Paper DH10

Data Transformation: Achieving Compliance with Irregular Data

Reyes Eduardo Elizalde Duarte, Syneos Health, Mexico City, Mexico

ABSTRACT

External data generated on clinical trials needs to be collected and submitted following Clinical Data Interchange Standards Consortium (CDISC) standards to ensure compliance with regulatory authorities. Early Phase studies face a particular issue due to the short duration of trial execution, so meeting CDISC standards becomes increasingly challenging.

The high number of data types, vendors, and external agreements hinders the standardization of data. Data Management works to ensure all data is transformed to meet CDISC standards and ensure SDTM programmers and biostatisticians can perform their needed tasks.

It becomes imperative, for early phase trials to plan, collect, organize, and analyze data as efficiently and correctly as possible.

This paper will go through the various challenges, steps, strategies, and insights to understand CDISC requirements, ensuring any delays are prevented and data meets all regulatory requirements.

INTRODUCTION

The pharmaceutical industry has changed greatly in the last decade. These changes, in turn, have increased the burden and necessity on data. Data has now become the key step for all trials to define success or failure, and with new emerging technologies, data is now as diverse as the drugs and people we study.

The integration of real-world data and big data, the advancements in precision medicine and health technologies, the importance of adaptive trials and real-time data monitoring, the emergence of Artificial Intelligence (AI) and Machine Learning (ML), and the trend towards decentralized clinical trials (DCTs), have all opened a new world of data and how we handle it.

One type of data that poses a particular challenge, is external data. External data, in terms of this paper, is defined as all data not directly recorded on the clinical database. External data is sourced outside the clinical database and can be later integrated or transferred to be included on the study's Study Data Tabulation Model (SDTM) datasets.

The main challenge we face is that external data will depend on the source system used to collect it, and in most scenarios, the Data Management (DM) team handling the clinical database does not participate on the built of the external database. This means that every external data can have its own format, structure and organization, hindering the standardization.

It becomes a critical task to transform all these "irregular" data into CDISC compliant data. However, transforming data is no easy task. It requires several steps during the planning and collection of data, but it also directly impacts the organization and analysis of this data.

DM needs to have proper knowledge of the Clinical Data Acquisition Standards Harmonization (CDASH), SDTM, and Analysis Data Model (ADaM) standards to know how to transform data. This positions DM as the critical point where data is defined as compliant or not usable; we are the first step in this domino effect.

A good analogy of data transformations is how our brain functions. In neuroscience, all neural networks exist in a critical state, which in short terms means there is a balance between stability and flexibility. Our brain is never too rigid nor too chaotic.

DM needs to be very similar to our brain. CDISC standards are clearly defined (too rigid) and the external data is always different (too chaotic), thus DM needs to maintain that balance, and our best tool is to mold or transform the data.

Our brains can make decisions by collecting, transforming, and analyzing data in an instant, and in Early Phase clinical trials, DM needs to work just as fast.

Early phase trials are always challenging due to their duration. There is less time to plan, collect, organize and analyze data, so handling external data as efficiently as possible will be a key factor for the success of the study.

The shortened timelines on early phase trials force us to consider external data sooner than usual. Vendor management, definition of systems and agreements, and data transformation specifications will be the most critical steps.

Teams will need to adapt, using CDISC standards as our weapons against chaotic data. Our final goal being regulatory compliant external data that can be used during all stages of the clinical trial. This paper will detail all the challenges, strategies to face them, and possible solutions to ensure success.

MAIN CHALLENGES WHEN HANDLING EXTERNAL DATA

Handling external data comes with several challenges. Prior to transforming any data, the hurdles and limitations must be identified to correctly assess the best strategy and reach regulatory compliance.

It is important to understand that all external data behaves differently. Different vendors have their own structure and Standard Operating Procedures (SOPs), and every study will have different timelines and milestones. Knowing this, it

may seem almost impossible to find a solution that fits all. However, we can use the CDISC standards and Good Clinical Data Management Practices (GCDMP) to create a plan that can fast track the processing of external data, prioritizing regulatory compliance.

First, lets deeply understand these challenges and the hurdles prior to start thinking about data transformations.

DIVERSITY IN EXTERNAL DATA

Clinical trials have evolved in parallel to technology. From paper studies to personal wearables, data management has changed greatly in the last years, adapting to new data.

Clinical trials now require different types of samples, tests, and assessments, all to meet study endpoints. This has, in consequence, created new types of data with new behaviors, formats, structures, and goals.

The main challenge when facing external data, is that any assessment can be externally sourced. Sponsors can decide to contract vendors to generate and analyze any kind of data, all depending on the study needs. Clinical trials can be so diverse that we will not use the same domain, datasets, variables or even units the same way in more than one study.

The different data categories, hinder our efforts on standardization. Clinical data can be categorized based on how we measure it:

- Nominal – Data can only be categorized (e.g. Gender).
- Ordinal – Data can be categorized and ranked (e.g. Cancer stages).
- Interval – Data can be spaced in equal intervals with no true zero (e.g. Temperature).
- Ratio - Data can be spaced in equal intervals with a true zero (e.g. Weight).

We can also categorize data based on how it behaves:

- Discreet – This type of data is countable, and any individual value can have meaning or be interpreted (e.g. 12-Lead ECGs, X-Rays or Pharmacokinetics).
- Continuous – This type of data is measurable and unspecific to any individual values; you can only give meaning to the whole (e.g. Holter monitoring, Echocardiography or Functional MRI).

In addition to these categories, data can have several more attributes that need to be considered when handling it.

1. Data latency
2. Volume
3. Velocity at which data is generated
4. How structured or unstructured data is
5. Veracity
6. Accuracy
7. Precision
8. Completeness
9. Consistency

These attributes must be included on our plans on how to manage external data and how we are moving from collection to organization and finally analysis.

External data for clinical trials can be a combination of any of these categories or attributes, and a study can have several combinations of external assessments, forcing us to have more than one strategy on how to handle these data.

EARLY PHASE TIME CONSTRAINTS

Early Phase clinical trials are characterized for having short durations. The trials can be as short as one month, removing all possibilities of updating or changing the data handling strategies throughout the study, exacerbating all challenges a trial could have.

These accelerated timelines impact the tasks related to external data vendors, especially contracts, budgets and agreements. There is a high demand to finalize all administrative prerequisites as soon as possible, so the data management team can begin working with vendors on the specifications related to external data.

The short duration of these early phase trials demands fast turnaround times for data collection, review and transformation. Handling several external data types that need to be either integrated or transferred puts additional pressure to the already short timelines.

Data quality is also an important factor that can be impacted during these trials. Both vendors and review teams have limited time to perform long and rigorous data reviews. Vendors may not have the required time to review all data records, as they are focused on sample processing and analysis. Review teams may jeopardize reconciliations and quality checks to meet milestones.

Consistency of data is closely related to the data quality mentioned above. Handling several data types may result in different structures and formats. The short timelines may prevent or leave little room for resolving any discrepancies between the data collected and data reported. Different vendors have their own standards and SOPs.

The long duration of late phase studies allows for more transfers and reviews to be performed. For example, a 3-yearlong study, with monthly external data transfers, will receive a total of 36 data transfers, allowing the detection of errors through long periods of time, even if an error is missed during month 10, it can still be corrected on month 12. Early phase studies do not have this privilege. This gives only 3 transfers throughout all the study to collect, review and transform the data. Any minor errors or issues can be easily missed if not handled correctly.

Finally, handling complex external data types, as the ones described on the section above, requires more effort, time, and resources. Many early phase trials study the safety characteristics of drugs, so complex assessments like imaging or genomic tests are common. Complex external data in accelerated timelines can result in major delays, if there is not a specific plan on how to manage this data once received. Teams will have to be familiarized with the CDISC standards used for each data type and understand why, how, and when data will be collected.

Early phase studies require proactive and forward-thinking management. When data collection strategies are being planned and specified, the team will need to start understanding and assessing how each of these strategies will impact the following steps, all the way to data analysis. All risks will need to be identified and mitigated as soon as possible.

MANAGEMENT OF VENDORS AND THEIR SYSTEMS

The huge diversity in external data comes in handy with a wide variety of vendors and technologies. All vendors will follow internal standards for data collection and processing, creating an ocean of possibilities. If we then consider that vendors now also have a plethora of technologies and systems to their disposal, our combinations become unique. Vendor management becomes a key skill to ensure that external data meets CDISC standards.

All data handled by vendors on external sources should come from validated systems. This ensures that data is already compliant to the Food and Drug Administration's (FDA) Code of Federal Regulations (CFR) Title 21 Part 11. However, this does not ensure data quality. There is still a regulatory obligation to ensure all data generated externally is CDISC compliant even if vendors cannot provide it as such.

These systems are frequently designed and customized to meet the vendor's needs, not the trials or DM's requirements. All systems will have a different structure, providing different insights with the same data.

Standard processes include creating Application Programming Interfaces (API) when integrating external data into the clinical database or, data import agreements (DIA) and data transfer agreements (DTA) when data is transferred through a secure method. However, each of these options come with their inherent challenges.

APIs main issues is the variety of external data. Having so many different data types make APIs very complex to set up, as format and data structure need to be standard between external sources or the API will fail. This means it is hard to create an API that can handle more than one data type. Any data inconsistency or misalignment between both systems will hinder data collection and all further steps.

On the other hand, DIAs and DTAs can easily be personalized. However, they require a lot of manual labor which can cause several errors. Creating these types of agreements requires constant communication and understanding of both systems (source and clinical database). While creating the transfer specifications and defining all needed data points, it is easy to face issues with data standardization. The vendor's data does not look like the clinical database nor does it have the required properties to be used during analysis.

Transferring data externally also requires an additional step called reconciliation. This is a key step to maintain accuracy, consistency and completeness of the external data and although new tools can be used to facilitate reconciliation, most of the work still falls under manual reviews.

The variety of systems and processes owned by vendors hampers the efficiency of API and DIAs/DTAs. Data cannot be handled equally between vendors if the type and system differ.

TRANSFORMING DATA TO REACH STANDARDIZATION

The challenges when handling data have been detailed and are now understood. However, part of DM's job is to find solutions and strategies for all these challenges and the main solution proposed on this paper is transforming data to reach standardization.

Data transformation can be understood as any change on the data attributes. These changes will allow data to be effectively used with different systems, processes, or structure. Data transformation will never mean changing the source value or source meaning of the data, only attributes of the data are updated. The traceability of data needs to be the core concept when transforming data. If traceability is lost, data is no longer compliant.

The recommended data transformations to be used are reformatting, aggregation, cleaning, mapping, normalization, enrichment, and filtering. All these transformation tools can be used throughout the data life cycle to reach standardization, ensure compliance and guarantee that data can be used for analysis.

During a clinical trial, four periods can be distinguished as key steps to ensure compliance with CDISC standards. First, the planning of the data, which is an important step to know what and how data is needed. Second, collecting all the data defined during the initial step. Third, organizing the data to ensure further standardization and compliance. Finally, analyzing the data to give it the desired meaning.

This paper will discuss two types of data transformations. One performed during data collection, and another performed during data organization. To facilitate the understanding of these concepts, data transformations during data collection will be called 'data shaping' and transformation during data organization will be called 'data conversion'.

These steps will be closely connected to each of the CDISC standards, CDASH, SDTM and ADaM. These three standards will be our greatest allies to transform the data.

PLAN: WHAT DATA AND HOW IS IT NEEDED?

The initial step in any data cycle is planning what data will be needed to perform the corresponding analysis; with the goal of confirming or assessing the trial's endpoints. The FDA and several other regulatory authorities require the

protocol to be written to maintain patient safety, scientific validity, and data integrity which is the key part for this paper.

The protocol is the first place where discussions around data collection need to happen. It is highly recommended to involve data managers, SDTM programmers, and biostatisticians, to detect any possible gaps or issues during protocol writing. These three teams are experts on each of the CDISC standards (CDASH, SDTM and ADaM) and can easily detect risks or issues the trial could have once data starts to be generated.

Understanding the goal of collecting data will facilitate regulatory compliance. It is important to understand what data needs to be collected. It is expected that the clinical database records the minimum amount of data required to maintain data privacy. Data managers are specifically trained to design and understand data needs. The protocol will guide data managers to know what electronic Case Report Forms (eCRFs) need to be built and what data will be generated externally.

Data managers are also trained to understand system limitations, and to address any issues related to external data. For example, the system used may not be able to handle APIs with external systems, so an external transfer is needed. Data managers understand the timeline constraints and resourcing limitations as well, so they can incorporate these limitations when creating the best plan to tackle the data of the study.

When APIs are not feasible to collect all needed data defined in the protocol, the data managers will create agreements with vendors to receive all data externally. This is our next crucial step.

Now that it has been defined what data we need, it is important to define how it will be collected. The most common method is through eCRFs, however, as previously defined, vendors need to be contracted to use their processes and systems. In this scenario, DM will create DIAs/DTAs to define in detail the data expectations.

These agreements will be a powerful tool to ensure data is CDISC compliant, even if not generated on the clinical database. The next section will detail this process.

COLLECT: HOW WILL DATA BE GENERATED AND COLLECTED?

The next step to consider, and where our data transformation strategies come into play, is the collection of data. Data collection is the main step of our domino effect. If data is not collected correctly, it cannot be categorized as needed and the analysis will not be able to be performed.

We now must define what the technical specifications of our data, i.e. format, structure and organization. However, during this step, it is important to understand we are still handling raw data. We should never transform data to give it a different meaning.

Data collected will come in all shapes and forms, this is part of our challenges; we need to ensure data can be shaped to meet CDISC standards even if it is not inherently compliant. By having standardization as our main goal, solutions can be generated to fast track and make collection more efficient. CDASH will help us reach this goal. CDASH standards were created to standardize data formats and structures. This facilitates data flow into the next step, organization.

The most common place where CDASH standards are present is during eCRF development. Data managers ensure all data collected on the eCRFs are required by protocol and will be usable during the next steps of the data cycle.

Nevertheless, CDASH standards are just as important during the collection of external data. It has already been established that not all data received externally will meet these standards due to the diversity of data, tests, vendors, and systems, so DM needs to strategize the best way to handle these huge amounts of data with the constrain of early phase timelines.

To standardize external data, the CDASH classes and domains can be used to our advantage. CDISC has already gone through the elaborate processes of categorizing all data types we can have on a trial. CDASH also contains controlled terminology that will aid on the standardization of data.

When the transfer specifications, included on DIAs/DTA or on its own, are built to collect external data, one of the key points will be to first identify the type of data per the CDASH classes, i.e. interventions, events, findings, and then if needed identify the domain, i.e. PC for PK data, LB for Lab data, etc. Once this is identified, it will become increasingly easier to know the variables required per the CDISC standards.

Unfortunately, as described on our ‘Main Challenges When Handling External Data’ section, not all vendors and systems will be able to meet these requests. This is where our data transformation will come in handy. The data transformation, or data shaping for data collection, will help us standardize the pending variables or datasets.

For example, if the safety laboratory vendor provides a test with category ‘Laboratory Chemistry’, but we require the category to be ‘Clinical Chemistry’ the data can then be shaped to match this expectation. However, the data still has the same meaning and structure.

The main techniques during data shaping will be reformatting, cleaning, and mapping. Let us explore each technique more in detail.

- Data reformatting refers to all transformations done on the data form. This can include changes like text to numerical data, changing date formats, changing missing data to ‘NA’ or ‘null’, etc.
- Data cleaning refers to transforming unprocessed data to refine data. This includes removing duplicates, correcting inconsistent data, handling invalid data, etc.
- Data Mapping refers to defining data correspondence between the clinical database and the external source This includes mapping the study visits between systems.

Combining the flexibility of CDASH to categorize data into their classes and using these data shaping techniques, we can create standards to almost every data type out there. Templates of DIAs/DTAs can be created to ensure CDASH

standards are met, and the needed data is collected externally. Then, for any data not compliant once received, data shaping specifications can be created to standardize the most common variable or records not compliant when collected by an external source.
Through these processes, we already have tackled and solved all our challenges, ensuring external data is CDISC compliant every time. However, we can still face some more complex challenges that cannot be resolved during this step. For these we will need to understand the next section to organize the data.

ORGANIZE: HOW WILL DATA BE PRESENTED FOR ANALYSIS?

We now understand the data collection process, its importance and the suggested solutions to standardize irregular data. However, it is also important to understand that not all challenges can be overcome during data collection. We can face complex issues that require more effort. Due to the limitation of each team's expertise, we require cross-functional collaboration to manage this more complex challenges.
Similarly to how data collection uses CDASH standards, data organization uses SDTM standards. SDTM provides a way to standardize data to facilitate its management and analysis. Following SDTM standards will allow us to warehouse, reuse, share, review and improve data.
SDTM follows the same structure as CDASH, data is classified in domains that can be described as logical groups of data. However, the classifications that SDTM handles are different that those used on CDASH. SDTM categorizes data into identifiers, topics, qualifiers and timing, all of these having specific domains.
SDTM has a machine friendly focus, while CDASH is more user friendly. This gives us an additional point of view of the data so additional data transformations are possible. The data organization transformations, or conversions, allows compliance when the planning and collection stages were not enough.
The main data transformation, or conversion, techniques we can use on this stage are aggregation, normalization, enrichment and filtering. Let's further detail these techniques to understand them:

- Data normalization refers to the process of adjusting data to a specific scale, without distorting the data. For example, converting Fahrenheit to Celsius, or pounds to kilograms.
- Data enrichment refers to adding additional data to a dataset to enhance it's meaning. For example, combining gender data to electrocardiogram (ECG) results.
- Data filtering refers to removing any data not relevant for the analysis. For example, not adding a subject that did not have Adverse Events (AE), to the AE domain or removing Screen Failures from specific domains.

In many scenarios, the clinical trial will require these 'conversions' in addition to the data shaping already performed during data collection. This is mainly due to the limitations of each standard and team. Data managers are not responsible for programming SDTM domains just as how programmers are not responsible to collecting external data.
These data conversions can be standardized by creating template programs or codes. This is possible because data has been previously planned, collected, and shaped, and is already CDASH compliant. These previous steps make data organization more efficient and quicker, avoiding unnecessary re-work or even defining data as non-usable, having to repeat the cycle again.
The data organization stage has given even more solutions to standardize the data and face any challenge, no matter how complex. Data is now ready to move to the last stage, analysis.

ANALYZE: GIVING MEANING TO DATA

The last stage on this data cycle is analyzing the data. During this stage there will not be any data transformations as we have already resolved our challenges and made it CDISC compliant. However, it is important to see the huge impact the effort of the previous stages will have during the analysis.
Data analysis is governed by ADaM standards. CDISC is involved through all the cycle, CDASH during collection, SDTM during organization and now ADaM for analysis.
ADaM standards provide a framework to facilitate analysis by having a clear understanding of the data. ADaM also categorizes data into data structures. These are very different to the previously discussed categories, as data is no longer grouped by type, but it is now categorized by its use for any required analysis.
During this stage ADaM standards no longer focus on ensuring data readiness, cleanliness, completeness, and accuracy. Therefore, if data is not compliant when shared for analysis, the trial results will not be correct, wasting all the time, effort and resources used.
All stages need to be considered with equal importance, as they are all part of the same cycle. If one stage is not done correctly, our challenges will not be overcome, and the trail will not have the expected results.

EXAMPLE SCENARIO

Let's think of a concrete scenario to understand the involvement of all steps. Safety laboratory data is commonly handled using DIAs/DTAs. The vendor will receive the blood samples from the clinical site to perform all analysis. Once the vendor has the analysis, they will share the data through an external data transfer to the data management team. DM already ensured most variables were CDISC compliant, but the vendor was only able to provide US units, could not match the visit names from the protocol, could not follow the needed test category names, could not remove tests not required per the protocol and could not provide average values.

This means that the DM team will have to 'shape' this data by cleaning all data inconsistent to the protocol by removing it, reformatting the test categories to give the expected names, and map the visit names to match the external system with the clinical database. However, during the data collection step, no conversions should be performed, as DM cannot manually update or convert the data. Hence, DM would not be able to normalize the units.

SDTM programmers will now have to step in and support with data conversion. The Statistical Analysis Plan (SAP) of this hypothetical study, requires US and SI units for the analysis. However, as defined above, the vendor could only provide the test results in US units. SDTM can then perform data normalization to convert the US units into SI units using established specifications.

Once all data is CDASH and SDTM compliant, the biostatistician will be able to provide the needed analysis, and for this hypothetical study, they require average values of laboratory tests. Using ADaM, they will be able to complete the analysis, including calculating average values.

After going through these steps data can be considered CDISC compliant and can be submitted to the FDA.

CROSS FUNCTIONAL TEAMWORK IS FUNDAMENTAL

The technical aspect and discussion around data transformations was defined on the section above, but an additional characteristic of the process can be pointed out. The need for cross functional teamwork and support.

All functional areas (FA) involved on the process detailed on this paper, have responsibilities and knowledge, but they also have limitations. One FA will not be able to plan, collect, organize and analyze all data, and no FA has the needed expertise on all three standards: CDASH, SDTM and ADaM.

The involvement of critical roles is needed throughout the process. Data managers, SDTM programmers and biostatisticians are particularly needed due to their expertise. Each of these roles needs to be involved during all stages, with varying levels of responsibility, based on the company needs.

CONCLUSION

The requirement to comply with regulations forces clinical trials to strategize and design plans to tackle the challenges of data diversity and vendor management, especially for early phase trials with accelerated timelines.

Through the use of CDISC standards, teams can make use of several data transformations techniques to switch from irregular external data to predictable and structured data. All these techniques need to be clearly defined within SOPs or processes to ensure the meaning of data is not changed at any point, a clear limit needs to be defined on how much data can be transformed based on the study needs and team capabilities..

Clear strategies can be defined to quickly and effectively tackle irregular data. These can be standardized to ensure all trials follow the same data flow. However, we need to remember the initial analogy of this paper. Teams need to be on a critical state of balance between stability and flexibility, to ensure these processes can be quickly adapted if needed without losing track of compliance and standards.

The final step of the process, and what will remain open for the reader to analyze, is how involved each FA or role needs to be during each stage. Should Data Management be the owner of the process, with the active support of programmers and biostatistics, to easily manage all stages under the same umbrella? Or maybe a joined management between other FAs is needed to meet regulatory requirements.

Nevertheless, any strategy or path taken should always ensure irregular data is handled correctly, through data transformation, to ensure compliance of regulatory and CDISC standards.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Eduardo Elizlade
Company: Syneos Health
Address: Mexico City, Mexico
Email: reyes.elizaldeduarte@syneoshealth.com
Website: https://www.linkedin.com/in/reyes-eduardo-elizalde-duarte-402041137