PharmaSUG 2015 - Paper PO22

Challenges in Developing ADSL with Baseline Data

Hongyu Liu, Vertex Pharmaceuticals Incorporated, Boston, MA
Hang Pang, Vertex Pharmaceuticals Incorporated, Boston, MA

ABSTRACT

For a CDISC compliant submission, the ADSL (Analysis Data Subject Level) domain is the minimum required analysis dataset. However, developing an ADSL dataset can be quite challenging. One of the challenges is how to acquire and organize the baseline data for ADSL creation. Should ADSL be developed using SDTM data as its source data so that all ADaM (Analysis Data Model) domains are programmed independently without recursion? Or should ADSL be developed partially depending on other ADaM domain's programming? This paper will discuss various models that a sponsor may consider during the ADSL dataset generation.

INTRODUCTION

ADSL is the subject-level analysis dataset for ADaM. ADSL and its related metadata are required in a CDISC-based submission of data from a clinical trial even if no other analysis datasets are submitted [1, 2]. The FDA defines that the core variables, including all covariates presented in the study protocol should be included in each ADaM dataset [3, 4]. These core variables are typically developed and included in the ADSL dataset. Examples of key core variables include: study/protocol number, center/site number, geographic region, country, treatment assignment information, sex, age, race, analysis population flags (e.g., Intent-to-Treat (ITT), Full Analysis Set (FAS), Safety, Per-Protocol (PP)), and other important baseline demographic variables.

Although CDISC Guidance does not explicitly specify that ADSL should include subject baseline characteristic variables (BL), the FDA states clearly in 'Study Data Technical Conformation Guide' that; "In addition to the variables specified for ADSL in the ADaMIG such as those previously listed in the core variables section …, the sponsor should include multiple additional variables representing various important baseline subject characteristics/covariates presented in the study protocol." [3, 4]

For some clinical trials, these baseline variables can be easily derived directly from SDTM domains without involving sophisticated programming. However and most likely, ADSL needs to include some baseline data that may require a complicated derivation and can't be simply transferred from SDTM data. Also, those baseline variables are usually programmed in other individual ADaM domains. It is very costly if these variables are programmed with the same logic or even a macro in both ADSL and individual domains. An alternative is to only program these variables once in individual ADaM domains and then map these values to ADSL. Now, the data flow relevant to the programming sequence and relationship between ADSL and individual domains may result in some processing issues, namely, recursion. Recursion, if not properly performed, may greatly increase the risk of errors in both ADSL development and validation. For ADaM data development, we also need to consider the following fundamentals: traceability, analysis-ready, and programming efficiency, as well as FDA requirement compliance etc., though these are not the discussion focus of this paper.

There is no clear consensus in the industry on how ADSL should be programmed. It is now obvious that ADSL development is very challenging [5], even five years after CDISC ADaM and ADaMIG was published and accepted by the industry. For this paper, four different models are discussed based on our programming experiences and understanding.

FOUR PROPOSED ADSL DEVELOPMENT MODELS

MODEL 1: ONE-STEP ADSL DERIVATION

This is the simplest development approach among the four models described in this paper for ADSL programming (see Figure 1 below). It is only meaningful when all core and important baseline characteristic variables that are needed in ADSL domain can be easily derived from SDTM/non-ADaM source data without any complicated procedure. For example, baseline values that will be included in ADSL are the SDTM flagged baseline values and the corresponding ADaM data use same definition for baseline values. These baseline values can be directly mapped into ADSL variables from SDTM/non-ADaM source data. After ADSL is developed, all other ADaM domains can be programmed with common variables (core and baseline) from ADSL. The ADSL development does not depend on other ADaM domain development.

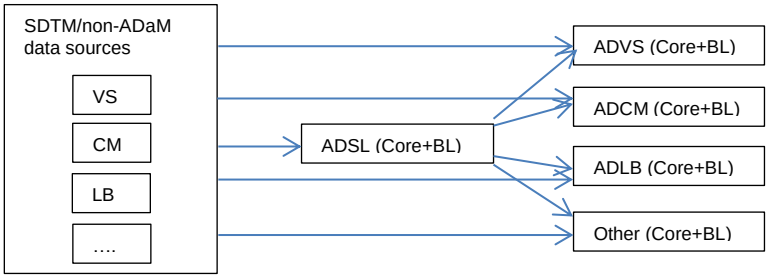

Figure 1 One-Step ADSL Derivation

MODEL 2: ONE STEP ADSL DERIVATION WITH AN ADDITIONAL SUBJECT LEVEL DATASET

For most of the scenarios in the late stage trials, baseline characteristic variables cannot be derived directly from SDTM data without further processing. Even derivations for some core population flags such as a protocol deviation flag are relatively complicated. Their derivations usually are performed in individual ADaM domains.

One of the options is to program these variables in both ADSL and their relevant individual ADaM domains using the same algorithm. Obviously, doing so will not only double the programming work but will also increase the risk of error. Thus, it is not a practical approach.

Some companies in the industry have a practice with adding another subject level ADaM dataset, such as ADBASE (different dataset name may be used) [6, 7]. In this model, ADSL and other ADaM domains do not include baseline characteristic variables. These variables will be programmed at the end of the ADaM development and in a separate ADaM dataset as shown in Figure 2.

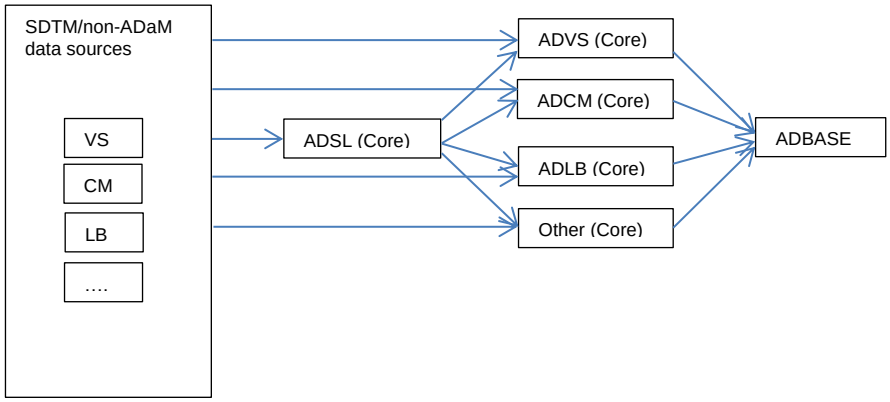

Figure 2 ADSL Derivation with an additional subject level dataset

MODEL 3: ADSL DERIVED LAST

This model develops an analysis dataset for all core variables first, say ADCORE for example, then develops all ADaM domains except ADSL. ADCORE is still a subject level dataset with one observation per subject structure. This dataset provides all core variables that are needed for all ADaM domains development. After all ADaM domains that are relevant to ADSL baseline data are programmed, the ADSL dataset will be created last. This model is similar to model 2, however, baseline variables are not included in the individual ADaM domains, only ADSL.

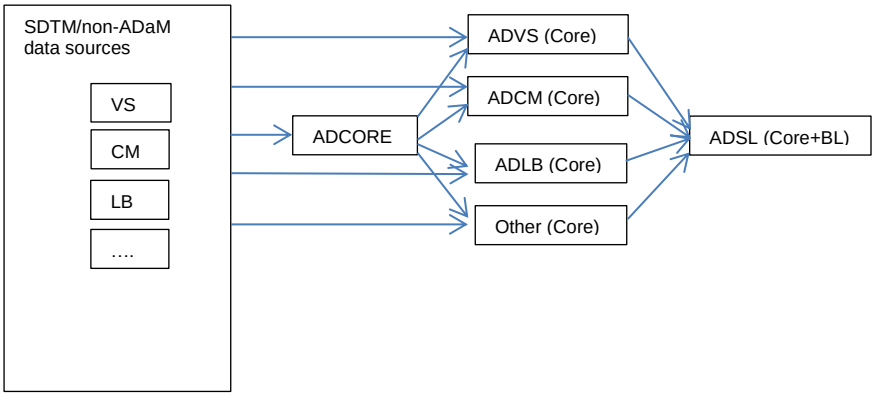

Figure 3 ADSL Derived Last

MODEL 4: TWO-STEP ADSL DERIVATION

This model includes two steps in programming ADSL: 1) to derive variables which can be programmed directly from SDTM/non-ADaM datasets for ADSL (draft). Most of the core variables can be derived at this step. And 2) to derive other ADSL variables which cannot be programmed without relatively complicated procedures from SDTM/non-ADaM datasets, including important baseline characteristic variables or per protocol population flag from the relevant ADaM domains as shown below in Figure 4. Once the ADSL is drafted out with core variables, other ADaM domains that contain important baseline characteristic data required for variables in ADSL will be programmed. Note that not all ADaM domains need to be drafted at this stage, only the ones with relevant baseline data are required. At the second step of ADSL development, the drafted ADSL merges with those drafted ADaM datasets to populate required values for baseline variables in the ADSL. After all ADSL variables are populated, individual ADaM domains will be developed based on SDTM/non-ADaM data and ADSL data for core and baseline variables.

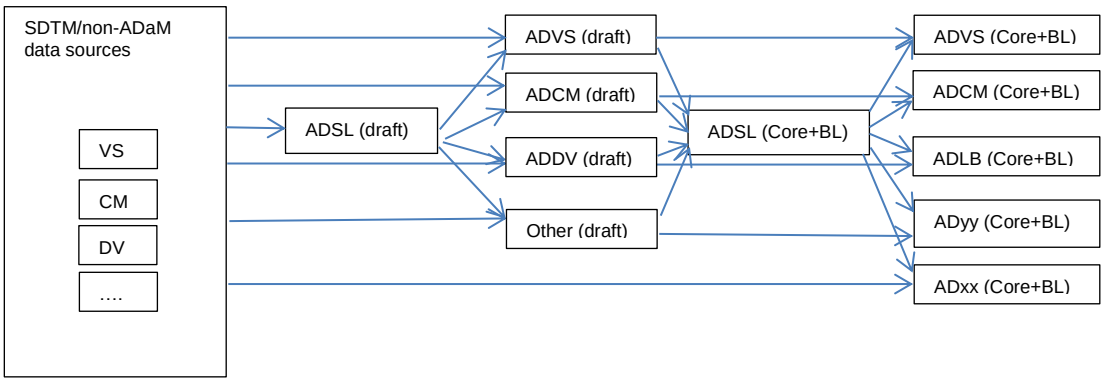

Figure 4 Two-Step ADSL Derivation

CONSIDERATION AND DISCUSSION

We may summarize the four ADSL derivation models and their advantages and challenges in table 1 below.

Table 1 ADSL derivation models and their advantages and challenges

Model	ADSL Content	Other ADaM Domain Content	Advantage	Challenge
1	All core and important baseline characteristic variables included.	All core and important baseline characteristic variables included.	1. Requirements by CDISC ADaMIG and FDA 'Study Data Technical Conformance Guide' completely met. 2. No dependency on any other ADaM domains. 3. Simple and clear data flow from source to ADaM. 4. Early completion of ADSL development. 5. No additional ADaM domain needed for fulfilling core and important baseline characteristic variable requirements. 6. 'Analysis-ready' principle met.	Application only to clinical trials that both core and baseline variables required for ADSL can be derived directly from SDTM/non-ADaM source data without substantial complicated algorithms. May be used in most of the phase 1 trials. For many of the late phase trials it will be very challenging to use this approach.
2	Core variables included, but not baseline variables.	Core variables included, but not baseline variables.	1. No dependence on any other ADaM domains. 2. Simple and clear data flow from source to ADaM 3. Early completion of ADSL development. 4. ADSL contains core variables that will be needed for other ADaM domain development.	1. Requirements by CDISC ADaMIG and FDA 'Study Data Technical Conformance Guide' not completely met. 2. Additional ADaM needed for important baseline variables that are not included in ADSL and individual ADaM domains. 3. An additional data merge step with ADBASE will be necessary for TFL creation. 4. CDISC ADaM "Analysis-Ready" principle not met.

CONSIDERATION AND DISCUSSION CONT.

Model	ADSL Content	Other ADaM Domain Content	Advantage	Challenge
3	All core and important baseline characteristic variables included.	Core variables included, but not baseline variables.	1. Requirements for ADSL by CDISC ADaMIG and FDA 'Study Data Technical Conformance Guide' completely met. 2. No dependency of other ADaM domain development on ADSL development. 3. Simple and clear data flow from source to ADaM.	1. Important baseline variables are not included in individual ADaM domains. An additional data merge step with ADSL will be necessary for TFL creation. 2. CDISC ADaM "Analysis-Ready" principle not completely met. 3. Early completion of ADSL development not practical. ADSL cannot be finalized before other ADaM domains relevant to baseline variables are done. 4. Additional subject level ADaM dataset needs to be programmed.
4	All core and important baseline characteristic variables included.	All core and important baseline characteristic variables included.	1. Requirements by CDISC ADaMIG and FDA 'Study Data Technical Conformance Guide' completely met. 2. "Analysis-ready" principle met. 3. No additional ADaM domain needed for fulfilling core and important baseline characteristic variable requirements. 4. Only draft ADaM domains relevant to baseline variables need to be developed earlier for ADSL.	1. Data flow needs to account for recursion. 2. Dependency of ADSL development on some other ADaM domains. ADSL cannot be finalized without drafting other ADaM domains relevant to baseline variables.

CONCLUSION

There are various challenges when developing the ADSL domain with baseline variable involvement. There are some fundamental factors that we need to take into consideration: traceability, analysis-ready, programming efficiency, clearness of connection between content and source data, and FDA requirements compliance, etc. It seems that there's no perfect solution or approach which currently exists or has been endorsed throughout the industry. Model 1 is the best option when ADSL is developed from SDTM/non-ADaM data sources (e.g. Phase 1 studies) as it has a simple process and clear traceability, meets FDA requirements and CDISC "Analysis-ready" principles. For late stage studies, we would prefer Model 4. This approach can be used in most of the programming scenarios and creates ADSL and other ADaM datasets that will meet both the CDISC guidelines and the FDA 'Study Data Technical Conformance Guide' recommendations with both core and baseline variables.

REFERENCES

[1] CDISC Analysis Data Model (ADaM) Implementation Guide, Version 1.0 Final. Published by CDISC December 17, 2009. Available at http://www.cdisc.org.

[2] CDISC Analysis Data Model (ADaM), Version 2.1 Final. Published by CDISC December 17, 2009. Available at http://www.cdisc.org.

[3] FDA Guidance for Industry: Providing Regulatory Submissions in Electronic Format — Standardized Study Data. Available at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm, December, 2014.

[4] FDA Technical Specifications Document: Study Data Technical Conformance Guide. Available at http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm, December, 2014.

[5] "The Biggest Challenges of ADaM" Terek Peterson and David Izard, from the Conference Proceedings for NESUG 2010.

[6] "Does All One-Record-per-Subject Data Belong in ADSL" Sandra Minjoe, from the Conference Proceedings for PharmaSUG 2012.

[7] "Interpreting CDISC ADaM IG through Users Interpretation" Angelo Tinazzi, from the Conference Proceedings for PhUSE 2013.

ACKNOWLEDGMENTS

The authors would like to thank Tracy Turschman, Jun Yu and the Vertex Statistical Programming group for their input, support and encouragement.

CONTACT INFORMATION

The authors can be reached at:

Hongyu Liu
Vertex Pharmaceuticals Incorporated
50 Northern Ave.
Boston, MA 02210
hongyu_liu@vrtx.com

Hang Pang
Vertex Pharmaceuticals Incorporated
50 Northern Ave.
Boston, MA 02210
hang_pang@vrtx.com