PharmaSUG 2024 - Paper DS-400

AI and the Clinical Trial Validation Process – Paving a Rocky Road

Steve Ross and Ilan Carmeli, Beaconcure

ABSTRACT

The validation of outputs in a clinical research environment acts as a guarantor process, confirming the accuracy and validity of the trial results, the investment of doctors, patients, and caregivers in the efficacy and utility of the trial, and the reputation of the sponsor and/or CRO conducting the validation. Double programming has done this heavy lifting for decades. The increasing application of AI (ML and NLP) gives statisticians and programmers unprecedented opportunity to apply this technology wherever validation takes place – during the development cycle to aid teams in getting to the right output sooner, and at the end of a study to check that tables in a package match what is expected for the deliverable.

This paper shows how to use AI and automation capabilities within Verify for activities including validating that the titles, footnotes, format, and content of the output matches the display in the mock shells, and whether big N's, small n's, and counts in the body of a display are logical and accurate within and across tables. This paper also illustrates a documentary audit trail that captures the end-to-end decision-making process and feedback from contributors as tables are revised from early deliverables to final. Automating critical iterative tasks can free up validation time and brainpower for statisticians and programmers to focus on the bespoke aspects of clinical trials, such as primary efficacy endpoints, complex algorithms and analyses, and ensuring that the truth of the data is told – paving the rocky road to the final product.

INTRODUCTION

Double programming combined with visual review have long formed the backbone of the validation process of clinical trial outputs, and are guarantors that the study is conducted and analyzed according to prespecified instructions designed to ascertain the safety and efficacy of medications under investigation.

This decades-old process was designed with then-current capabilities in mind, as well as then-current limitations. Paper submissions dominated the NDA submission process. In 1995, when this author entered the workforce, state of the art meant Vax terminals for programmers, and paper outputs for reviewers.

U-Hauls loaded with pallets of paper outputs arrived on FDA's doorstep every day, with submissions that had been manually reviewed at the very last stage by teams of statisticians, programmers, clinicians, medical writers, et al. "PROC EYEBALL" was used to ensure that titles and footnotes matched the body of the tables, the table numbering schema matches what's in the table of contents, that the body of the tables accurately reflects the intent of the mock shells, and that the tables are consistent both within a display, and across displays.

Similarly, twenty years ago, the DATA _NULL_ step in SAS was the de facto standard for generating the actual display output files (Tables, Listings, and Figures). Generating a table involved a lot of "PUT @ ZZZ" statements, which had to be tweaked and updated each time new data made it into a table. Today, with the advent of PROC REPORT, thousands and thousands of programmers have at their fingertips a vastly superior technology that revolutionized how displays were created and output. ODS in 2004 was yet another SAS "great leap forward." Other languages such as R, SPSS, and Python have made similar gains over time. With each of these advances, efficiencies were created, accuracy improved, and time to market (and to patients) has decreased.

And yet, the *process* of validation remains much as it was decades ago. We have tools in our toolkit (macros, batch files, etc.) that automate pieces of this process, but it is largely still a very manual undertaking, particularly as a deliverable reaches completion. Programmers still seek the elusive "`No unequal values were found. All values compared are exactly equal`". Statisticians and other reviewers visually review for formatting, internal consistency within tables, and cross-table consistency across tables.

Herein lies a fundamental disconnect in the ideal 'division of labor' between man and machine. We program some checks, but leave many others to the human eye. We know that humans are better at understanding, synthesizing, and reasoning. We also know that machines are better at high-volume, repetitive tasks. Advances in computing technology place us at an inflection point. A new platform for validation called Verify has been developed for the purpose of rebalancing and reimagining the TLF validation process and workflows, and is an AI-enabled platform that unifies this critical piece of the drug development cycle. Verify conducts elements of both programmatic validation and visual review, taking on the hard but repetitive tasks that can be done programmatically, leaving programmers and statisticians time to focus on the higher-order tasks such as ensuring that primary endpoint derivations are accurate, that tables make sense in context, and that inferences drawn from the TLFs are clearly conveyed. From the first deliverable in a study to the last, Verify creates a digitized database that facilitates linking of like information across table sets, identifies 'discrepancies' (places where metadata align, but the target data points differ), and brings together all reviewers into the same environment for seamless collaboration.

THE HIGH COST OF MANUAL VALIDATION

Despite efforts to streamline, downsize, and harmonize the process of analyzing clinical trials data, the fields of statistics and medicine continue to develop new measurement tools that tax our ability to accurately represent the meaning of the data we collect. However, the advent of 'learning technologies' such as the misnamed "Artificial Intelligence" have given rise to myriad new use cases in pharma research. AI has enabled advances throughout clinical research, from adaptive study design to advanced statistical analysis techniques.

Whether you work in Big Pharma, Small Pharma, CRO, or Biotech, the significant cost associated with bringing a new therapy to market demands that what you submit to regulatory agencies is of the highest quality, the first time. The FDA can, and does, delay or reject an application if it deems that the submitted package of analyses is not of sufficient quality to evaluate the safety and efficacy of the treatment under study. Such setbacks can cost companies millions in lost revenue, and perhaps more in reputation. It may even mean that companies run out of money before they can get their drug submitted. The FDA does not specify what form this validation should take. The onus on quality deliverables falls to sponsors to define, and we all fall back on familiar processes which can be inefficient and ignore basic differences in the way people and machines process data.

Statisticians and programmers everywhere can relate to this scenario: a deliverable nears completion and the database is locked. There is now a fixed window in which to extract data, rerun (and revalidate) SDTMs, AdaMs, and TLFs, and neatly package them. At some point during the review process, an error is discovered that requires a rerun of, say, an AE table, a series of AE tables, or perhaps the whole table set. The whole process needs to be repeated, even more quickly, and under even greater time pressure. Types of validation that are typically done at the end such as format checking, title and footnote review, internal consistency, and cross-table consistency, take on a heightened urgency as a large volume of checking is compressed into a smaller and smaller window. It is in exactly this scenario that a process that includes AI (machine learning, natural language processing, etc.) shines. In reorienting validation processes to take advantage of the natural strengths of man and machine, humans (statisticians, programmers, other reviewers) can focus on higher order processing such as design, understanding and meaning, while allowing Verify to churn through a 300 table set for inconsistencies in titles and footnotes, internal consistency within counts in individual tables, or cross-checking that the counts in overall AE summaries are repeated in subsequent tables in the AE set.

FIT FOR PURPOSE: WHERE MACHINE LEARNING FITS INTO A NEW VALIDATION METHODOLOGY

The volume of data collected in clinical trials is a natural target for taking advantage of AI-enabled computing power. In harnessing this technology, high volume, repetitive tasks can be done quickly, efficiently, and are infinitely repeatable to a high degree of accuracy. The iterative nature of validation and revalidation (and revalidation!) of the package in the scenario above is time consuming, exhaustive, and can be made much easier using current machine learning methodologies. Best of all, machine learning algorithms become better over time, and are not limited by the parameterization that current macro-

enabled validation employs. As a table set grows through the duration of the study, the breadth of the display types also grows. With this, an AI model is able to ‘understand’ more of the connections that link outputs.

Figure 1 illustrates how the structure and content of each display contains contextual information that facilitates understanding.

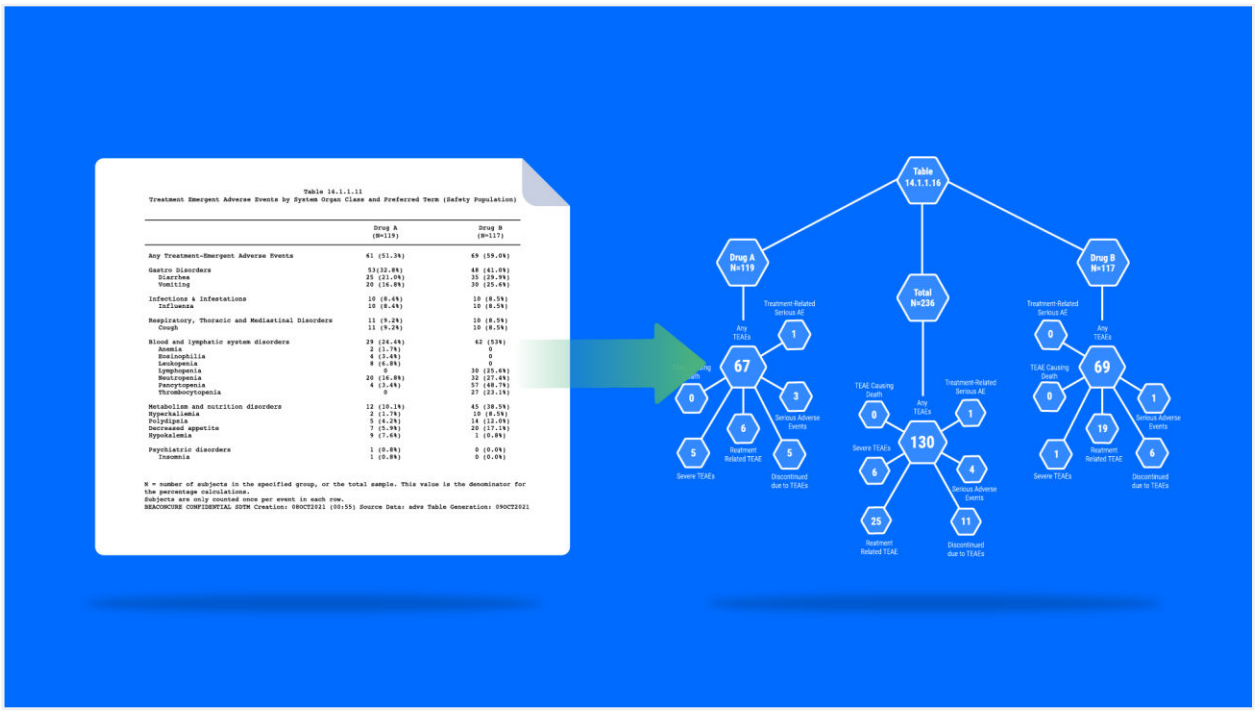

Figure 1. A table is parsed by Verify for inclusion in the AI-enabled digital database.

When this table is uploaded to Verify, the system first digitizes, or parses, the table to ‘harvest’ relevant metadata from the headers and footers, such as populations, datasets, treatment information, datetime stamps, etc. This metadata is attached to the cells in the body of the table, and is compared to referents such as a table of contents or a mock shell, or is used to search for like entities in the digitized representation of tables in the set.

Figure 2 shows how metadata comparisons across tables are facilitated by the digitization process.

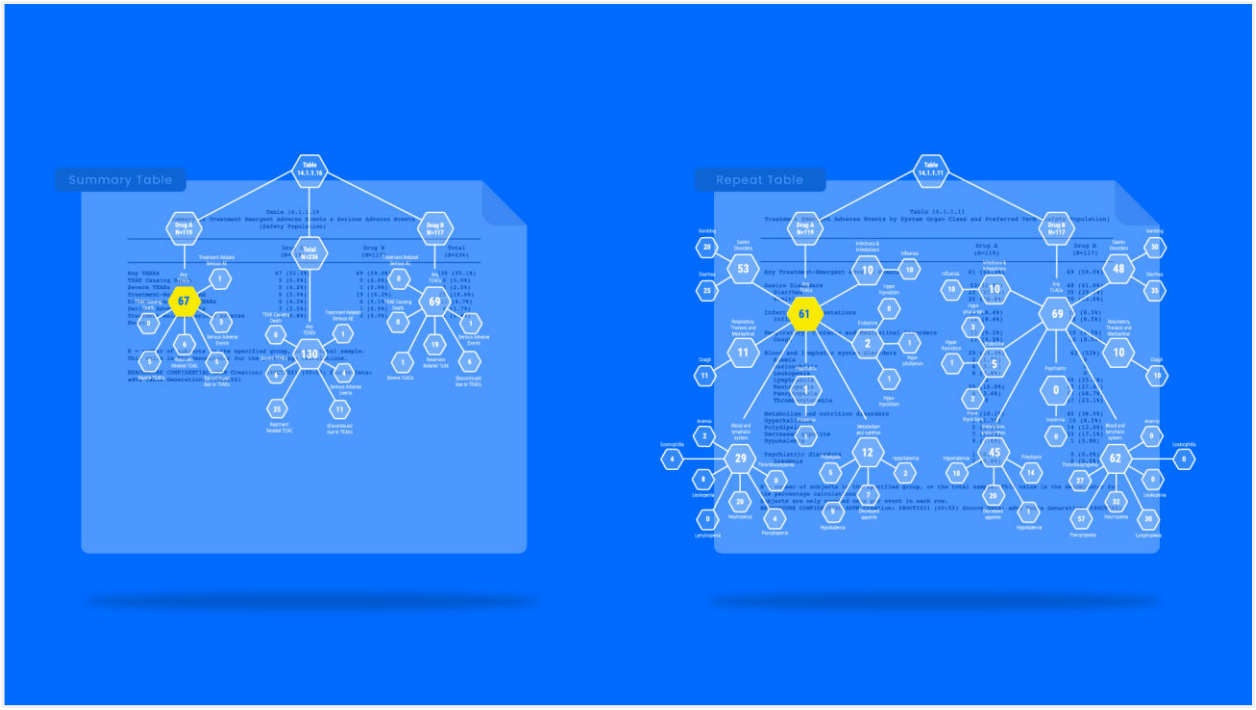

Figure 2. Two displays with shared metadata. Highlights indicate a discrepancy between values.

In the above case, the AI-enabled database has found ‘entities’ in two tables that share metadata, but do not match, flagging a discrepancy to be investigated. The table data are ‘self-referencing’, meaning that the comparison occurs between the two tables in the output package, i.e., the machine performs a check on cross-table consistency that is usually left to human visual validation. In traditional programmatic validation, other referents are used as well – the most obvious candidate being the study ADaM datasets. With its most recent release, Verify has also added ADaM datasets to the list of referent data it uses in validation. In the following example, tabular metadata extracted during the digitization of the TLF set yields information that allows Verify to check that the n’s in the column headers match ADaM data subsetted in the same way.

Figure 3 illustrates how table metadata is used to obtain the appropriate subset of the ADaM data for comparison purposes.

Figure 3. The digitized metadata facilitates the link between the tabular output and the ADaM data, highlighting a discrepancy.

Still another example of AI-enabled validation enabled by the digitization of .rtf files is hierarchy checks for decreasing n's. Traditionally, this validation is performed visually, not programmatically, but Verify automates this validation check.

In general, counts in visit-based displays should drop over time as subjects exit the study for various reasons, and while not unheard of, should be investigated.

Figure 4 shows two values that are flagged for attention. In this example, the post-baseline n count is greater than the number of subjects at baseline.

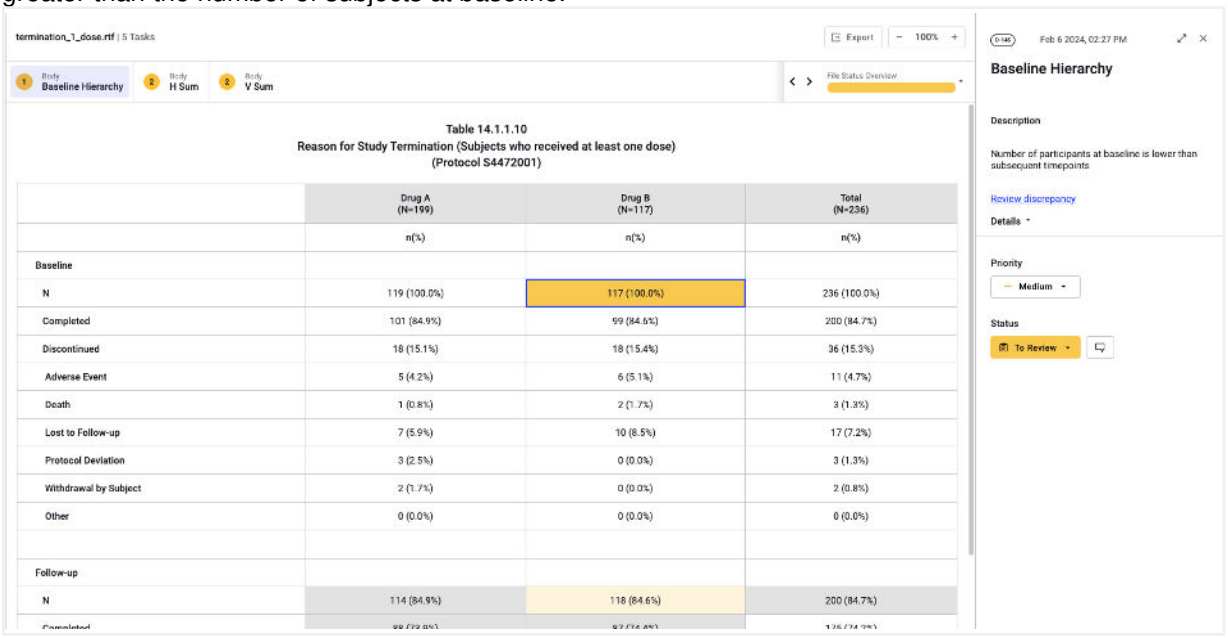

Figure 4. The Verify review interface highlights two values for investigation.

Finally, the review process is complex enough as it is. Current processes still have multiple review streams that work in isolation from each other – from the programmer and statistician who document their progress on a 'QC spreadsheet', to the internal reviewer who logs comments on a second spreadsheet, to other external reviewers who might provide comments via yet another spreadsheet, or a series of emails, or even a marked up copy of a .pdf file. Verify is a unified platform that brings together the entire review process into one place, to streamline and facilitate the accurate capture and review discussion from all sources. It also increases transparency, as different reviewers can interact with one another directly within the platform, and completes the process with an audit-ready output suitable for capturing the end-to-end process of TLF validation.

Figure 5 shows the ability to work within a single platform, which helps facilitate review, remove ambiguity in decisions, and preserve the conversation for future deliverables.

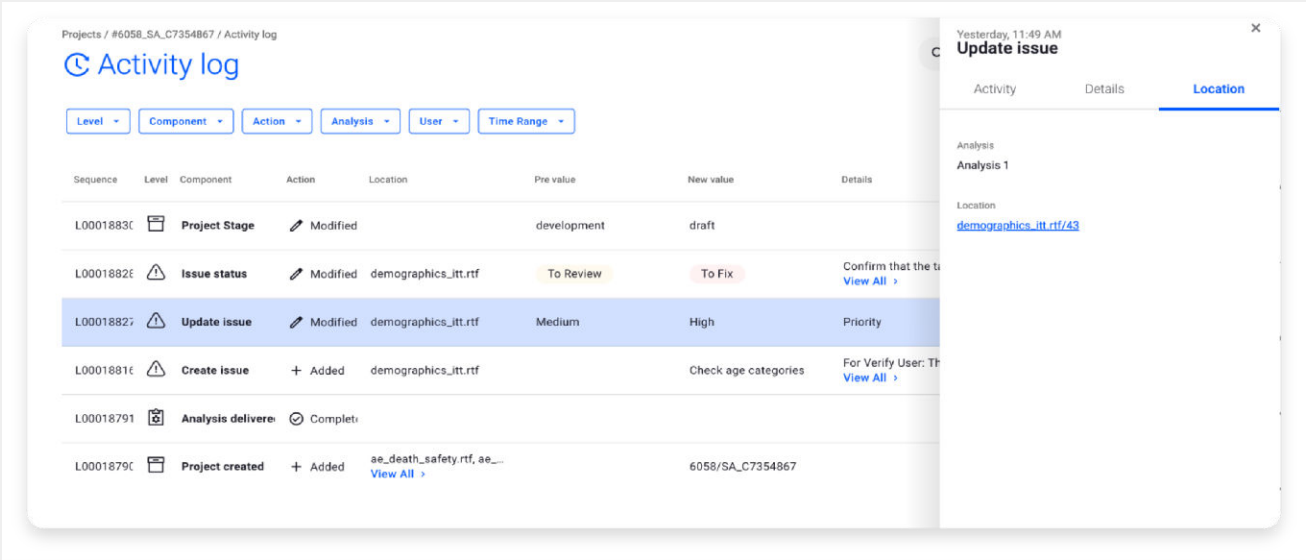

Figure 5. Verify documents and captures the full decision-making and update process for each display, in one interface.

CONCLUSION

The process of TLF validation is complex, time consuming, iterative, and always short on time. A reimagining of the workflows in the validation process allows for the full functional alignment of human and machine power for tasks to which they are best suited, particularly in the high speed and high stress environment of a clinical trial. Tools such as Verify facilitate this transformation through three technology pillars: digitization (the creation of a digitized database); smart validation using AI-enabled checks developed for users; and online collaboration, facilitating a fully transparent discussion of all aspects of deliverables. Paving a very rocky road.

ACKNOWLEDGMENTS

The authors would like to thank Sarah Michaels, Achinoam Ravet Perel, and Berber Snoeijer for their indispensable contributions in the development and production of this paper.

CONTACT INFORMATION

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

Steve Ross
Beaconcure
Steve@beaconcure.com
www.beaconcure.com

Ilan Carmeli
Beaconcure
Ilan@beaconcure.com
www.beaconcure.com