Paper TT12

Pooling Studies Through Code Traceability

Hamdi Chouikha, Bayer AG, Wuppertal, Germany
Kornelius Neufeld, Bayer AG, Wuppertal, Germany

ABSTRACT

Study pooling is a critical component of regulatory submissions, specifically for Integrated Summaries of Safety and Efficacy (ISS/ISE). Yet the process remains largely manual, error-prone, and constrained by non-standard study structures, legacy formats, and inconsistencies in derivations and value-level encodings. At Bayer we are reimagining study pooling through a traceability-driven approach. Enabled by the Verisian platform we are taking a blue ocean approach by shifting away from traditional programming towards code traceability-driven transparency. Here we demonstrate how we map variables, compare derivations, and harmonize value-level data, all independent of CDISC compliance. Using graph and AI technologies we can identify and explain differences across studies, understand underlying logic, and resolve discrepancies based on a chosen reference study. We walk through the implementation of this technology, highlighting gains in efficiency, quality, and demonstrate how traceability combined with AI is transforming the way pool studies at scale.

INTRODUCTION

An Integrated Analysis (IA) combines data from multiple clinical studies of a single compound or project into a Global Integrated Analysis Database (GIAD) - also called a “pool” - where datasets are standardized and harmonized for cross-study evaluation. This pooled database supports key deliverables such as Integrated Summaries of Safety (ISS) and Efficacy (ISE) for regulatory submissions, as well as inputs to Risk Management Plans (RMPs), labeling updates, authority requests, publications, investigator brochures, and broader clinical research support.

Pooling can be achieved at different levels: input studies can be aggregated at the level of SDTM(+), based on which the pool ADaM and TLFs are subsequently created (Figure 1).

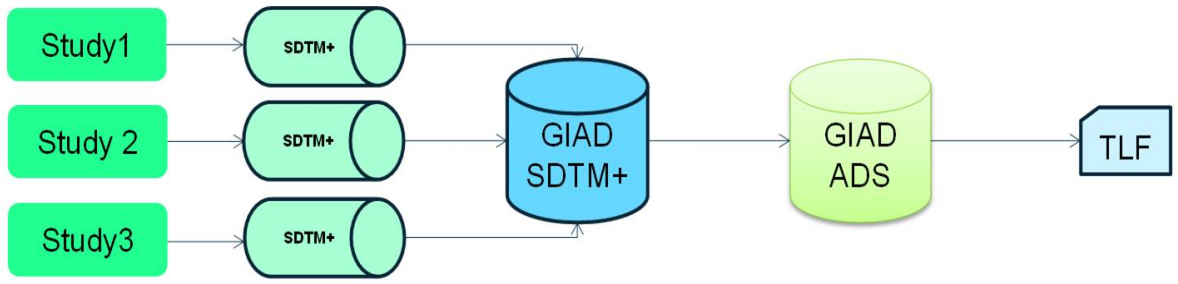

Figure 1: SDTM(+) pooling example

Alternatively, input studies can be pooled at the ADaM (ADS) level (Figure 2).

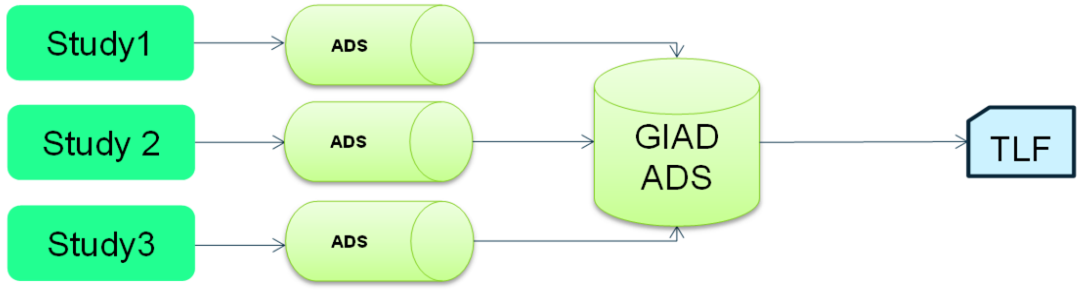

Figure 2: ADaM (ADS) pooling example

Pools can include previously created pools and combine them with additional studies, where the input pool and new study can be combined at different levels, leading to various complexities that need to be resolved for successful data integration (Figure 3).

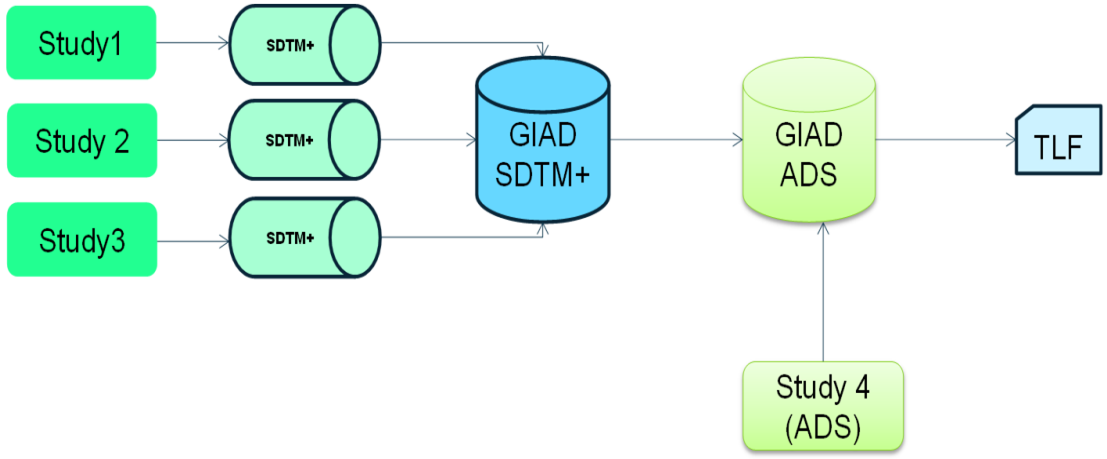

Figure 3: An example of complex pooling, where the final pool is composed of an input SDTM(+) pool that is combined with an additional study at ADaM (ADS) level

When creating clinical trial pools for integrated analyses we face many challenges. We have to consider varying data and quality standards as studies might have been prepared by different stakeholders (internal programming, CROs, other partners) and account for variations in study design, metadata, and data structures caused by evolving standards and codelists. Inconsistent derivations of variables, differing approaches to key definitions (e.g., baseline, treatment-emergent events, handling of missing data), and the absence of detailed derivation logic in metadata require extensive review of individual study programs. Moreover, multiple types of pools (e.g., monotherapy, placebo-controlled, combination therapy) are often needed, further increasing complexity.

THE POOLING PROCESS

Pooling consists of three main stages: Planning, Programming, and Validation. For the programming stage, the industry has a variety of solutions in the form of Scientific Computing Environments (SCEs) that vary in functionality and maturity. However, there is no meaningful support for planning and validating pools. This paper and accompanying presentation focus on how Verisian, in tight collaboration with Bayer, has built software to speed up the planning stage. We have also developed a solution to validate pools but will present these results at a later time.

POOL PLANNING

The same information sources are required for planning a clinical trial pool, regardless of whether it is based on SDTM or ADaM data. For all input studies, access to the Statistical Analysis Plan (SAP), specifications, data, and programming code is essential to ensure a comprehensive understanding of the study inputs and their relationship to the integrated pool being developed.

The planning process should be structured, reproducible, and make optimal use of all available information sources, such that all meaningful information is available at the right time. Equally important, this process should yield two key outputs: first, the finalized metadata and specifications for the target pool; and second, a comprehensive and trackable ToDo list detailing the required implementation tasks, no matter if simple (variables can be directly mapped) or complex (variables require custom derivation or harmonization). All of these goals are achieved with the Verisian Pool Planner.

VERISIAN POOL PLANNER

Once the input studies for a pool have been defined in the platform, their summary domains can be easily compared. The system highlights structural and content differences between studies, providing a clear, high-level understanding of their variability and facilitating an informed approach to integration (Figure 4).

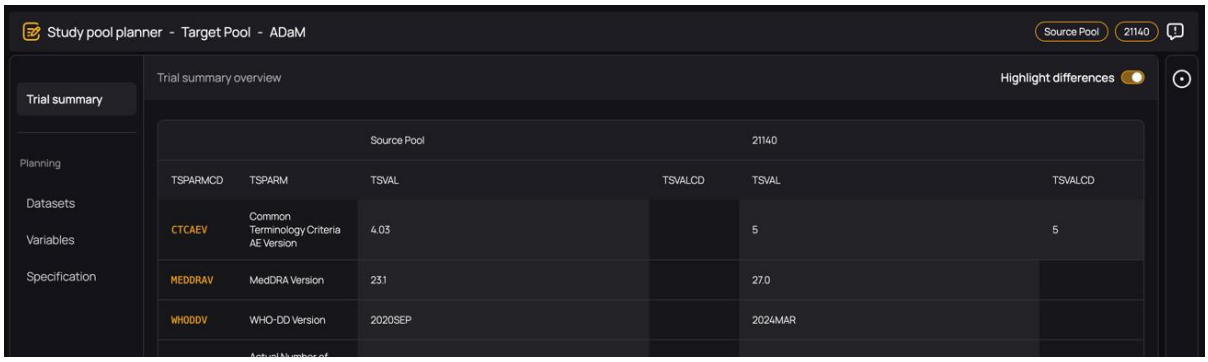

Figure 4: Example Trial Summary comparison across input studies

The process of designing the pool begins at the domain level, where all domains present in the input studies are displayed. The interface highlights all domains that exist in some studies but not others (mismatches). At this stage of the planner, users decide which domains to include in or exclude from the pool (Figure 5).

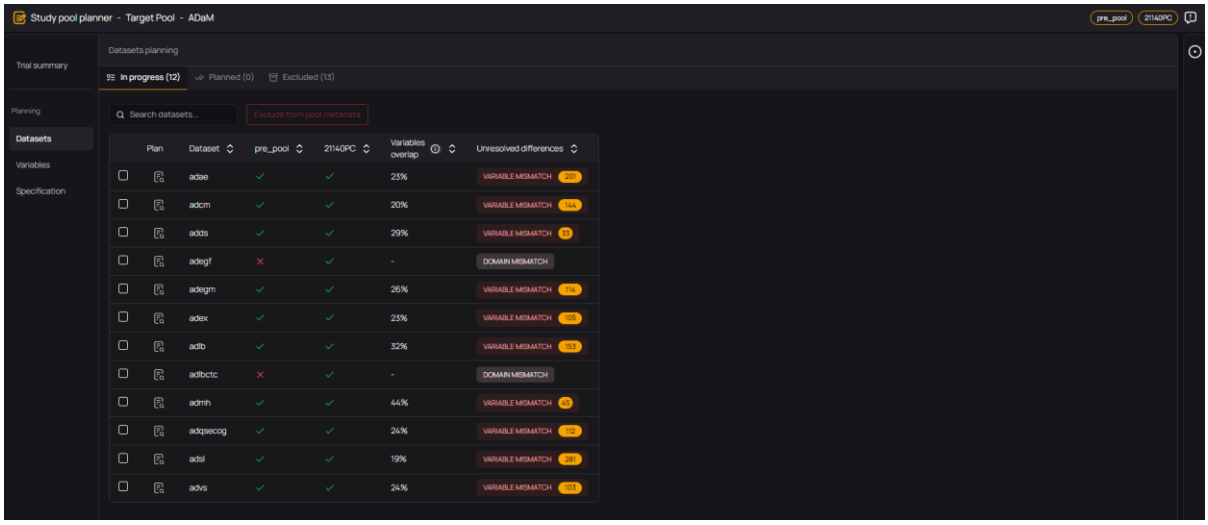

Figure 5: Choose the datasets to include in your pool based on pool input studies

As the next step, all variables for each included domain are displayed across studies, highlighting mismatches. At this stage of the planner, users decide which variables to include in the current domain of the target pool (Figure 6). Once this step is completed, the set of variables for that domain in the target pool is defined, and the process is repeated for all remaining domains.

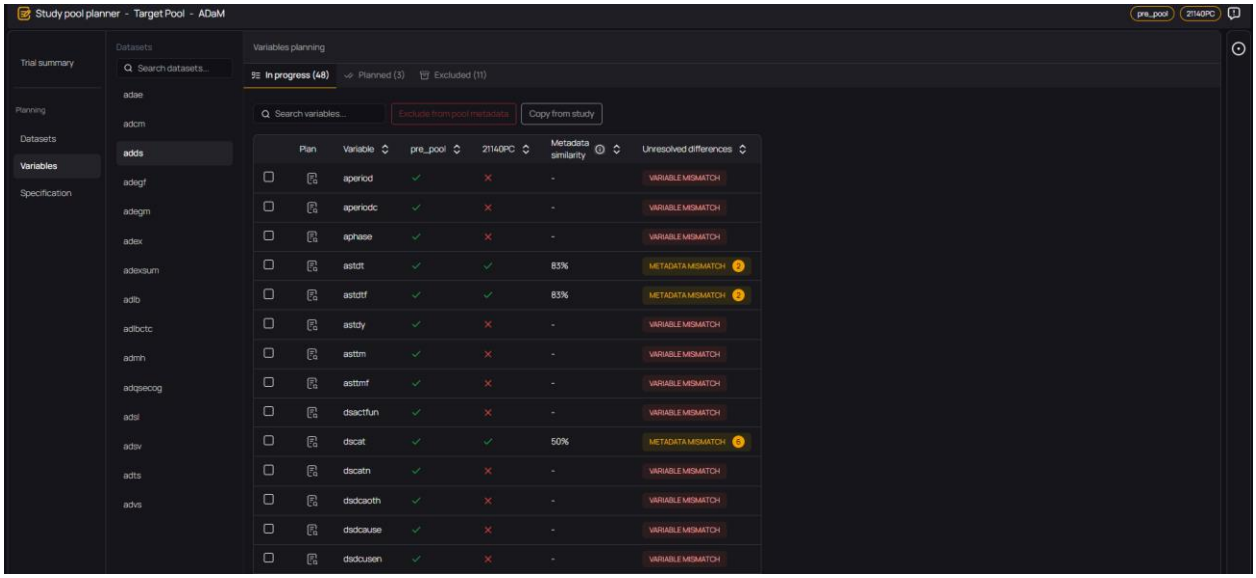

Figure 6: Choose the variables to include for each of your pool's datasets based on the pool input studies

Finally, for each variable in the pool, you determine the specifications based on the input studies or decide on a custom definition for the target pool. To make these decisions, you can refer to the metadata, data, and exact derivations of the variable of each of the input studies (Figures 7-9). Note that the code is filtered across your entire study to only those lines of code and associated comments that are required for the variable of interest, eliminating the need for manual tracing entirely, saving vast amounts of time and effort.

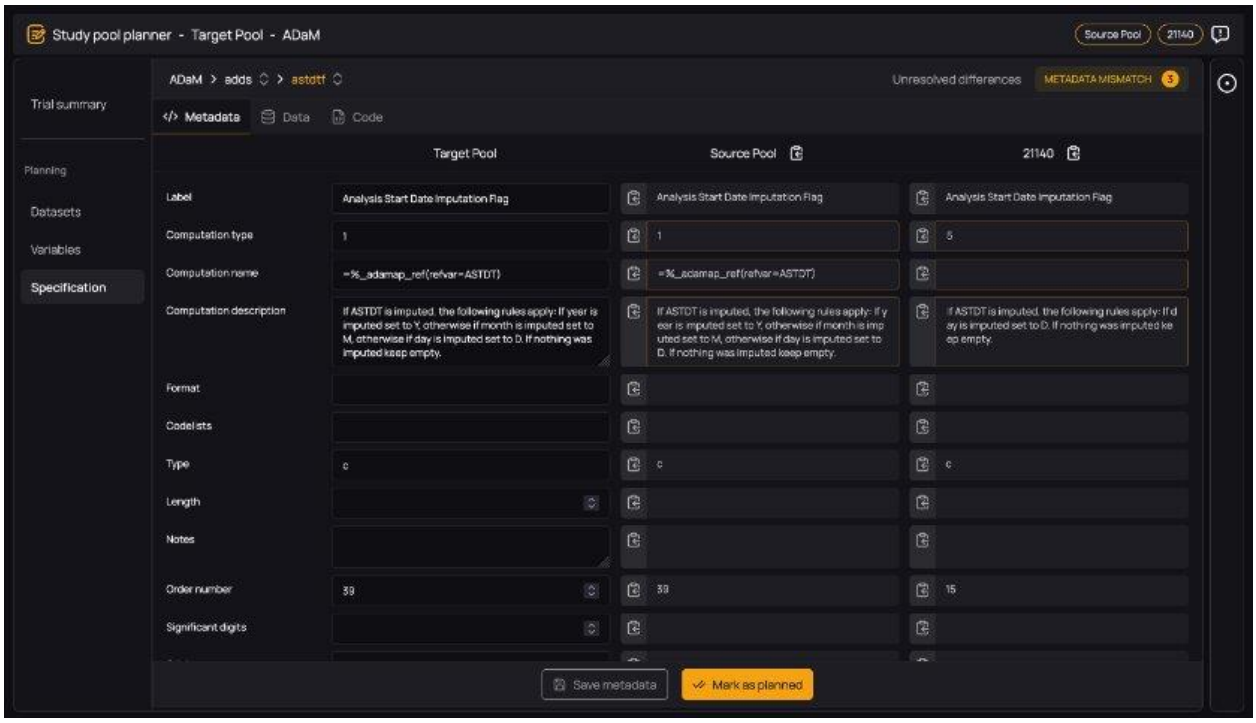

Figure 7: Easily compare a variable's metadata across pool input studies

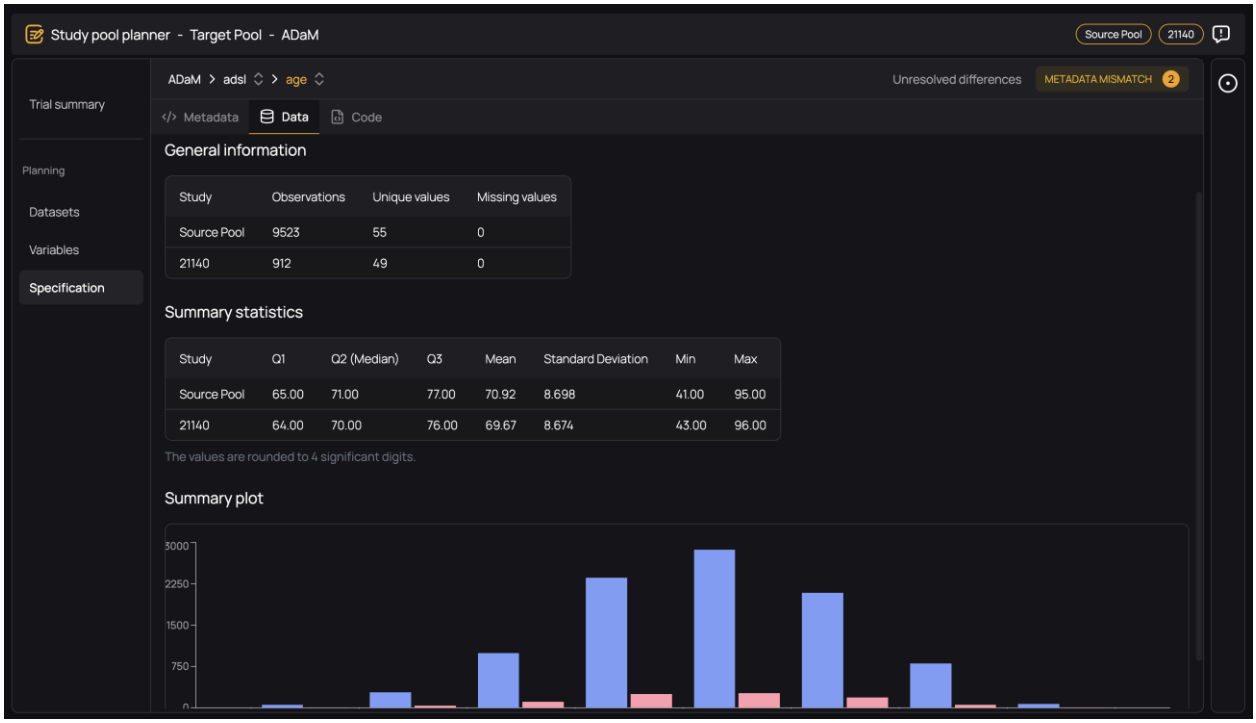

Figure 8: Compare a variable's data across pool input studies

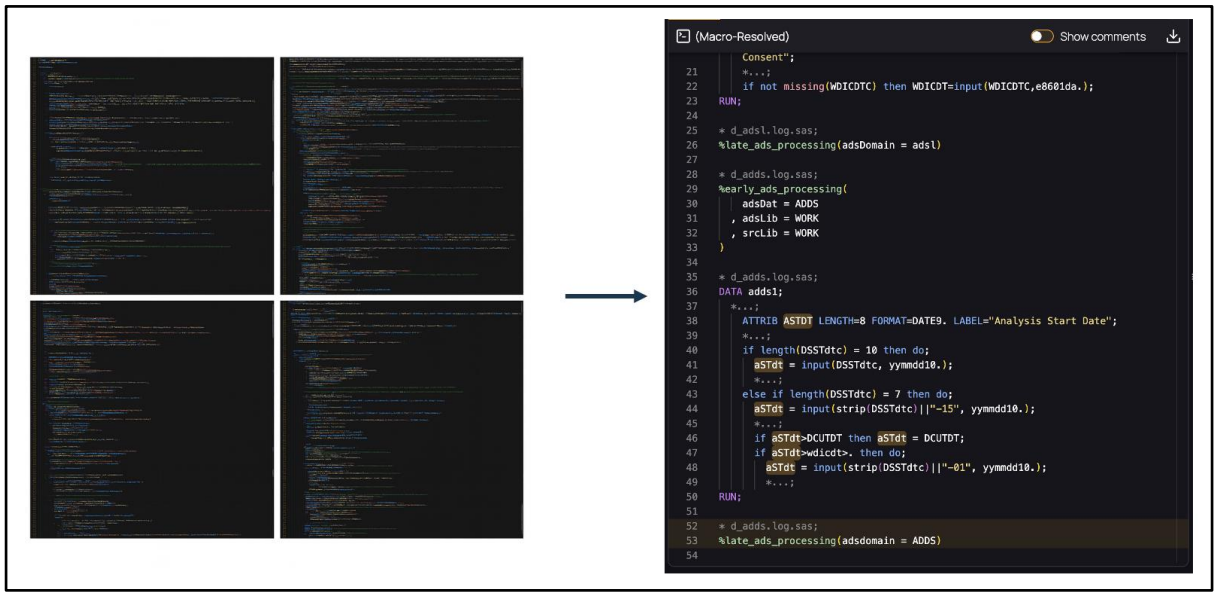

Figure 9a: retrieve the exact code used to derive the variable across each trial's code base, showing only the lines of code required for its calculation while hiding any other code

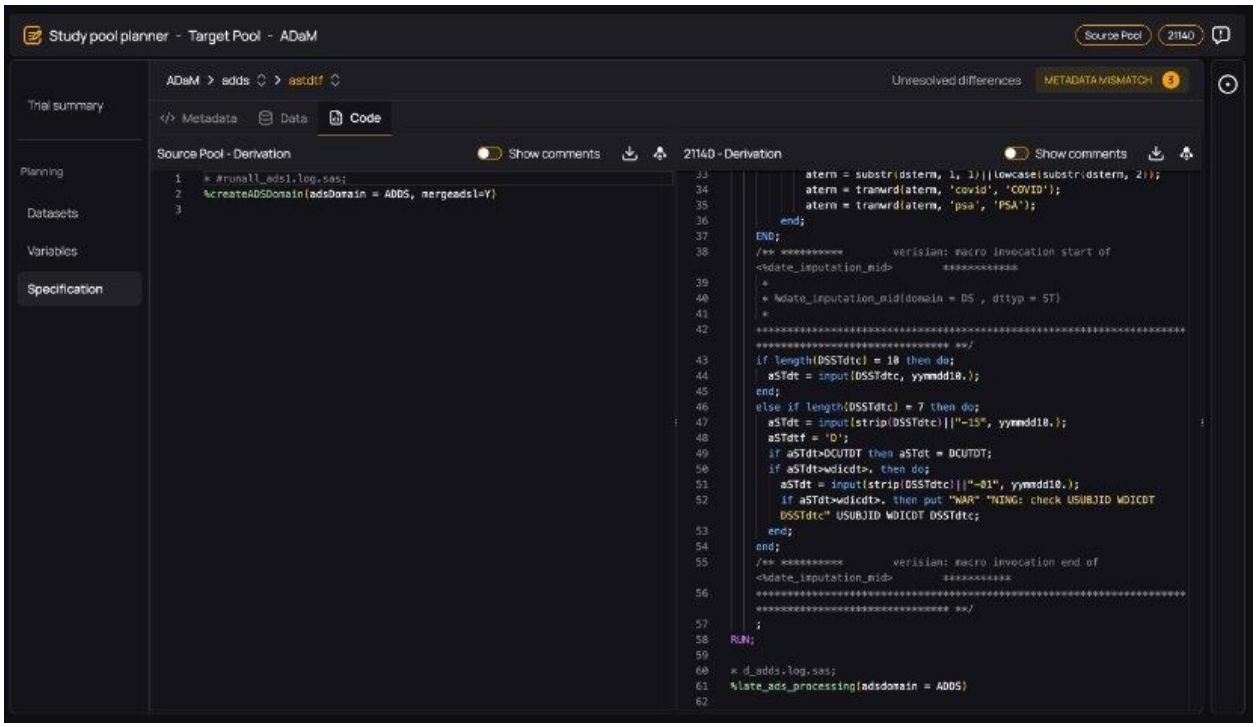

Figure 9b: easily compare the logic of derivations across input studies, saving hours of manual tracing

By retrieving only the relevant information and displaying it side by side, users can understand each variable at the metadata, data, and derivation levels, gaining a complete view across all input studies. This comprehensive perspective enables the confident creation of the target pool's metadata and specifications. For each variable, a corresponding task is automatically generated, allowing users to track the specific actions required to map that variable from each input study, resulting in the desired outputs: complete metadata specifications for the target pool and a trackable ToDo list (Figure 10).

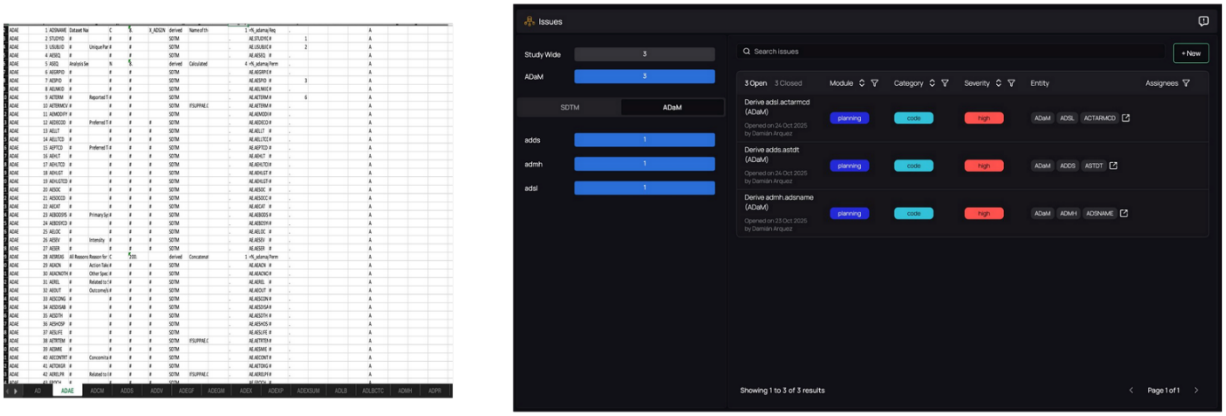

Figure 10: Specifications for the target pool are downloadable in a variety of formats, tasks can be assigned, tracked, discussed, and resolved in the Verisian Pooling module

POOL ADAM PLANNER

Once an SDTM+ pool is completed, one could use the input studies and their existing metadata and derivations to plan our pool's ADaM derivations since they are likely to be very similar (Figure 11).

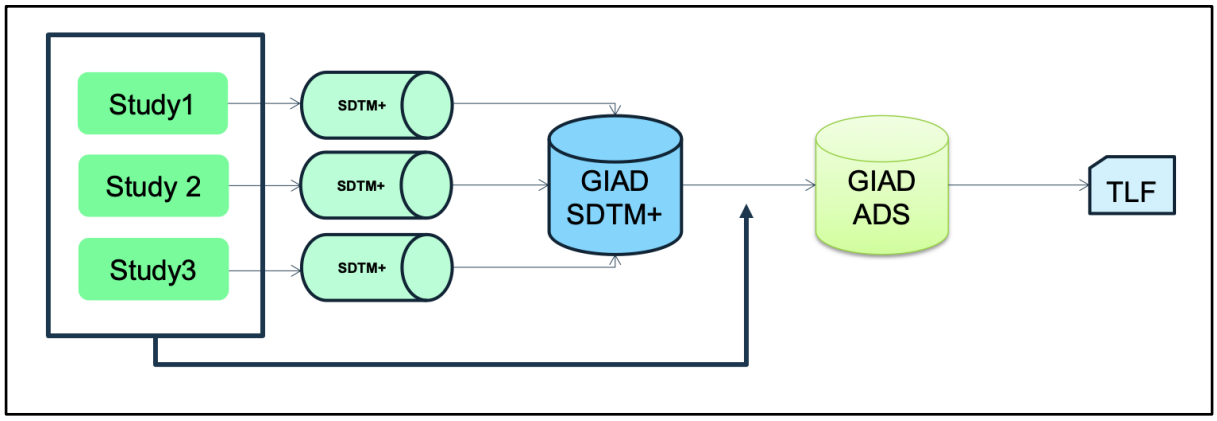

Figure 11: Planning the ADaM derivations in an SDTM(+) pool

Using Verisian's ADaM Planner, we can compare input studies not only at the level at which they are pooled (e.g., SDTM for SDTM pools), but at any level. For example, when working with SDTM pools, it is possible to compare the corresponding ADaM datasets of the input studies in the same way as previously shown. A user can use the cross-study domains, variables and within variable information (associated metadata, data, and implementation in code) to plan the pool's ADaM layer. Crucially, Verisian's code traceability allows the extraction of reusable code templates from input studies to be attached to the specification of the pool, providing exact implementation guidance for pool programming (Figure 12).

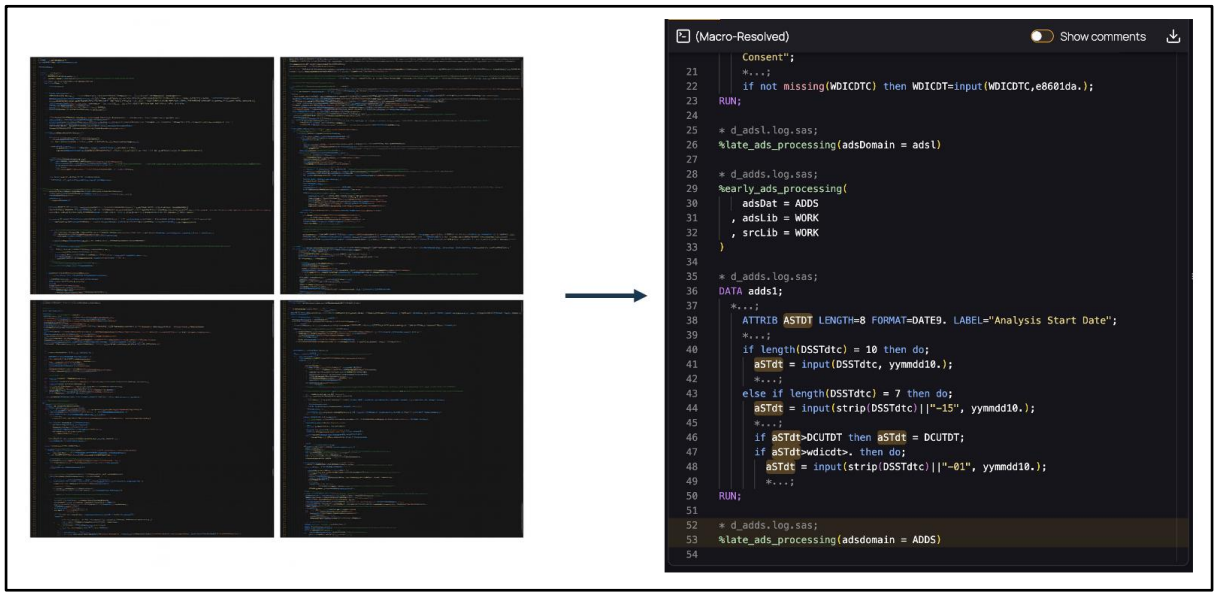

Figure 12a: Code templates are extracted automatically from pool input studies

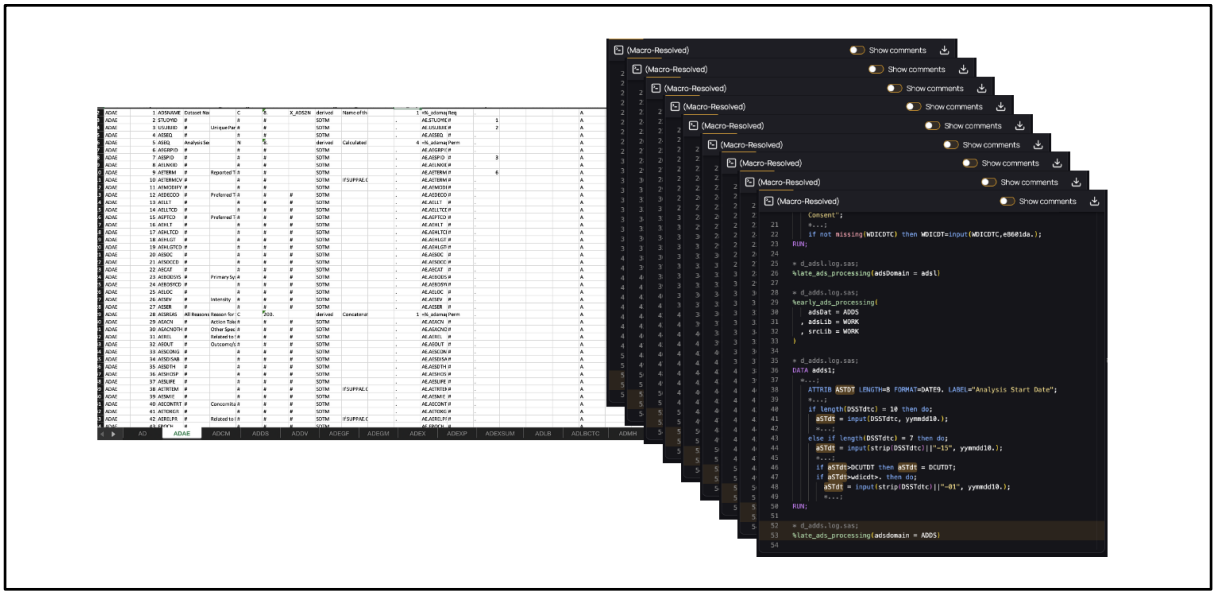

Figure 12b: Code templates can be attached to the pool metadata to guide pool programming

CONCLUSION

Integrated analyses (IAs) combine data from multiple clinical studies into a standardized, harmonized database (GIAD) to support regulatory submissions, safety and efficacy summaries, labeling updates, and broader research activities. Pooling such data is complex due to differences in study design, standards, metadata, and derivations, requiring a structured and reproducible process for planning data integration. The Verisian Pool Planner addresses these challenges by enabling systematic comparisons of study domains and variables, highlighting inconsistencies, and guiding the creation of metadata specifications and a trackable ToDo list for all required programming work. Through code traceability, Verisian allows users to review derivations and extract reusable code directly from input studies, streamlining and standardizing pool programming. Together, these capabilities improve the efficiency, transparency, and reproducibility of integrated analyses for SDTM and ADaM-level pooling, as well as ADaM planning.

ACKNOWLEDGMENTS

We would like to extend our gratitude to the entire Verisian Team for their dedication and invaluable contributions, and for their support in preparing the slides and this paper. Their traceability-driven solutions have been instrumental in shaping this work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Hamdi Chouikha and Kornelius Neufeld
Company: Bayer AG
Address: Aprather Weg 18a, 42113 Wuppertal
Work Phone:
Email: hamdi.chouikha@bayer.com or kornelius.neufeld@bayer.com