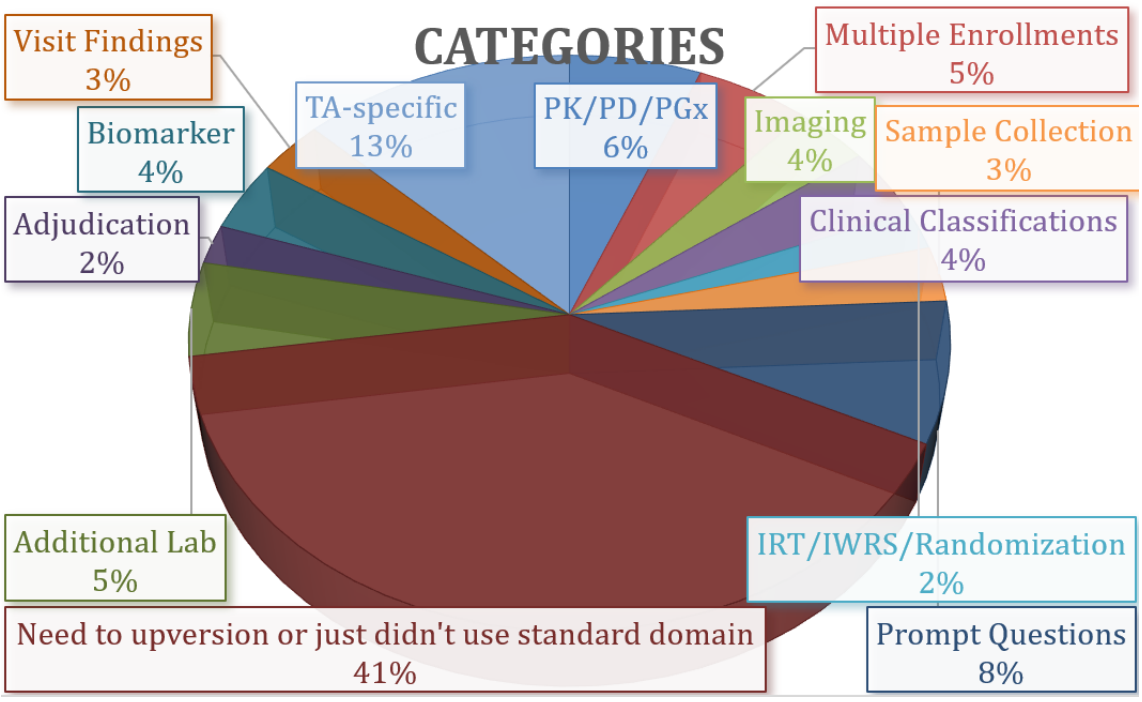
At the time of this writing, we're still waiting for the FDA to add SDTMIG v3.4 to the Data Standards Catalog. SDTMIG v3.4 was published in November of 2021, so no studies are currently using it. So, again, there is really slow adoption of new versions of the standards.

CUSTOM DOMAINS USED PER VERSION OF SDTMIG



This chart shows, for each version of SDTMIG, how many studies are using the version (in blue), and how many of those studies contain custom domains (in red). We can see that the use of custom domains is trending downwards with each subsequent version of SDTMIG, which is good and what we'd expect.
71% of the SDTMIG v3.1.2 studies contain at least one custom domain, for v3.1.3 it is 63%, for v3.2 its 45%, and for then v3.3 it is 25%. So again, we are moving in the right direction, but there's always room for improvement.

CATEGORIES OF CUSTOM DOMAINS



This pie chart shows my attempt to categorize the custom domains used. We can see that the majority of custom domains created were for data where a standard domain exists; if the preparer had up versioned to a newer version

of SDTMIG, the use of a custom domain wouldn't have been necessary; also, there are numerous cases where a standard domain exists in the same version of SDTMIG that the preparer used, however they created a custom domain instead, which in most cases would be considered an incorrect implementation. I'm not going to discuss these in this presentation, as it is likely just an incorrect implementation. The next biggest of my categories is what I labeled 'Therapeutic Area specific', with 13% of the custom domains being used. This was stuff like breast feeding data, wound assessment, stuff like that. I'm not going to discuss these categories either.

DETAILS OF CUSTOM DOMAIN USE

Next, I'll pick some of those categories of custom domain use, and dig a little deeper.

MULTIPLE ENROLLMENTS

Custom Domains - Demographics for Multiple Enrollments	
Dataset	**SDTMIG**
XD - Demographics Support	3.3
DC - Demographics as Collected	3.2
XM - Multiple Participation	3.2
XM - Demographics (Multiple Screenings)	3.2
XD - Demographics for Re-Screened Subjects	3.2
XD - Demographics	3.2
XM - Additional Demographics	3.2
XD - Demographics for Re-Screened Subjects	3.2
XM - Demographics Re Screened Subjects	3.2
DC - Demographics as Collected	3.2
XI - Previous Subject Identifiers	3.2
XI - Previous Subject Identifiers	3.2

The table above shows the names and labels of some of the custom domains used to handle multiple enrollments. If the same exact domain name/label exists on multiple rows, it means that the same exact domain name/label was used in two different studies (perhaps by the same preparer, or perhaps a coincidence).

Regarding multiple enrollments, the FDA's Technical Conformance Guide, October 2022, section 4.1.1.3, states: *For subjects with multiple enrollments within a single study, the primary enrollment should be submitted in DM. Additional enrollments should be included in a custom domain with a similar structure to DM. Clarifying statements in the RG would be helpful.*

For subjects with multiple screenings and no subsequent enrollment, include the primary screening in DM with additional screenings in a custom domain with a structure similar to DM.

For subjects with multiple screenings and subsequent enrollment, include the enrollment in DM with screenings in a custom domain with a structure similar to DM.

There is currently no CDISC guidance for multiple enrollments, however it is likely that this guidance will be added in the next version of SDTMIG that is published. At the moment, preparers of the data are using custom domains, named whatever they choose, to contain this data.

VISIT FINDINGS

Custom Domains - Visit Data	
Dataset	**SDTMIG**
ZO - Office Visit Log	3.2
ZV - Missed Visits	3.2
XC - COVID-19 Missing Visits	3.2
XV - Subject Visit Findings	3.2
XV - Subject Visit Findings	3.2
EE - Epidemic Events	3.2
XU - Unforeseen Impact	3.2

The table above shows the names and labels of some of the custom domains used to handle visit findings. Again, if the same exact domain name/label exists on multiple rows, it means that the same exact domain name/label was used in two different studies (perhaps by the same preparer, or perhaps a coincidence).

Visit findings became more important during the COVID-19 pandemic, missed visits became a common occurrence, and in some cases remote visits took the place of in-person visits. CDISC created their Guidance for Ongoing Studies Disrupted by COVID-19 Pandemic guidance in which they introduced an events domain to capture these visit findings. However, the FDA, in their Technical Conformance Guide, October 2022, Appendix D, states: *It is the*

current preference of the Agency that for all clinical studies, not limited to those impacted by COVID-19, subject visit data for scheduled (whether or not they occurred), and unscheduled visits be submitted in one single dataset structured as the current CDISC Subject Visits (SV) domain. It is also Agency preference that three non-standard variables (NSVs) for missed visits, --REASOC (Reason for Occur Value), --EPCHGI (Epi/Pandemic Related Change Indicator), and –CNTMOD (Contact Mode), outlined in the CDISC document "Guidance for Ongoing Studies Disrupted by COVID-19 Pandemic" be included within the SV domain and not within the supplemental SUPPSV domain or in other SDTM datasets. Submitting subject visits information in one single structured dataset allows both the human and technology consumer of this information to operate efficiently and with confidence that all visit data are considered during regulatory review.

This likely has led to confusion, as preparers are directed to update their SV dataset by the FDA, in a way that conflicts with what CDISC allows for SV, and what CDISC has proposed. Note that in SDTMIG v3.4, the variables requested by FDA have made their way into the SV domain, which resolves the situation hopefully, although the FDA hasn't added SDTMIG v3.4 to their Data Standards Catalog and therefore this version of SDTMIG is not yet in use by the industry.

PROMPT QUESTIONS

Custom Domains - Prompt Questions	
Dataset	**SDTMIG**
XA - Reference Information	3.3
XM - Clinical Monitoring	3.3
XA - Reference Information	3.2
XD - Debriefing Questions	3.2
XI - Screening Questions	3.2
XQ - Questions Dataset	3.2
QU - Questions Dataset	3.2
PQ - Proxy Question	3.2
ZZ - Leading Questions	3.2
XX - CRF Findings	3.2
XQ - Findings	3.2
ZA - PROTOCOL REQUIREMENTS	3.2
ZN - External Data	3.2
PQ - Proxy Question	3.1.3
XX - Study Maintenance	3.1.3
MQ - Monitoring Questions	3.1.2

The table above shows the names and labels of some of the custom domains used to handle prompt/indicator/probing questions. There currently appears to be no regulatory guidance on handling this type of data.

CDISC does provide a little guidance on this type of data. In the Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.4 (Final), section 4.5.7, it is mentioned that: *Questions such as "Did the subject have any concomitant medications?" or "Did the subject have any medical history?" should not have records in an SDTM domain because (1) these are not valid values for the respective topic variables of CMTRT and MHTERM, and (2) records whose sole purpose is to indicate whether or not a subject had records are not meaningful.*

This indicates that prompt/indicator/probing questions should not be mapped to SDTM datasets. In another section of the same implementation guide, Domain section TU, Assumption #9, it states however: *If indicator questions for specific types of tumor or lesions are collected (e.g., Does the subject have target tumors? Does the subject have any non-targets? Did the subject have metastatic disease at screening?), then these TUTESTs will be included in TU. If indicator questions are not collected, do not introduce them into TU.*

This seems to indicate that in some cases you may want to retain answers to prompt/indicator/probing questions, and if so, to include them in the domain as a separate record.

The reason why preparers may be creating custom domains for this type of data, as speculation, may be that 1) some preparers are extremely cautious in dropping collected data from their SDTM datasets, or 2) this data may be needed in analysis for some reason, or even just the possibility of it might being necessary may lend caution to dropping the data.

Without guidance on how to consistently map this data in a standard way, if the need to keep the data exists, results in preparers creating whatever custom domain name/label, and possibly even structure, that they choose.

ADDITIONAL LAB DATA

Custom Domains - Additional Lab Data	
Dataset	**SDTMIG**
XB - Laboratory Test Results (Conventional)	3.3
ZT - Other Test	3.3
XB - Laboratory CRF Data/Local Laboratory Data	3.2
XB - Laboratory Test Results (Conventional)	3.2
XL - Non-Safety Laboratory Results	3.2
ZL - Laboratory Test Results - US Units	3.2
XL - Laboratory Test Results XL	3.2
XL - Laboratory Test Results XL	3.2
XD - Other Tests Results	3.2
XF - Hormones	3.2
ZL - Laboratory Test Results - US Units	3.1.3
ZM - Lab Miscellaneous	3.1.3

The table above shows the names and labels of some of the custom domains used to handle additional lab data. There currently appears to be no regulatory guidance on handling this type of data. Again, if the same exact domain name/label exists on multiple rows, it means that the same exact domain name/label was used in two different studies (perhaps by the same preparer, or perhaps a coincidence).

The use of custom domains to handle additional lab data (instead of just putting everything in the same LB dataset), seems to be mostly due to two situations. The first is when there is lab data that is used for other purposes, that is not associated with typical safety labs, and the preparer decided to just to keep it separate from the data in LB for ease of use in analysis. The second situation is when results for different sets of unit, for the same lab tests, are provided. This may be due to the difference in requirements of the regulatory agencies, or just wanting to provide the regulatory agencies everything to avoid having to resubmit data.

The FDA stated in their Technical Conformance Guide, October 2022, section 4.1.1.3: *FDA may require laboratory data using conventional units for reviewing submissions and labeling. Sponsors should discuss with the review divisions what laboratory data should utilize conventional units prior to submission.*

The PMDA stated in their FAQs on Electronic Study Data Submission, June 2022, answer to question 4: *The use of SI units is required for all variables and parameters of test results to be stored in the Findings class domain of the SDTM dataset, as long as SI units are applicable…In such cases, it is possible to store data in uniform units other than SI units into "SUPP—" if necessary.*

Without a standard way to handle this not-uncommon situation, we see that preparers put this similar data across the industry into custom domains named whatever they choose.

ADJUDICATION DATA

Custom Domains - Adjudication	
Dataset	**SDTMIG**
XA - Adjudication	3.3
ZA - Adjudication	3.3
ZA - Adjudicated Results	3.2
EA - Event Adjudication	3.2
XA - Adjudication Results	3.1.3

The table above shows the names and labels of some of the custom domains used to handle adjudication data.

The FDA's Technical Conformance Guide, October 2022, Appendix D, states: *There are no existing standards or best practices for the representation of adjudication data as part of a standard data submission. Until standards for adjudication data are developed, it is advised that sponsors discuss their proposed approach with the review division and also include details about the presence, implementation approach, and location of adjudication data in the SDRG. Whenever adjudication data are provided, they should be clearly identified so that the reviewer can distinguish the results of adjudication from data as originally collected.*

In the Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.4 (Final), section 4.5.4, for one of their examples, mentions that …*evaluations of the adjudication committee were represented in SUPPAE.* Certain domains show the use of a variable –ACPTFL (Accepted Record Flag) that *identifies those records that have been determined to be the accepted assessments/measurements by an independent assessor.*

PHUSE has released a Best Practices for Submission of Event Adjudication Data (October 2019) guidance in which they recommend keeping event adjudication in a separate domain and have proposed a new EA domain for this purpose. We see that it appears that at least 1 study is already using this proposed new domain, while other studies are using their own custom domains.

CHALLENGES IN KNOWING WHERE DATA SHOULD BE MAPPED

It can be challenging knowing which SDTM domain some collected data should be mapped to. Some/most is straightforward, however some less common data might be a challenge to know where to put it. Two examples are given, of actual data found in actual custom domains:

EXAMPLE 1: APGAR SCORE

If you search the Implementation Guides for the term 'apgar', you likely will not find anything to assist in determining where to put the data. One approach is to search CDISC controlled terminology…there you will see that related values are found in two (at the time of this writing) codelists: Apgar Score V1 Clinical Classification Test Name and Category of Clinical Classification. Then we would need to see where Clinical Classification data is mapped to (the RS domain), and we have our answer.

EXAMPLE 2: ZO – CLINICAL OUTCOME DATASET

This custom domain used values from the Disease Outcome SDTM codelist (C102578), such as: CLINICAL CURE, CLINICAL FAILURE, etc. However, we don't see any link between this Disease Outcome codelist and an SDTM domain or variable. It's unclear what this CDISC codelist is used for, and while this method resulted in an answer in Example 1 (Apgar Score), we are no closer to knowing the answer as to where this data should definitively be mapped to.

These challenges are attempts to defend preparers that use custom domains in some instances. There should be a better way to easily, quickly, determine where a type of data should go in SDTM.

IMPACT OF NOT USING STANDARD DOMAINS

AUTOMATION

A goal of the industry should be to make drugs available to patients as fast as possible. In order to meet this goal, increased automation is necessary. Manual review is inefficient, and review time is always limited. Manual review is also usually insufficient in evaluating and enforcing data quality. Since automation relies on high data quality and high data standardization, the use of custom domains should be kept to a minimum.

VALIDATION

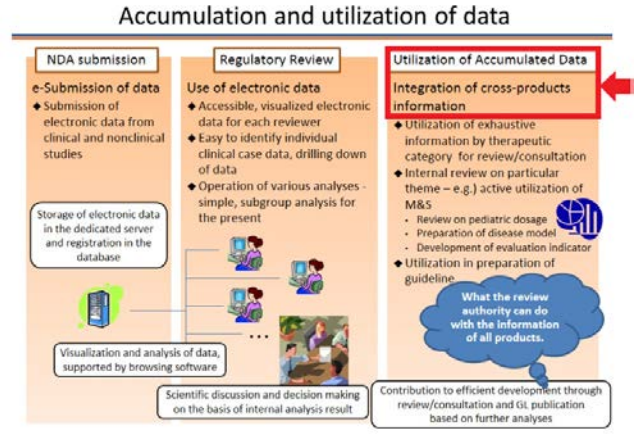
When custom domains are used, a very general set of validation rules must be run, since we don't know what the data is. It is not possible to run domain-specific checks, as well as controlled-terminology checks. The result of this is that data quality suffers, as the preparer doesn't see validation issues alerting them to non-conformance of their data to the standard.

We can see this using the example of Additional Lab Data:

- In SDTMIG v3.3, there are 155 validation rules run on the LB domain, but there are only 144 validation rules run on a custom findings domain.
- Variables such as –TESTCD and –TEST would not be checked against the LBTESTCD and LBTEST codelists, even though the controlled terminology should have been used.

In some cases, for multiple enrollments for example, validation might not even run on the domain, as custom special-purpose domains are not allowed per CDISC guidance, and with preparers naming these whatever they want, validation cannot be adjusted to run on these domains.

CROSS-STUDY/CROSS-PRODUCT ANALYSIS



This graphic is from the PMDA's website (https://www.pmda.go.jp/english/review-services/reviews/0002.html) showing that part of the reason for requiring electronic submission of standardized data is to be able to integrated data and analyze across products. With usage of significant numbers of custom domains across the industry, this ability to analyze across products or studies becomes limited.

CONCLUSION

Use of custom domains is going down with implementation of each new SDTMIG, which is a good sign, however there is always room for improvement.

More precise guidance should be created for some of these patterns of custom domain use, to increase standardization.

An easier way to determine where data should be mapped would be very useful to preparers, and may help reduce the usage of custom domains somewhat.

Reducing custom domain usage = more standardization = more automation**.**

REFERENCES

[1] FDA. "Study Data Technical Conformance Guide" October, 2022. https://www.fda.gov/media/153632/download
[2] PMDA. "Technical Conformance Guide on Electronic Study Data Submissions" April, 2022. https://www.pmda.go.jp/files/000247157.pdf
[3] PMDA. "FAQs on Electronic Study Data Submission" June, 2022. https://www.pmda.go.jp/files/000247170.pdf
[4] CDISC. "Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.4 (Final)". November 2021
[5] CDISC. "Guidance for Ongoing Studies Disrupted by COVID-19 Pandemic". Version 1.0. April 2020
[6] CDISC. "Subject Visits and COVID-19". https://www.cdisc.org/kb/articles/subject-visits-and-covid-19
[7] PHUSE. "Best Practices for Submission of Event Adjudication Data". October 2019. https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Optimizing+the+Use+of+Data+Standards/Best+Practices+for+Submission+of+Event+Adjudication+Data.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Michael Beers
Pinnacle 21
Blue Bell, PA, USA
mbeers@pinnacle21.com