

Source Data Quality Issues in PopPK/PD Dataset Programming: a Systematic Approach to Handle Duplicates

Ibrahim Priyana Hardjawidjaksana; Els Janssens; and Ellen Winckelmans,
SGS Health Science, Mechelen, Belgium;

ABSTRACT

In a perfect world, source data used for PopPK/PD dataset programming would be complete and seamless, but practice teaches us that this is rarely the case. One of the most prevalent reasons affecting data quality is data duplication. This issue can present a challenge for programmers when preparing the datasets as it causes merging errors, failure of standard programming logic, and could increase the complexity of the program, resulting in more time and effort spent on programming.

The standard approach for PopPK/PD datasets is commenting out records containing data quality issues and record its reason in a specific variable. However, the decision of omitting data from the dataset and how to handle the data quality during programming requires careful consideration since it can have a significant impact on the accuracy and integrity of the analysis. It's therefore crucial for both the PK analyst and PKPD programmer to be aware of the risks associated with the decisions taken. Good coordination and understanding between both parties are necessary and reaching a mutual agreement is imperative.

We will share our view on how dataset programming prior to database unblinding can be advantageous for early issue detection, and how standardization of assumptions to tackle a certain data quality issue will enhance completeness of the popPK/PK dataset and accelerate the analysis process.

INTRODUCTION

Population Pharmacokinetic (PopPK) modeling is widely used in drug development and regulatory submissions. It longitudinally analyses all relevant pharmacokinetic (PK) and pharmacodynamic (PD) information across a range of investigational drug doses and populations to identify factors that can affect a drug's exposure or effect. The NON-Linear Mixed Effects Modeling (NONMEM®) software, which runs in Fortran90/95 is currently the most used tool for PopPK/PD modeling. NONMEM allows PopPK/PD mixed effect modeling, which is useful when there are only a few pharmacokinetic measurements sampled for each subject in the population, or when the data collection design varies considerably between these individuals. This model considers a combination of fixed and random effects as predictor variables.

The input dataset used for NONMEM analysis (further referred to as PopPK/PD dataset or just dataset) is required in a specific pre-defined format. These datasets which typically include dosing lines, PK lines, and PD lines can encompass biomarker results, adverse events, response observations etc. The vertical structure of the PopPK/PD dataset is arranged as one line (or record) for each subject, type of line, date, time, visit and planned point time (TPT). Identifiers for each type of lines are differentiated by a combination of Event ID (EVID), Compartment Number (CMT) and DVID values.

The horizontal structure of the datasets, in addition to the record key identifier, typically consists of additional covariates customized to the requirement of the study. These covariates can be:

- exclusive for dosing lines, i.e., amount of dosing, duration, and rate of infusion for intravenous (IV) studies, as well as Interdose Interval and additional doses (ADDL) for study with daily dose.
- exclusive for PK/PD lines, i.e., such as a below limit of quantification (BLQ) flag, dependent variable (DV) or Log DV.
- subject-level information that is applicable for all lines, like baseline weight, BMI or other demographic data, Creatinine clearance, and biomarkers baseline values.
- time varying information, which is applicable for all lines i.e., immune response and, concomitant medication taken.

To program a high quality PopPK/PD dataset for NONMEM analysis, all data in the dataset must be correct and consistent. This requires that the source data used for programming (SDTM and ADaM datasets) are also complete and seamless. In practice however, this is rarely the case. One of the most prevalent issues affecting data quality is data duplication. This problem can present a challenge for PKPD programmers when preparing the datasets, as it can cause merging errors, failure of standard programming logic. Furthermore, handling these issues, could increase the complexity of the program, resulting in more time and effort spent on programming.

To remediate the source data quality issue affecting the vertical structure of PopPK/PD dataset, variable C is used to mark which records should be excluded from the analysis. **Table 1** provides a non-exhaustive overview of the most used reasons for commenting out “C” observations and how often they are used across studies. As shown below, some of these reasons are related to source data quality issues.

Reason	Occurrence Across Studies			Data Quality Issue
	PK Sample Lines	PD Lines	Dosing Lines (in most cases dosing lines are not commented out)	
Concentration or result is below limit of quantification (BLQ)	Most Studies	Some Studies	-	-
Sample line or PD line precedes first dose	Most Studies	Rarely used	-	-
Planned predose sample isn't taken before or at the same time of drug administration	Rarely used	-	-	-
Sample record is a duplicate (same date and time)	Some Studies	Some Studies	-	Duplication
Sample record is a duplicate (same VISIT and TPT)	Some Studies	Some Studies	-	Duplication
Concentration or result is missing	Most Studies	Most Studies	-	Missingness
Time since last dose cannot be calculated because previous dose date/time is unknown.	Most Studies	Most Studies	-	Missingness
Sample date or time is missing or incomplete	Most Studies	Some Studies	-	Missingness
Related dosing information is missing for that visit	Rarely used	Rarely used	-	Missingness
Treatment is placebo	-	-	Most Studies	Missingness
AMT=0 and treatment is active drug	-	-	Some Studies	Missingness
Dose date or time is missing or incomplete	-	-	Some Studies	Missingness
AMT is missing	-	-	Some Studies	Missingness

Table 1. Reasons for commenting out “C” observations.

Commenting out records that contain data quality issues is the standard approach in creating PopPK/PD datasets (see section 3.4 Standard Flags and Record Identifiers in Population PK Dataset in the Basic Data Structure for ADaM PopPK Implementation Guide Version 1.0) and an additional explanatory variable can be created to record the reason for exclusion. However, the decision of omitting data from the dataset and how to handle the data quality (i.e., duplicate records) during programming requires careful consideration, as it can have a significant impact on the accuracy and integrity of the analysis.

As a PopPK/PD dataset can potentially be modified to accommodate various types of analyses, the decision on how to handle data quality issues can vary across studies, study teams, and companies. This often results in highly customized programs for PopPK/PD datasets that are not reusable for other studies, leading us further away from the automation of file creation. Additional inefficiencies occur due to time spent discussing how to tackle an issue, followed by a rigorous QC process with a lengthy code to process to remedy the problem. This paper will further focus on the preferred approach to handling duplicate records in the source data.

DUPLICATE RECORDS

Duplications can happen for many reasons i.e., accidental double data entry, mistakes in datetime recording, or assigning results to the wrong visit or timepoint. Another factor that can increase the risk of duplicate records is the involvement of multiple data sources, such as when different labs are used for

sample analysis. These multiple data sources are prone to merging mistakes or incorrect references to the result. Determining whether two or more records are real duplicates needs to be evaluated on a case-by-case basis and is highly dependent on the key variables used to evaluate it.

DUPLICATES AFFECTING THE VERTICAL POPPK/PD DATASET STRUCTURE

Within the vertical structure of a PopPK/PD dataset, duplications are identified when there are multiple dosing or sample lines or PD lines within a set of keys. The primary SDTM domains contributing to the building of the vertical structure are the EX and PC domain. If a PD analysis is performed, then LB, AE, IS, etc. domains may also be included. Duplicate records in these source datasets will be directly translated as duplicates in the PopPK/PD dataset lines.

The most common rule to determine whether certain lines are duplicates or not, is based on combinations of the following key variables:

1. STUDYID, Subject/PAT, EVID, DVID, CMT, **Date, (and Time)** as shown in **Table 2**.
2. STUDYID, Subject/PAT, EVID, DVID, CMT, **VISIT, CTPT** as shown in **Table 3**.
3. STUDYID, Subject/PAT, EVID, DVID, CMT, **Datetime, VISIT, CTPT**.

	STUDYID	PAT	EVID	CMT	DVID	DATE	TIME	VISIT	CTPT	NTPT	MDV	CONC	CONCU	PCSPEC	count_dup
1	ABC123	777	0	1	4	2021-06-08	09:31:00	CYCLE 01 DAY 1	PREDOSE	0	1	0 ug/mL	SERUM		2
2	ABC123	777	0	1	4	2021-06-08	09:31:00	CYCLE 01 DAY 1	POSTDOSE	0.001	1	0 ug/mL	SERUM		2
3	ABC123	777	0	1	4	2021-06-09	09:55:00	CYCLE 01 DAY 2	PREDOSE	0	0	110.66 ug/mL	SERUM		1
4	ABC123	777	0	1	4	2021-06-09	14:30:00	CYCLE 01 DAY 2	POSTDOSE	0.001	0	264 ug/mL	SERUM		1
5	ABC123	777	0	1	4	2021-07-06	10:08:00	CYCLE 02 DAY 1	PREDOSE	0	0	468.8 ug/mL	SERUM		1
6	ABC123	777	0	1	4	2021-07-06	14:12:00	CYCLE 02 DAY 1	POSTDOSE	0.001	0	1132 ug/mL	SERUM		1
7	ABC123	777	0	1	4	2021-11-23	09:05:00	CYCLE 07 DAY 1	PREDOSE	0	0	67.14 ug/mL	SERUM		2
8	ABC123	777	0	1	4	2021-11-23	09:05:00	CYCLE 07 DAY 1	POSTDOSE	0.001	0	75.57 ug/mL	SERUM		2
9	ABC123	777	0	1	4	2022-01-18	15:35:00	CYCLE 09 DAY 1	PREDOSE	0	0	79.74 ug/mL	SERUM		1
10	ABC123	777	0	1	4	2022-01-18	17:00:00	CYCLE 09 DAY 1	POSTDOSE	0.001	0	89.51 ug/mL	SERUM		1
11	ABC123	777	0	1	4	2022-03-15	16:10:00	CYCLE 11 DAY 1	PREDOSE	0	0	17.35 ug/mL	SERUM		1
12	ABC123	777	0	1	4	.	.	CYCLE 13 DAY 1	POSTDOSE	0.001	1	0 ug/mL	SERUM		1
13	ABC123	777	0	1	4	2022-05-10	10:00:00	CYCLE 13 DAY 1	PREDOSE	0	1	0 ug/mL	SERUM		1
14	ABC123	777	0	1	4	2022-06-03	15:00:00	END OF TREATMENT			1	0 ug/mL	SERUM		1

Table 2. Example of duplicate lines based on Date and Time.

	STUDYID	PAT	EVID	CMT	DVID	DATE	TIME	VISIT	CTPT	NTPT	MDV	CONC	CONCU	PCSPEC	count_dup
1	XYZ789	101	0	2	3	2022-04-21	17:44:00	CYCLE 01 DAY 1	POSTDOSE	0.001	0	131.89 ug/mL	SERUM		1
2	XYZ789	101	0	2	3	2022-04-21	12:38:00	CYCLE 01 DAY 1	PREDOSE	0	1	0 ug/mL	SERUM		1
3	XYZ789	101	0	2	3	2022-04-22	14:36:00	CYCLE 01 DAY 2	POSTDOSE	0.001	0	345.03 ug/mL	SERUM		1
4	XYZ789	101	0	2	3	2022-04-22	09:40:00	CYCLE 01 DAY 2	PREDOSE	0	0	91.94 ug/mL	SERUM		1
5	XYZ789	101	0	2	3	2022-05-18	13:56:00	CYCLE 02 DAY 1	POSTDOSE	0.001	0	966.4 ug/mL	SERUM		1
6	XYZ789	101	0	2	3	2022-05-18	11:04:00	CYCLE 02 DAY 1	PREDOSE	0	0	338.29 ug/mL	SERUM		1
7	XYZ789	101	0	2	3	2022-06-16	14:27:00	CYCLE 03 DAY 1	POSTDOSE	0.001	0	939.89 ug/mL	SERUM		1
8	XYZ789	101	0	2	3	2022-06-16	11:20:00	CYCLE 03 DAY 1	PREDOSE	0	0	149.83 ug/mL	SERUM		1
9	XYZ789	101	0	2	3	2022-08-11	13:56:00	CYCLE 05 DAY 1	POSTDOSE	0.001	0	747.31 ug/mL	SERUM		1
10	XYZ789	101	0	2	3	2022-08-11	10:50:00	CYCLE 05 DAY 1	PREDOSE	0	0	142.63 ug/mL	SERUM		1
11	XYZ789	101	0	2	3	2022-10-06	13:35:00	CYCLE 07 DAY 1	POSTDOSE	0.001	0	834.63 ug/mL	SERUM		2
12	XYZ789	101	0	2	3	2022-09-22	14:02:00	CYCLE 07 DAY 1	POSTDOSE	0.001	0	768.91 ug/mL	SERUM		2
13	XYZ789	101	0	2	3	2022-10-06	10:30:00	CYCLE 07 DAY 1	PREDOSE	0	0	147.43 ug/mL	SERUM		2
14	XYZ789	101	0	2	3	2022-09-22	11:00:00	CYCLE 07 DAY 1	PREDOSE	0	0	139.43 ug/mL	SERUM		2
15	XYZ789	101	0	2	3	2023-01-26	13:45:00	CYCLE 11 DAY 1	POSTDOSE	0.001	0	805.49 ug/mL	SERUM		1
16	XYZ789	101	0	2	3	2023-01-26	10:40:00	CYCLE 11 DAY 1	PREDOSE	0	0	131.2 ug/mL	SERUM		1
17	XYZ789	101	0	2	3	2023-03-23	13:55:00	CYCLE 13 DAY 1	POSTDOSE	0.001	0	724.69 ug/mL	SERUM		1
18	XYZ789	101	0	2	3	2023-03-23	10:50:00	CYCLE 13 DAY 1	PREDOSE	0	0	145.83 ug/mL	SERUM		1

Table 3. Example of duplicate lines based on VISIT and CTPT.

DUPLICATES AFFECTING THE HORIZONTAL POPPK/PD DATASET STRUCTURE

When programming the horizontal structure of PopPK/PD dataset, duplications can cause a variable to remain unpopulated. This might occur, for example, when programming the time varying variable of concomitant medication (CM) taken. **Table 4** below shows examples of concomitant medications with overlapping date ranges from the SDTM CM domain. These lines will eventually be expanded for each date so that the information can be merged into PopPK/PD dataset based on the date value. The overlapping

date ranges will cause duplicate records. Subsequently, the question arises as to whether these duplicate records are the result of a data quality issue (such as an incorrect start date/time of medication entered in the eCRF) or an expected duplicate due to the subject taking a double dose on that day.

	Study Identifier	Unique Subject Identifier	Standardized Medication Name	Dose per Administration	Dose Units	Dose Form	Dosing Frequency per Interval	Route of Administration	Start Date/Time of Medication	End Date/Time of Medication	Study Day of Start of Medication	Study Day of End of Medication
1	CM77DUP	637	EPINEPHRINE		1 mg	INJECTION	ONCE	INTRAVENOUS	2022-08-26	2022-08-26	618	618
2	CM77DUP	637	EPINEPHRINE		1 mg	INJECTION	ONCE	INTRAVENOUS	2022-08-26	2022-08-26	618	618
3	CM77DUP	637	EPINEPHRINE		4.2 mg	INJECTION	QD	INTRAVENOUS	2022-08-26	2022-08-26	618	618
4	CM77DUP	637	EPINEPHRINE		1 mg	INJECTION	ONCE	INTRAVENOUS	2022-08-26	2022-08-26	618	618
5	CM77DUP	637	PROPRANOLOL		20 mg	TABLET	BID	ORAL	2021-04-18	2021-09-14	123	272
6	CM77DUP	637	PROPRANOLOL		40 mg	TABLET	BID	ORAL	2021-05-04	2021-08-17	139	244
7	CM77DUP	637	VALPROATE SODIUM		450 mg	INJECTION	QID	INTRAVENOUS	2021-01-01	2021-01-04	16	19
8	CM77DUP	637	VALPROATE SODIUM		1800 mg	INJECTION	ONCE	INTRAVENOUS	2021-01-01	2021-01-01	16	16
9	CM77DUP	637	VALPROATE SODIUM		600 mg	TABLET	BID	ORAL	2021-01-04	2021-01-27	19	42
10	CM77DUP	637	VALPROATE SODIUM		500 mg	TABLET	BID	ORAL	2021-01-28	2021-02-01	43	47
11	CM77DUP	637	VALPROATE SODIUM		600 mg	TABLET	BID	ORAL	2021-02-02	2021-02-23	48	69
12	CM77DUP	637	VALPROATE SODIUM		450 mg	TABLET	QD	ORAL	2022-03-18	2022-08-24	457	616
13	CM77DUP	637	VALPROATE SODIUM		300 mg	TABLET	QD	ORAL	2022-03-18	2022-08-25	457	617

Table 4. Example of duplicated source concomitant medication data used to populate a time varying variable.

	Study Identifier	Unique Subject Identifier	Parameter	Parameter Code	Analysis Value (C)	count_dup
853	MUT789DUP	5932	Mutation Type	MUTTYPE	Exon 19del	1
854	MUT789DUP	5941	Mutation Type	MUTTYPE	Exon 21 L858R	1
855	MUT789DUP	5946	Mutation Type	MUTTYPE	Exon 21 L858R	1
856	MUT789DUP	5950	Mutation Type	MUTTYPE	Exon 21 L858R	2
857	MUT789DUP	5950	Mutation Type	MUTTYPE	Exon 19del	2
858	MUT789DUP	5955	Mutation Type	MUTTYPE	Exon 19del	1
859	MUT789DUP	5964	Mutation Type	MUTTYPE	Exon 21 L858R	1
860	MUT789DUP	5968	Mutation Type	MUTTYPE	Exon 19del	1
861	MUT789DUP	5973	Mutation Type	MUTTYPE	Exon 19del	1
862	MUT789DUP	5977	Mutation Type	MUTTYPE	Exon 21 L858R	1
863	MUT789DUP	5982	Mutation Type	MUTTYPE	Exon 21 L858R	1
864	MUT789DUP	5986	Mutation Type	MUTTYPE	Exon 21 L858R	1
865	MUT789DUP	5991	Mutation Type	MUTTYPE	Exon 21 L858R	1
866	MUT789DUP	5995	Mutation Type	MUTTYPE	Exon 19del	1

Table 5. Example of duplicated Mutation Type from the PF dataset.

Duplication issues can also be detected when programming subject level variables, some examples among others:

- Multiple records of immune response and titer value with the same datetime, in a subject who tests positive for antibody to investigational medication.
- Multiple records of EGFR mutation type while only one mutation type per subject is expected as shown in **Table 5** above.
- Multiple bodyweights recorded during the screening visit, and both are marked as VSBLFL = 'Y'.

HANDLING DUPLICATE DATA – THE PREFERRED APPROACH

NONMEM analysis needs to be ready in time for making further drug development decisions. As such, data quality issues, particularly duplicate values, can hamper the ability to achieve this goal. If not properly addressed, duplicates can distort the analysis by increasing data counts, and affecting other statistical measures. Therefore, it is beneficial to implement a dependable approach for handling data issues to ensure timely results.

Our preferred approach in handling duplicate data quality issues is shown in **Figure 1** below.

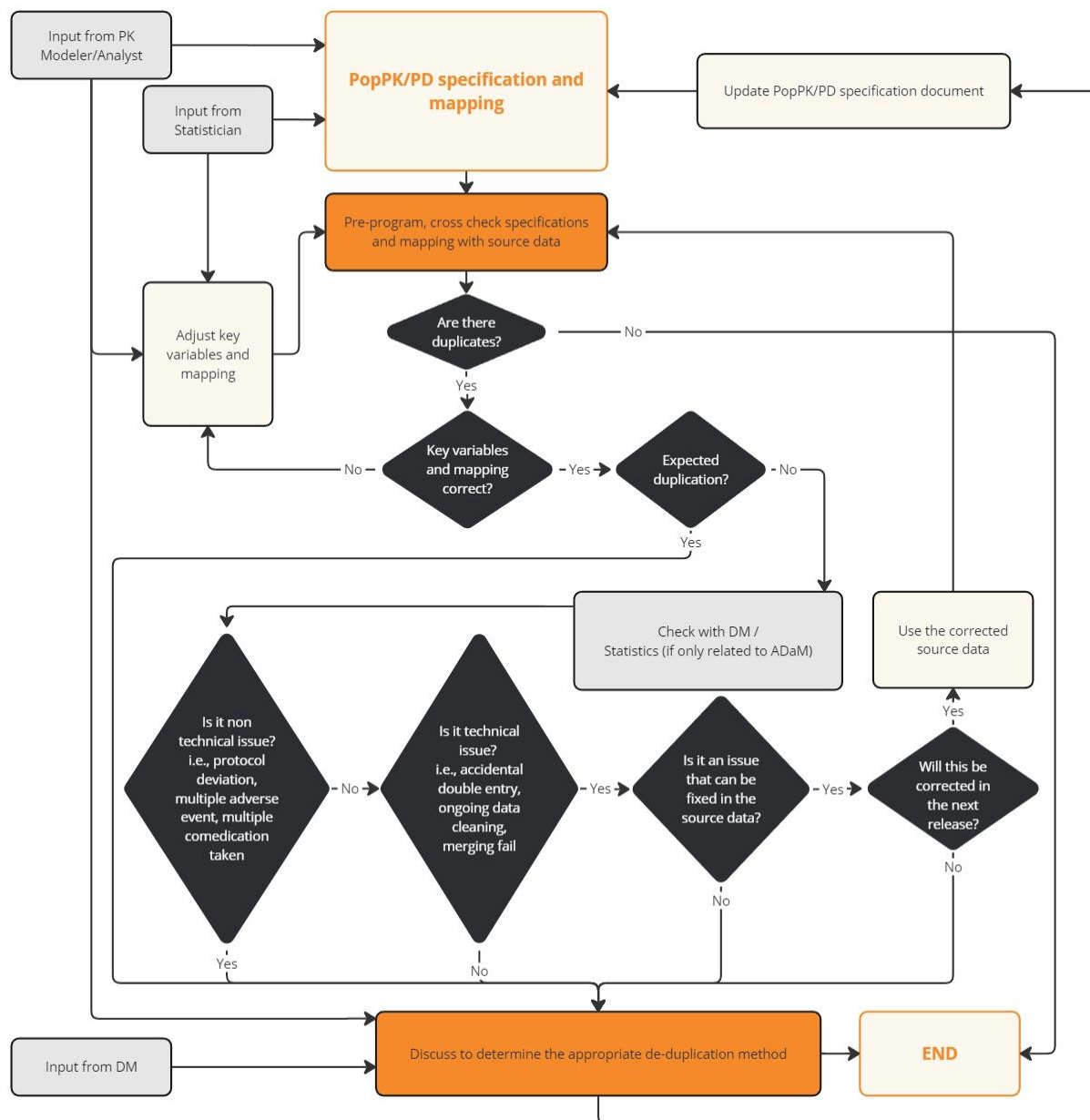


Figure 1. Preferred approach in handling PopPK/PD input data quality issue

SPECIFICATIONS OF THE POPPK/PD DATASET

A complete and clear PopPK/PD dataset specification, including accurate source dataset mapping, is an essential starting point in the programming process. This dataset specification provides PKPD programmers with clear directions on how to proceed and what information the PK modeler/analyst expects in the dataset. Additionally, a comprehensive specification assists in establishing duplication criteria, subsequently identifying these issues, and determining the appropriate variables to use as keys.

Next, the PKPD programmer can pre-program the PopPK/PD dataset and check the input SDTM and/or ADaM data to identify any potential issues, including duplications.

DETECTING DUPLICATES IN SOURCE DATA

During programming of the popPK/PD dataset, it is preferred to assess the issue at hand. Before indiscriminately removing duplicates, consider the following example of simple SAS® proc SQL code to check for duplicate data in PC dataset:

```
proc sql;
    create table _nonmem1_mark_dup as
        select *, count(*) as count_dup
        from nonmem1
        group by STUDYID, PAT, EVID, DVID, CMT, DATE, TIME;
quit;
```

The code above creates a dummy variable in a temporary dataset that gives the count of duplicated records based on STUDYID, PAT, EVID, DVID, CMT, DATE, TIME as key factors (example **Table 2** above). These records are no longer considered duplicates, when VISIT and PCTPT are included as keys. This raises questions about the accuracy of the datetime recording. In the case of **Table 3** question arise whether a backup sample was taken at that particular VISIT and PCTPT and one of the lines should be marked as an unscheduled visit Alternatively, it could be a sample that is associated with a different VISIT and PCTPT but mistakenly assigned to the wrong VISIT and PCTPT.

Marking the duplicate within a temporary dataset itself created during programming instead of outputting the duplicated line in a separate temporary dataset makes the investigation for duplication reasons easier. The option to have an oversight of all records per subject and inspect the sequence of datetime or VISIT and TPT become possible. In a thorough investigation of data duplication reasons, the first step is to verify the correctness of the mapping in the PopPK/PD dataset specifications and the appropriateness of the key variables used. This is also applicable when dealing with source datasets from different domains.

It's also important to check available information in other datasets (including SUPP-- domains) that could shed light on duplication. The PKPD programmer should consult to PK modeler or analyst and statistician to decide if changing of the mapping, additional filtering, altering the key variables will yield the correct information for their analysis.

In some cases, duplication is expected due to the study setup or protocol design. For instance, if lab samples are taken both in a central lab and a regional lab, the PK modeler/analyst and statistician must collaborate to determine which information should be utilized for consistent analysis.

COLLABORATION WITH DATA MANAGEMENT AND STATISTICS TEAM

When an unexpected duplication is found, DM should be contacted to investigate if this is caused by a data entry issue that needs to be queried or whether this duplication will not be resolved. A collaboration between DM and the PKPD programmer is in these cases beneficial for data cleaning process, particularly when DM has not yet detected the issue. Duplicate records can be challenging to identify, and in this manner, early detection becomes possible, enabling DM to address and resolve the duplication before database lock. Its preferable to handle and clean data quality issues upstream, so the corrected input data can be used and handling the issue in the PopPK/PD dataset itself is minimalized.

In addition to a collaboration with DM, one between the PKPD programmer and the STAT team, responsible for the ADaM datasets can also be beneficial. If certain data issues are already tackled in ADaM dataset, then ADaM dataset can be used as source information complimentary to the SDTM dataset. If the duplication issue persists in the input datasets, then the PKPD programmer needs to discuss and come into agreement with PK modeler/analyst on which deduplication method suits best for their analysis.

MOST COMMONLY USED DEDUPLICATION METHODS

There are many methods that can be used to handle duplicate records, among others:

1. Take the average value:
 - If the duplicate is related to the vertical structure / observation lines of the PopPK/PD dataset, create extra lines with average AMT/CONC/DV values. Delete or comment out “C” the original records.

- If it's a value for a certain covariate (horizontal structure of PopPK/PD dataset), calculate the average value (i.e., see example in **Table 4** red box)
- 2. Take the sum value:
 - If it's related to the vertical structure / observation lines of the PopPK/PD dataset, create extra lines with sum AMT/CONC/DV value. Delete or comment out "C" the original records.
 - If it's a value for a certain covariate (horizontal structure of PopPK/PD dataset), calculate the sum value (i.e., see example in **Table 4** Green and Blue Box across the overlapping dates)
- 3. Keep all lines that have non-missing values / --STAT not equal to 'NOT DONE' value. If the duplicates are NOT DONE lines, then keep the one with the lowest --SEQ.
- 4. Choose one line (or value if it's a covariate) based on datetime, either the earliest or the latest
- 5. Choose one line (or value if it's a covariate) based on --SEQ value, either the lowest or the highest.
- 6. Keep all lines in the file and add comment "C" on the duplicate lines. The method to determine which line is the duplicate refer to rule 4 or 5.
- 7. Create helping variable/covariates to identify duplicate lines.

DEDUPLICATION: EXAMPLES

It is essential to proactively communicate any identified duplication issues to the PK modeler/analyst. A mutual agreement between the PK modeler/analyst, statistician and PKPD programmer on which deduplication method chosen for each duplication case is imperative, as each deduplication method may have different impact on the NONMEM analysis.

This is also to ensure that the chosen deduplication methods do not compromise the accuracy and miss-aligned of their respective analyses. Any decision regarding how duplication is handled needs to be recorded, and the PopPK/PD dataset specification document should be updated accordingly.

Following are examples of deduplication methods applied in programming PopPK/PD vertical structure:

- The highest risk for dosing records duplications in PopPK/PD vertical structure is found in studies involving IV dosing and daily dosing.

In IV studies, most duplication issues occur when dosing is interrupted. In such cases, it would be advisable to conduct a thorough examination of the EC SDTM domain. The most common action taken in this situation is to leave the lines as they are, with two (or more) lines for each interrupted dosing line. DOSN does not increase for each part of interrupted dosing lines. Another approach is to summarize the lines by populating AMT as the sum of EXDOSE within the complete dosing administration. Start infusion datetime is populated by taking the EXSTDTC of the first part of interrupted infusions, and the end of infusion datetime is populated by taking the EXENDTC of the latest part of interrupted infusions (method 2).
- Conversely, in studies involving daily dosing, there can be cases where the EX domain contains overlapping date ranges. If the overlapping dates have the same EXDOSE, it likely indicates a mistake, and the most common approach is to remove one of the duplicate lines. However, if the overlapping dates have different EXDOSE values, further investigation is needed.
- Duplications can occur in PK/PD sample lines due to the collection of backup samples for the same subject, VISIT, and --TPT. Additionally, there may be instances of duplicate samples within the same subject, VISIT, --TPT, and DATETIME, particularly when some of the duplicated lines have a --STAT value of 'NOT DONE' as part of ongoing data cleaning. Duplicate lines with a --STAT value of 'NOT DONE' are then removed from the PopPK/PD dataset (method 3).
- In situations where there is a problem with the initial sample, the lab may opt to collect and analyze a backup sample. As a result, duplicated lines with a filled ORRES value can occur. To address this scenario, it is advisable to select the line with the highest SEQ value and "C" comment out or remove the other duplicated lines (or other option between method 4, 5, or 6). Another option is to leave the duplicated data as is and create a helping variable (as in method 7) to indicate which line is duplicated without "C" commenting out the line itself.

With regards to horizontal structure of PopPK/PD dataset, deduplication is imperative. Some variables will only be impacted by having a biased value. However, for some variables, duplication in the source data can cause variables to remain unpopulated. Following are examples of deduplication application in programming PopPK/PD horizontal structure:

- Referring to the case presented in **Table 4**, if the purpose of the time varying variable is solely to indicate whether a specific concomitant medication was taken or not, then removing duplicates would be adequate. In such cases, the conflicting overlapping information can be resolved by considering the status as "taken". However, if the dose details are crucial, such as when populating a text variable by concatenating CMDOSE, CMDOSU, CMFREQ, and CMDOSTXT for additional information, further action needs to be determined after consulting with DM, the PK modeler/analyst, and the Statistics team. Depending on the decision taken, the lines may be removed, or the overlapping information may be processed (e.g., by summing or averaging the dose amount).
- Multiple bodyweights recorded during the screening visit, both marked as VSBLFL = 'Y'. A deduplication can be performed by selecting the value with the highest VSDTC or the value with the highest VSSEQ (method 4 or 5).
- In the example **Table 5**, mutation type is a stratification factor for randomization and, also for the log-rank test analysis in a later stage. The decision on which value to select is important because it determines the stratum group to which the subject will be assigned. Deduplication is crucial in this case as it indicates a genuine data issue. If the duplication impacts many subjects, it may suggest an underlying issue. Why is it possible that such categorizing variable can have many duplicates? This can also lead to mistakes in assigning the subject to the wrong stratification group and can make the analysis severely invalid. Thus, one needs to be careful in deciding which deduplication method to take in case DM cannot provide explanation to the issue.

If there's a severe and abundance issue of duplication in both vertical and/or horizontal PopPK/PD structure, the parameter estimation from PK modelling can be compromised. The analysis may become biased if a certain duplicate data handling is applied consistently across re-occurring similar issues. In such cases, it becomes crucial to thoroughly clean up the data and carefully consider which method of duplicate handling should be implemented.

ADVANTAGES OF INCLUDING AN UNBLINDED TEAM

A key challenge in NONMEM analysis is its dependency on the final SDTM and ADaM datasets. Particularly in blinded studies, these datasets only come available after database unblinding. This can delay the delivery of the PopPK/PD dataset if data issues like duplications are only detected at this stage. Programming the PopPK/PD dataset prior to unblinding can enhance the likelihood of timely analysis delivery after database lock.

This is where including an unblinded team can induce a significant advantage on data quality and timelines. This team can create and clean the unblinded SDTM datasets prior to database unblinding, allowing for early preparation of the PopPK/PD dataset program and proactive issue detection and resolution. Additionally, dummy subject IDs can be used to create a blinded dataset, enabling the modeling team, to prepare their analysis earlier.

Upon unblinding, the PKPD programmer can rerun the program with minimal or no changes using the final source datasets. This results in a prompt delivery of the PopPK/PD dataset to the PK modeler or analyst. The unblinded result can be used for the actual analysis, which should closely align with the preliminary analysis using the dummy subject IDs.

CONCLUSION

In today's fast-paced environment, the desire for on-time completion of each milestone in clinical study, is more significant than ever. For PK/PD analysis specifically, one contributing factor that aids in achieving this objective is providing the PK modeler/analyst with a high-quality and modelling ready PopPK/PD

dataset in a timely manner. This necessitates not only technical expertise but also an efficient and consistent streamlined process.

The recently published Basic Data Structure for ADaM PopPK Implementation Guide Version 1.0 introduces a standard and guideline for creating PopPK/PD datasets. One of the chapters in this guide focuses on handling missing data, providing advice and techniques to address this issue. However, it is worth noting that there is no specific section dedicated to duplicate handling in the guide. This paper aspires to contribute to the ongoing discussion by offering insights and recommendations on the best practices for handling duplicates in the PopPK/PD dataset.

As most of the duplication issues are isolated cases, a streamlined approach on to how handle duplicates will most probably have a minimal impact on the quality of the NONMEM analysis. On the other hand, it will reduce programming time by eliminating the need for case-by-case handling. In addition, the code to create PopPK/PD datasets will become more reusable bringing us closer to an automated process. The greater the level of automation and standardization, the lower the risk of inducing human errors, thereby impacting the quality of the analysis positively. Cross-functional collaboration, establishing common understanding and mutual agreement on modeling purpose and programming requirements, plays a pivotal role in the successful execution of these efforts.

REFERENCES

- Cheng, A.M. 1998. "The Causes, Impact and Detection of Duplicate Observations" PharmaSUG 1998. Pfizer, Inc., New York, New York.
- Koukuntla, R.K. 2014. "NONMEM® – A Programmer point of view" PHUSE 2014 Paper PP12. Cytel, Hyderabad, India
- Sirichenko, S. and Kanevsky, M. 2016. "Duplicate records - it may be a good time to contact your data management team." PharmaSUG 2016 – Paper-TT18. Pinnacle 21, Plymouth Meeting, Pennsylvania.
- Radtke, D.B. and Moser, B. 2018. "How PK and PD analyses drive dataset designs," PharmaSUG 2018 - Paper AA-11. Eli Lilly and Company.
- Su J and Kang J. 2018. Challenges and Strategies in PKPD Programming. PharmaSUG 2018. Available at <https://www.pharmasug.org/proceedings/2018/AA/PharmaSUG-2018-AA13.pdf>
- Thanneer, N. 2018. "Data Standards for Population Pharmacokinetic Analysis (POPPK) and Non-Compartmental Analysis (NCA)." PharmaSUG 2018 - Paper AA-16. Princeton, NJ: Bristol-Myers Squibb Company.
- Pokhrel, N. 2020. "Population PK Data Sets – Basic Concepts and Challenges from the Perspective of a SAS Programmer." PHUSE US 2020 Paper AS06. GSK, Pennsylvania, USA
- CDISC. 2023. "Basic Data Structure for ADaM PopPK Implementation Guide Version 1.0." Available at: <https://www.cdisc.org/standards/foundational/adam/basic-data-structure-adam-poppk-implementation-guide-v1-0>

RECOMMENDED READING

- ***Population Pharmacokinetics Guidance for Industry***
U.S. Department of Health and Human Services Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
February 2022
<https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs>
<https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances>
- ***NONMEM® Tutorial Part I: Description of Commands and Options, With Simple Examples of Population Analysis:***
<https://ascpt.onlinelibrary.wiley.com/doi/10.1002/psp4.12404>

- ***Implementing NONMEM POP PK Specifications and Data Standards through the Use of a Browser-based Order Form***
Robert Fox, Huan Liu, Nidal Al-Huniti, Rosado-Carpenter, Olga Quantitative Clinical Pharmacology, AstraZeneca, Waltham, Massachusetts, USA https://www.lexjansen.com/css-us/2016/PP12_Final.pdf
- ***Population PK/PD Analysis – SAS® with R and NONMEM® Make Customization Easy***
Sharmeen Reza, Cytel Inc., Cambridge, MA PharmaSUG 2017 - Paper SP04

CONTACT INFORMATION

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

SGS Health Science

Generaal De Wittelaan 19A bus 5
2800 Mechelen – Belgium

Work Phone:

EUROPE: +32 15 27 32 45

AMERICAS: + 1 877 677 2667

Email: clinicalresearch@sgs.com

Website: <https://www.sgs.com/cro>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Any brand and product names are trademarks of their respective companies.