PhUSE EU Connect 2018

Paper SI10

Implementation of a Metadata-based Approach to Statistical Planning, Analysis & Reporting

Frank Freischläger, Estimondo GmbH, Bad Nauheim, Germany
Hanspeter Schnitzer, Estimondo GmbH, Bad Nauheim, Germany

ABSTRACT

Statistical programming of analysis datasets and outputs and statistical writing of analysis plans and reports are time-consuming activities in clinical research. Some source codes and standard texts are reusable, based on CDISC data standards and efficient programming, but individual project work is highly dependent on study indication, design, objectives, assessments, estimands, and the statistical methods applied.

In line with the CDISC Protocol Representation Model (PRM), we defined metadata datasets determining all aspects of a study relevant to biostatistics. The metadata feeds both the statistical programming environment and an expert writing framework, enabling partial automation of programming and writing activities while accelerating biostatistics processing.

Metadata is turned into macro parameters and, with the clinical data, into results datasets, formatted displays, and textual descriptions. We present the components of the approach and their interactions together with examples of auto-generated method description, data description and analysis processing, all based on the metadata.

INTRODUCTION

In clinical research, the workload in biostatistics and statistical programming is gradually increasing. The reasons for that range from more diversified portfolios and increased regulatory requirements to expanded concepts like the estimands framework, innovative clinical study designs, and advanced statistical approaches, all of which are accompanied by high-level qualification needs. At the same time, the capacities of trained associates are limited. To master these challenges, experts in the field should focus on the more complex parts in statistical planning, analysis and reporting, while, thanks to initiatives of standardization, the routine work should ideally be automated.

Processes of statistical planning, analysis, and reporting at a pharmaceutical company, the sponsor, or a contract research organization (CRO) are based on a clinical trial protocol (CTP), a raw clinical database, as well as policies, guidelines and further instructions by the sponsor, that may vary greatly across and even within sponsor companies. Also, although there are helpful initiatives like PRM or TransCelerate's Common Protocol Template, CTPs are still in need of standardization, particularly regarding those elements relevant to biostatistics [1]. Service providers like us come in to deliver statistical analysis plans (SAP), analysis datasets, and statistical outputs (tables, figures and listings; TFL), and to contribute to the reporting of clinical study results. On order fulfillment, based on templates and standards, activities of statistical writing and statistical programming are accompanied by supportive instruments and detailed documentation including mock TFLs, statistical programming specifications, analysis dataset specifications, and reviewer's guides.

On interventional studies, the raw and derived data is predominantly following CDISC standards, SDTM and ADaM, respectively, and the clinical study reports (CSR) follow ICH E3 standards [2]. However, in late-phase non-interventional studies, this may not be the case. Those may require a flexible approach, covering most of the contingencies. Throughout the years, the community has achieved efficiency in making use of standard programming libraries and standard computing environments, but even the more routine tasks still require significant hands-on resources in writing source code as well as text documents.

In this paper, we discuss the concept and implementation of a system in statistical planning, analysis and reporting which aims to separate reusable standards from variable components of the service processing. The system is based on defining a widespread set of metadata that, once collected, allows to semi-automate project delivery to reduce resource needs on routine work. Ultimately, we aim to shift manpower to tasks that cannot be automated, while, at the same time, assuring outcomes of high quality by replacing manual steps. The system is implemented with SAS, Microsoft Office and Visual Basic for Applications.

As a part of statistical outputs, listings and figures respectively document individual data and visualize individual or aggregated data. Generally, these are descriptive outputs. Likewise, most tables, describe aggregated data, whereas only a minor (but increasing) portion of the tables display the results of statistical modelling and inference. Based on a personal investigation of six recent CSRs of different clinical phases, indications, study designs and sponsors, we estimate the portion of descriptive outputs to be roughly 82% of all TFLs. Certainly, the CSR selection is not representative, but the little survey underpins that streamlining processes of data description can have a significant impact on productivity. With such a system in place, trained associates should be able to concentrate on

the 18% of outputs reporting results of statistical model-based inference. The latter are the drivers of decision-making in drug development.
In the following, we describe the metadata repository, its handling and application in statistical programming, reporting, and planning. It is a flexible, learning system. We conclude with a summary of the system and an outlook for extensions to the system and applications outside biostatistics operations.

CLINICAL STUDY METADATA

The key component of the system is a metadata repository that is interacted with via a control center. It directs the data processing and analysis as well as some writing activities in reporting and planning. Metadata objects include, but are not limited to,

- general objects like clients, stakeholders, projects, studies, semantics, tools, and references
- protocol-related objects like documents, objectives, selection criteria, and analyses,
- data-related objects like data sources, variables, groups, and special-interest specifications,
- design-related objects like epochs, visits, disposition, and exposure,
- statistics-related objects like methods, method options, descriptors, and time definitions,
- analysis-related objects like populations, intercurrent events, estimands, and evaluations,
- deliverable objects like derived datasets, outputs, reports, and plans
- programming-related objects like derivations, templates, styles, settings, as well as library, directory, and dataset specifications, and
- deliverable-related objects like options, invocations, controls, structures and sentences.

A project may comprise one or more studies (meta-analysis, integrated summaries, pooled analyses). On a study, one or more analyses (final, interim, central monitoring, ad-hoc, post-hoc) may be required. Semantics define wording that is to be used consistently throughout project lifetime. Tools include utilized software and applied standards and dictionaries including SDTM, ADaM, MedDRA, and WHO-DD. Documents cover input, process documentation and deliverables. Groups are treatment arms as well as subgroup definitions.
The metadata repository contains objects which are project-specific characteristics, but also general definitions of data derivation, description, and analysis including the formatting of outputs. It provides all ingredients to generate the trial domains that SDTM and PRM have in common and represent study metadata. To that regard, the system is compliant with PRM, which (beyond trial domains) does not lend much support to the biostatistics contribution to CTP production.

FLEXIBILITY IN APPLYING STANDARDS

The metadata repository is organized in such a way that selected metadata is mandatory while other metadata, if not defined per individual deliverable, is automatically taken from a higher level of definitions. As visualized in **Figure 1**, intermediate layers may include specifications on a study level, sponsor standards or company standards as a default approach to prepare for metadata-based processing. Further layers may be added at a compound level or for all tables, for example. Not all metadata is given on all levels. If metadata makes sense on more than one level, it is only required to fill one of those levels with metadata. If specifications are given in more than one level, starting from individual, the lowest level is considered. Levels above the project level are reused across projects of, e.g., the same sponsor.

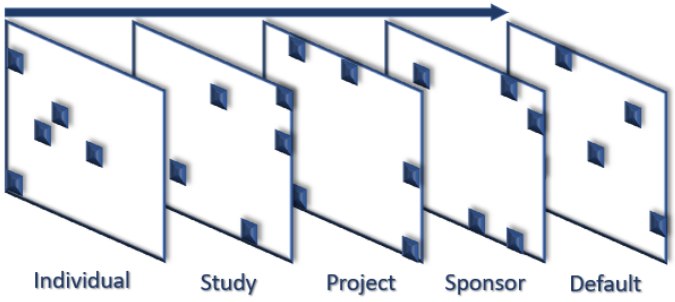

FIGURE 1: Layer approach of the metadata repository

If there is a standard metadata setting on a sponsor level and the sponsor wishes a table to deviate from that, the exception-from-the-rule details would be given on the individual level for that particular table, while other tables would not be affected and follow the sponsor-level standards.

IMPORT, ENTRY & QUALITY CONTROL

Metadata is either imported from other sources or entered based on the information at hand. E.g., SDTM trial domains could be imported and transformed, if provided with raw data. Some metadata items are selected from a list of valid values. Edit checks ensure that mandatory items are not missing, and that data is consistent across the

metadata repository. Metadata entry requires subject-matter experience and diligence. It is recommended to arrange for independent duplicate metadata entry as an additional quality step.

STATISTICAL PROGRAMMING

The metadata relevant to statistical analysis programming enables the production of TFLs in a streamlined way, based on derived datasets (preferably ADaM datasets). Templates and styles are generated based on the metadata identified on the relevant levels.

Given the metadata, including the list of deliverables and associated settings as nominated per deliverable, a library of macros ensures the production of TFLs for data description without the necessity of further efforts beyond the metadata specifications and the initiation via the control center. The deliverables are combined into booklets with a table of content and bookmarks for easy navigation. The system is open to additional programming, e.g. for more complex model-based statistical analysis.

LAYOUTS AND FORMATTING

For documentation, original statistical outputs of the software are documented, while the TFLs are following layouts and formatting as previously approved by the client via mock TFLs. Based on test study data or simulated data, the system is able to produce mock TFLs at an early stage and on short notice via an add-on routine. All the relevant specifications are in the metadata including styles, structures, titles, footnotes, margins, columns, and other settings per output. The layouts are automatically transformed into programming instructions via control functions as part of the metadata. Among other areas of application, this procedure works well for the standard layouts as specified by the PhUSE CSS Analyses and Code Sharing Working Group [3].

As illustrated in **Figure 2** below for a mixed demographics table, some examples of elements of standard formatting are

- the formatting of counts and percentages, e.g. "`xx (xx.x)`",
- the syntax of label and unit as either "Label [unit]", "Label (unit)", or "Label, unit",
- the syntax of label and frequency explanation as either "Label, n (%)", "Label n (%)", or "Label - n (%)",
- the rules for displaying and considering categories including missing values,
- the display of zero counts as either "`0`", "-","." or "`0 ( 0.0)`",
- the display of "100%" with or without decimal places.

Demographic Parameter		PL (N=XXX)	T1 (N=XXX)
Sex n (%)	n*	xx	xx
	Female	xx (x.x)	xx (x.x)
	Male	xx (x.x)	xx (x.x)
	Missing	xx	xx
Age (years)	n*	xx	xx
	Mean	xx.x	xx.x
	SD	xx.x	xx.x
	Median	xx.x	xx.x
	Q1, Q3	xx, xx	xx, xx
	Min, Max	xx, xx	xx, xx
	Missing	xx	xx
Age Categories n (%)	n*	xx	xx
	<65	xx (x.x)	xx (x.x)
	≥65 and <75	xx (x.x)	xx (x.x)
	≥75 and <85	xx (x.x)	xx (x.x)
	≥85	xx (x.x)	xx (x.x)
	Missing	xx	xx

FIGURE 2: Exemplary mixed Demographics table layout with formatting details highlighted

RESULTS & OUTPUT DATASETS

The determined sample characteristics are stored in results datasets. Results datasets contain numerical results. These are converted to character variables and formatted according to the settings as discussed above. The resulting texts are then stored in output datasets and used for further processing in the creation of tables.

MACRO-FREE SOURCE CODES

The system is based on macros, but it produces macro-free source codes that are executable, documented, and provided to the client along with the respective deliverables. Per run, particularly on production runs, outputs are produced twice: with macro-based programs and with macro-free programs. Results are compared to assure reliable outputs from macro-free programs.

CHANGE CONTROL AND VALIDATION

With each production run and delivery, a snapshot of the project metadata is archived. The system is exclusively accessed by the development programmer. Execution is logged, and all log files scanned. Activities of validation programming are independent of the system. However, observing the quality steps described above, the metadata is maintained and used by the entire team. Results of validation programming are compared with system-based results of development programming by means of results datasets, output datasets, and outputs.

STATISTICAL REPORTING

The table layouts define how numerical results will appear in a table. For example, point estimates and confidence limits are prepared for display as "`xx.x (xx.x, xx.x)`". When this is used for tables, it can also be used for reports describing results in texts. While it is recommended to prepare in-text tables with aggregated data and results of statistical analyses by the system in the same way as described before for post-text tables, textual description of results should be limited and automated.

To create sentences with results and documents composed of those sentences, methods of data-to-text conversion technology, also known as natural language generation (NLG) are applied [4]. There is a variety of applications to healthcare data [5]. Examples include the auto-generation of patient safety narratives [6], and the reporting of regression analyses [7]. Our system applies NLG to the composition of document templates, the definition of document structures and content, and the production of respective reports that serve as a basis for further expert writing.

Key metadata tables to support statistical reporting are those of structures and sentences. Structures define the order and sequence of sections including paragraphs, in-text tables and in-text figures. Exemplary CSR structures are found in ICH E3 standards [2]. Paragraphs comprise sentences. Sentences define how results data is to be put into words Sentences are referenced by an identifier, preselected from a great variety of standard sentences and linkable with the data and metadata via codes. These codes serve as placeholders and are associated with the respective text snippet representing results or metadata. These snippets are prepared and coded by the system in a vertical structure so that each code is unique. The following examples give an impression of various options of reporting based on automatically created and organized sentences. For ease of display, some codes are simplified.

EXAMPLE: TOTAL NUMBER OF PATIENTS SCREENED

The results section of a CSR typically starts with a sentence like this one below (resulting sentence). Placeholders are highlighted, and respective codes are displayed (linked sentence), The number of screened patients, and the number and percentage of screen failures are taken from respective results datasets on study disposition. The semantic of patients, the post-text table identifier and the way post-text tables are referred to are taken from the metadata.

Resulting sentence:	A total of 500 patients were screened, of whom 50 (10.0%) failed screening prior to randomization (see Table 14.1.1.1).
Linked sentence:	A total of [DISP_SCR_TOT_N] [SUBJ_TXT_PL] were screened, of whom [DISP_FAIL_TOT_N] ([DISP_FAIL_TOT_PCT_SCR_TOT]%) failed screening prior to randomization ([REF_TXT1] Table [REF_PT_DISP]).

EXAMPLE: SHORT DESCRIPTION OF BASELINE CHARACTERISTICS

A short statement sentence like the following would require research by the author. Likewise, the system not only creates the sentence from data and metadata like in the previous example. It identifies the most frequent categories of gender and race, and labels and displays those along with their frequencies.

Resulting sentence:	Most patients were female (62.1%) and white (44.3%).
Linked sentence:	Most [SUBJ_TXT] were [DEMO_GENDER_MOSTFRQ_LBL] ([DEMO_GENDER_MOSTFRQ_PCT]%) and [DEMO_RACE_MOSTFRQ_LBL] ([DEMO_RACE_MOSTFRQ_ PCT]%).

EXAMPLE: WITHIN-GROUP PRE-POST CHANGES IN A CONTINUOUS VARIABLE

Per treatment arm, the mean baseline and endpoint values of systolic blood pressure are described, and the direction of change is interpreted. The mean change could also be displayed. Without the need for manual writing, the system is able to identify the appropriate word to describe the change by comparing mean baseline and endpoint values.

Resulting sentence:	In Group1 subjects, mean sBP increased from 108.3 mmHg (SD 43.1) at baseline to 111.0 mmHg (SD 64.2) at Week 24.
Linked sentence:	In [ARM_1_LBL] [SUBJ_TXT], mean [SBP_LBL] [SBP_CHG_BL_24__MEAN_1_INTERPR] from [SBP_BL_MEAN_1] [SBP_UNIT] (SD [SBP_BL_SD_1]) at baseline to [SBP_24_MEAN_1] [SBP_UNIT] (SD [SBP_24_SD_1]) at [VIS_24_LBL].

These examples illustrate the dynamics of auto-generated sentences. Once selected and provided with data, the sentences are automatically created by the system (via the control center), arranged in the template according to the

structure, and combined with in-text output. This is achieved by a function replacing codes in the linked sentences with associated text snippets.
The resulting documents are editable, updateable with new data, and reproducible if invisible codes are not contaminated, which is controlled for by a routine. On document revision, if review comments are to be incorporated, the author should balance the time and effort of editing in the document versus editing in the metadata.

STATISTICAL PLANNING

The same principles as used for statistical reporting and described above are applicable to statistical planning as well. An SAP and associated planning documents can be created. Although, when selecting and producing the sentences for an SAP, there is no dependence on the clinical data and derived results but on metadata alone, the approach reaches a more challenging level of complexity due to the various aspects to consider when planning the statistical analysis of a clinical study. Therefore, the degree of automation possible is lowest in this stage, but automation can still save a significant amount of valuable expert's working time and, at the same time, ensure a standardized, consistently metadata-based approach.
While descriptions of more complex statistical methods and other specific details require manual writing, as a first step, the system creates the document based on SAP-related metadata such as protocol-, design- and data-related objects as well as populations, estimands, endpoints, evaluations and methods.
The structure of an SAP, together with instructions about its content, is typically established in client or sponsor templates. Mandatory fixed texts are part of those templates. The templates alternatively can be created by the system in a pre-defined structure. Sentences for the SAP are embedded and may include the following metadata-based examples with highlighted placeholders and respective replacements.

EXAMPLE: DEFINITION OF TREATMENT-EMERGENCY

As the following sentence defines treatment emergency, the lag time, time unit, and semantics of intake and study treatment are taken from the metadata.

Resulting sentence:	An adverse event is treatment-emergent, if onset or deterioration of the event appear after the first study treatment intake and no later than 45 days after the last study treatment intake.
Linked sentence:	An adverse event is treatment-emergent, if onset or deterioration of the event appears after the first [TRT_LBL] [TRTAPPL_LBL] and no later than [TE_LAGDEF] [TE_LAGUNIT] after the last [TRT_LBL] [TRTAPPL_LBL].

EXAMPLE: DEFINITION OF AN ESTIMAND

When an estimand is to be defined, the patient population, the endpoint, which is a derived variable, a specification of consideration of intercurrent events, and the population-level summary measure are populated with respective metadata information [8].

Resulting sentence:	Difference in means between treatment conditions in the change from baseline to Week 24 in sBP in the targeted population regardless of intercurrent events.
Linked sentence:	[EST_1_SUMY_LBL] in the [EST_1_ARR_LBL] in [EST1_PARAM_LBL] in the [EST_1_TAPOP_LBL] [EST_1_INTEVSPEC].

The system is not designed to write an entire SAP automatically, but it can significantly assist with the routine tasks of writing definitions, method descriptions, and specifications. This also holds true for further planning documents such as analysis dataset specifications and statistical programming specifications. Among other content, the latter contains details of the level of validation per program associated to each output, which is defined in the metadata as well.
Once the planning documents have been generated by the system based on the metadata, further content may be added by the authoring statistician. Later, the same content is reusable as part of reporting requirements by linguistically transforming the description of planned methods into past tense and feeding the methods sections of CSRs accordingly.

CONCLUSION

We are implementing a system that supports statistical programming of TFLs with a conceptual computing environment and assists statistical writing with a flexible neural approach. We successfully implemented it to reduce the overall workload while increasing productivity and reliability of the deliverables. It changes the landscape of biostatistics by placing a metadata repository in the center of our process (see **Figure 3**). Time savings outweigh the efforts required in feeding and maintaining the repository, especially when standards can be consistently applied across projects.

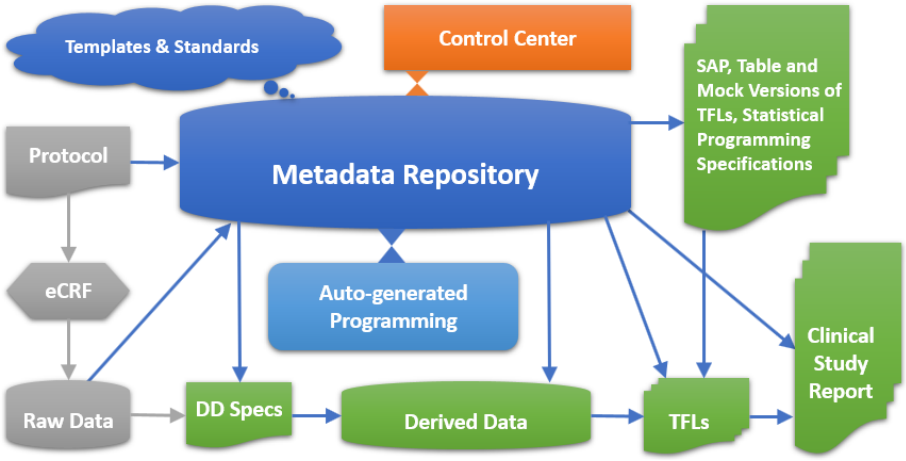

FIGURE 3: The metadata-based approach to biostatistical planning, analysis and reporting

Metadata collection can be optimized at the frontend and backend of the database to further improve the cost-value ratio. By adding standard templates, styles, layouts, structures, and sentences, the metadata repository is growing and of increasing benefit. Further developments are useful to extend the coverage of standard texts, also towards some inferential statistical methods, but only up to a meaningful degree. The system was not designed to and will not replace statistical or medical writing activities completely, but it allows experts to concentrate on the description and interpretation of more complex data, while the system provides the basics and beyond.

It is conceivable to apply the system in other areas such as the production of quality documents (standard operating procedures, forms, templates) or service proposals. Some of the metadata are useful for service proposals and could be reused once a project would be awarded.

The most challenging aspects lie in the transformation of metadata and results data into a vertical data structure to be linked with sentences during the data-to-text conversions. Here, technologies such as Microsoft Office OpenXML or Linked Data (RDF/XML, SPARQL) may have the potential to increase the power and usability of the system. In summary, it is a flexible, fast, reliable, efficient, and evolving system that is not intended to be comprehensive. It requires experience and accuracy and yet has limitations with complex study designs and advanced statistical methods.

REFERENCES

[1] CDISC (2010): Protocol Representation Model v1.0.

[2] ICH (1995): Structure and content of clinical study reports (Harmonized Tripartite Guideline E3, Step 4).

[3] PhUSE Computational Science Standard Analyses and Code Sharing Working Group (2018): Analyses and Displays Associated with Demographics, Disposition, and Medications in Phase 2-4 Clinical Trials and Integrated Summary Documents, White paper, v2.0.

[4] Reiter E, Dale R, Feng Z (2000): *Building natural language generation systems*. Cambridge University Press, Cambridge, UK.

[5] Pauws S, Gatt A, Krahmer E, Reiter E (2018): Making effective use of healthcare data using data-to-text technology. In Consoli S, Recupero DR, Petkovic M (Eds.), *Data Science for Healthcare: Methodologies and Applications*, Springer, New York, in press.

[6] Ledade SD, Jain SN, Darji AA, Gupta VH (2017): Narrative writing: Effective ways and best practices. Perspect Clin Res 8, 58-62.

[7] Lloyd JR, Duvenaud D, Grosse R, Tenenbaum JB, Ghahramani Z (2014): Automatic construction and natural-language description of nonparametric regression models. In *Association for the Advancement of Artificial Intelligence*, July 2014.

[8] ICH (2017): Estimands and sensitivity analysis in clinical trials (Harmonized Guideline E9(R1), Step 2).

CONTACT INFORMATION

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

Frank Freischläger
Estimondo GmbH
Hintergasse 22
D-61231 Bad Nauheim
Work Phone: +49 151 6772 0571
Email: frank.freischlaeger@estimondo.com
Web: www.estimondo.com

Hanspeter Schnitzer
Estimondo GmbH
Hintergasse 22
D-61231 Bad Nauheim
Work Phone: +49 177 745 9759
Email: hanspeter.schnitzer@estimondo.com
Web: www.estimondo.com