Paper AS05

Improve programming and traceability with datasets for statistical analysis result

Jeppe Rømer Juul, Novo Nordisk A/S, Copenhagen, Denmark
Henning Pontoppidan Föh, Novo Nordisk A/S, Copenhagen, Denmark

ABSTRACT

The result of a single statistical analysis is often presented in several different tables and figures. Performing the same analysis several times in different programs increases execution time and the risk of inconsistencies in the presented results. To avoid this issue, we propose to divide the programs producing tables and figures into two batches. Programs in the first batch perform the statistical analyses based on ADaM data and save the results in datasets, which we will call analysis results datasets. Only these programs will be submitted to regulatory authorities. Programs in the second batch read the analysis results datasets and present the results in tables and figures. This approach reduces execution time and ensures consistency in the presented results. We will propose how best to document and provide traceability for the analysis results programs and datasets, and share feedback received from FDA and PMDA on this approach.

INTRODUCTION

Programming the statistical analyses for a clinical trial may be a complex and computationally demanding task. This is in particular the case if the statistical analyses include imputation of missing data using for instance multiple imputation or mixed-effects models for repeated measures. Challenges arise when a single analysis result needs to be presented several times, for instance in a table and as part of a forest plot. Performing the same analysis several times in different programs increases execution time and the risk of inconsistencies in the presented results. Also, the programs become complex to read and understand when they include both code to perform the statistical analysis and code to present the results, such as a proc template procedure including specification of axes, labels, line colours, footnote definitions etc. This added complexity is especially problematic since these programs are key deliverables in regulatory submissions.

In this paper we present a programming strategy to address these challenges. The strategy includes the creation of a new type of datasets, which we call analysis results datasets.

WHAT ARE ANALYSIS RESULTS DATASETS?

The programs performing the statistical analyses traditionally read ADaM datasets such as ADLB and output a table or a figure presenting the analysis results. These programs contain code for performing the statistical analysis as well as code for displaying the result with the appropriate layout (see Figure 1). Typically, the analysis and layout parts are entangled within the programs, which may therefore become fairly long and complex.

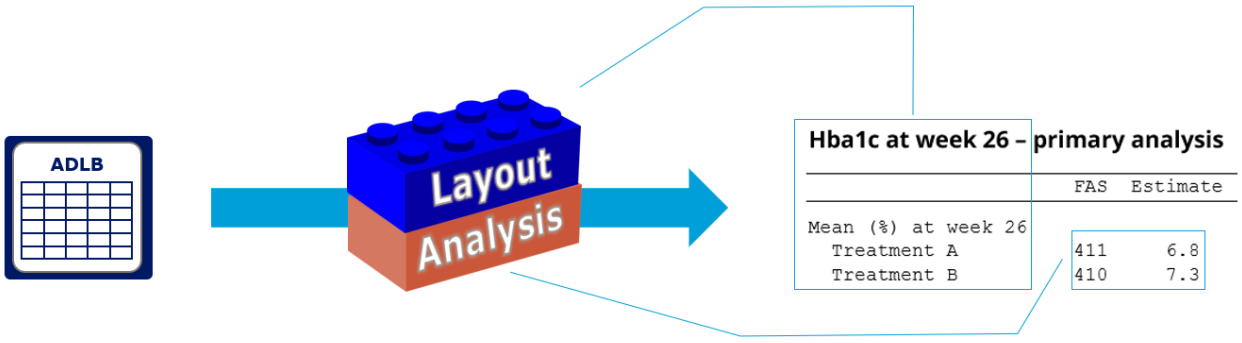


Figure 1 Standard programming approach for statistical analysis outputs. A single program performs the statistical analysis and displays the result in a table or a figure with the appropriate layout.

Instead of the traditional approach, we propose to split the statistical analysis and the layout part into two separate programs. Doing this, we introduce a new type of dataset that contains the statistical analysis results and metadata associated with the analysis – this dataset we denote an *Analysis Results Dataset* (see Figure 2 Alternative programming approach for statistical analysis outputs. One program performs the statistical analysis and stores the result in an Analysis Results Dataset. Another program reads this dataset and displays the result in a table or a figure with the appropriate layout.). Since this is not an ADaM dataset we suggest naming it with the two-letter prefix "AX".

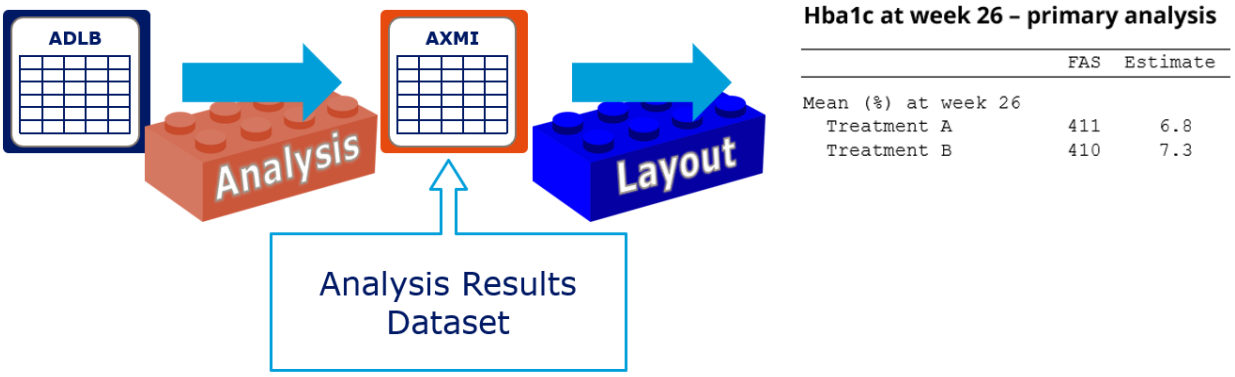


Figure 2 Alternative programming approach for statistical analysis outputs. One program performs the statistical analysis and stores the result in an Analysis Results Dataset. Another program reads this dataset and displays the result in a table or a figure with the appropriate layout.

WHAT ARE THE BENEFITS OF INTRODUCING ANALYSIS RESULTS DATASETS?

Many benefits are gained by separating the statistical analysis and layout from each other and introducing analysis results datasets. Below we will outline some of these.

REUSABILITY OF STATISTICAL ANALYSIS RESULTS

In our suggested approach, the statistical analysis runs only once in a dedicated program. The results can subsequently be used for several tables, figures, and listings in the clinical trial report. This reduces the execution time and ensures consistency between the presented results.

Likewise, the analysis results can easily be extracted and re-used when comparing the results of several studies within a clinical programme, for instance for integrated summary of safety or integrated summary of efficacy documents.

Finally, the results of one statistical analysis might be re-used when performing other statistical analyses. This could for instance be relevant for sensitivity analyses such as tipping point analyses.

SEPARATION OF STATISTICAL ANALYSIS AND LAYOUT PROGRAMS

Programs performing statistical analyses must be programmed by a statistician and will typically be double programmed by a different statistician. These programs are often key deliverables in a regulatory submission and are referred to in analysis results metadata. In contrast, the programs that merely extract the results and present them in tables or figures with the appropriate layout can be done by a statistical programmer and are not important for regulatory authorities.

Separating these very different tasks into separate programs allows statisticians and programmers to work in parallel and utilize their core competencies. Also, it simplifies the programs that are submitted to regulatory authorities, since these do not include code relating to layout of outputs.

STORAGE OF STATISTICAL ANALYSIS RESULTS INTO STRUCTURED DATASETS

Having the results of the statistical analyses stored in well-defined and structed datasets is in itself beneficial. It makes it easier to browse the results within a trial or between trials and to refer to the results of old trials. Also, results or result qualifiers that are not presented in the clinical trial report can be stored in the analysis results datasets for future reference. This may for instance be relevant if a statistical analysis yields a point estimate with a standard error and a 90 % confidence interval, and only the point estimate and confidence interval are presented in the clinical trial report. The standard error can then still be stored in the analysis results datasets.

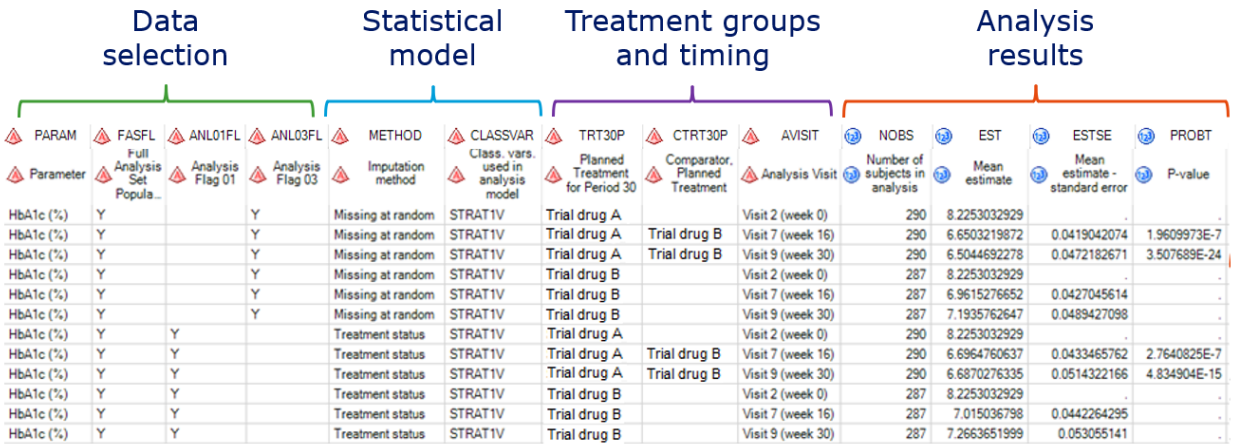
HOW SHOULD ANALYSIS RESULT DATASET BE STRUCTURED?

The structure of the analysis results datasets should facilitate that the results of several different analyses, including sensitivity analyses, can be stored in the same the same format.

The analysis result datasets are not ADaM datasets and cannot be made to follow the ADaM implementation guide, since they are not subject centric. That is, they do not contain the USUBJID variable. Nevertheless, our experience has shown that it is most convenient to follow a Basic Data Structure (BDS) like structure, where the composite key includes both PARAMCD and AVISITN. One notable difference is that the analysis result will not be a single value that can be stored in AVAL. Instead, the result can be a combination of a mean estimate, change from baseline estimate, estimated treatment difference, standard error, confidence interval, p-value, and even the exponentiated versions of all of these values. Since these values are both derived and used together, we suggest keeping these values in a “broad” dataset structure with different variables for each of these results instead of transposing the dataset into a “long” format.

We propose to structure the analysis results dataset so that the variables can be dived into 4 main groups (see Figure 3):

1. Data selection, e.g. parameter, population and analysis flags used to select data for the analysis
2. Statistical Model, e.g. imputation method, covariates, and other model parameters
3. Treatment groups and timing
4. Analysis results, e.g. estimates, standard errors and p-values



PARAM Parameter	FASFL Full Analysis Set Popula...	ANL01FL Analysis Flag 01	ANL03FL Analysis Flag 03	METHOD Imputation method	CLASSVAR Class. vars. used in analysis model	TRT30P Planned Treatment for Period 30	CTRT30P Comparator, Planned Treatment	AVISIT Analysis Visit	NOBS Number of subjects in analysis	EST Mean estimate	ESTSE Mean estimate - standard error	PROBT P-value
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug A		Visit 2 (week 0)	290	8.2253032929		.
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug A	Trial drug B	Visit 7 (week 16)	290	6.6503219872	0.0419042074	1.9609973E-7
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug A	Trial drug B	Visit 9 (week 30)	290	6.5044692278	0.0472182671	3.507689E-24
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug B		Visit 2 (week 0)	287	8.2253032929		.
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug B		Visit 7 (week 16)	287	6.9615276652	0.0427045614	.
HbA1c (%)	Y		Y	Missing at random	STRAT1V	Trial drug B		Visit 9 (week 30)	287	7.1935762647	0.0489427098	.
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug A		Visit 2 (week 0)	290	8.2253032929		.
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug A	Trial drug B	Visit 7 (week 16)	290	6.6964760637	0.0433465762	2.7640825E-7
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug A	Trial drug B	Visit 9 (week 30)	290	6.6870276335	0.0514322166	4.834904E-15
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug B		Visit 2 (week 0)	287	8.2253032929		.
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug B		Visit 7 (week 16)	287	7.015036798	0.0442264295	.
HbA1c (%)	Y	Y		Treatment status	STRAT1V	Trial drug B		Visit 9 (week 30)	287	7.2663651999	0.053055141	.

Figure 3 Proposed structure of analysis result datasets. Only a selection of variables is included.

HOW TO PROVIDE TRACEABILITY AND DOCUMENTATION?

The traceability between ADaM datasets and the analysis results presented in the clinical trial report is normally provided through analysis results metadata in define-xml. We suggest following this approach to documentation also when creating analysis results datasets.

Specifically, for a regulatory submission we suggest to provide the following (see also Figure 4):

- The programs performing the key efficacy and safety analyses
- Analysis results metadata linking to these programs in the "programming statements" section
- A description of the programs in the ADRG section 7: "Submitted programs"

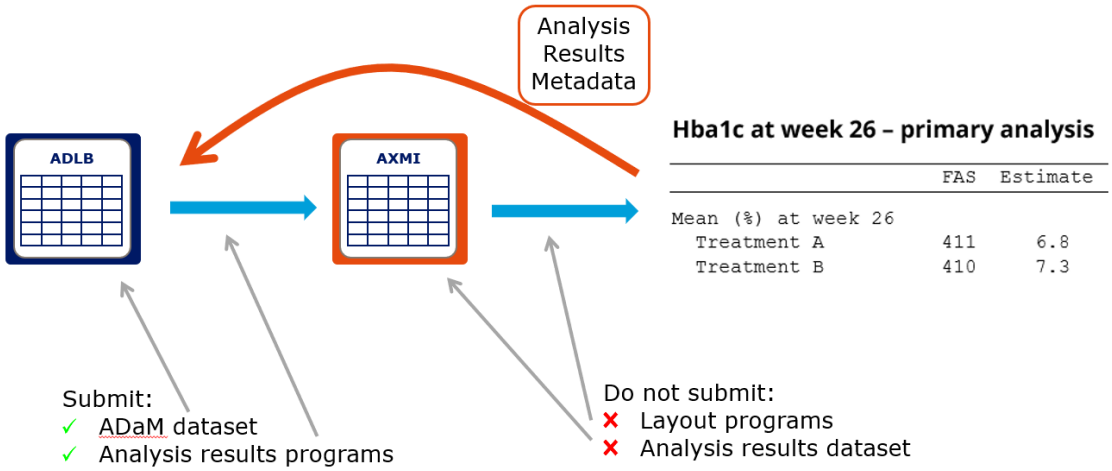


*Figure 4 Overview of elements that will be and that will **not** be submitted to authorities.*

Note that we do not suggest providing the analysis results datasets themselves in xpt format, nor to include these datasets in define-xml along with the ADaM datasets. The reason for this is that it is the programs performing the analyses and not the resulting datasets that are the main interest of the statistical reviewers. Reviewers should browse the results in the clinical trial report, where context such as titles and footnotes are included, and not in separate datasets. Also, the analysis results datasets are not easily described in the variable-by-variable structure of define-xml. A better description of the statistical analysis can be provided in the "submitted programs" section of the ADRG as well as in the statistical analysis plan. The analysis result metadata for statistical analysis can structured in a very simple format as illustrated in Figure 5.

table 14.2.3

Display	HbA1c at week 26 - primary analysis
Analysis Result	
Analysis Parameter(s)	PARAMCD = "HBA1C"
Analysis Variable(s)	ADLB.AVAL (Analysis Value)
Analysis Reason	SPECIFIED IN SAP
Analysis Purpose	PRIMARY OUTCOME MEASURE
Data References (incl. Selection Criteria)	ADLB [PARAMCD = "HBA1C" and ANL02FL = "Y"] .
Programming Statements	[SAS version 9.4] ax-mmrm2-mre-prim.txt

Figure 5 Example of how easy analysis result metadata in the define.xml can be made when using analysis results datasets. Note the reference to the submission ready program.

WHAT ARE OUR EXPERIENCES WITH AGENCIES AND THE ANALYSIS RESULTS DATASETS?

We have followed the above strategy of creating analysis results datasets in several submissions to both PMDA as well as FDA. Both agencies have accepted the datasets and their feedback has in general been positive. In the FDA statistical review and evaluation document for one submission, FDA wrote “Data and analysis programs provided by the applicant were consistently well organized”. We see this as a confirmation that the strategy of creating analysis results datasets is beneficial for clinical programmes for which data and programs will be submitted to authorities.

CONCLUSION

In conclusion we claim that analysis results datasets improve performance and consistency of statistical analysis programs. Furthermore, it enables programmers and statisticians to work in specialized units, as well as potentially reuse the analysis results for later additional analysis and outputs.

We would recommend storing the analysis results in a BDS-like structure, but with one row containing several numerical results such as estimates, standard errors as well as confident intervals. We are aiming to create analysis programs which will be short and specialized in order to make them submission ready. This means that one analysis program should not loop over several parameters, but only create one desired analysis.

With the proposed strategy the analysis results become fairly simple to document in the analysis results metadata and we have until now received positive feedback from both FDA as well as PMDA in data submission where we have applied this approach.

ACKNOWLEDGMENTS

The described work above is the consolidated task of several Novo Nordisk colleagues as well as consultants. A special credit should go to Lars Lehmann Andersson, Gitte Dam Jepsen, Camilla Mondrup Andreassen, Julie Kjærulff Furberg, Jan Skowronski, Mads-Jeppe Tarp-Johansen and Jacob Bonde Jacobsen.

CONTACT INFORMATION

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

Jeppe Rømer Juul
Novo Nordisk A/S
Novo Allé 1
2880 Bagsværd
Denmark
Work Phone: +4530750940
Email: jxju@novonordisk.com

Henning Pontoppidan Föh
Novo Nordisk A/S
Novo Allé 1
2880 Bagsværd
Denmark
Work Phone: +4530794981
Email: hpf@novonordisk.com