PharmaSUG China 2024 - Paper SA- 10005

Programming Activities of Conducting Data Masking in Open-Label Study

Xinyun Gao, Dizal Pharma

ABSTRACT

As we all know that randomized double-blinded study is the most ideal design to reduce systematic bias during clinical trials, however it is impossible to conduct at certain cases. For example, when distinguishable treatment schedules are required, open-label design is an unavoidable option. In this situation, to make the conduction and outcomes of the study more convincible and reliable, it is better for sponsor to choose 'Sponsor-Blinded' method to limit bias during study management, particularly when interpreting analysis results. This article is to introduce data masking method when considering blinding study team members at aggregated summaries during conducting of your study. It will go through basic principle of data masking, guidance of working procedure in conducting data masking, examples of how to implement blinding considerations by programming, etc. The examples show what kind of masking need to be done when data collection easily reveals treatment information and simply using a dummy treatment group may not suffice. Additionally, when it comes to cross-over study design, different masking considerations may be required for two treatment periods. Apart from programming related methods, it also includes related programming project management, experience and suggestions of study communication, pathway access control of different data transfer and dealing with integrated analysis includes masked study members. Although SAS was used for programming in example study, no SAS programs will be included in this paper's contents as its focus on providing comprehensible rationales to any statistical programmer with basic understanding of SDTM/ADaM data mapping.

INTRODUCTION

When study described as "open-label" in protocol, it may be open from a regulatory, investigator and participant perspective, but not fully open to sponsor study team. Especially in confirmatory studies, keeping sponsor-blinded will strength study integrity. Health authority also has guidance suggesting using some blinding method to reduce bias and prevent sponsor from summary analysis of treatment group whilst the study is ongoing. Since sponsor still has the obligation to monitor and management study, in particular of participants' safety signals, sponsor-blinded mainly focus on aggregated level of study data. In this case, data masking should be performed on all study data from any source which could reflect treatment group differences.

This paper will introduce some basic considerations when conducting data masking. One data masking plan or masking specification should be generated to document masking rationale and specify masking method of each data point. It is recommended that programming lead of the study to drive corresponding discussions since he/she is most familiar with data processing and how the masking will affect it. After the document finalized, all in-scope data should be done properly following the specification. Process related to conducting, transferring and validating data masking should also be clarified and aligned with involved study members in advance.

Programmers will face more challenges when using masked data for regular programming activities, such as SDTM and ADaM mapping. Study programming lead need pay more attention on communication and other project managements with other functions since there is more risk of leaking unblinded information than typical double-blinded studies. Activities of preparing pooling analysis will also be different when including this kind of internal-blinded study. Some suggestions and notes will be given in the paper.

DATA MASKING PROGRAMMING DEVELOPMENT

PRINCIPAL RULES

When considering data masking, the first thing to think about is the dummy method of treatment group. Unlike blinded studies, CRF of open-label study are usually designed to serve different data collection requirements due to different schemes of treatment groups and group information will be collected in EDC.If simply dummy the group information, such as replace group information as "Drug A/Drug B", will

still have other data points leaking real treatment group. Also, the principle of data masking is to prevent any kind of summarizing real aggregated results. Not only the typical way of statistical programming workflow should be taken into consideration, but also some unusual data summary method, such as one study team member using excel data to perform some analysis and get the real treatment group results or close to.

One efficient way is to break the true connections between treatment group and other data. As subject ID is the basic key variable when merging different raw datasets, proper masking subject ID will make proper blinded data mapping.

There are two methods of masking subject ID: one is to simply replace real ID with dummy values, the other is to add re-ordering in the dummy IDs.

Table 1 is an example of connections among the real subject ID and two masked IDs.

Real Subject ID	Dummy Subject ID	Re-ordered Subject ID
AAA	002	001
BBB	006	002
CCC	001	006

Table 1. Mapping Data for Subject ID Masking

When applying different masked subject IDs for different raw datasets, two true subjects' information will constitute as one fake subject. To make sure summary by treatment group is not disclosing the true result, mask ID in raw dataset which collects treatment group information must be different with the others. The better way is to use dummy subject ID in treatment group and re-ordered subject ID for other safety or efficacy datasets. This method can reduce logic issues due to CRF designed with different treatment schemes as much as possible.

Figure 1 shows using upper method of masked subject IDs as one example, SDTM will be dummy at subject-level. Subsequent ADaM and TFLs are not able to reflect true group analysis results based on these masked data.

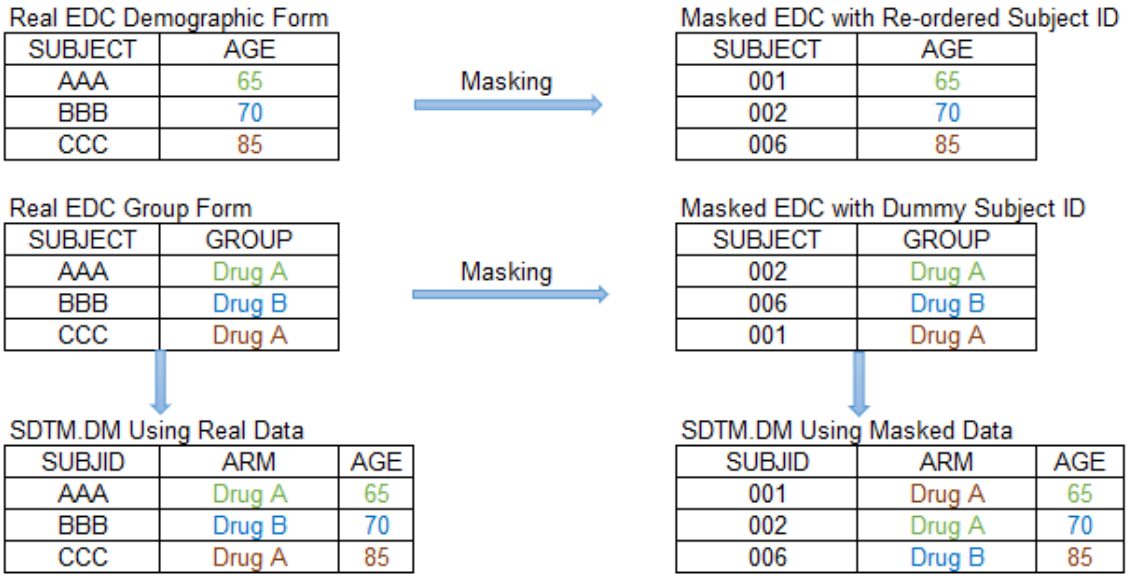

Figure 1. Example of How Masking Works in SDTM Mapping

As mentioned before, when study designed as open-label Randomized Controlled Trials (RCT), treatment schemes of each group are different most of the time. For example, when one group dosing route is oral and the other is IV, exposure forms can be designed quite differently. In this case, if masked data has one subject assigned as oral treatment group but exposure information linked as IV, SDTM and ADaM

mapping can be quite difficult with logic issues. Thus, it is better to use the same subject ID masking method in treatment group dataset and other group-sensitive datasets, such as exposure, treatment discontinuation, etc. based on CRF design. Then left raw datasets use the other method.

To facilitate programming, it is recommended to generate one mapping file like Table 1 at first. Then merge each raw dataset with it to replace with dummy subject ID or re-ordered subject ID as needs. Re-ordering should consider separate screen failed and enrolled subjects. Other blocks can also be added according to study design.

SPECIFIC MASKINGS FOR EDC AND EXTERNAL DATA

After the general masking of subject ID, next step is to evaluate which specific data point need masked. Study team should go through CRF page by page, pick every data field which may reflect group information. Some may be very straightforward, such as adverse event causality with different treatments. Some may become meaningful after connecting with other data, such as the date of treatment discontinuation. According to the criteria defined in protocol, this time point may reflect some information related to efficacy. In this part, opinions from each cross function are all important. Not only statistics, but also make sure physicians, pharmacologists and other applicable team members cannot summary the real treatment differences based on masked data.

Two options of masking these specific data are removing or replacing the values. Removing all the values will eliminate the risk of accidentally unblinded, but cause programming problems during data mapping. Too much work has to be left after data release. More efficient way is to dummy values with some random or consistent values, such as using protocol-defined dose for every records. But programmer must always pay attention to data mapping logic when dummy values, such as last dosing date in exposure page and end of treatment date in discontinuation page should still be consistent after masking.

External data should follow the same masking way after EDC. Since most of the time external data types are typical and have few groups information, like central lab or IRC, subject ID masking same as EDC would be enough. But if includes PK results, masking should be very careful or even consider limited receivers in study team.

If some data is only collected for one group, such as samples for mechanism biomarkers, should be removed from each raw data transfer during blinded period. In the other word, statistical programmer should not start data mapping for these data until unblinding.

MASKING FOR CROSS-OVER PERIOD

In many phase III open-label studies, protocol allows the comparative group participants cross to experimental group treatment. Usually this cross-over period is not included in primary analysis but still have analysis purpose. It is obvious only one group can cross to the other, vice versa, all cross participants belong to the specific group. This makes the masking of cross-over period data quite challenge.

One option to remove all subjects' data after cross-over, then no treatment group can be conjectured. But this will cause a lot of burden of programming after unblinded. If you have plenty resource, the choice is to dummy almost all the data in this period so no information can be summarized real, and programming team can conduct the data mapping as normal.

The compromise way is to keep the minimum amount of cross-over data and mask critical values. It is very sensitive of how much cross-over subjects' data can be kept and transfer to programming for data mapping. It should be discussed thoroughly in study team. For specific data related to cross-over criteria, masking should also be conducted, similar to previous replacing the values.

Making of subject ID should also be careful when cross-over subjects exist. Crossed and non-crossed subjects should be re-ordered as sperate blocks. Otherwise, logic issue will come again when mapping cross-over period data.

ORGANIZE DATA MASKING DOCUMENTS

Rationales of conducting data masking should be documented as one masking plan or specification. The principal rules of masking subject ID, dummy specific data values or remove from transfer should all be written down in the document. This document is the basic guidance of conducting data masking, so it is better for statistical programmer to organize it. The general principles can be written in descriptive wordings with some flow charts to help understand. For specific dummy data points, the format can be similar as SDTM or ADaM mapping specs, which is more straightforward to programmers.

Apart from how to conduct programming, there should always have a document of data transfer. No matter the masking is conducted by a third-party vendor or sponsor's internal unblinded team, access to different kinds of data should be strictly controlled. Usually, Data Management (DM) team will be responsible of raw data extraction and transfer, and they can hardly be blinded. The lead study programmer, usually blinded, should discuss with DM to set up separated data transfer pathways. One for real data, only masking conduction team can access. The other is for masked data transfer, which can be accessed by blinded study programmers for subsequent data mapping.

If validation of masking is required, this role must be chosen carefully. The masking validation programmer must be independent from study team, and all the work is to validate whether the masking has been conducted properly before transfer to study team. The communication and data transfer pathway between masking conduction team and validation programmer should also be documented.

When study is ongoing, there may be amendments of protocol or data collection requirements. At that time, the lead study programmer should evaluate whether the updates will affect data masking. If masking rationale need update, related document should be updated first.

PROGRAMMING CHALLENGES WITH MAKSED DATA

REGULAR PROGRAMMING WORK

During internal blinded period, even though the data are nor real, programming team still needs to start data mapping. Since the basic masking rationale is to mix two subjects' data, time logic issues will happen in masked data. For example, in VS domain, there can be records with visit name "Post-baseline Visit 2" on study day -36. Because the first dose date and this assessment do not belong to the same subject. Baseline may be flagged on different records when derived based on visit name or study day.

When dealing with real data, some logic has already been embedded in data collection, such as visits and study days, either derivation makes sense. But when it comes to masked data, programmer should use the most conservative logic in derivation, and this will save effort after unblinded. In this case, baseline should be derived using SDTM.XX. XXDTC /XXSTDTC (*XX refers to domain*) <=SDTM.DM.RFSTDTC, instead of visit name equals protocol-defined baseline visit, such as "Visit 1 Day 1" or "SCREENING".

Another logic issue related to first dose date is Epoch derivation. There can be dummy subject's data with ICF date later than first dose date and cause problems either in SE domain or EPOCH variable mapping. The easier way is to develop SE domain as usual and leave EPOCH as unmatched during dry run.

POOLING ANALYSIS

It can be more challenging when internal blinded study needs to be included in some pooling analysis. At set-up stage, study team should clarify whether this study's results can be revealed when it is ongoing. Normally, aggregated results by treatment group of the study should not be available until unblinding, but some health authority annual reports may require it because it has been stated in protocol that the study is open label, such as DSUR.

To guarantee the blindness of study team, reports for such pooling analysis should be generated by an independent team. Team of conducting data masking is a good option. However, they may only be capable of generating reports for the blinded study, especially when it is a third-party vendor and your company has no intention to disclose additional data, in the other hand internal study programmers will deal with other pooling needed studies. In this situation, any programmers participated in the blinded study should not have access to the final aggregated results of pooling report. Otherwise, there will be chances to accidentally achieve the summarized safety or efficacy results of the blinded study from the

pooling outputs. Of course, if there is no budget limits or resource limit (for example, company has both local and global programming team), the ideal method is to let a whole other internal programming team or third-party vendor to perform all the pooling analysis including the blinded study together.

When the report of blinded study and reports for other study are separate, it is recommended to set up one independent team from other function, who will be responsible of pooling the results of blinded study and the others then have it validated. Programmers should discuss with the person about the importance of keeping study team blinded and set up data transfer pathways. The access control can be similar as previous masking data transfer.

PROGRAMMING PROJECT MANAGEMENTS IN INTERNAL-BLINDED STUDY

DISCUSSION OF MASKING RATIONALE

As the conductor of data masking, programmer knows the best of whether the logic can be implemented ideally and how this could affect following data mapping. Therefore, it can be more efficient when lead study programmer drives the discussion of masking. Of course, the principal rules of masking usually led by statistician and statistical programmer should evaluate the practicality, even do some tests. Not only technique issue need be taken into consideration, but also reasonable resource allocation. Compromise is sometimes necessary when the rule is too complex to program and has few contributions on affecting study team's judgements. Statistical programmer should play an important role in balancing these.

After general logic has been confirmed programmable between study statistician and programmer, other study team members can be involved in specific details.

COMMUNICATION PATHWAY

When internal blinding is agreed by study team, statistician may develop a blinding plan to document managements of study blinding and specify blinding status of each study member. Although all functions are required to be blinded at aggregated level, functions related to participants monitoring will always be unblinded at subject-level. Communication with these functions must be careful because they may not be as sensitive as programmers and statisticians to unblinded data, especially when communicating with DM. Because programmer and DM work very closely during study and will talk about problems in data collection, see how these will affect data mapping. The lowest requirement is to hide treatment group and subject ID in discussion. The better way is to blur time frame of the issue and only share general description of it. Sensitivity data should be predefined and notified to each team member in case other members may leak by incidence during regular communication.

There may be problems in masking programming, or the masking validator find some issues when QC masked data. Any discussion related to programming issues in masking must be restricted between the masking conduction team and unblinded validator. Only if the problem is related to masking rationale and validator confirmed no sensitivity data need be included during the discussion, study lead programmer can be involved.

If statistical programming team is going to support regular medical monitor, such as MDR, the producer and data transfer should be different from other studies. Although MDR only includes subject-level data, it still should be generated by person out of study programming team. Data delivery of raw data and MDR also need access control, separated transfer pathway must be set up. Only DM, medical and MDR programmer can have access to raw data and delivered reports. Discussion of MDR modification only involve medical team and MDR programmer.

DELIVERY TIMELINE AFTER DATA BASE LOCK

Every company has standard timeline of TFL delivery after database lock (DBL), but internal blinded studies will be different from the others. It would be necessary to keep internal blinding until DBL of primary analysis, statistical programming will face much challenge in final delivery. In truly open-label or double blinded studies, programmers can get most programs ready during dry run. Conformance rules checked by Pinnacle 21 may reveal some data issue, and statistical programmer could send them to DM team to help data clean. But all these works cannot be done in masked data, especially when some raw

data are removed for blinding. Some domains and TFLs can only be done after unblinding, such as PK and protocol deviation.

Study lead programmer should notify team of these effects and propose reasonable delivery timeline. As TFL delivery is always on the critical path, it is understandable other functions want to receive as soon as possible. But we also have principles to uphold.

CONCLUSION

Data masking in open-label study make programming work quite different from other studies. Although health authority has not strictly required internal blinded, but it is always better to have good compliance. Some may think determine masking rationale is a more statistician's work, but we programmer are executors, and our opinion will make activities more efficient.

Develop programs to mask data is the basis of data masking. The principal rules of prevent study team from treatment aggregated analysis results must be understood and accepted by each related function. All raw data should be evaluated carefully, make sure no sensitivity data being ignored. Programmer should always keep data mapping flow in mind to find if any data may leak important information after linked with others.

Documentation is also very important in data masking. Confirmatory studies usually last years and study member transition will happen. The consciousness of keeping blinded may be faded over time or new members are not familiar with the rules. Good documentation makes conducting activities more organized. Keep the documents up to date also help the rigorousness of data masking.

Getting raw data masked is the first step of programming work in data masking, following regular activities dealing with masking data affect the eventual masking quality. Regular data mapping derivation may need adjustments when using masked data. Discussion with unblinded functions must be careful in case unblinding by accident. But study lead need more timely communication with other study members since situation is different from other typical study designs.

In a word, data masking cause more challenges in programming regular work. It has higher requirements for data mapping proficiency, logical thinking, communication sense and project management skills. Programmers of internal-blinded study should know the importance and rules in data masking better than anyone else.

ACKNOWLEDGMENTS

I would like to thank Dizal study statistician for sharing masking rationales and statistical programmer Zhiping Yan, Han Cheng for supporting data masking validation.

RECOMMENDED READING

- *药物临床试验盲法指导原则*
- *Considerations for open-label clinical trials: design, conduct, and analysis, Higgins, Karen, FDA/CDER/OTS/OB/DBIII; Gregory Levin, FDA/CDER/OTS/OB*
- *Blinding Sponsors for Open Label Studies: Challenges and Solutions，Quan Jenny Zhou, Shengyan Hong, and Zhengping Ma，Eli Lilly and Company, Indianapolis, IN*

CONTACT INFORMATION

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

Xinyun Gao
Dizal Pharma
Xinyun.Gao@dizalpharma.com