PharmaSUG 2012 - Paper DS13

"Analysis-ready" – Considerations, Implementations, and Real World Applications

Ellen (Zihui) Lin, Amgen Inc, Thousand Oaks, CA

Beiying Ding, Amgen Inc, Thousand Oaks, CA

ABSTRACT

"Analysis-ready" is one of the widely accepted key principles that have been applied in clinical trial analysis database design. It is also emphasized in CDISC ADaM v2.1 and CDISC ADaM Implementation Guide v1.0. However, there still remains ambiguity in some aspects of "analysis-ready" concept and implementation. This is largely due to the often very diverse nature of collected data and planned analyses. The lack of consensus in the industry on the concept of analysis and the application scope of "analysis-ready" principle has also led to drastically different interpretations and implementations. In some situations, programming logic such as complicated data derivation dependency may affect implementation of this principle as well. In the past years working on analysis database design and regulatory submissions, we have encountered numerous challenges and decision makings trying to maximize "analysis-ready" potential in practice. In this paper we share our experiences and interpretations of "analysis-ready" for various purposes such as descriptive summary, statistical modeling, manual data review, and special considerations for ad hoc analysis and data integration. For the purpose of illustration and facilitation of discussion, we also provide working examples to highlight approaches employed in some typical data and analysis scenarios.

1. INTRODUCTION

"Analysis-ready" is one of the ADaM design principles that are emphasized in CDISC ADaM v2.1 and CDISC ADaM Implementation Guide v1.0. In these two CDISC ADaM documents, "analysis-ready" dataset is defined as "analysis datasets that have a structure and content that allows statistical analysis to be performed with minimal programming". The brief definition clearly states the need for adequate information and reasonable structure in the analysis datasets in order to alleviate the burden of data manipulation at analysis stage. Understandably there is a lack of explicit guidelines beyond this given the multifarious nature and objectives of clinical trials and research projects as well as other practical considerations such as programming logic conventions. However, because there is so much left to speculate on the concept of analysis and the application scope of "analysis-ready" principle, drastically different interpretations and implementations have been observed in the industry and there doesn't appear to be much consensus even on some of the most commonly encountered data analysis scenarios.

Having worked extensively on the analysis database design and successful regulatory submissions in the past few years, we have come to realize and appreciate the importance and benefit of the "analysis-ready" concept. However, in trying to maximize its potential in practice we have also had to overcome a lot of obstacles and faced many challenges before reaching reasonable decision making and establishing working models. In this paper we share our interpretations and experiences on that and what might work in the "analysis-ready" framework for various typical purposes such as descriptive summary, statistical modeling, manual data review, ad hoc or exploratory analysis, and data integration. For the purpose of illustration and facilitation of discussion, we also provide working examples to highlight approaches employed in some typical data and analysis scenarios.

All discussions are in the context of SAS® programming, but the concepts and principles apply to all statistical software packages. Dataset and variable names referenced in this paper are solely for the demonstration of ideas and designs and are not intended to reference any particular versions of the industry data standards.

Any views or opinions presented in this paper are solely those of the authors and do not necessarily represent those of Amgen Inc., regulatory agencies, or data standard organizations (e.g., CDISC).

2. WHY ANALYSIS-READY DATASETS

Statistical analysis of clinical trials (especially phase 2 or 3 trials) typically requires a lot of complex data manipulations and / or derivations from collected data to the generation of analysis reports. In general, we do not consider collected data or tabulation data (such as SDTM) analysis-ready. For example, bone mineral density (BMD) is collected at scheduled visits; in order to summarize the percent change from baseline in BMD for each visit, the

collected data need to be further manipulated and processed in order to determine the analysis set (i.e., including all subjects who are randomized and have baseline and at least one post baseline BMD assessments), select the baseline value, derive percent change from baseline, define analysis window for each visit and select the record for analysis when multiple assessments exist within an analysis visit window.

Innate in the nature of analysis-ready design is the possibility of a unified framework that is more reflective of the intended analysis plan and, in turn, the fundamental objectives of the clinical or research project. It allows better documentation of database design, algorithms, data derivations and manipulations, which further enables simplicity and efficiency for statistical source programming as well as transparency in quality control process.

In addition, the concept of a well-designed analysis data structure that is readily adaptable becomes even more critical and appealing when it comes to combining data across projects / databases. In our experience with recent data filing, the implementation of analysis-ready data process across phase 3 trials has proven pivotal in successful integrations of efficacy and safety analyses.

Last but not least, an analysis-ready data package greatly facilitates the regulatory review process for statistical and medical reviewers. It offers easier interpretation of data and analysis results and eliminates or minimizes the effort and need for repeating data manipulations and derivations for the purpose of reproducing results especially those complex in nature. In fact, providing analysis-ready datasets is also a regulatory recommendation aimed at promoting clear and unambiguous communications between sponsors and regulatory agencies.

Although not meant to be comprehensive, the above mentioned merits of analysis-ready data certainly warrant a closer look at the concept, implementation and applicable scope. In Sections 3 and 4 we explore the definition and attributes of analysis-readiness. Sections 5, 6, and 7 provide more details on special considerations for manual review, ad hoc or exploratory analysis, and integrated database. In Section 8, we offer our thoughts on where the boundary should be for analysis-ready data design before a brief concluding summary in Section 9.

3. OUR SCOPE OF "ANALYSIS"

As far as we are aware, there is not a universally accepted definition of "analysis" or the scope of "analysis dataset". It is noted in ADaM v2.1 that "Within the context of ADaM, at a minimum analysis datasets contain the data needed for the review and re-creation of specific statistical analyses. It is not required that the data be collated into analysis-ready datasets solely to support data listings or other non-analytical displays, although some may choose to do so". In the "Study Data Specifications" document issued by the Food and Drug Administration (FDA), it is defined that "analysis datasets are datasets created to support results presented in study reports, the ISS and the ISE and to support other analyses that enable a thorough regulatory review. Analysis datasets contain both raw and derived data". Based on these guidelines and our filing experiences, we have in our practice employed an expanded definition of "analysis" i.e. "any use of the study data to make medical / clinical conclusions or statistical inferences, which may include but are not limited to descriptive summary, statistical modeling, manual data review, data listings, ad-hoc or exploratory analysis, and integrated analysis". This strategy is based on careful review of regulatory submission needs as well as our internal programming and manual data review processes and has been tested in numerous projects. In terms of implementation, we use a linear approach where we create SDTM datasets from collected data first and the SDTM data will then serve as the input data for creation of the analysis datasets. We believe SDTM data, a form of tabulation data, cannot support the above defined analysis efficiently especially when there is no automatic tool to facilitate the data review. Analysis-ready datasets are needed to support the purposes.

There are a lot of debates on whether or not creating data listings should be considered "analysis". We believe it deserves a closer look on a case by case basis. Our data listings usually include both collected and derived data. For example, in a typical listing of treatment-emergent adverse events, in addition to collected data such as subject ID, adverse event verbatim / coded terms, we also need to include derived data, e.g.,

- Actual treatment group
- First dose date
- Last dose date
- Number of days since previous active dose of the investigational product (IP)

Some collected data may also need to be converted from SDTM format back to the original collected format for display (such as the Adverse Event On-going flag collected in CRF). If we use SDTM datasets directly to produce the listings, we would have to handle potentially complex data manipulations (e.g. complex data merging between SDTM AE and EX domain in order to derive "number of days since previous active dose of the IP") and conversions in the listing programs, which is inefficient, hard to document and increases the risk on quality. In this example, our solution is to create an adverse events analysis dataset which includes the above derived or converted data to support both adverse event summary analyses and data listings.

4. OUR INTERPRETATION OF "ANALYSIS-READY"

The most ideal "analysis-ready" design is to have "one-proc away" analysis database, which means that a dataset has all variables and records to support a specific analysis using one or multiple SAS statistical procedures directly without any data pre-processing. In fact, ADaM v2.1 uses a "one-proc away" example to illustrate the concept of analysis-ready. While we strive to have one-proc away analysis datasets, due to the very diverse nature of clinical trial data and analyses, it is not always achievable or practical. As defined in CDISC ADaM documents, under the concept of analysis-ready, we allow minimal data manipulations or derivations in table, figure, listing (TFL) creation programs in general.

Listed below are examples of simple data manipulations commonly seen in TFL programs:

- Selecting data records using one or more variables already derived in the dataset

  Data subsetting may be done in a data step or within a statistical procedure. This kind of data manipulations are commonly seen in TFL programs because it is inefficient and impractical to always create permanent data subsets that support only one specific analysis.

- Sorting data records

  An analysis dataset usually supports multiple analyses including data listings (data display), which may require different sorting orders of the data records.

  In addition, PROC SORT with option NODUPKEY is often used to count number of unique records based on a set of criteria (e.g., count number of unique subjects who have at least one treatment-emergent adverse event).

- Simple data merging among analysis datasets

  When analysis dataset is not one-proc away and variables required for an analysis belong to more than one analysis dataset, data merging among two or more analysis datasets has to be done in TFL programs. For example, in a phase 3 pivotal study, we conduct an analysis for "Lumbar Spine BMD Percent Change from Baseline by Visit" based on ANCOVA model adjusting for treatment group and covariates such as age group, androgen deprivation therapy (ADT) duration at study entry, etc. Due to complex derivation dependency, instead of including ADT duration at study entry in the BMD analysis data set (ADBMD), we decide to derive it in the baseline analysis dataset (ADBASE, one record per subject) which has a higher derived order than ADBMD. In the analysis program, we then merge ADBMD and ADBASE by Subject ID to get ADT duration at study entry before ANCOVA modeling.

  That being said, we strive to design an analysis-ready database which only requires TFL programs to handle simple data merging using only one variable (i.e., Subject ID) as the merging key. Whenever possible, we handle more complex data merging at analysis data creation step to ensure quality and improve efficiency. Below are a few examples of complex data merging between:

  - Two analysis datasets using more than one variable as the merging key
  - An analysis dataset and an SDTM domain (except DM domain)
  - An SDTM standard domain and a SDTM suppqual domain
  - Two SDTM domains using RELREC (e.g., AE and AE finding domain)

  In general, we do not expect programmers, medical or statistical reviewers to carry out complex data merging without detailed instructions or well-documented algorithms. We find it cost saving when complex data merging is documented and done in analysis data creation step, especially when an automatic data review tool is not available.

- Data transposition

  In general we discourage performing data transposition in TFL programs especially when the transposed data structure is expected to support multiple analyses. For example, in a recent filing, patient reported pain scores were collected at scheduled visits. An analysis dataset for these endpoints (ADQOLBPI) was created accordingly in a one record per subject per visit structure to support the analysis of pain by visit and pain change from baseline by visit. In addition, a transposed dataset (ADSLQOL) was used to support multiple imputation analysis

which requires data to be in a one record per subject format, e.g. worst pain (BPI3) at Week 5 and Week 9 in ADQOLBPI was converted to two separate variables (BPI3W5 and BPI3W9) in ADSLQOL.

However, if a transposition of an existing analysis dataset is only going to be needed one-off, e.g. a quick listing to assist line listing review, we choose to do the data transposition in the listing program instead of creating a separate permanent analysis dataset. For example, our laboratory analysis dataset (ADLB) is a one record per subject per analyte per visit layout. We routinely transpose ADLB in the listing program in order to create lab listings in the form of one record per subject per visit where multiple analytes from the same subject and same visit are displayed in the same row. Of course, ADLB contains all collected and derived data that are needed to create the listing.

5. SPECIAL CONSIDERATIONS FOR MANUAL REVIEW

As we consider manual review by medical and statistical reviewers to be part of the analysis, some additional design details should be implemented to better serve this purpose especially if an automatic data review tool is not available. A review-friendly analysis database is manifested by:

- A reasonable / convenient data structure
- Coherent data content
- Consistent and logical ordering of data variables within and across datasets
- Data sorted in a desirable order
- Numeric data values properly formatted, i.e. with desired number of decimals
- Manageable file size

6. SPECIAL CONSIDERATIONS FOR AD HOC ANALYSIS

While planned analysis addresses our main study objectives, ad hoc or exploratory analyses frequently arise. It is challenging to adopt the same analysis-ready principle in this setting since requirements for these analyses are rarely available during the database design phase.

Our routine is to keep as much data from the collected data (or CRF data) as possible in the analysis datasets. When additional data derivations are needed for ad hoc analysis, programming can be done using the existing analysis datasets without having to resort to SDTM. For example, adverse events CRF has a question of "action taken for adverse event" which is usually not used in planned tabular summaries. However, we still include it in the analysis dataset, ADAE, along with other data collected on the adverse events CRF. When an ad hoc analysis on summary of adverse events by action taken is requested, some data derivation using ADAE can serve the purpose and we do not need to merge ADAE back with SDTM AE domain, AE suppqual domain, or AE finding domain or re-derive analysis data from SDTM.

7. SPECIAL CONSIDERATIONS FOR INTEGRATED DATASETS

Although integrated datasets can be created from analysis dataset only, from SDTM only, or from a combination of analysis and SDTM datasets for individual studies, we choose to use the analysis dataset only approach whenever possible. In other words, we design the analysis datasets for individual studies to be analysis-ready so that only minimal programming is required for the creation of integrated datasets, which should also be analysis-ready for the integrated analysis. Here are a few basic principles that can be applied to achieve this goal:

- Allow new variables to be added in integrated datasets

Sometimes integrated analysis has different focus or requirements from those for the individual studies. For example, for integrated safety analysis we might need to add variables for several adverse events of interest which may not be included or analyzed at each individual study level. In general, our study level analysis database should provide sufficient data to support the derivation of new variables in integrated datasets in a relatively straightforward manner.

- Use current versions of coding dictionaries in integrated datasets

Individual studies are usually conducted and analyzed at different time points hence may use different versions of medical dictionaries such as MedDRA and WHODRUG. When we create integrated datasets, we apply the most current version of the dictionaries for all studies that are included in the integrated database while leaving the individual study databases unchanged.

- Consider re-deriving categorical values as part of minimal programming

Due to different study populations or analysis requirements between individual studies and integrated analysis, some categorical variables may need to be derived again in integrated datasets. In the example below, age group are defined differently for study 001, study 002, and the integrated analysis:

Analysis	AGEGRP (Age Group)
Study 001	55 - 69 YEARS 70 - 79 YEARS >= 80 YEARS
Study 002	< 18 YEARS 18 - 40 YEARS 41 - 60 YEARS > 60 YEARS
Integrated analysis	< 50 YEARS 50 - 59 YEARS 60 - 69 YEARS 70 - 79 YEARS >= 80 YEARS

- Require a consistent design of analysis database across studies

There are a lot of considerations on designing analysis database consistently across studies, such as data structures, data contents, variable name associated with compatible definition, and use of controlled terminology wherever applicable, etc. In particular we consider assigning each variable a meaningful name and specific data content good practice. Using generic variable names such as "ANLFLAG1 – analysis flag 1" or "AGE – age" (i.e., without unit) which can represent different definitions in different studies makes interpretation and integration difficult.

8. DO NOT OVER-EMPHASIZE ANALYSIS-READY

Although less common, we find that over-emphasis of analysis-ready sometimes could complicate or clutter the analysis database design and programming process. Below we discuss a few examples where analysis-ready may be over-done:

- Do all variables required for a specific analysis need to be in the same dataset?

As discussed in previous sections, this is part of the requirement of 'one-proc away' which is something nice to have but may not be always practical. Analysis database is a complex system and has to be derived based on a carefully designed deriving order (i.e., derivation dependency). Naturally, we want to derive each variable only once in the entire database; and variables belonging to different derivation order would need to be stored in different datasets. A simple data merging can be performed in TFL programs to get variables from multiple sources.

- Do we always need to create new variables which are a simple function / transformation of other variables?

In many scenarios, this may not be needed. For example, there are two variables, NADOS01 and NADOS02, in ADSL, where NADOS01 is "number of active doses received in period 01" and NADOS02 "number of active doses received in period 02". An analysis summarizes these two items plus "total number of active doses received in period 01 and period 02". We suggest the arithmetic addition (NADOS01 + NADOS02) be done in table program instead of creating a new variable for total dose in ADSL. Having many unnecessary or redundant variables often complicate the naming conventions or the interpretation of the data.

- Should "composite" flag variables be created?

Record selection that can be achieved by using a combination of a few variables does not warrant an additional flag. For example, we do an analysis using a subset including observed data on or before month 12 visit and one unique record from each visit window among subjects who have received at least one dose of IP. Instead of creating a new flag to perform this data selection and analysis, we have a few variables designed for the purpose:

- A population flag variable for subjects who received at least one dose of IP
- A visit variable to choose records on or before month 12
- A record flag variable to indicate the unique record selected from each visit window
- A data type variable to choose observed data

In this design, each variable has a unique meaning, and can serve multiple analysis purposes when used individually or in conjunction with other variables.

9. SUMMARY

"Analysis-ready" is one of the most important principles that are applied in the often very complex clinical trial analysis database design process. Different interpretation of "analysis" scope and specific programming process or design may lead to different decisions or approaches in its implementation. Our goal has been to build a high-efficiency and high-quality analysis database that strongly supports data analysis and regulatory submissions. We believe that it is generally cost effective to accommodate complex data manipulations and derivations in the analysis data creation step and minimize data programming during TFL production stage. On the other hand, too much emphasis on the analysis-ready concept might complicate the analysis database unnecessarily and may eventually prove to be burdensome. By sharing our views and some of our experiences, we hope to attract more attentions to this topic and encourage further discussions on finding the delicate balance in achieving a flexible, robust and content-rich "analysis-ready" structure in practice.

REFERENCES

1. CDISC Analysis Data Model (ADaM) version 2.1
2. CDISC ADaM Implementation Guide version1.0
3. FDA Study Data Specifications version 1.6
4. CDISC SDTM Implementation Guide version 3.1.1

ACKNOWLEDGEMENTS

The authors wish to thank the Bone Therapeutic Area statistical programming and biostatistics teams at Amgen Inc. Many great ideas, considerations, and examples mentioned in the paper come from the two IND submissions that the teams delivered in the past few years.

TRADEMARKS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact one of the authors at:

Ellen Lin
E-mail: zlin@amgen.com
Phone: 805 447 1018

Beiying Ding, Ph.D.
E-mail: bding@amgen.com
Phone: 805 313 4309