PharmaSUG 2026 - Paper PK-282

From Chaos to Clarity: A Programmer's Perspective on Standardizing Population Pharmacokinetic (PopPK) Data for Regulatory Success

Naveen Muppalla, Exelixis Inc, Senior Statistical Programming Manager
Shibani Harite, Exelixis Inc, Sr. Director, Statistical Programming

ABSTRACT

Population Pharmacokinetics (PopPK) modeling typically follows a multi-stage process encompassing base, intermediate, and final models. This paper emphasizes the critical role of statistical programmer in streamlining the PopPK data workflows and preparing data sets and documents for regulatory submissions. Integrating Phase I dose-escalation studies with Phase III pivotal trials require robust data harmonization and adherence to regulatory standards. We explored two strategies for New Drug Application (NDA) submission. The legacy approach included creation of NONMEM ready PopPK csv files, define.pdf and a reviewer's guide. The CDISC approach which included creation of adppk.xpt data set leveraging the ADaM PopPK implementation guide (Version 1.0) with machine-readable define.xml metadata. This paper compares these two approaches from a statistical programmer's perspective by highlighting key challenges such as handling complex dosing schedules, imputing missing/partial exposure data, reconstructing Pharmacokinetic (PK) timepoints, and deriving interval-based concomitant medication covariates. Source data included SDTM domains for ongoing studies and ADaM data sets from completed studies. This approach minimized dependencies and allowed early model development. This paper also outlines a blueprint for the Analysis Data Reviewer's Guide (adrg) and eCTD Module 5.3.3.5 structure ensuring data and model traceability. By mastering these two approaches, statistical programmers can streamline the regulatory submission workflow and align with FDA technical conformance expectations thus reducing the submission review time.

INTRODUCTION

PopPK analyses are critical for any regulatory submissions. These analyses support dose selection and optimization, evaluate special population such as renal/hepatic impairment and pediatrics and support labeling. The objective of the statistical programmer is to build a reproducible, traceable, and automated pipeline that transforms raw SDTM data into a model-ready format.

BACKGROUND

The data collection in oncology trials can be complex. While infusion dosing typically follows cycle-based dosing schedules, oral medications with daily dosing are collected at intervals aligned with scheduled visit dates. The time dependent variables such as relative times and actual times used for PopPK analyses often include complex algorithms. Any partial and missing data can lead to missing data imputation further adding to the complexity. All efforts need to be made to collect complete data to allow meaningful correlation between covariates, exposure and PK concentration.

Population pharmacokinetic analyses face several data challenges. Patients often use concomitant medications that may affect the PK profile and require careful covariate identification. High patient attrition due to disease progression or toxicity can result in missing data. In addition, PopPK analyses frequently require pooling data from early First-in-Human studies to later Phase III global trials.

DATA COLLECTION AND HARMONIZATION

Early involvement of statistical programmers during the CRF design phase is essential to ensure accurate integration of pharmacokinetic, dosing, and bioanalytical data. Pharmacokinetic samples are collected at

protocol specified time points and their actual collection date and time are captured in EDC. The bioanalytical laboratory provides concentration results, including LLOQ and BLQ qualifiers as an external data set. Dose administration records, including dose interruptions and holds are captured in the EDC. Proper synchronization of these data sources is critical for the integrity of downstream population PK analyses.

HANDLING INCONSISTENCIES AND COMPLEX IMPUTATION

Simple and complex imputation methods have been applied to address missing values in baseline data and missing dosing times. All imputed values are clearly labeled to ensure transparency and to allow the pharmacometrician to perform subsequent analyses.

COMPLEX VARIABLE DERIVATIONS

Some of the critical domains used in the variable derivation for PopPK data set are described below.

Exposure (EX) timing variables: These are complex because they rely on correct sequencing across multiple records and must follow study-specific rules and predefined guardrails, such as before-and-after adjustments, overlap resolution, and record removal. Oral dosing is often collected as dosing episodes rather than as individual daily records. When exposure records are partially missing, expanding these episodes to align with PK sampling dates adds further complexity.

PK/event datetime derivations: These derivations require accurate alignment between PK and exposure (EX) records using planned timepoint logic. They are complex because they involve datetime imputation, timepoint-based backward and forward calculations, removal of duplicate records, and propagation of collection times into exposure records when dosing times are missing. Dose reference date-time values are commonly derived relative to specimen collection times, such as Predose or post dose measurements.

Concomitant medication (CM) effect covariates: These depend on the interventional drug and are based on medication lists provided by the pharmacometrician. Their derivation is complex because it requires filtering across multiple domains, partial date imputation, external drug mapping, consolidation of overlapping medication periods, handling of ongoing treatments, and time-dependent merging with PK and exposure event-time data sets. Time-varying binary flags are then derived based on these date-time relationships. Additional complexity arises when defining consolidated medication start and end intervals. In some cases, the pharmacometrician may recommend extending the medication end date by one day to account for a potential carryover effect.

DATA CLEANING CHALLENGES AND SOLUTIONS

Pooling amplifies data quality issues because inconsistencies become cross-study artifacts. Table 1 summarizes high-impact issues and robust mitigation strategies.

Table 1. Common PopPK data issues and recommended resolution patterns (from define/complex algorithms)

Issue	Impact	Resolution pattern
Partial/missing dosing dates (EXSTDTC/EXENDTC)	Invalid dose intervals; non-chronological sequences	Milestone-anchored imputation using prior/next dosing milestones; apply chronological and logical consistency rules
Missing dosing times	AFRLT/APRLT missing; misaligned sample times	Infer time from PK timepoints; carry-forward last valid time when justified
Dose overlap/back-dating after imputation	Negative or zero time-since-dose artifacts	Adjust EXSTDTC to prior EXENDTC+1 day; drop artificial intervals when warranted
Symbolic lab qualifiers (> / <)	Covariate bias in	Convert to numeric via small scaling

	renal/hepatic groups	(1.01/0.99) then apply unit conversions
Conmed overlaps and ongoing therapy	Incorrect time-varying covariate flags	Consolidate overlapping intervals; treat missing end as ongoing; extend end by 1 day
Study-specific BLQ text	Inconsistent DV handling	Standardize DV derivation rules and BLQFN across studies

SUBMISSION APPROACHES: LEGACY VS. CDISC

APPROACH 1: THE LEGACY METHOD

In the legacy submission approach, initial version of data set is sent to pharmacometrician to identify exclusion records. These are then identified as flags and excluded from the final data set.

Missing numeric values is presented as a period (.) in the csv file to allow NONMEM to identify missing records. Supporting documentation also included, such as a define.pdf and a comprehensive Analysis Data Reviewer's Guide (ADRG), with an optional complex algorithms document.

This approach offers direct compatibility with NONMEM, requires no data conversion, and avoids additional adjustments. As a result, regulators can easily run the models and verify the results.

APPROACH 2: THE CDISC ADPPK STANDARD

Following the ADPPK Implementation Guide (IG), this approach uses the ADSL.xpt data set for subject-level information and the ADPPK.xpt data set based on the ADaM Basic Data Structure. The associated metadata is defined using a machine-readable define.xml file. An intermediate NONMEM-ready CSV data set is first created to allow the pharmacometrician to begin model development. At the same time, the final ADPPK.xpt data set is prepared and finalized for inclusion in the eCTD submission. Missing numeric values remain missing (blank) in XPT and the missing will need to be treated as 0 to identify the missing values.

This approach supports compliance with the CDISC ADaM ADPPK standard, enables the use of conformance checks, and allows reviewers to easily review the metadata.

Table 2 Legacy Vs CDSIC Standard variables

Dimension	Legacy (NONMEM CSV)	CDISC ADPPK (XPT)
Subject ID	ID	USUBJIDN
Record exclusion	C (custom)	EXCLF (Standard)
Subject with Evaluable PK Flag	PKFLAG	PKFL
Dose planned	DOSENOM (custom)	DOSEP
Dose date/time Imputed flag	DOSIMPFL	FLGREAS
PK date/time Imputed flag	PKIMPFL	FLGREAS
Analysis Date time of event/observation	DTC (character)	ADT, ATM
Covariates	Custom names (BILBL, etc.)	Standardized: TBILBL
Sex, Race	N/A	SEX, RACE
Date and Time of the Event	N/A	UDTC- U stands for domain name of SDTM, e.g. PC-> PCDTC
Dose Duration	DUR(Custom)	DOSEDUR

Traceability	Manual documentation	UDTC, USTRESC

END-TO-END DATA FLOW IN BOTH SUBMISSION STRATEGIES

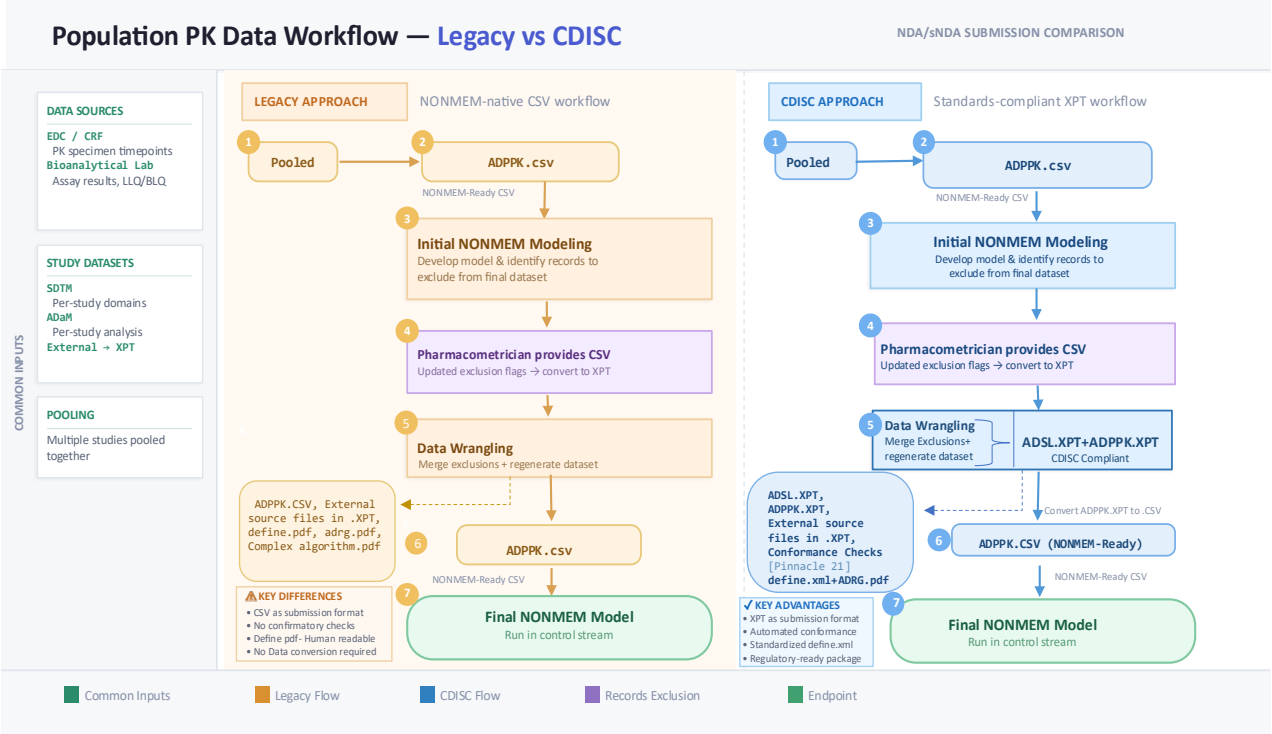

Figure 1. End-to-End PopPK Data Flow of Legacy and CDISC Approach: From Collection to NDA-Ready Submission

DATETIME IMPUTATION AND TRACEABILITY

EXPOSURE DATES IMPUTATION (EXSTDTC/EXENDTC)

Exposure date imputation is anchored to within-subject dosing milestones. Start dates are constrained to occur after the previous dosing milestone (+1 day). End dates are constrained to occur before the next dosing milestone (-1 day), with safeguards for cutoff/death dates.

EVENT DATETIME RECONSTRUCTION FROM PK TIMEPOINTS

When time is missing, PK planned timepoints can be used to infer dosing times (e.g., PRE = sample time + 30 minutes; X = sample time − X hours) and to populate missing time components in dosing records via date-matching logic across expanded dosing days.

Table 3. Date and Time of the Event (DTC) Algorithm

Applies when	DTC algorithm (in order)
DVIDN=0 (Dosing Records)	If PCRFTDTC time missing: PRE => PCDTC + 00:30; other PCTPT present => PCDTC - X (X from PCTPT; e.g., 2H => -2h); sort by USUBJID, PCRFTDTC, PCTPT and keep first PCRFTDTC; expand EX days (EXSTDTC to EXENDTC) and merge PCRFTDTC; when EX date matches PC date and EX time missing => assign PC

	time; remaining missing time => carry forward last non-missing time; keep records present in SDTM.EX.
DVIDN=1 (PC Observations)	If PCDTC time is missing and PCTPT present => impute from dose time for that day: PRE => -00:30; X hours => +X; if PCTPT contains 'UNSC' => set to PCDTC; else leave missing. (X = numeric hours, e.g., 1, 2, 3.)

Abbreviations: PCRFTDTC = Reference dosing datetime; PCDTC = PK sample collection time; PRE = Predose; PCTPT = Sample collection time point; PC = Pharmacokinetic sample collection; UNSC = unscheduled.

TIME-VARYING CONCOMITANT MEDICATIONS INTERVALS

Concomitant medication dates are partially imputed only when year is present; intervals are consolidated within medication category and mapped to each event date-time to derive time-varying flags. End dates are extended by one day to reflect residual PK effects.

Table 4. Algorithm to derive time-varying concomitant medication numeric flag variable

Variable	Label	Algorithm (in order)
CMINH1	Concomitant CYP3A inhibitor (strong)	**Partial date imputation:** Follow the study specific imputation rules **Identify category:** Use external list provided by Pharmacometrician and perform keyword search on CMTRT and/or CMDECOD. **Build intervals:** Per subject, create unique medication intervals from start/end; if overlaps within category, then collapse to 1 interval (start=min, end=max); if any medication is ongoing (missing end or end missing relative to CMENRTPT) consider interval ongoing. **Assign flag:** Merge intervals to target records (e.g., PK/EX) by subject; for each record at collection DTC: on medication = 1; else 0. **Optional:** After date imputation and interval construction, extend CM end date by +1 day (PK impact may persist up to 1 day beyond reported end date). The requirement depends on the analyte of interest determined by Pharmacometrician.

Inputs / abbreviation's: CMTRT =Concomitant treatment reported verbatim; CMDECOD=WHO drug preferred term; CMENRTPT=CM end relative time point.

NDA SUBMISSION-READY PREPARATION (FORMATS, METADATA, AND PACKAGE)

SUBMISSION PACKAGE INVENTORY

Table 5. Study data package inventory of PopPK data sets (submission-ready)-Legacy approach.

Artifact	Format	Purpose / reviewer value
Pooled PopPK data sets (ADPPK,)	CSV	NONMEM-ready integrated data sets for base and final PopPK models
define.pdf	PDF	Human-readable metadata (reviewer convenience)
Reviewer's guides (ADRG / integrated guides)	PDF	Context, derivations, data issues, traceability, navigation
Complex algorithm	PDF	Complex algorithms description

Programs, macros, scripts	TXT/ASCII	Reproducibility: creation programs, run order
Analysis Output Programs	MOD/LST	Relevant control streams and analysis outputs

Table 6. Study data package inventory of PopPK data sets (submission-ready)-CDISC approach.

Artifact	Format	Purpose / reviewer value
SDTM data sets (study-level)	XPT	Standardized tabulation data for each study (e.g., PC, EX, CM, LB)
ADaM data sets (study-level)	XPT	Analysis-ready data sets supporting analyses and pooled derivations
ADSL, ADPPK	XPT	CDSIC compliant data sets
define.xml	XML	Machine-readable metadata for data sets/variables
Reviewer's guides (ADRG / integrated guides)	PDF	Context, derivations, data issues, traceability, navigation
Annotated CRF (aCRF)	PDF	Traceability from collection to SDTM mappings
Programs, macros, scripts	TXT/ASCII	Reproducibility: creation programs, run order
Analysis Output Programs	MOD/LST	Relevant control streams and analysis outputs

eCTD MODULE STRUCTURE: WHERE POPPK FITS

Proper placement in the eCTD ensures the Clinical Pharmacology reviewer can find your data sets without delays.

eCTD Module 5.3.3.5 Population PK Study Reports	
Legacy placement m5/5335-population-pk-study-reports/data sets/ • adppk.csv (NONMEM-ready) • define.pdf • adrg.pdf • (optional) complex_algorithms.pdf • (optional) model control streams: .mod /.lst	**CDISC ADPPK placement** m5/5335-population-pk-study-reports/analysis/adam/ • **data sets/** adsl.xpt, adppk.xpt • **metadata/** define.xml • (optional) define.pdf (rendition) • adrg.pdf • (optional) interim NONMEM-ready .csv for early modeling
Reviewer navigation (recommended order) 1) adrg.pdf → 2) define.xml / define.pdf → 3) data sets (CSV or XPT) → 4) optional programs & model files for reproducibility	

KEY RECOMMENDATIONS

- **Standardize early across studies** (units, routes, visit/cycle/day conventions, PK timepoint naming) before pooling starts.

- **Treat date/time as critical data**: Use imputation for incomplete/partial data when deemed necessary, retain originals, and flag derived vs observed values for traceability.
- **Create a chronological dosing**: Expand dosing episodes when needed, resolve holds/overlaps consistently, and enforce chronology to avoid invalid time-since-dose results.
- **Build covariates as time intervals** (especially concomitant meds): consolidate overlaps, handle ongoing therapy, and document carryover assumptions (e.g., end date + 1 day).
- **Use a hybrid workflow when timelines are tight**: Produce a NONMEM-ready CSV for early modeling while finalizing ADPPK XPT + define.xml.
- **Optimize the submission for reviewer navigation**: clear data set inventory, correct eCTD placement (Module 5.3.3.5), and an ADRG that explains key derivations/decisions.
- **Automate and validate**: Reproducible run order, log checks, and conformance check to reduce late rework and reviewer questions.

CONCLUSION

PopPK results are only as trustworthy as the harmonized data supporting the analyses. The statistical programmer's role is to ensure a consistent and reliable approach is followed for harmonization in accordance with the data standards guidelines. By harmonizing data across studies, fixing tricky dosing and PK timing gaps, using careful datetime imputations, and building clear, traceable time-varying covariates (like concomitant medication intervals), we turn messy clinical data into model ready data sets that a pharmacometrician and a regulator can easily follow. Although the legacy approach is still well received and accepted by the agency. It is advisable to follow the standard approach to take advantage of the conformance checks which makes it easier not only for the programmer but also the reviewer to identify data inconsistencies and limitations.

REFERENCES

https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/model-data-format
Population PK - Multiple Subjects, Multiple Studies | CDISC
[1] FDA. Study Data Technical Conformance Guide (Technical Specifications Document).
[2] FDA. Study Data Standards Resources (Data Standards Catalog and related guidance links).
[3] https://www.cdisc.org/standards/foundational/adam/basic-data-structure-adam-poppk-implementation-guide-v1-0

ACKNOWLEDGMENTS

We thank Mun Sang Yue (Clinical Pharmacology & Pharmacometrics Director Clinical Pharmacology & Pharmacometrics, Exelixis Inc, Alameda, CA.) for valuable insights and feedback that supported the preparation of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:
Naveen Muppalla
Email: nmuppalla@exelixis.com
Exelixis Inc, Alameda, CA, USA.

Shibani Harite
Email: sharite@exelixis.com
Exelixis Inc, Alameda, CA, USA.