PharmaSUG 2024 - Paper LS-304

Translation from statistical to programming: effective communication between programmers and statisticians

Diana Avetisian, IQVIA

ABSTRACT

Statisticians and programmers who are working in clinical trials know that effective communication between the groups is a key factor for a study to succeed. But do we really know what statistician is expecting from programmers when they are working on complex statistical models and what programmers need to be efficient in the process of efficacy outputs creation?

Programmers have a complex task when it is come to statistical analysis implementation, they are creating ADaMs based on CDISC standards which is allowing them to create outputs and on the next step producing TLFs to display results of statistical analysis. The truth is that it is not always obvious how data should look like to be suitable for a particular statistical analysis that's why programmers and statisticians are working together to help each other to achieve the best results.

At the same time, statisticians have their own challenges. What kind of information is needed for programmers to create TLFs? Is it enough to have a programming note in the shells and some code examples or do programmers need more?

To answer all these questions, it will be useful for each group to have some guidance to understand each other needs that's why in this paper the author discusses:

- the process of efficacy output creation from the raw data to final table;
- some common questions from statistician and programmers;
- some tips and tricks for programmers how to understand the requirements of statistical analysis even without statistical background;
- programmer's expectations from statisticians.

INTRODUCTION

This paper is dedicated to collaboration between programmers and statisticians which is critically important for every clinical trial to succeed and be objective. On one hand, it is well known that communication in the team is a key factor not only in clinical trials but in any area. On the other hand, it is not always easy to find a common ground for effective communication especially when it comes to conversations within people with a different background and roles. Conducting the study is a complicated process where multiple groups communicate with each other, trying to explain their challenges/concerns, solve some problems together but biostatistics group is combining people with two different roles: programmers and statisticians. They supposed to work together but at the same time each of them has their oven area of responsibility and can have a different background especially when it comes to a complex analysis that's why it is important to help and support each other.

In order to create an effective communication between programmers and statisticians we have to understand each other better, keep in mind our strong sides and areas of expertise, understand the requirements to the process and results.

To provide some useful recommendations to both groups at first we will consider a process overview of efficacy output creation. In this section we will identify steps which every team member is performing and specify areas of responsibilities. At the same time, we will analyze on which phase we have to support each other the most, identify our similarities and differences in the process.

Based on the process overview in the next section we will create a list of questions form programmers and statisticians to themselves and to each other to simplify the communication process. Such questions can be applied to almost every study, but it is also important to identify the right time for every question. If

concerns are transparently discussed in timely manner the team can reduce the stress level, number of quality issues and achieve the best quality.

Even with the list of questions related to analysis and study phases it is still can be difficult to trust each other, don't hesitate to raise concerns and to know when the best time for discussions is. Based on experience it is possible to come up with some tips and tricks to make people a little bit more comfortable in the process of building an effective communication.

Finally, it is important to know strong sides and responsibility areas of each role and based on it to understand what kind of support we can expect from each other. We have to know that there are no expectations that just one person will know each and every aspect of the study in all details. By aligning assumptions and performing effective communication we can efficiently split the assignments, responsibilities, reduce a personal pressure, feel supported and build a good and trustful environment which is the big goal for every team. It is also allowing us to be more productive during the whole process, save a lot of time and budget for every study, avoid rework and additional costs after study finalization.

THE PROCESS OF EFFICACY OUTPUT CREATION FROM THE RAW DATA TO FINAL TABLE

To understand statisticians and programmers expectations from each other we have to know how the process of statistical output creation looks like for each side. Obviously, we all know that in the end of the process we will have a table or figure and listing which is supposing our analysis but do we really know how different the process looks like for programmer and statistician? To discuss our needs let's talk about the usual study where we are following CDISC standards and try to describe the output creation step by step. Here we are considering study with CDISC standards as it contain all possible steps: aCRF(annotated case report form) creation, SDTMs(Study Data Tabulation Model), ADaMs(Analysis Data Model), TLFs(tables, listings, figures) programming. For studies which are not following CDISC standards some steps can be skipped (for example for some studies ADaMs or TLFs can be programmed based on raw data, etc.) but the general logic still will be applicable.

On the very first step the whole biostatistics team is starting with documentation: protocol, SAP creation, any other study related documents which may help us understand the purpose of the study and how it will be conducted. At this step statisticians already start their work by creating the SAP or reviewing it, depend on the roles on study. At the same time programmers' input to the SAP might be needed but is not a requirement on this stage.

For programmers the work starts when database is created, and their first task is CRF annotation and review of data specification for vendor data. At this step every programmer is familiarizing themselves with the data, trying to understand it and see what data manipulation will be required and if it is possible to perform them according to CDISC standards and GPP (Good Programming Practice). On the next step programmers are creating SDTMs. It worth to mention that help from statisticians for such tasks can be useful especially to confirm if all the data for future analysis is available and collected properly but at the same time, we have to say that the biggest part of communication are happening between programmers and data management. At the same time when SAP is approved and at least test data is available statisticians are working on shells creation, they are putting together the list of outputs which should be produced for the study and it is a key factor for our next step which is ADaMs creation.

When we are talking about SDTMs it is good to mention that on this level data is standardized but there are no derivations or complex algorithms. On ADaM level every programmer will ask a question "What variables, rows should be derived in addition to collected data to create outputs?". It can look like a very simple question, and everybody can say that all answers can be obtained from SAP and shells which is true but actually it is a point where statisticians and programmers should have an effective communication to understand the requirements for analysis. For safety outputs data manipulations are pretty standard for every study so programmers can create a lot of derivations by themselves, but efficacy analysis are very specific and different, so it requires more investigations from both sides.

From programmers perspective it is very important to understand how data should looks like to be sufficient for planned statistical analysis. At the same time for statisticians, it is important to make sure that collected data will satisfy all assumptions for planned statistical procedure to work correctly. At this

point statisticians and programmers have to align their expectations, statistician supposed to be able to explain the requirement for input data and based on it programmers supposed to create an ADaM dataset according to CDISC standards, GPP and structurally sufficient for statistical procedures.

Finally, when ADaMs are ready programmers have to produce TLFs, usually the main outputs are tables which can be illustrated by figures and all of them are supported by corresponding listings. The easiest outputs to create are listings since it is just the representation of the data, which is used for statistical analysis, the only thing for programmers to check is if any intermediate steps should be represented in the listings. Each ADaM datasets should be created based on level of details which is required for listings and to not miss any intermediate calculations even if they are not used directly in table, it can be good for any reviewer and traceability. Here is steps that should be performed for listings creation:

- Subset the data from ADaM based on shells;
- Format the data to fit the requirement;
- Report subsetted and formatted data in rtf/pdf file.

The process of tables and figures creation is more complicated since it is requiring more steps. Sometimes figures can represent the results of statistical analysis and sometimes it is just a graphical representation of the data. The process of figures creation is similar to tables, the main difference is report part. Since report is mainly technical thing let's talk more about tables creation process (it can be easily adjusted for figures if it will be needed):

- Subset the data from ADaM based on shells;
- Format the data for statistical procedure;
- Make sure that data correspond to requirement of specific procedures, it can be either technical requirement if some variables supposed to have a specific type or format or statistical requirement like distribution of selected data, balanced treatment groups, etc.;
- Use required statistical procedure like proc mixed, proc reg, proc mi, proc mcmc, proc glm, proc genmod, proc lifetest, etc. for selected data based on type of analysis which supposed to be performed;
- Select the results from the output of statistical procedure;
- Format selected results for report;
- Report formatted statistical results in rtf/pdf file.

Based on specified we can conclude that some of specified steps can be tricky for programmers without statisticians support especially if programmer doesn't have statistical background. That's why it is very important for programmers to proactively communicate with statisticians and raise questions if something is not clear and for statisticians to provide a required information to programmers in advance with some basic explanations.

Based on the steps specified above we can summarize the general process of efficacy output creation for programmers in Figure 1

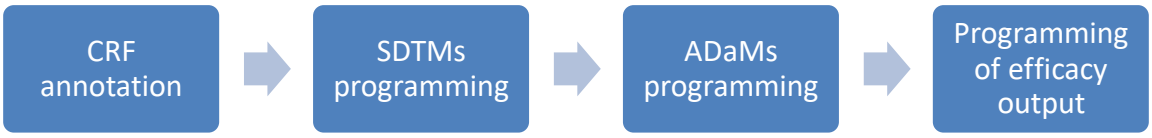

Figure 1. Creation of efficacy output from programming perspective

and for statisticians in Figure 2

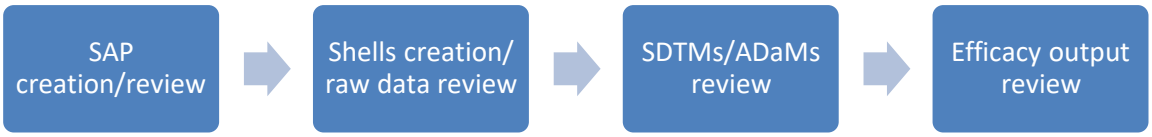

Figure 2. Creation of efficacy output from statistical perspective

Obviously, steps in Figure 1 and Figure 2 are general and don't describe all the items of the process but it is a good starting point to see what kind of details can be required and moreover when such details should be discussed within the team to be efficient and avoid rework in future.

SOME COMMON QUESTIONS FROM STATISTICIANS AND PROGRAMMERS

The key to effective and high-quality work is obviously the hard skills of each team member but it is not enough for the team to succeed. Within each task every person is raising some questions, and the good team player is usually trying to answer those questions by themselves first and find some hints in study documentation and other sources of the data. Unfortunately, not all the questions are easy to answer and sometimes we may need some help from professionals in a specific field. For programmers the first person to ask is usually their team lead but if question is related to SAP, shells or some details of statistical analysis then the first point of contact is statistician. Simultaneously statisticians are working a lot with study data and sometimes may need help with some data checks, investigations, technical issues or CDISC standards and for such cases the best person to ask is programmer who can provide support and advice how to achieve the desirable results. Let's talk about the most common questions for programmers and statisticians and try to determine what can be done to make the task easier for both sides on each step of the process which were described in previous section.

QUESTIONS FROM PROGRAMMERS

Programmers are starting the study by familiarizing themselves with documentations: protocol, CRF, SAP, shells and any other related documents. On this stage there are some common questions which every programmer has to ask:

- Does SAP include all analysis specified in the protocol? Are these documents consistent?
- Do shells include all planned analysis specified in SAP?
- Do we have all required data to perform analysis specified in SAP and shells?
- Is it possible to link the data between different CRF forms and vendor data to perform required analysis? (Such question is applicable to the raw data and SDTMs)

All such questions can be classified as consistency checks which ideally should be answered before SDTMs creation or the latest prior to ADaMs creation. Usually all of them can be answered by programmer with some help from data management but it is not involving statisticians. The reason why such questions are so important is hidden in the answers. In case the responses are not ideal and data or documents don't satisfy our needs to them we have to talk with statistical team and raise more questions:

- If certain analysis can't be performed, do we need to collect data in different way? Should database be changed to allow programming team to perform required analysis, or do we need to change SAP, shells and consider different statistical approach?
- Do we need to include more details to SAP and shells to be able to perform all consistency checks and statistical analysis?
- Do we need to know more about the data and have a better understanding of the statistical analysis?
- What assumptions for data should be checked for a specific statistical procedure?

It is important to have answers to such questions before ADaM specifications and datasets are created as it is preventing rework, can help with database design and will give the team some time to think about possible challenges/issues without extra timeline pressure. Obviously such questions will be answered

eventually during the study but if programmer will ask them in advance then the solutions can involve data management, statistical team, clinical team and other team members which is allowing study to achieve better results rather than identify problems after DBL(database lock) or just prior to it when the solutions are much more limited since the biggest part of the data are already collected. The proactive communication from programmers in the beginning of the study can prevent rework of study documentations and codes, at the same time it will increase the quality of analysis and increase team efficiency.

Based on Figure 1 such concerns should be raised on first 2 steps during CRF annotation, SDTM programming and partially during ADaM specifications creation. When we are moving to the next step of ADaM programming more questions have to be raised based on SAP and shells:

- How should data look like to be suitable for TLFs creation? What data structure is applicable to your study?
- Do you need to rederive some variables collected in raw data and presented in SDTMs to have a more accurate data description? For example, it can be analysis timepoints based on actual time difference instead of collected timepoint.
- How many records per key analysis variables are expected? For example, it can be one record per subject per analysis per parameter. How to select a required number of records for analysis? Do you need to derive analysis flags to select the correct subset of the data?
- How to analyze “Not Done” records? Are they part of analysis or should you exclude them completely?
- If a few sources of the same information are available, then which one should be used as the primary source? How data inconsistency will be handled if data can’t be reconciled on DM level?
- Does the imputation of missing data applicable to your study? Can it be applicable under certain conditions or as additional sensitivity analysis?
- If missing data supposed to be imputed, then where imputation should be done? Is it ADaM level or table level?
- Do you need to check if data from ADaM with required filters are satisfying assumptions applicable to a specific statistical tests or procedures? Are such assumptions specified in SAP?
- Do you have enough data for analysis? Such question can be very important on early stages of the study when not all the data are collected but it is expected to perform a Dry Run or on late phases of the study in case treatment groups are unbalanced and enrollment didn’t meet the initial expectations.
- Do we have a back-up plan/analysis in case the main analysis will not work (e.g. model will not converge, not enough subject in a specific treatment group, not enough responders, etc.)? Do the back-up analysis will require another ADaM structure or more derivations on ADaM level?
- Does complex algorithm cover all possible cases presented in the data? Should complex algorithms be modified to cover some common deviations in the data?

Such list can be extended even more but those questions are the one which is applicable to almost every study and needs to be answered during ADaM programming. All such questions can be answered by statistical team but all of them should be raised proactively by programmers to avoid problems and additional challenges during TLFs programming.

On the last step of Figure 1 programmers supposed to create outputs. As it was specified above there are couple of steps for efficacy table creation, some of them are obvious and technically simple but another one may require more discussions such as:

- Do you know all the filters which supposed to be applied for each output? Do you have a required variables in ADaM dataset to subset the data?
- Is there a requirement for type of variables in statistical procedure? Should it be numeric, or it doesn’t matter? How will a numeric order (for example numeric version of treatment variables) affect

analysis? Sometimes it can be important how numeric version of variables is derived especially if during statistical analysis the comparison between groups supposed to be performed.

- Do you know which statistical procedure should be used for a specific output? Some type of analysis can be performed using different procedure (e.g. some regression models, simple linear models), every procedure can give slightly different results especially in decimals so it is important to have an agreement between production, qc side and statistical team.
- How to select results from output datasets of statistical procedure?
- If statistical procedure is giving warning or error how to determine is it a problem with statistical concept, data structure or technical issue with the code?

Such questions can help create an efficacy output quickly and do self-qc of such output even without special statistical knowledge. Moreover, if such questions is raised during programming phase instead of review phase it is allowing the team to identify problems much earlier which is giving the team more time to fix any possible issues and achieve better quality.

QUESTIONS FROM STATISTICIANS

Since programmers are performing a lot of data checks and creating codes as soon as they are receiving raw data it may look like all questions supposed to be raised by programmers only but in fact it is not true. Statisticians need to know a lot of different facts about data, and it is meaning, the purpose of the study to be able to create SAP, shells and check the results of analysis. Statisticians are working with the whole study team, not only programmers. Some of statistical questions require support from medical writers, clinical team and other groups but a lot of questions can be answered by programmers especially during the review process. In Figure 2 we described steps of efficacy output creation for statisticians so let's consider which questions can be raised to programmers on each of those steps.

When statistician has to create or complete an initial review of SAP data is usually not available since it is an early stage of the study. As it was specified before programmer's review of SAP without data can have some benefits but it is not a strict requirement so there are not too many questions to programmers during this study phase (first step of Figure 2) but still some of them can be useful:

- Do we need to include code examples to SAP for a complex statistical procedure? Can such code be developed by programmers independently or help from statistician is required?
- Does SAP contain an information about preferable procedures for a specific analysis? It can be required to avoid mismatches in decimals due to applying different procedures on production and qc sides.
- Does SAP language is clear to people without statistical background? Is it understandable for programmers who don't know all the details about statistics?
- Are data assumptions specified clearly for each type of analysis?
- Is there a back-up plan which can be programed in advance in case some criteria is not met for a main statistical analysis?

Such kind of questions is easy to answer/implement on early stages of the study when SAP is not signed and approved but can save a lot of time for team starting from SDTMs codes creation till final delivery.

On the next step statisticians are creating shells and checking raw data to see how data is collected for the study, step 2 of Figure 2. During this phase statisticians might need some help from programmers with data checks, algorithms to combine some raw datasets and CDISC standards for SDTMs/ADaMs creation to have a better understanding of future data structure.

- What structure of the data is expected based on CDISC standards?
- What ADaM datasets will be created for the study to cover analysis?
- How raw data will be transformed in SDTMs and ADaMs level comparing to raw data?

- Was all raw data transferred to SDTMs and ADaMs?
- Is efficacy endpoint fully covered on ADaM level? For example if analysis supposed to compare responders and non-responders, usually criteria and derivation for responders are clear to programmers but how non-responders are covered on ADaM level?

The last but not least step for statisticians is the review process of ADaMs and TLFs, steps 3 and 4 of Figure 2. Sometimes during the review it is not fully clear for the reviewer how results was obtained by programmers especially when spot checks are performed based on the raw data that's why some questions can be raised:

- Is it possible to track all derivations from the raw data to final output?
- How to determine what is wrong when results are not matching between statistician and programmers?
- How to explain what results can be expected after analysis and what can be considered as a major programming issue?
- Is it possible that the wrong source of the data is used?

Sometimes it can be difficult to answer such questions but with proper communication between statisticians and programmers their knowledge of statistics, CDISC standards, codes creation can be combined to achieve better results and conduct statistical analysis correctly.

TIPS AND TRICKS TO UNDERSTAND REQUIRMENTS FOR STATISTICAL ANALYSIS

In previous section we considered a lot of common questions from programmers and statisticians, the main goal of specifying those concerns is to get answers and determine when such problems should be raised. When it is too late to raise questions? Is it too late if nothing can be done already since database was locked or even study was closed and submitted to FDA/EMEA/PMDA or is it too late when it is not enough time prior to delivery to fix the issues? Probably the answer is yes to both scenarios, the first one is a bit more difficult to fix than another since it is easier to extend the timelines rather than unlock the database or perform an adhoc analysis. Ideally, we would like to avoid such situations and the goal is to build an effective communication between programmers and statisticians, learn how to identify and raise important questions on time which will allow us to reduce the stress level for the whole team, perform planned analysis with better quality and spend less time on it by being more efficient and productive.

Sometimes even when all the questions are known it is still not guaranteed that concerns will be raised on time and provided answers will be helpful, miscommunication can happen even when it looks like that everything is clear. The problem is that statistical and programming approaches to the output can be different, for programmers the focus in process is codes and for statistician it is results and their meaning. Both sides have some common assumptions e.g. codes with warning/error is only programming issue and not a statistical problem, specified in SAP analysis supposed to work without checks of main data assumptions based on real data, expected data structure is clear based on specified type of analysis, etc. It is not totally wrong to have such assumptions until every team member will keep in mind that it is not always true especially because for the biggest part of cases such assumptions will work but it is always better to double check such things as early as possible. In order to be effective, we actually have to communicate all our assumptions clearly to each other and if it is needed "translate" them. For example for a person with statistical background it is obvious that t-test can be applicable only for data with normal distribution, for linear mixed effect models the errors supposed to be normal distributed and explanatory variables are linearly related to the response, to perform compare analysis between responders and non-responders by treatment it is a requirement to specify not only criteria for responders but also define non-responders sample but such things are not that obvious for programmers without statistical background. That's why all such things should be clearly specified and moreover "translated" to programmers to clearly specify expectations and requirements to the codes and data structure. It is good to keep in mind that it is working in both ways, statisticians don't always know all CDISC standards and algorithms of derivations from technical perspective that's why it can be useful to "translate" the logic from programming language,

make it clear for an external reviewer without code knowledges/GPP knowledge and explain a specific standards which were applied to each dataset and output.

The team can have a lot of benefits if statisticians and programmers will be more transparent with each other and to do it we can specify some tips and tricks which will simplify the process of such communication:

- If something is not 100% clear then do some research, try to identify what exactly is a problem but don't waste too much time on it. In the team there are people who will help and support you. Do not afraid to ask too many questions.
- Be proactive and try to think about every step in advance, remember that all steps are related and depend on each other so think about purpose and consequences of every decision.
- Identify the right time to ask questions. It is always better to perform a review not only when outputs are created but also before coding to be sure that sponsor expectations can be satisfied, specified in protocol analysis can be performed using methods from SAP. Don't wait till last steps of outputs creation to see that specific statistical procedure is not working for some reason.
- Set-up regular calls between programmers and statisticians to align study needs and expectations.
- Be transparent in communication, if something is not working and it is not a syntax error then raise a question, ask for additional help.
- Be sure that all derivations are well documented and tracible. For ADaMs ask to add more rows/variables for all intermediate derivations for a better traceability. For outputs add programming notes or annotations to shells with all the details of analysis: required procedure, variables, timepoints, back-up analysis.
- It is possible to ask the same questions multiple times during the study, data is changing all the time so if question was resolved for some data cut it doesn't mean that the same concern will not be applicable to new data. It is always better to have a data-driven approach to all problems.
- Don't afraid to acknowledge that some type of statistical analysis is new for you. Do some investigation, read corresponding documentation and ask for support and explanations with coding and/or statistical purpose and expectations.

To summarize all points above it will be good to say:

- Keep in mind all common questions/checks specified in previous section and ask them in the beginning of the phase to which they are related. Start with some investigation and escalate question to the study team if it is needed.
- Extend the list of common questions based on study, always keep in mind that a lot of things can be study specific.

EXPECTATIONS

Based on process overview, specifically Figure 1 and Figure 2 it is possible to see that programmers and statisticians are working to achieve the same goal but in a different manner. Statisticians are responsible for statistical documents and review process, algorithms creation, results of analysis and programmers are trying to make such goals come true by developing codes keeping in minds different standards. What is our expectations from each other to make our tasks easier and provide more support?

For programmers the expectations from statisticians can be summarized as follows:

- Provide all details for each algorithm, every SAP algorithm supposed to cover every possible situation from the data in order to create efficient code and don't update code every time when new data are added;
- Help programmer to develop logical checks of the data based on complex algorithms;
- Specify statistical procedures which should be used for complex analysis, code example can be

helpful but not a requirement, it depend on complexity of analysis but at least it is good to mention the procedure;

- If it is possible provide annotation for shells with complex statistical analysis or at least specify what kind of results are expected;
- Help with data selection for a specific output, specify the criteria and provide such details in advance, it is possible that some criteria is easier and better to program on ADaM level rather than TLFs level;
- Specify ADaM data structure, at least what kind of rows and variables it has to contain;
- Explain the logic and goals of statistical analysis;
- Specify data checks which supposed to be performed prior to statistical analysis. Even if such checks are manual and not a part of ADaM dataset or output explain the goal of such checks. Explain under what condition it is possible to proceed with analysis;
- Help programmer to select results from statistical procedure output which supposed to be displayed in final TLFs.

For statisticians the expectations from programmers is a little bit different:

- Perform data checks based on Pinnacle 21 report, study documentation (cross check that actual value corresponds to their descriptions from documentation, identify abnormal values which can be data issues) and raise them to data management team;
- Transfer all available raw data to SDTMs to be able to use it at any time even if initially it wasn't needed for analysis, every detail and piece of the data is important;
- Create SDTMs, ADaMs based on CDISC standards, avoid Pinnacle 21 issues, create data structure which will be suitable for planned analysis;
- Check the consistency of data in case multiple data sources are available;
- Create all codes following GPP (Good Programming Practice), perform all steps of programming validation prior to statistical review;
- Transform analysis logic into codes/algorithms;
- Make sure that all transformations of the data and derivations are tracible, easy for review and well documented;
- Do some investigation of logic/codes issues prior to raising a question but don't hide any problems, raise reasonable concerns and be transparent;
- Explain how to combine the data from different sources for analysis and cross-checks from technical point of view.

These lists of expectations can be extended even further but it is a very good start which can be used to develop a specific list on your own based on personal experience and check if each of us are meeting others' expectations in order to be efficient, comfortably work together and provide a good level of support. Obviously, such lists includes some points which are specific to our job role but effective communication is not based only on some of our hard skills which is allowing us to raise some good questions and concerns in timely manner but it is also important to keep in mind how we are communicating our findings. It is possible that even with good findings the environment is not comfortable due to ways of how it is communicated. To avoid such problem, we have to keep in mind that first of all we are a team and every issue is not somebody's else problems, every study concern or even quality issue is owned by a team and not by a specific person. Our main goal is to help and support each other, reduce stress level and achieve the best quality results and to do it we have to be polite, always respect each other and be accountable for all our actions personally and as a part of a team.

ACKNOWLEDGMENTS

The author is grateful for all the support from IQVIA Bios team and valuable experience which encouraged the creation of this paper.

CONTACT INFORMATION

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

Diana Avetisian
IQVIA
+380639608781
diana.avetisian4@iqvia.com
https://www.linkedin.com/in/diana-avetisian-801a3813a/