PharmaSUG 2025 - Paper LS-360

Data management and biostatistics synergy: how to achieve and what can be expected

Diana Avetisian, IQVIA

ABSTRACT

Conduct of clinical trials can be a compilated process especially when it comes to data collection and statistical analysis. There are a lot of standards to consider for data collection, analysis and representation of the results such as CDISC standards and regulatory recommendation including different guidances for specific therapeutic area. To achieve high quality results, meet all expectations and have a successful study submission all functions have to work in sync. Each group have their own rules, limitations, and standards to consider so in order to succeed we have to find a way and right time to communicate and help one another to avoid any unnecessary rework. Specifically, when it is coming to data collection and analysis data management and biostatistics have to work together and understand each other needs. Obviously, the key point to have best results is effective communication but is it enough? What background knowledge should each of us have to understand other function and speak the same language?

Even though both data management and biostatistics departments are working with data and actively involved in data collection process and analysis the responsibilities are different and sometimes it can be challenging to understand each other logic/language. The synergy of data management and biostatistics can help achieve tremendous progress in data collection, reduce number of potential data issues, save study budget, prevent a lot of unnecessary questions during submission, save a lot of time and reduce stress level for all groups. So how all of us can help one another to reach such ambitious but nevertheless necessary goal?

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 database creation and input from different departments;

- some common questions from data management and biostatics departments;

- some tips and tricks for programmers how to understand the requirements of data collection and possible downstream effects ahead of time even without specific background;

- mutual expectations between data management and biostatistics for efficient collaboration.

This paper is a sequel to paper "Translation from statistical to programming: effective communication between programmers and statisticians". Both papers are dedicated to the topic of effective communication between different groups and people with different background but their collaboration is a key point to successful studies.

INTRODUCTION

This paper is dedicated to collaboration between data management and biostatistics departments 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 as a team. 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 data collection process that's why it is important to help and support each other.

In order to create an effective communication between data management and biostatistics 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 database creation and lock. 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 data management and biostatistics 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 results.

Even with the list of questions related to data 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 DATABASE CREATION AND LOCK, INPUT FROM DIFFERENT DEPARTMENTS DURING THIS PROCESS

To understand and improve collaboration between different departments we need to know the details of the data collection process from different sides. Obviously, we all expect that in the end of the process we will have to lock the database with clean data and analyze it but do we really understand the differences for data management and biostatistics in the database creation, cleaning and lock? To discuss our needs and find out what can be improved let's consider the details of the procedures for the usual study where we are following CDISC standards and try to describe database related activities step by step. Here we are considering study with CDISC and other therapeutic area specific standards as it contains all possible steps and checks: designing of eCRF, creation of database, data cleaning process including query management, lock of the database, analysis of collected data. For studies which are not following CDISC standards some steps can be omitted (for example for some studies where SDTMs are not used there will be less limitations since there will be no need to check Pinnacle 21 report and resolve issues related to the data in that report, etc.) but the general logic still will be applicable.

On the very first step every team is starting with documentation: protocol review, 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 data management starts their work with eCRF design, creating documentation of data collection, reconciliation plan, data transfer guidelines, data acquisition specifications, data coding if applicable and database creation with support from cross-functional team including but not limited to medical and biostatistics departments. Statisticians start their work by reviewing protocol, creating the SAP or reviewing it, depend on the roles on study and providing input to data management about draft eCRF to see if all data required will be collected and can be used to perform all planned analysis. At the same time programmers' input to the eCRF design and SAP might be needed to confirm that study can be conducted according to the standards.

Quite often statistical programmers start their work with CRF annotation when database is created and review of data specifications for vendor data. In the beginning of every study each programmer is familiarizing themselves with the data that are planned to be collected, checking what kind of data manipulation will be required and if it is possible to perform them according to CDISC standards and GPP

(Good Programming Practice). However, it might be beneficial for both data management and biostatistics departments if programmers especially CDISC standards specialists will perform review of draft eCRF to confirm if it corresponds to all standards especially for studies where therapeutic area specifics might be applicable. In addition to standard compliance, it is important for programmers to check the structure of CRF forms to confirm that they will be able to correctly map data to corresponding SDTMs, merge CRF data with vendor data and merge data from multiple SDTMs to ADaMs if it will be needed for analysis.

On the next step statistical programmers are working on SDTMs. In order to start specification preparation they need annotated CRF, vendor data specification and ideally it is a good practice to have at least test data transfer to check dataset's structure and formats. During review of test data transfer feedback from statistical programmers is crucial for data management to avoid any problems when actual data will be collected. It can be comments about data transfer process, number of raw data datasets, content of those raw datasets such as formats of variables, number of variables, name and labels of those variables and ways of data representation in datasets for different cases, etc. Such feedback can help build efficient collaboration between data management and biostatistics from the beginning and reduce number of possible problems for both departments. On this step it is important to not just identify any possible issues but also communicate them in timely manner. The earlier any possible questions will be identified the earlier it is possible to discuss them, find solutions together and avoid any timelines extension, rework and save study budget.

Next important milestone for both departments is first transfer of actual data. It means that data management is finalizing their process of data extraction and transfers and biostatistics are starting with data analysis, mapping data to SDTMs, creating first compliance report such as Pinnacle 21 report for SDTMs and starting the process of data issue identification on their side. At the same time data management already have in place some edit checks to verify data entry on their side. Edit checks are prepared and monitored by data management but to create them sometimes it might be beneficial to have input from medical team. In addition, it might be good to develop such checks together with biostatistics department as a part of data cleaning plan for the study. Obviously, the biggest part of the work here is performed by data management and clinical data programmers who are creating database and adding edit checks there but suggestions of additional edit checks and review of planned edit checks by statisticians and statistical programmers can make data cleaning much more efficient for all departments. It can help to avoid double work on data management and biostatistics sides. It is important to mention that specifics of edit checks can have some limitations and it is crucial for data management to explain the capability of it and check if suggestions are possible to implement and if they are beneficial in the end. Conversations must be driven by both parts and can't be just blind requests; it is possible that some ideas will be better to implement on biostatistics side rather than adding them as edit checks.

Once database and all related documentation are created and process of data collection is started it is a beginning of data cleaning phase of the study. Obviously edit checks mentioned above are important part of data cleaning but it is not enough. Regular and thorough data cleaning should be an ongoing effort throughout the study from different teams, not just a task reserved for before the lock during the final stages. Data management is responsible for raising queries based on feedback from medical and biostatistics teams. By conducting frequent data reviews, cross-functional team can promptly identify and resolve data discrepancies, minimize number of issues before database lock and more importantly make sure that there will be no issues after database lock, ensure that data will be reliable and accurate. A well-organized system for managing queries is crucial. To have an efficient process of data cleaning it is important to clearly define the roles and responsibilities to ensure that all queries are resolved efficiently and in a timely manner. This proactive approach to query management helps to maintain high quality of the data and ensures that no critical issues are overlooked, facilitating a more seamless transition to the database lock. Sometimes it is not easy to find inconsistency in data especially between different forms so to help with such task biostatistics department is conducting dry run to create ADaMs and TLFs. During dry run it is crucial to verify all analysis algorithms and make sure that collected data is suitable for analysis. In case of any problems data issues should be reported to data management department as soon as possible, quite often it is beneficial to include medical team for support to make sure that results of analysis will make sense from medical point of view and data is corresponding to medical expectations (such as ranges for some values, definitions of derived parameters, etc.).

The last but not least is the process of database lock itself. First step is to define clear criteria for database lock such as required data quality standards, the completion of data cleaning, and resolution of queries. It is important to know what defines database lock, it can be the completion of eCRF activities (data collection, cleaning, validation, reconciliation, SDV'd (Source Data Verification), and sign-off), or the approval of the SDTMs. Obviously collaboration between data management and biostatistics is crucial here. On one hand the lock of database is clearly the responsibility of data management once all forms are cleaned and there are no pending queries or issues. On the other hand final data should be accepted by statistical programmers and statisticians to verify quality especially if the process of database lock is defined by lock of SDTMs. In order to have smooth process of database lock it is important to establish effective communication from the very beginning of the study, make sure that discussions between key stakeholders regarding data issues or other data related concerns are regular and not postponed till last moment. For this process to be successful it is important to encourage real-time data entry on sites, timely data entry at clinical sites can help to prevent backlogs that could delay the lock. Regular data quality audits and interim reviews (conducting interim analysis in addition to dry runs) are vital for early detection of potential issues. These practices reduce the burden before the final lock by addressing problems as they arise and reducing stress level for the team. It is important to understand that each team brings something to the table so each practice specified above require constant input from different groups, it was mentioned that communication between data management and biostatistics is important however it should also include close collaboration with monitors (CRAs) as they have to complete monitoring visits and resolve any outstanding issues within planned timelines. Once data are fully checked by multiple groups it is good to go through a checklist and verify that each step is completed:

- All expected subject data are available in the database.
- Data review listings were checked and all issues had suitable actions.
- All queries are answered and resolved (sometimes answers should be double-checked to make sure that response is filling expectations).
- Medical coding for adverse events, medical history, concomitant medications is completed and approved.
- Any vendor/external data is reconciled with clinical database.
- Any issues that are not possible to resolve are documented properly.
- Final SDTM package is approved (if applicable).

SOME COMMON QUESTIONS FROM DATA MANAGEMENT AND BIOSTATISTICS

Obviously hard skills are important for any task, it doesn't depend on type of work or department as to complete assignment each of us need certain type of knowledge in our field. But is it enough? Every milestone in clinical trials is a team effort, it can be team from our department that we know and understand better as we have some common ground, or it can be cross-functional collaboration. For study to succeed we need to achieve synergy between departments. To do that we need much more than just our hard skills, soft skills sometimes are much more important. It is starting from leadership of each group as they represent their whole team and should be supported by each team member to help the leads, team in general, cross-functional collaboration and achieve better quality of data and study in general. 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 especially it is applicable to data review process as it can be difficult to differentiate between data issue and some real case medical scenario that are not a part of usual pattern. For statistical programmers the first person to ask is usually their team lead and statistician but if question is related to data itself, the way how CRF is designed, how it is collected, what scenarios are acceptable and what can be expected then it can be data managers or medical team. At the same time if some unusual case was found in the data medical team and data management may need some assistance from biostatistics to understand how to document it and what consequence it may have for analysis. It is important to note that it is not just statisticians who will determine how to approach

any problematic scenario, it is also statistical programmers who will check how it is affecting programming algorithms and if there are any conflicts with CDISC standards. Let's talk about the most common questions for biostatistics and data management and try to determine what can be done to make the task easier and more efficient for both sides on each step of the process which were described in previous section.

QUESTIONS FROM BIOSTATISTICS

Programmers are starting the study by familiarizing themselves with documentations: protocol, draft eCRF, SAP once it is available and any other related documents. If programmers are involved to the study early enough then the first step for them is to review draft eCRF and provide their input, in order to do that they can raise some questions:

- Does eCRF include all data required for future analysis? Does it contain all forms to collect data based on protocol?
- Is there any vendor data required for the study? Is there data specification for vendor data?
- In case vendor data is planned for the study, is there reconciliation plan between vendor and clinical database?
- What documents are available related to data? eCRF is just a first step, is there data transfer specification? Is there data acquisition specification?
- If medical coding is expected, then who is responsible for it? On what level coding will be performed? How it will be included in the data? Is it data management or biostatistics responsibility to include to datasets? Is there data coding guideline available?
- 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)
- If there are specific therapeutic area guidance then does planned data correspond to those standards? Does data collection process, eCRF forms satisfy to CDISC standards?

All such questions can be classified as preliminary checks and it is ideal to answer them before data collection, it is easier to make changes and updates to the database during creation period and avoid unnecessary updates during data collection process. Usually all these questions can be answered by programmer and statisticians if all documents are available. Since it is better to raise those questions early on it is possible that some documents are not finalized, in this case any possible concerns can be discussed with data management. 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 data management and medical teams to discuss details:

- 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 and consider different statistical approach?
- Do we need to update any of out processes for a specific study to accommodate certain needs?
- Do we need to know more about the data and have a better understanding of the statistical analysis or medical background behind it?
- What assumptions for data should be checked for a specific statistical analysis? Do we need to perform these checks once data will be collected or there are some expectations that should be clarified for a better design of eCRF?

It is important to have answers to such questions, it 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 statistical team will ask them in advance then the solutions can involve data management, 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 biostatistics 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.

Once database is created, test data transfer and first real data transfer are checked there is some next steps in the process for biostatistics. Obviously once the data collection process started all groups should be proactive with data review, raise issues if it is needed and work on data cleaning process. To support it statisticians and statistical programmers are creating SDTMs and checking compliance reports, performing consistency checks between different forms and conducting dry run to check if planned analysis can be performed. During these processes it is good to check some more questions:

- After SDTM creation are there any issues in compliance report (such as P21 report)? If there are such issues then it is important to check the source of it and make sure that corresponding data issues are raised to data management.
- How should data look like to be suitable for TLFs creation? What data structure is applicable to your study? Is it possible to transform raw data to satisfy those requirements on SDTM and ADaM level?
- 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? Can reconciliation be performed on DM level or statistical programmers have to check data on regular basis and raise their concerns?
- Are there any specific scenarios in data that were not expected and not a part of expected algorithms? Should such cases be considered as data issues and expected to be fixed or is it logical to have some exceptions and it should be analyzed and accounted during statistical analysis?

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 data cleaning process. These questions should be raised internally first in biostatistics department but if there are no answers then it should be extended to data management.

On the last step of the process which is lock of the database biostatistics department supposed to accept data, verify that there are no pending issues and create SDTMs if it is required. During this process it can be good to check a few things:

- Are all data issues closed in bios data issue log? Are resolutions provided to raised issues are suitable and meet expectations?
- Are all errors and warnings from compliance report related to data issues resolved? If something is not resolved check if such issues were properly documented and can be explained in reviewer guide.
- Is the proper documentation available for any deviations from the process and expected results?
- Is medical coding completed for all records? Is it consistent? Is it done based on latest version of dictionary?

Such questions can help to achieve smooth database lock process without extra pressure and stress. Moreover, if such questions is raised some time prior to database lock it is allowing the team to identify any pending issues earlier and avoid possible consequences not just for data quality but also for timelines and avoid any rework for other groups especially when SDVs are signed.

QUESTIONS FROM DATA MANAGEMENT

From the very beginning of the study data management is driving all activities related to data collection including database creation and corresponding documentation. Next step is all activities related to data cleaning and finally database lock. Even though data management is performing all main activities they still need input and support from other groups. During data collection to resolve any queries, verify final data CRAs support is crucial. In case of any protocol deviations, unusual scenarios it is important to have input from medical team to fully understand possible consequences for the study. To have high quality data constant review from medical and biostatistics are necessary. So, it is safe to say that every group is related to data collection process and that's why data management must initiate discussions when it is needed and raise concerns to move on with data cleaning without extra problems.

During database creation it is obviously crucial to work with medical team to make sure that data will have sense from medical perspective and wouldn't be incomplete. However, when we are talking about CDISC standards it is important to involve biostatistics as soon as draft eCRF is available, work together on corresponding study documentation and make sure that study will be fully compliant. To achieve better results, it can be useful to raise some questions on early stages of the project:

- Are there any specific standards applicable to the study? Should biostatistics specialist be involved in review process before database go live?
- Are there possibilities, deviations that will be difficult to record using current forms?
- Is there reconciliation plan created for vendor data?
- Who is responsible for medical coding if it is needed?
- How protocol deviations will be reported and who will lead that activity? Is it clinical team who will report it in their system or would it be data that will be reported and monitored by data management?
- Is it possible to link all collected data with each other if it will be needed?
- Is there a possibility to have duplicates in data? Are there enough distinct fields/variables to avoid duplicates? For example, if few samples are collected within the same day it is important to make sure that time component is included as a field in CRF form to differentiate assessments.
- Is there any input required from medical or biostatistics teams for certain questions in addition to standard reviews?
- Is there enough edit checks created for most common issues? Is it possible to add some automatic checks to avoid manual work in future?
- How often data transfers should be performed to different groups for review?

Such kind of questions is easy to answer/implement on early stages of the study when data are not collected yet, it can save a lot of time for team in future and help to avoid unnecessary database updates in the middle of the study. Some of those questions can be answered fully only with some help from cross-functional team.

Once database go live and data are added to database it is time to start data monitoring, review and cleaning process.

- How often data review is performed outside of automatically generated queries?
- Is there enough information in data issue logs to raise a query in database?
- If any specific situation is reported by clinical or medical team do we need biostatistics input to understand possible consequences for analysis?
- Is there regular communication from the beginning of the study with cross-functional team to review data on regular basis and discuss ongoing challenges?
- Is there any issues that are applicable for multiple subjects and raised more often than others? Is there a way to track and avoid similar issues?
- Are ongoing issues closed on regular basis? The process of verifying and closing issues are important as well as raising data issues. Who is responsible for closing data issues in different logs?
- Are there issues that require changes in database?

Query management is always performed by data management but issues leading to those queries can be raised by multiple people outside of DM so questions above can help to navigate data cleaning questions related to other groups.

The last but not least step is database lock. It can be a smooth process if data checks and reviews were performed on constant basis during study but certain aspects still should be considered especially when people outside of data management is involved:

- Are timelines allowing enough time for cross-functional team to provide their approval of data? Especially for data collected in the very end of the study.
- Are criteria for database lock clearly defined? Who should be involved and what steps need to be completed? Who will be representing each department and sign database lock forms?

Involving the key stakeholders early in the discussion of database lock process will help to ensure that the requirements for each function are clear from the beginning and will simplify DBL.

Sometimes it can be difficult to answer some of the questions by yourself but with proper communication group knowledge of CDISC standards, medical background, data collection requirements, statistics can be combined to achieve better results and have smooth database lock.

TIPS AND TRICKS TO UNDERSTAND REQUIRMENTS OF DATA COLLECTION PROCESS

Previous section is dedicated to common questions from data management and biostatistics and the reason behind it is to determine most common problems/concerns, figure out when the best time during study is to raise them and how to get answers which will be understandable and beneficial for both groups. When it is too late to raise questions? Obviously the answer can be when study is closed and submitted to FDA/EMEA/PMDA or database lock is completed however sometimes even before database lock it can be too late to voice some concerns especially if solution to fix it will require some drastic measures such as a change in database design/CRF form or if information can't be retreated from closed sites for some subjects or if needed information wasn't collected and there is no way to add or restore it. Is it too late to raise question if there is not just enough time for required action based on timelines? Sometimes the answer is yes but it is always a tough choice between the possibility of fixing the issue and moving timelines especially quite often straight forward answer to fix the issue doesn't exist and it is difficult to tell if serious issue can be fixed for all subjects. With all specified unfortunate scenarios, the problem is usually possible to fix however sometimes there is no easy way to do it. If database was locked or study was closed, then ad-hoc analysis can be performed or database can be unlocked. If there is not enough time to fix the issue, then timelines can be extended. Only if information wasn't collected properly and there is no way to go back and retrieve it then there is not much that can be done especially on data management level but in such cases biostatistics department can try to help and derive required milestones from existing data if it is possible, add another layer of calculations to their algorithms to have more accurate results. Each of specified cases is not pleasant for any department, it can be stressful to deal with such problems and can put a lot of pressure on team members, affect study budget. Ideally we would like to avoid such situations as much as possible and to do it our goal is to build effective communication and corresponding actions between different groups, learn how to identify possible issues on early stages of the study which will allow us to reduce stress level for the whole team, achieve better quality of data/analysis and spend less time on it by being efficient and productive.

Sometimes even when all the general 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 most common problem is that approaches to data collection, issues and lock are different for data management and biostatistics. Data management and medical team usually are reviewing queries one by one focusing on subject information and certain forms. At the same time statistical programmers and statisticians are more focused on patterns of issues and not just certain subjects that's why if some data issues were resolved for some subjects previously it doesn't guarantee that the same type of issue will not appear for new records in future. Obviously, all groups are focusing on reviewing data but the way of doing it and logic behind it can be different. While data management is usually verifying data form by form for a specific subject, have edit checks to confirm that data entrance corresponds to certain expectations, biostatistics are more focused on consistency checks between different forms. So, to raise data issues we have to keep in mind that describing just logical pattern of some problem is not enough we have to provide enough details such as subject id, name of the form, field with a problem and our actual expectation of what should or shouldn't be there. The same is applicable to data management when they need input from statisticians, it is not enough to just ask question about certain subject, it is important to specify how many similar cases are currently in the database and what are the options to fix it, such information is vital for statistical team to see possible

effect on analysis and find a solution to accommodate it on SDTM or ADaM level. Both groups have some assumptions and expectations about data from the very beginning of the study and it is not wrong to have them but we have to keep in mind that real world can be different from it. Usually for the biggest part of data our expectations can be met but it is always good to perform checks on regular basis. To be effective, we actually have to communicate all our assumptions clearly to each other and if it is needed "translate" them. For example if there are solicited adverse events for the study and specific form is collecting results for a certain period it is good to know what supposed to happen after it. What form should be used if solicited adverse event was extended beyond specified period? How duration of such event should be calculated? How to merge data for the same event from different forms? What if solicited adverse event was not reported by subject and it has appeared later that subject forget to report some information? Answers about duration and merge of data are usually coming from biostatistics and it is important to translate them in terms of data entrances and specific variables, specify our requirements to make it happen. Some other answers to these questions may seem obvious but in fact each study is different, so it is good to discuss such things as early as possible and align our expectations. Some people may say that certain situations are rare and may not happen, but it is better to be prepared and know how to handle unusual scenarios rather than identify a bunch of data issues when information is entered and didn't have availability to fix it.

The team can have a lot of benefits if data management and biostatistics 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:

- Be proactive and try to think about every step in advance especially during database creation, don't wait for the data to be collected to investigate it and find some issues, try to prevent problems by reviewing draft eCRF, test data transfer and checking it for standard compliance. Remember that all steps are related and depend on each other so think about purpose and consequences of every decision.
- If something is not 100% clear or there is a concern then do some research, try to identify what exactly is a problem but don't waste too much time on it. It is good to raise a question to the team, there are people who will help and support you.
- Identify the best time to ask questions. With data issues obviously it should be raised as soon as it is identified but try to prevent new issues by doing review and asking questions in advance. Don't wait till last steps to check if data will satisfy required algorithms or till the time of lock to raise issues.
- Set-up regular calls between data management and biostatistics to align study needs, expectations and review data issue log together.
- Be transparent in communication, if something is not working and it can be related to data then raise a question, ask for additional help.
- 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 data and corresponding analysis especially if it is study specific are new to you. Do some investigation, read corresponding documentation and ask for support and explanations.

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 as soon as possible. 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 and data driven.

EXPECTATIONS

Based on process overview it is obvious that data management and biostatistics are working to achieve the same goal which is collecting and analyzing data with high quality and eliminating all possible data issues but the way of doing it is a bit different between groups. Data management is leading a way of data collection and lock but input from monitors, medical team and biostatistics are crucial to this process. While data management is responsible for database creation and corresponding documentation biostatistics are responsible for review of it as early as possible, verifying it for analysis purposes and standards compliance. During data cleaning phase data management is raising, tracking and closing queries based on edit checks, feedback from medical and statistical teams. At the same time biostatistics department is working on consistency checks between different forms and sources of data, creating dry run to verify that data correspond to requirements of planned analysis and raising any possible concerns about specific cases. During database lock process once all data are cleaned and verified data management is driving the whole procedure step by step, but biostatistics department is the one who is accepting data after the lock (either just raw data or generating SDTMs for approval as the last step of database lock). What are our expectations from each other to make our tasks easier and provide more support?

For data management the expectations from statistical team can be summarized as follows:

- Review draft eCRF, any documentation to verify that analysis expectations are met and collected data will be represented according to CDISC and therapeutic area standards;
- Verify that data from clinical database and vendor data can be linked and mapped correctly to SDTMs;
- Verify that all planned analysis can be performed using data from eCRF, provide feedback if there are any concerns and/or lack of details is a concern;
- Provide feedback about data transfer process, number of raw data datasets, content of those raw datasets such as formats of variables, number of variables, name and labels of those variables and ways of data representation in datasets for different cases, etc. based on test data transfer;
- Provide input about edit checks, suggest some edit checks that will benefit both groups and reduce rework on both sides;
- 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;
- Raise data issues as soon as they are identified especially issues identified during analysis algorithm testing, any inconsistencies between forms;
- Make sure that all issues reported in understandable way, not just general patterns that can be concerning but specific variables such as subject id, name of the form, specific field (variable in raw dataset) and expectations of what should or shouldn't be there to be able to raise corresponding queries in database;
- Support from statistical team for any questions from data management. Specifically, it is related to any unusual scenarios in data and their consequences for analysis;
- Support in closing raised issues from data issue log including but not limited to the issues raised by biostatistics;
- Provide support during database lock process by following timelines to clean data on time and accepting data once database is locked or creating SDTM package for approval (depend on study definition of DBL).

For biostatistics the expectations from data management are a little bit different:

- Create database/eCRF design;
- Create documentation of data collection, reconciliation plan, data transfer guidelines, data acquisition specifications, data coding guideline;

- Provide test data transfer to biostatistics to test the way of transferring data and have a feedback about content of raw datasets;
- Create edit checks in database to increase data quality and provide greater efficiency during data review and cleaning activities;
- Review data regularly and own cleaning process by managing queries and communicating with monitors on sites;
- Involve monitors, medical and biostatistics teams in decision making process about data collection or any updates in database design;
- Update design of database during study if it is required, transfer all data from one version of the database to another without data loss, make sure that any additional/new data entered properly after updates;
- Conduct a final comprehensive review of all data to ensure completeness, accuracy, and consistency. Involve all stakeholders in this process including clinical operations, data management, and biostatistics teams to make sure that every aspect of the data is checked.
- Define criteria for database lock, navigate this process and create timelines to achieve it. Lock the database once data is fully cleaned.

Obviously lists specified above are just a start and can be extended even further, moreover they can be customized based on personal experience. The point is to use specified lists to make sure that each of us are meeting others' expectations in order to be efficient, comfortably work together and provide a good level of support. It is easy to see that specified expectations are specific to our roles, but the common ground is effective communication and teamwork. All our hard skills and ability to perform certain tasks will not make sense if we will not be able to raise questions, talk to each other and request support where it is needed in timely manner. We can't expect miracles from other departments that will be performed in no time so we have to remember how the process is working for other group; it is good to understand what steps and how much time can be required to solve any issue even if it requires actions from a different department or collaboration between them. 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. Synergy between departments is achievable only through combination of hard skills, proper communication, support and creation of trusting environment to be comfortable to ask for help and express any concerns.

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/