Paper TT16

Taking down the fence between Biostatistics and Medical writing

Rob Wartenhorst, GSK, Amsterdam, The Netherlands

ABSTRACT

Even though we distribute files digitally, biostatistics still share the results with medical writing in the same format as we did decades ago. The only change being a transition from paper to PDF. Medical writers then need to interpret and turn this into text for the CSR. How would it look if biostatistics could share the analysis results as data instead? By leveraging a scientific writing copilot, it would eliminate the potential for manual errors in the CSR text and remove repetitive writing tasks.

INTRODUCTION

This paper will showcase how the data flow does not need to stop at the biostatistics group, instead it keeps traveling to the next consumer, the medical writer. Increasing the automation is enabling us to deliver faster with better accuracy. Even though advanced natural language processing is used by the scientific writing copilot, at the biostatistics side some adjustments have to be implemented to enable the automation.

ADJUSTMENTS ON BIOSTATS SIDE

The initial scope of the scientific writing copilot proof of concept was to consume RTF tables, identify the desired numbers from the RTF table(s), to be fed into the automated interpretation process. While this allows for independent development of the copilot product, we are creating from highly structured analysis data, some vaguely structured data in RTF format, which then in turn has to be converted back into highly structured data. What if biostats and scientific writing would collaborate? Whereby scientific writing could focus on configuring the copilot, and biostats would focus on making the data available? Historically there has not been a need for consumption of results data in an automated way. As long as we could generate a table in the same display format, that we have been creating them in for decades, all was good.

The tools we had in place did create a kind of pre-display dataset (ready for a simple proc report for example), mainly for QC activities. This was a good starting point. For automation purposes having a defined set of columns and variability in the number of rows is much more desired than having variability in the number of rows. The modifications that needed to be made, came down to transposing the pre-display data, from a wide format to a skinny-er format using standardized column names and additional columns that identify the table the data belongs to. This is currently an additional task, taking about 2 days, of which most can be done before database lock. The learnings from this additional task are feeding into our future standard Analysis Results Data standard, which will remove the need for the transformation and will allow for fully automated passing through of the results data.

In the example below the transformed data has additional metadata in the cells with the grey background. This is the main difference. The remaining labeling information is a more structured and less PROC REPORT ready.

Example pre "transpose"

ANALYSISVAR	GRP1_LABEL	GRP1_DENOM	GRP1_PERC	GRP1_NUM	GRP2_LABEL	GRP2_DENOM	GRP2_PERC	GRP2_NUM
Age	GROUP1				GROUP2			
n	GROUP1	200		200	GROUP2	100		100
Mean	GROUP1	200		51	GROUP2	100		50
Median	GROUP1	200		52	GROUP2	100		49
Minimum	GROUP1	200		3	GROUP2	100		1
Maximum	GROUP1	200		98	GROUP2	100		100
Age category	GROUP1				GROUP2			
>=50	GROUP1	200	55	110	GROUP2	100	40	60
<50	GROUP1	200	45	90	GROUP2	100	40	40

Example post "transpose"

TITLE	POP	ANALYSISVAR	GROUP1	GROUP2	RESULTVAR	RESULTVAL
Summary of Basline Characteristics	Exposed Set	REF	Table 14.1.1	Table 14.1.1		
Summary of Basline Characteristics	Exposed Set	N	200	100		
Summary of Basline Characteristics	Exposed Set	n	200	100	Age	
Summary of Basline Characteristics	Exposed Set	Mean	51	50	Age	
Summary of Basline Characteristics	Exposed Set	Median	52	49	Age	
Summary of Basline Characteristics	Exposed Set	Minimum	3	1	Age	
Summary of Basline Characteristics	Exposed Set	Maximum	98	100	Age	
Summary of Basline Characteristics	Exposed Set	n	110	60	Age category	>=50
Summary of Basline Characteristics	Exposed Set	%	55	60	Age category	>=50
Summary of Basline Characteristics	Exposed Set	n	90	40	Age category	<50
Summary of Basline Characteristics	Exposed Set	%	45	40	Age category	<50

WHAT IS GENERATED BY THE COPILOT

The scientific writer configures the style of the report in the studio of the copilot tool. This ensures consistency within the company standard. The tool then generates the interpretation, like the example below, which is inserted into the CSR using a WORD Add-In.

Example interpretation text

> **Unsolicited AEs** were reported in **123** (**12.3%)** of the patients in the **study drug** and **123** (**12.3%)** of the patients in the **control group**.
> **Grade 3 adverse events** were reported in **1** (**1%**) of the patients in the **study drug** and **1** (**1%**) of the patients in the **control group**.

The content in bold is provided by the data. This allows for configurations be set up long before study completion, with completion of the content automatically from the data without any chance of typo's.

IS IT A SUCCESS?

The copilot has been rolled out for the vaccines therapeutic area. By nature of the therapeutic area, the vaccines studies are quite similar and a good fit for automation. As expected, the tool the tool produces accurate and standardized results. No errors in numbers are found. Obviously, constancy across studies is very high with machine generated text. Most importantly the CSR is generated much faster. Using the sections available now, there is a 50% time saving in authoring the first draft. Following the positive outcomes from the vaccines studies, the other therapeutic areas in GSK are also adopting the copilot. For now, until we have our new analysis results data standard in place there is some additional programming work needed for the data transformations, which is more than offset by the savings of time in the medical writing team.

CHALLENGES

On the biostatics side there are no technological challenges. The challenge is, until we have a standardized analysis results data structure in place, there is the potential for delay at Statistical Analysis Complete time due to the extra activity of creating the data for the copilot. Even if using the tool saves multiple days for scientific writing, biostatistics can be hesitant to commit to adopting this automation as there is a risk of missing their KPI.

On the scientific writing side, people can fear of the automation taking over their job. It however takes away the tedious and error prone task of writing the interpretation, giving the writer more time to focus on the much more interesting parts of the CSR.

A technology challenge that our tech department is working on, is passing along the data in a controlled environment to make this a seamless integration between all the tools.

WHAT IS NEXT

As GSK is transitioning from their legacy SCEs into a new SCE, and modernizing the reporting tools, an important aspect of this modernization is the implementation of a proper analysis results data structure. Not only will this separate the rendering of the analysis results from the generation of the analysis results, providing benefits to the clinical programmer, such as using a different programming language for rendering than for creating the analysis results, it also opens a wealth of opportunities for producing the CSR much faster. A CSR can be completely drafted before database lock, using mock data. With automated flow of data through all the steps, the CSR can then be refreshed with final data within hours after database lock. Clinical programmers also don't have to worry about formatting the tables or deal with the last-minute change requests on table format, rounding or precision of results. Finally, although this is more challenging than generating interpretations, the process would be even more efficient if the copilot tool could create and insert the intext tables automatically.

CONCLUSION

It is very exciting to see the information from clinical studies flowing in the form of data, instead of being handed off in the form of digital paper. For clinical programming, it can be facilitated with little or no additional effort, while providing a significant time saving for the scientific writer.

ACKNOWLEDGMENTS

The author would like to thank Eishita Agarwal from the medical writing innovation team for driving this innovation. Also the author would like to thank Karine Bourgeois for managing the data transformations on the biostatistics side.

CONTACT INFORMATION

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

Rob Wartenhorst
GSK
Amsterdam, The Netherlands
Email: rob.x.wartenhorst@gsk.com