PharmaSUG 2016 - Paper SP04

Scrambled Data – A Population PK/PD Programming Solution

Sharmeen Reza, Cytel Inc., Cambridge, MA

ABSTRACT

Population pharmacokinetics/pharmacodynamics (pop-PK/PD) modeling and simulation is a necessity in the drug discovery process. It allows PK scientists to evaluate and present safety and potency of a drug. Also regulatory agencies require population analysis results as part of submission package. Scientists' involvement in mainstream clinical study team is essential in aligning analysis timelines with study conduct activities. In order to support analyses, pop-PK/PD programmers create NONMEM®-ready data sets at different stages of a trial. It is critical to deliver data sets to PK scientists in a timely manner enabling them to prepare models, and optimize based on updated data at each stage. Upon receiving final data, pop-PK/PD programmers produce NONMEM-ready data set in a short window after a study database lock. Due to the sensitivity of PK data, accessibility is a major difficulty that programmers face during the development phase. Since blank concentration results is not a feasible option for data set creation and in turn PK analyses, a reasonable solution is to build and test code on scrambled data at intermediate stages. At present, formal data requests need to be in place and takes several weeks to process. The idea is to have scrambled data available throughout a trial with pre-planning and required approval as necessary. Careful measure needs to be taken for scrambling PK related variables where sample collection method is not standardized and regular randomization process is not in effect. Suitable SAS® techniques are discussed in this paper with clear advantages of scrambling in research and development.

INTRODUCTION

In recent decades population pharmacokinetics/pharmacodynamics (pop-PK/PD) analysis has become an integral part of drug development. Pop-PK/PD programmers create NONMEM®-ready data sets (hereafter referred to as NONMEM data sets in short) using SAS® for PK scientists to use. NONMEM data sets vary in size consisting of more than one study, where each has a minimum of dosing and PK components/domains in it. Pop-PK/PD programmers pool these domains from various input sources such as SDTMs and ADaMs. The number of domains expands depending on how PK specs are defined by scientists, driven by the PK/PD features of the drug. For this reason the length of a pop-PK/PD program becomes exponentially larger and complexity increases compared to typical SDTM and ADaM.

Per PK scientists' needs, a NONMEM data set is produced at different stages of drug development cycle: interim, primary, final, ad hoc, etc. Input SDTMs and ADaMs typically come in 'blinded' fashion from Study Programming Team to pop-PK/PD programmers before database lock (DBL). In blinded data treatments are randomized and results are left blank; this is purposefully done by clinical data management (CDM) to maintain data integrity since information like concentration samples reveals treatment assignments. It takes considerable amount of time for scientists to develop models and generate simulation results, which are dependent on available data (Collins et al., 2011). In case of scientists requiring early un-blinded data in ad hoc scenario, maybe due to safety reporting, only restricted access is granted to pop-PK/PD programmers so that non-missing PK samples can be utilized in NONMEM data set generation.

For performing timely analyses, PK scientists expect NONMEM data set within a week after a study DBL. Meeting this timeframe becomes a challenge for pop-PK/PD programmers taking care of all subjects' data in programming after DBL. Although use of blinded data helps out programming initiation, creating programming code to incorporate all possible data scenarios for the final snapshot (data received) can become a daunting task. When 'scrambled data' are supplied instead of blank samples, both expected and unforeseen data cases can be considered in PK specs and coded ahead of time. Scrambling refers to a technique where subject information is strategically shuffled and is a proven way to address current obstacles of sharing protected data in research and development; examples are discussed in this paper to illustrate where long-term benefits are achieved.

BACKGROUND

PK analysis involves various stakeholders and contributions from diverse functions in creating NONMEM data set (Reza, 2015). It takes collaboration with various groups for gathering specs and data to meet production timelines. Major challenges encountered include delayed data release due to PK sensitivity and spec updates based on data refreshes, which makes the programming activities go through complex looping and sometimes re-work. Pop-PK/PD analyses happen at different stages of the drug development process. Program preparation and testing earlier than study DBL require dummy or scrambled input sources, whereas full validation occurs upon receiving final data.

Receiving NONMEM data set created on scrambled results allows scientists to review the data set and adjust PK specs as needed, although for preliminary model preparation interim look at the live data is required. Requests for intermediate input sources take lengthy data preparation steps and weeks to approve, this means pop-PK/PD programmers are unable to support PK scientists at desired stages in modeling and simulation.

CURRENT PROCESS AND DATA TRANSFER

The collection methods of PK, PD, biomarker and other specialized efficacy components are not the same as for general safety data. CDM has standard process for receiving, cleaning, and loading data into a clinical database from case report forms (CRFs), local and/or central labs and transferring them to different departments while securing blind of a trial (Collins et al., 2010). Unless there is early un-blinding to NONMEM programmers treatment codes are released at DBL; until then programmers use data where random treatments are assigned to patients. For PK and such specialized components, however, samples arrive from different vendors, biobanks, biorepositories; each vendor has an individual data transfer plan (DTP). Upon receipt of biological data CDM sends them to sponsor companies or other functions within the company removing sensitive information as need be. Also samples can be in crude form (example: in a simple excel file) which require cleanup and processing before transforming into SAS data sets. The timing of PK transfers is often not in sync with other database activities, thus it takes extra time for CDM to fit it in.

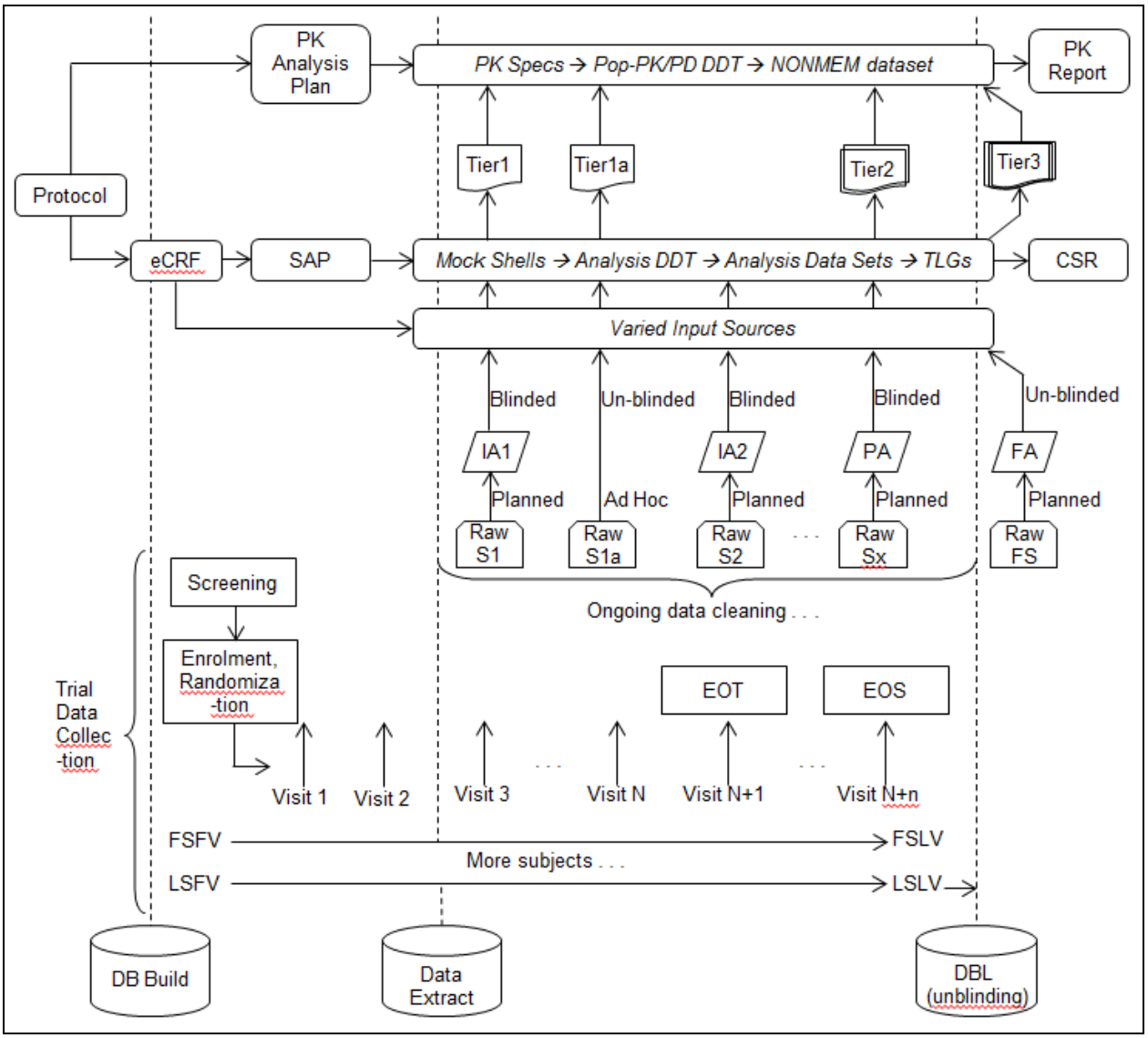

Figure 1. Current Process and Data Transfer (note: Sx is Snapshot x, acronyms are in Appendix A)

In Figure 1 dashed lines are approximate time points in data flow; it shows PK analysis plan comes much after data collection starts. Scientists' data expectations are typically not known by CDM early enough, and blinded snapshots contain blank PK results at Tier1 and Tier2. For ad hoc request at Tier1a, restricted access is needed to receive un-blinded data and is granted to certain groups. At Tier1a one group of programmers is un-blinded, whereas at Tier2 another is blinded. Since restricted access is rather group specific, ad hoc snapshot (S1a) before DBL is beyond the scope of this paper. Also, NONMEM data set may not be required for all snapshots, as shown at S2 taken for the second Interim Analysis (IA2). During data cleaning – specs, input sources, and NONMEM data set creation all go through multiple iterations involving changes displayed in italics. PK concentration arrives only after DBL and leads to more changes in specs and code at Tier3.

The planned analyses are laid out in the PK analysis plan in detail including: list of studies in NONMEM data set to be pooled, components to be included, covariates, handling of un-blinded data, modeling methods and strategies, definitions, milestones, roles and responsibilities, handling of missing data, etc. Once PK specs are defined and input data sets are ready, pop-PK/PD programming starts; back and forth discussion occurs between programmers and PK scientists during the development phase. Depending on the maturity of data programming triggers questions for scientists leading to specs update (Lu, 2015; Reza, 2015).

Pop-PK/PD programmers can initiate program preparation using blinded or dummy data to a certain extent; there is still significant work involved incorporating full-set of data after DBL which impacts delivery time to PK scientists. This is due to the fact that PK specs do not accommodate all data values or vice versa, i.e., contrary to the ideal case where no program tweaking is needed upon un-blinding. Dummy PK data does not help in defining rules in PK specs since correlation of dummy samples and dosing cannot be established, and this is not discussed further.

DESIRED PROCESS AND DATA TRANSFER

Sample collection, modeling and simulation are rarely tracked in standard clinical trial timeline chart, where every step is planned, formalized and documented (Wahid, 2015). Clinical Pharmacology representative is not part of a clinical study team historically in most organizations, since PK analysis is somewhat a recent phenomenon in drug submission (Collins et al., 2010). It is now essential for a PK scientist to be part of study team, understand processes, review study documents, be trained (overview of CDISC, for example), pre-plan and present PK analysis needs, and line up milestones/timelines with other functional areas and collaborate accordingly.

Data management itself is a time-consuming activity, requires significant resources of a drug company; specialized sample collection adds another layer of complexity to general CDM paradigm as different vendors, transfer plans, data formats are involved. At present PK data take several weeks to reach to programmers. It requires formal approval from clinical team. Processing of PK data must be streamlined, maintained as part of standard data flow and study activities to reduce cost, resource use and time (Collins et al., 2010).

The purpose is to support PK/PD modeling by producing regulatory submission/eSub compliant NONMEM data set to PK scientists in a week after DBL (Cuijpers et al., 2007). Models are tested with all study data at this time. However, scientists start model preparation way earlier than DBL requiring NONMEM data sets at strategic time-points as per the PK analysis plan. Models are constructed in a 'tiered' approach based on different individual time-points, as in snapshots illustrated in Figure 1. In each tier gathering more data, running through the models, and making sure they fit to data, enables scientists to come up with optimal design (Waterhouse, 2011). Also diverse data scenarios allow testing of possible rules provided in PK specs. All this is only possible when data transfer at any step is prompt upon PK scientists' requests.

Timely filing of analysis reports to regulatory agencies increases the chances of meeting drug's acceptance criteria (Collins et al., 2010). To facilitate this submission goal pop-PK/PD program needs to be semi-ready before DBL. This means code design, development, and testing should be nearly completed using well formatted input sources. The idea is to use standard sources like SDTMs and/or ADaMs for populating a NONMEM data set from database snapshot taken at each time-point, with additional (or cleaner) data for testing programming modules and checking their robustness. This does not fully eliminate the need for code adjustments when un-blinded data come at DBL because unforeseen cases can always arise; however, using the same input structure reduces code customization and validation time.

PK specs should not vary significantly from study to study, and must not be data-driven so that programmers can create modularized and reusable code (Lu, 2015). Standardization of PK specs at least for certain components can be based on therapeutic areas (example: Oncology) and its implementation guides can emphasize rules for individual domains and variables – as is done for SDTMs.

As data from all tiers build on top of one another and are used in testing code, NONMEM data set delivered at each tier must contain primary components (dosing and PK) with randomized treatment information and sample values. It is expected that real data will not be released until un-blinding at final DBL. Pop-PK/PD programmers and in turn scientists need to have scrambled data populated for the purpose of firming up code and PK specs. In scrambled

data – treatment assignments are kept blinded and concentration results are mixed up so that subjects are unidentifiable (Warton et al., 2014).

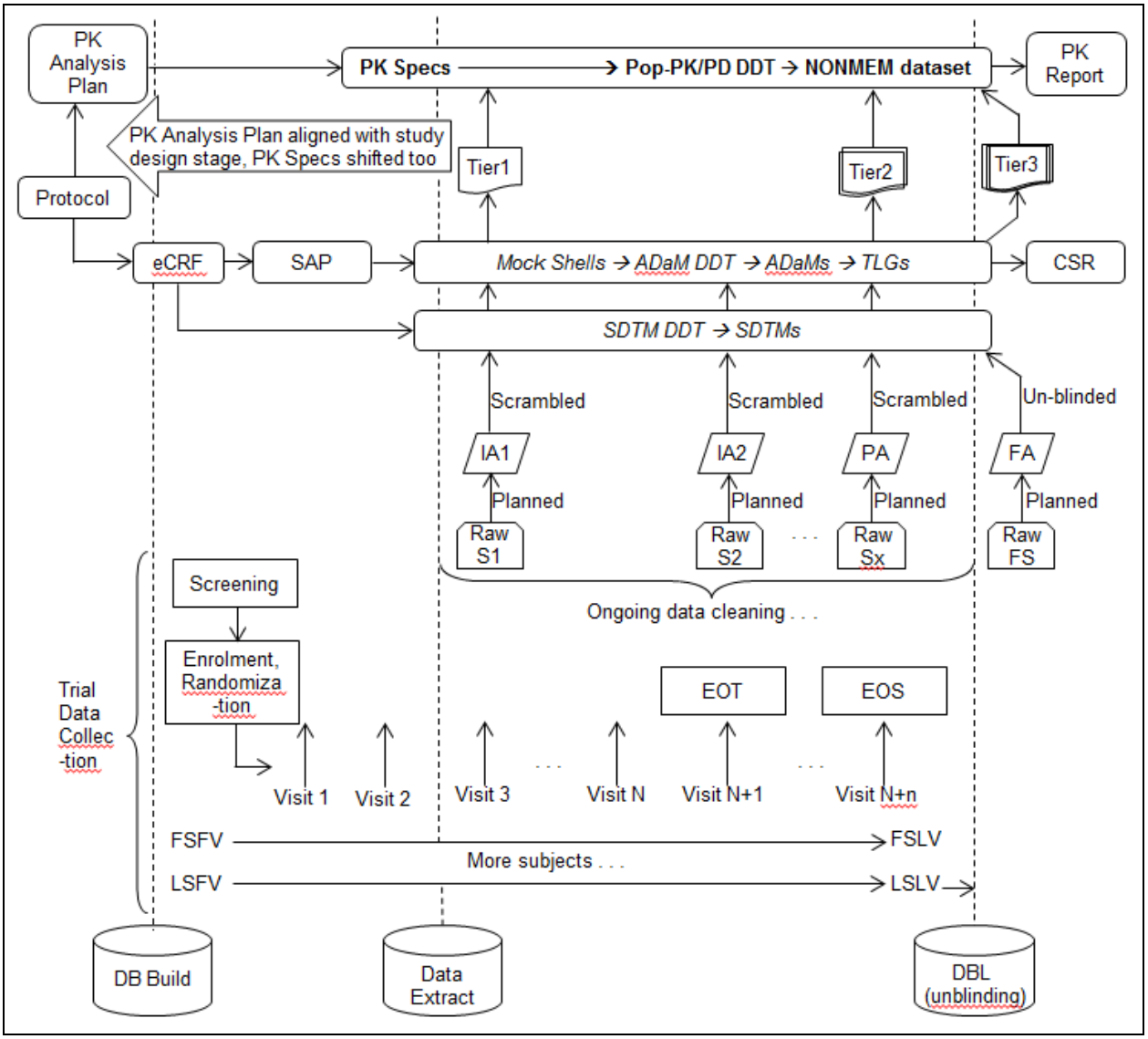

Figure 2. Desired Process and Data Transfer (note: Sx is Snapshot x, acronyms are in Appendix A)

In Figure 2 the main changes from the previous figure are: PK analysis plan is aligned to study design stage before data collection starts and PK specs follow. Also SDTMs/ADaMs are used as input sources and scrambled data are received by NONMEM programmers rather than blinded (without values). During data cleaning – specs, input sources, and NONMEM data set all go through multiple iterations like before. However, adjustments to PK related specs and data sets are handled with scrambled data at Tier1 and Tier2 before DBL. It allows code to be in semi-ready status and only rerun is needed for final data, with minimal changes at Tier3 which is after DBL.

Real data scrambled, after subjects are de-identified and/or "anonymized" (discussed in next section), is a practical consideration where scrambling algorithms are decided by qualified statisticians (Shostak, 2006; TransCelerate, 2013; Warton, 2014). It must be thought through at the study design stage when informed consent forms are written and before data collection. That way data collection itself can proceed while satisfying de-identification (Mattern, 2011). Since PK data request can arise frequently or on an ad hoc basis given business needs, shuffling logic will presumably have to be rigorous to protect sensitive information, along with standardization and automation to facilitate an efficient way to conduct clinical trial research and development (TransCelerate, 2013). Standard scrambling algorithm will lead to requiring standard PK specs in the process.

SCRAMBLING TECHNIQUES

For PK additional level of protection is needed. Scrambling logic works well for PK data sharing policies to support modeling and simulation. Sensitive information is shuffled through scrambling such that subject identification or treatment assignments cannot be determined. There are various ways this can be achieved. Two concepts: De-identification and Anonymization are currently in practice for sharing data among different research groups.

The following are US regulatory directives and guidance with options for protecting privacy of personal clinical data (Shostak, 2006; Mattern, 2011):

A. The Common Rule (45 CFR 46) defines anonymized data in which re-linking keys are destroyed, this requires Institutional Review Board (IRB) approval.
B. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (45 CFR 160 and 164) specifies 18 identifiers in Private Health Information (PHI) that need to be removed or randomized.
    1. De-identified data set (DDS)
        a) Safe Harbor method which requires removal of all identifiable health information.
        b) Statistical De-Identification (SDD) by applying statistical principles and methods.
    2. Limited data set (LDS) has many identifiers removed, but it retains some. LDS data transfer requires Data Use Agreements (DUA).

For clinical analysis Common Rule is not suitable because in anonymization destroying keys breaks relational database connectivity and attributes (Poulis et al., 2013). In choosing a de-identification technique it is critical to protect privacy. Safe Harbor is not feasible as it strips subject information that PK modeling requires, like dosing dates. LDS is not a solution either, as it is customized per project and retains some identifiers which risk exposing subject confidentiality; additionally its DUA takes months to put in place (Warton et al., 2014).

SDD can be effective by using sophisticated randomization, coding, encryption, and aggregation methods compliant with HIPAA. It should be utilized in combination with scrambling criteria designed by a qualified statistician and in agreement with reviewers: clinician, PK scientist, pop-PK/PD programmer and data management (Shostak, 2006; Mattern, 2011; TransCelerate, 2013). Thus, for de-identifying PK data the applicable concepts are: 1) Apply well-designed scrambling algorithm, and 2) SDD.

Techniques used to shuffle data should not alter the structure, attributes and formats of a data set for pop-PK/PD programmers to use; else it would defeat the purpose of expediting code development and testing before running on real data. It is usually at the discretion of a statistician to decide on the intensity of scrambling. To facilitate this there are various SAS functions and procedures, such as RANUNI, RANPERM, PROC FCMP, PROC IML. Their usage is described with example code in references given below (Baviskar, 2012; Warton, 2014).

Scrambling approach needs to keep linkable characteristics in relational database intact, retain data structure and formats in revised data sets exactly the way it is in the original. Other than dates, Private Health Information (PHI) is not collected in PK data. For de-identification, variables from PK data must be carefully chosen by experts investigating study documents, input sources and PK specs. In this effort statisticians and scientists work collaboratively and design standard algorithm taking into account PK specific rules (example: partial or missing date imputation).

While applying randomization techniques and re-ordering, one needs to avoid introducing a disproportionate number of missing values, with unrealistic patterns or ranges. Distribution of variables should be comparable between original and scrambled data by obtaining unchanged marginal but randomized conditional distribution (Warton et al., 2014). PHI should be masked before re-ordering original data. Setting random offsets is a typical way of altering dates. SAS macros, statements, formats, procedures, Output Delivery System (ODS) can be effectively utilized to shuffle. In that paper example code is given showing long-term benefits by using scrambling logic and for a NIH-funded diabetes study; the following steps are performed.

1) Variables are grouped. An offset is applied to dates, which is generated using RANUNI function. Also a random ScrambleID is created.
2) Each group is assigned a random number using RANUNI function with a seed value.
3) Maximum group number is determined, and %SCRAMBLE macro goes through that many number of iterations.
4) For testing results, descriptive statistics are used via UNIVARIATE, MEANS, FREQ procedures to validate that marginal distribution on non-offset variables are unchanged.
5) Metadata in scrambled data set should be intact and verified by generating data dictionary using PROC CONTENTS.

WHAT TO SCRAMBLE IN BLINDED RANDOMIZED TRIAL

As mentioned before, removal of identifiers in Safe Harbor is not suitable for PK analyses. Actually PHI is not quite collected in clinical trials (Shostak, 2006). Amendments can be made for "enhanced Safe Harbor" applying recoding with randomly generating identifiers (TransCelerate, 2013). For general safety data (example: AE, CM) and other textual variables various de-identification techniques are available including "own in-house dictionaries". In blinded data scenario when pop-PK/PD programmers need to work with realistic concentration results, treatments are already randomly assigned to subjects. Thus, there is no further need to de-identify safety data associated with a subject. Scrambling should be applied to raw PK data from which input SDTM (PC) and ADaM (ADPC) are populated. Manual review is needed to identify variables which are to be scrambled.

Before determining scrambling algorithm key PK variables need to be identified. Below is a list of common variables in raw PK data that have potential to reveal subject information.

Variables to Consider for Scrambling			
Assay	Cohort	Sample Type	Subject ID
Assay No	Comments	Sample Unit	Time Point
Assay Group	Planned Visit	Site	Visit Date
Below Quantification Limit (BQL)	Sample Result	Study ID	Visit Time

Table 1. Possible raw PK Variables for Scrambling

Similarly for other specialized components in NONMEM data set, like PD and biomarker, the variable lists would need to be carefully picked for scrambling.

PROS AND CONS OF SCRAMBLING DATA

Pros are mainly:

- Scrambled PK data supplied at different time-points allow pop-PK/PD programmers to develop and test code before final production, supports PK scientists to build and optimize specs/models and accelerate submission of analysis reports to regulatory agencies.
- Data transfer can be quick using standard and automated scrambling techniques.
- Scrambled data can be shared among different functional groups.
- There is no need to have controlled access to scrambled data location.
- Meeting appropriate criteria for scrambling avoids lengthy IRB approval and DUA.
- Creation and testing of pop-PK/PD data set as well as for PC and ADPC will be streamlined.
- Address un-clean data faster before DBL.

Cons can be:

- Initial setup is resource and time expensive as it requires multiple functions to work collaboratively.
- Quality of data is not as per PK scientist's expectation, in other words not realistic after scrambling.
- PK scientists' and statisticians' time and involvement to establish scrambling decisions and techniques.
- Resource needs to run scrambling code at CDM side.

CONCLUSION

Pop-PK/PD programming development occurs throughout the clinical study life cycle. Adding more data to the NONMEM data set from subjects and/or observations at different stages contributes to modeling and simulation efforts. Models are prepared well in advance so that analysis reports can be submitted to regulatory agencies on time after the DBL. This increases chances of meeting drug registration timelines and helps obtain usage-license faster, thus establishing safety, potency and future prospects. PK modeling time-points need to be aligned with study team's timelines. At any stage if input sources come blinded with blank concentration results, it is challenging for pop-PK/PD programmers to prepare code adequately for final delivery. Also it does not serve the purpose of supporting model development in stages. Instead of missing data points, real data scrambled is an option for programmers to work with before final data set production.

Scrambling algorithm needs to be robust applying a combination of techniques (scrambling itself and SDD) suitable for sensitive PK data. There are various SAS functions and procedures available to aid such implementation. Standardization and automation are suggested for both PK specs and scrambling algorithm which bring efficiency in making a NONMEM data set that is submission ready (TransCelerate, 2013). Appropriate data sharing approvals are still applicable for compliance and additional protection. Advantages of scrambling outweigh the difficulties if cross-functional teams establish initial setup collaboratively, ultimately helping a company reach its goal.

REFERENCES

- Collins, A., Peterson, M., Silva, G., 2010. Streamlining the PK/PD data transfer process. *Pharmaceutical Programming*, 3(1):24-28.
- Collins, A., Silva, G., 2011. Streamlining the PK/PD data transfer process — 1 year later. *Pharmaceutical Programming*, 4(1-2):28-30.
- Cuijpers, A., Wuyts, H., Van de Vliet, I., 2007. Pharmacokinetic data file(s) creation process. PhUSE, Paper PO08.
- Lu, G., 2015. Programming strategically for PK/PD data. PhUSE, Paper SP01.
- Baviskar, J., 2012. Scrambling of Un-Blinded Data without 'Scrambling Data Integrity'! PharmaSUG, Paper PO16.
- Warton, E., Moffet, H., Karter, A., 2014. Breaking Eggs to Make Omelets: Distributing Analytic Effort with Scrambled Datasets. WUSS (http://www.wuss.org/proceedings14/64_Final_Paper_PDF.pdf)
- Shostak, J., 2006. De-Identification of Clinical Trials Data Demystified. PharmaSUG, Paper PR02.
- TransCelerate, 2013. Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach. TransCelerate BioPharma Inc.
- Mattern, J., 2011. Degrees of De-identification of Clinical Research Data. *Journal of Clinical Research Best Practices*, 7 (11).
- Waterhouse, T., 2011. Experiences in Optimal Design for Population PK/PD Models. http://www.maths.qmul.ac.uk/~bb/PODE/PODE2011_Slides/TimWaterhouse.pdf
- Wahid, R., 2015.Tools for Conduct of Early Phase Clinical Trials. http://www.who.int/phi/DAY1_09_Wahid2_PM_SaoPaulo2015.pdf
- Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S., 2013. Anonymizing Data with Relational and Transaction Attributes. Volume 8190 of the series Lecture Notes in Computer Science pp 353-369, Springer.
- Reza, S., 2015. Growing Needs in Drug Industry for NONMEM Programmers Using SAS®. PharmaSUG, Paper SP07.

ACKNOWLEDGMENTS

The author would like to acknowledge Cytel Inc. for providing the opportunity to work on this paper. The author thanks the following colleagues/managers from present and previous workplaces for thoughtful comments.

Arsenault, Patti (Director, Global Head of Clinical Data Management, Cytel)
Baker, Jim (Senior Vice President of Clinical Research Services, Cytel)
Eschenberg, Michael (Director Biostatistics, Amgen)
Hendricks, Lisa (Global Head CP and Methodology Biostatistics, Novartis)

CONTACT INFORMATION

Sharmeen Reza
Cytel Inc.
Work Phone: 269-743-7221
E-mail: Sharmeen.Reza@cytel.com

TRADEMARKS

APPENDIX A – ACRONYMS

ADaM	Analysis Data Model	IA	Interim Analysis
CRF	Case Report Form	LSFV	Last Subject First Visit
CSR	Clinical Study Report	LSLV	Last Subject Last Visit
DBL	Database Lock	NONMEM	Nonlinear Mixed Effects Modeling
DDT	Data Definition Table	PD	Pharmacodynamics
EOS	End of Study	PK	Pharmacokinetics
EOT	End of Treatment	SAP	Statistical Analysis Plan
FA	Final Analysis	SDTM	Study Data Tabulation Model
FS	Final Snapshot	Sx	Snapshot x
FSFV	First Subject First Visit	TLG	Table, Listing, Graph
FSLV:	First Subject Last Visit		