Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PKPy2: A Python framework for joint population pharmacokinetic and pharmacodynamic estimation, diagnostics, and uncertainty assessment

Hyunseung Kong
Seoul National University
Inyoung Kim
inyoungkim@korea.kr
Korea National Defense University

Research Article

Keywords: Population pharmacokinetics, Pharmacokinetic-pharmacodynamic modeling, Nonlinear mixed-effects models, Marginal likelihood, Software validation, Python

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11218379/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PKPy2: A Python framework for joint population pharmacokinetic and pharmacodynamic estimation, diagnostics, and uncertainty assessment

Hyunseung Kong[1], Inyoung Kim[2]

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[2] Department of Defense Science, Korea National Defense University, Nonsan, Republic of Korea

Corresponding Author: Inyoung Kim[2]
Department of Defense Science, Korea National Defense University, Nonsan, Republic of Korea
Email address: inyoungkim@korea.kr
Telephone: +82 41 831 5383
ORCID: Hyunseung Kong, https://orcid.org/0000-0001-7681-7429; Inyoung Kim, https://orcid.org/0000-0003-3062-6576

Abstract

Population pharmacokinetic (PopPK) analysis usually relies on specialized, often commercial, software. PKPy introduced a Python workflow, but it derived population parameters from separately fitted individual models, so individual estimation error entered the estimated interindividual variability. We developed PKPy2, a standalone Python package that estimates structural parameters, interindividual variability, residual error, and covariate effects jointly by marginal likelihood, with optional fixed values and bounds, numerical convergence checks, and confidence intervals. An event-record interface extends estimation to NONMEM-format dosing records, nonlinear and pharmacokinetic-pharmacodynamic (PK/PD) models, correlated and interoccasion random effects, time-varying covariates, and censored observations, with standard diagnostics and uncertainty tools. In 200 simulated one-compartment intravenous datasets, all PKPy2 fits converged, and under sparse sampling the relative root mean squared error of the interindividual variance of volume was 32.49%, compared with 51.66% for PKPy. Recovery was similar to that of nlmixr2 and saemix on the same datasets, and coverage of nominal 95% intervals ranged from 90% to 100%. Predictions, likelihoods, and diagnostics of the event-record interface agreed with rxode2, independent quadrature, and nlmixr2. In clinical theophylline and warfarin data, PKPy2 estimates differed from published expert NONMEM estimates by at most 3.37% and 4.56%. In tobramycin data, the published model had a different maximum-likelihood solution; when the expert's pharmacological judgment was entered as fixed values and bounds, PKPy2 reproduced the expert estimates within 6.0%. PKPy2 provides joint PopPK and PK/PD estimation, diagnostics, and uncertainty assessment in Python.

Keywords Population pharmacokinetics · Pharmacokinetic-pharmacodynamic modeling · Nonlinear mixed-effects models · Marginal likelihood · Software validation · Python

Introduction

Population pharmacokinetic (PopPK) models quantify typical drug disposition, interindividual variability, and covariate effects relevant to dose selection [1]. These analyses are commonly performed with specialized nonlinear mixed-effects software. NONMEM uses NM-TRAN control streams and a commercial license [2,3]. Open-source R packages such as saemix and nlmixr provide alternative estimation environments [4,5]. Each platform has its own model syntax, data conventions, and diagnostic workflow. A native Python implementation can connect PopPK estimation directly with widely used data-processing and scientific-computing tools.

PKPy was developed to provide a Python-based workflow for pharmacokinetic model construction, parameter estimation, covariate analysis, and diagnostic evaluation [6]. Its published implementation supported one- and two-compartment models with intravenous administration or first-order absorption. Population estimates were obtained by fitting each subject separately and summarizing individual estimates with geometric means and the covariance of log-transformed parameters. This two-stage workflow is straightforward. The dispersion of individual estimates includes biological variability and uncertainty from individual fits. The effect becomes pronounced under sparse sampling, when individual parameters are weakly informed even though the pooled dataset contains population-level information.

Nonlinear mixed-effects models address this issue by representing individual parameters as deviations from population values and integrating over latent individual effects. Likelihood approximations, stochastic approximation expectation-maximization, and importance sampling are established approaches for estimating these models [2,4,7,8]. A practical implementation also requires robust concentration prediction across dose histories, accurate integration of subject-level likelihoods, clear convergence criteria, and parameter uncertainty calculations.

We developed PKPy2 to extend PKPy from separate individual fits to joint population inference under explicitly specified models. Users declare the structural model, fixed and estimated parameters, interindividual variability, residual error, and covariate effects in Python code. PKPy2 performs marginal-likelihood estimation, provides analytical concentration predictions for single and repeated dosing, applies independent numerical checks before reporting convergence, and calculates local confidence intervals. The package implements established nonlinear mixed-effects methods within a standalone Python environment. Two interfaces share this estimator. A compact interface declares the closed-form one- and two-compartment models of the primary evaluation, and an event-record interface reads NONMEM-format analysis data, as used in routine population analyses, and extends the model set to multi-compartment, nonlinear, and pharmacokinetic-pharmacodynamic (PK/PD) models with correlated and interoccasion random effects, time-varying covariates, and censored observations. Both provide

model diagnostics with plots and tools for uncertainty assessment and covariate selection. Figure 1 summarizes the transition from individual fits to joint population inference.

We evaluated the numerical calculations against independent implementations and examined parameter recovery and interval coverage in repeated simulations under rich and sparse sampling. The original PKPy fitting components and a Gaussian-likelihood two-stage estimator served as development comparators, and the established open-source nonlinear mixed-effects programs nlmixr2 and saemix were applied to the same simulated datasets. Smaller simulation experiments assessed three additional compartmental structures. We validated PKPy2 with real clinical data by applying it to the public theophylline, warfarin, and tobramycin datasets and comparing the estimates with published expert NONMEM analyses and with nlmixr2 and saemix fits of the same models. The event-record interface was verified against rxode2, an independent quadrature implementation, nlmixr2, and the npde package, and its estimation was evaluated in simulated datasets and a warfarin PK/PD analysis. All evaluations used prespecified models.

Methods

Population model and software scope

For subject i, each positive pharmacokinetic parameter p_{ik} was represented as

$$\log p_{ik} = \log \theta_k + X_{ik}\beta_k + \eta_{ik},$$

where θ_k is the typical population value at the covariate reference, $X_{ik}\beta_k$ is the covariate contribution, and η_i follows a multivariate normal distribution with mean zero and covariance matrix Ω. Ω is diagonal in the compact interface used for the primary evaluation; the event-record interface described below also allows correlated blocks. Conditional on the individual random effects, the observed concentration y_{ij} was modeled as

$$y_{ij} \mid \eta_i \sim N(f_{ij}, \sigma prop^2\, f_{ij}^2 + \sigma add^2),$$

where f_{ij} is the model-predicted concentration. Structural parameters, interindividual variance components, residual standard deviations, and covariate coefficients could each be fixed or estimated. Fixed components were excluded from the optimization coordinates and the count of free parameters, and estimated components could be restricted by lower and upper bounds on the reporting scale, as with parameter bounds in NONMEM. Power covariate relationships were expressed relative to a declared reference value. This parameterization follows standard nonlinear mixed-effects PopPK formulations [1,2,4,5,7].

The population likelihood was obtained by integrating each subject-specific likelihood over the random effects:

$$L = \prod_i \int \{\prod_j p(y_{ij} \mid \eta_i)\}\, p(\eta_i)\, d\eta_i,$$

The objective function value was defined as OFV = -2 log L [2,4,7].

Normal-density constants were retained. PKPy uses a squared log-concentration residual objective. PKPy2 uses the integrated population likelihood. OFV values from the two packages were analyzed within their own objective definitions. Oral models assumed unit bioavailability; clearance and volume were interpreted as apparent parameters, CL/F and V/F. The compact interface comprises one- and two-compartment structures with intravenous bolus or first-order

oral input, deterministic absorption lag, explicit dose histories, and diagonal random effects. Dose histories followed standard event-record conventions [9].

Joint estimation and numerical assessment

PKPy2 used a staged estimation procedure. A deterministic Laplace approximation of the marginal likelihood was first minimized from the starting values and from quasi-random perturbations of the structural parameters; the best Laplace optimum served only as the starting point for the subsequent stages, and distinct optima were recorded as a multimodality diagnostic. Initial marginal-likelihood refinement followed, with stochastic approximation expectation-maximization when required and subsequent importance-sampling refinement. Estimates that satisfied the independent numerical criteria after initial refinement proceeded directly to final reporting. Other fits entered the SAEM stage, which updated the distribution of individual effects and eligible population parameters before marginal-likelihood refinement resumed. Conditional regression updates were used for eligible log-parameter mean structures, and diagonal variance components were updated from conditional moments. Other coordinates, including residual-error parameters, remained in numerical optimization. These components follow established likelihood, EM, SAEM, and importance-sampling methods for nonlinear mixed-effects models [2,4,5,7,8]. The estimation path was recorded for every dataset.

Importance proposals combined a component from the population distribution $N(0, \Omega)$ with a subject-specific component [2,7]. When pilot sampling produced an effective sample size below 50 or a maximum normalized importance weight above 0.1, the subject-specific proposal was recalculated from a verified conditional mode and positive local curvature. Importance weights continued to target the marginal likelihood, and the conditional mode improved the proposal distribution.

Final convergence required agreement between two independently generated integration banks. The absolute OFV difference between banks had to be no greater than 0.05, the maximum absolute projected score no greater than 0.2 in both banks, and the effective sample size at least 100 for every subject. Positive parameters were optimized in logarithmic coordinates, and the projected score accounted for declared bounds and the lower bound on interindividual variance. These criteria defined numerical acceptance. Fits outside the criteria retained their estimates and termination reasons.

Analytical prediction and repeated dosing

Concentrations and parameter sensitivities were calculated analytically for one-compartment intravenous, one-compartment oral, two-compartment intravenous, and two-compartment oral models. Oral models combined first-order absorption with exponential disposition terms. Limiting expressions and series expansions were used when absorption and disposition rate constants were equal or nearly equal. Analytical solutions for linear compartmental models are established in pharmacokinetic analysis and are also used in tools such as rxode2 [1,10].

Multiple doses were handled by linear superposition [1,10]. For the one-compartment intravenous and oral models and the two-compartment intravenous model, exponential state components were propagated between chronologically ordered dose and observation events. This state representation removed repeated summation of every preceding dose at each observation. The two-compartment oral model used analytical direct summation. Dose amounts and times were taken from the input records.

Local uncertainty assessment

Local parameter uncertainty was evaluated at an accepted estimate using the full marginal-likelihood Hessian. The Hessian was constructed from changes in the score using two independent randomized quasi-Monte Carlo integration banks [11] and two finite-difference step sizes. If H denotes the Hessian of the OFV, the local covariance matrix was approximated by

$$\mathrm{Cov}(\hat{\psi}) \approx 2H^{-1}.$$

Wald intervals for positive parameters were calculated in logarithmic coordinates and transformed back to the reporting scale. Interindividual variability was reported as a variance, and the coordinate transformation accounted for the log-standard-deviation coordinates used during optimization. The observed-information covariance and reporting-scale transformations follow standard likelihood and delta-method conventions [2,12,13].

Intervals were reported when the information matrix was positive definite and met prespecified checks for integration-bank agreement, finite-difference stability, objective consistency, effective sample size, and Hessian symmetry. The reported intervals quantify local uncertainty under the declared model. Interval availability and empirical coverage were evaluated separately. Intervals were not calculated for estimates on a declared bound.

Event-record interface

Models beyond the four closed-form structures are declared through an event-record interface. Analysis data follow NONMEM conventions: each record carries the subject, time, event type, dose amount and compartment, infusion rate or a modeled rate or duration, steady-state flag and dosing interval, additional doses, the observed output, and an optional censoring flag [9,12]. A structural library provides one- to three-compartment disposition with intravenous bolus or infusion and first-order, zero-order, or transit-compartment absorption [14], bioavailability and lag time, effect-compartment models with linear or Emax-type responses, parent-metabolite models, Michaelis-Menten elimination, target-mediated drug disposition in full and quasi-steady-state forms [15,16], and indirect-response models I-IV [17]; linear compartmental systems and ordinary differential equations (ODEs) can also be declared directly. Linear systems were advanced exactly between events through the eigendecomposition of the system matrix or, when the matrix was not diagonalizable, the matrix exponential [18], and their steady states were obtained in closed form. Nonlinear systems were integrated with an adaptive Dormand-Prince method [19] or, for stiff systems, a Rosenbrock method [20], and their steady states were obtained by repeating the dosing interval to convergence.

For parameter k of subject i at record j, the individual value was

$$\psi_{ikj} = g_k^{-1}(g_k(\theta_k) + \Sigma_m f_{km}(z_{imj}) + \eta_{ik} + \kappa_{iko}),$$

where g_k is the log, logit, or identity function, f_{km} is a power, exponential, linear, or categorical effect of covariate m, whose value z_{imj} can change within a subject, η_i follows $N(0, \Omega)$ with correlated blocks allowed, and κ_{iko} is the interoccasion effect of the occasion o that contains record j [21]. Each observed output had an additive, proportional, combined, or log-normal residual model, the last corresponding to a log-transform-both-sides analysis. An observation below the quantification limit (LLOQ) contributed the censoring probability $\Phi((LLOQ - f_{ij})/s_{ij})$, with residual standard deviation s_{ij} and both terms on the log scale for log-normal outputs (M3 method) [22]. Estimation followed the procedure described above. A Laplace approximation at the conditional modes was minimized from several starts, and the importance-sampled marginal likelihood was then refined in stages by quasi-Newton steps validated on independent integration banks. Each subject's importance proposal was adapted to the weighted mean and covariance of a pilot sample. Convergence required the two-bank OFV agreement and effective sample sizes of the compact interface, and a projected score of at most 0.2 in each coordinate, or 0.2 divided by the local standard error for coordinates with standard errors below one, which places the estimate within about 0.1 standard errors of the stationary point in these coordinates (Online Resource 1, Section S11).

Diagnostics, uncertainty tools, and covariate selection

Both interfaces share the diagnostic and inference tools. PKPy2 reports conditional modes (empirical Bayes estimates) and conditional means of the random effects, population and individual predictions, weighted, conditional weighted [23], and individual weighted residuals, normalized prediction distribution errors (NPDE) [24,25], η- and ε-shrinkage [26], simulations from the fitted model, and visual predictive checks (VPC) with optional prediction correction [27] and censoring. In addition to local Wald intervals, parameter uncertainty can be assessed with a sandwich covariance, profile likelihood, a nonparametric case bootstrap [28], and sampling importance resampling (SIR) with iterative proposal updates [29,30]. Nested models are compared by likelihood-ratio tests, and covariate effects can be selected by stepwise forward inclusion and backward elimination [31]. A plotting module draws goodness-of-fit, individual-fit, and VPC plots from these results with matplotlib [32]; it produced the visual predictive checks (Fig. 7) and the diagnostic plots and code example in Online Resource 1, Section S7.

Numerical validation and simulation design

Analytical predictions were evaluated in 12 parameter conditions for each of the four closed-form structural models. The conditions included single and repeated dosing and cases with equal or nearly equal absorption and disposition rate constants. Predictions were compared with an independent matrix-exponential implementation, and selected conditions were also checked with DOP853 numerical integration as implemented in SciPy [33]. Analytical sensitivities were compared with finite differences of the independent predictions.

Marginal-likelihood calculations were compared with an independent Gaussian-quadrature implementation [7,34] in four one-compartment intravenous settings. The settings combined rich or sparse sampling with the presence or absence of an estimated body-weight effect on clearance. Each model contained one clearance random effect. Optimization used 4,097 quadrature nodes and was checked with 8,193 nodes. The reference implementation used its own prediction, likelihood, and finite-difference Hessian calculations.

The primary simulation comprised 200 independently generated datasets. Each dataset contained 30 subjects receiving a single 100 mg intravenous bolus. The generating values were CL = 4 L/h, V = 40 L, $\omega^2(CL) = 0.09$, $\omega^2(V) = 0.04$, and proportional residual SD = 0.15. Rich sampling included measurements at 0.25, 0.5, 1, 2, 4, 6, 8, 12, 18, and 24 h. Sparse sampling included measurements at 1, 8, and 24 h. One hundred datasets were generated for each design using exponential equations implemented separately from PKPy2. Structural starting values were 80% of the generating values, each estimated interindividual variance started at 0.08, and proportional residual SD started at 0.20.

Each dataset was analyzed with PKPy2, the unchanged model and individual-fitting components of PKPy, and a Gaussian-likelihood two-stage control. The Gaussian control used the generating residual SD. A two-stage result was accepted when every individual optimization returned successful status.

Three additional structures were examined in ten simulated datasets each: a one-compartment oral model, a two-compartment intravenous model, and a two-compartment oral model. Each dataset contained 30 subjects receiving a single 100 mg dose and 16 sampling times from 0.1 to 48 h. The one-compartment oral model used CL = 4 L/h, V = 40 L, and Ka = 1.2 h^{-1}. The two-compartment models used CL = 4 L/h, V1 = 25 L, Q = 8 L/h, and V2 = 50 L; the oral model also used Ka = 1.2 h^{-1}. Proportional residual SD was 0.15. These simulations extended the evaluation to additional compartmental structures.

Relative bias and relative root mean squared error were calculated against the generating values. All planned replicates were retained in the denominators for convergence and interval availability. Coverage was reported for available intervals, together with the fraction of all planned replicates that produced an interval containing the true value. Wilson score intervals were used for proportions [35], and Monte Carlo standard errors were calculated for relative bias. Differences in relative RMSE were evaluated using 10,000 paired bootstrap resamples [28] of datasets accepted by both methods.

Real-world clinical data and expert NONMEM reference analyses

PKPy2 was applied to the public theophylline, warfarin, and tobramycin datasets [36,37]. The theophylline analysis used a one-compartment oral model with first-order absorption, fixed weight scaling of clearance and volume, and interindividual variability on clearance, volume, and the absorption rate constant. The warfarin analysis used the published one-compartment oral model with lag time, weight effects on clearance and volume, and the reported fixed and

estimated variability terms. The residual-error models reported for the expert analyses were used as the primary specifications, and alternative error models were evaluated as sensitivity analyses. PKPy2 parameter estimates were compared with expert NONMEM results reported in the PKGPT study [37]. The PKPy2 models matched the expert NONMEM models. Comparisons included point estimates and relative standard errors. Detailed dataset processing, model correspondence, and the NONMEM covariance settings [12] are provided in Online Resource 1. The tobramycin dataset of the PKGPT study contains 236 concentrations from 97 patients who received 858 intravenous bolus doses [37]. The expert model is a two-compartment model with power effects of creatinine clearance on CL and of body weight on the central volume (V1), interindividual variability on CL, and a combined residual error with a fixed additive term (Online Resource 1, Section S10). PKPy2 fitted this model from the documented initial values of the expert analysis and from the expert estimates. Because these fits reached a different optimum (Results), the model was refitted with the modeling judgment underlying the expert solution expressed as fixed values and bounds: V1 proportional to body weight (exponent fixed at 1), a small central compartment (V1 $\leq$ 10 L), a peripheral volume consistent with the extracellular distribution of aminoglycosides (V2 $\leq$ 30 L), and the remaining parameter bounds of the expert analysis. These expert-judgment constraints were used for the comparison with the other programs.

Comparison with established open-source NLME software

To compare PKPy2 with existing tools, the 200 primary simulation datasets and the three clinical datasets were also analyzed with nlmixr2 [5] using first-order conditional estimation with interaction (FOCEi) and stochastic approximation expectation-maximization (SAEM) [38], and with the saemix package [4]. All programs received the same observation records, exported from the PKPy2 analysis data, together with the same structural and statistical models, fixed terms, bounds, and starting values. nlmixr2 used analytical linear-compartment solutions, and ordinary differential equations for the warfarin model with absorption lag. Default estimation and covariance settings were used: the sandwich covariance for FOCEi, the stochastic-approximation covariance for SAEM, and the linearized Fisher information for saemix. The combined residual-error variance was defined as in PKPy2. nlmixr2 SAEM does not apply parameter bounds. saemix was applied to the simulated and theophylline data; it does not support the fixed variance terms of the warfarin model or the bounds of the tobramycin model.

Simulation estimates were summarized with the metrics used in the primary evaluation. Wald 95% intervals for CL and V were formed on the logarithmic scale from the standard errors reported by each program. Paired differences in relative RMSE between PKPy2 and each program were evaluated with 10,000 bootstrap resamples of datasets [28]. Because the objective functions reported by the programs are defined differently, the clinical estimates of every program were also evaluated with a single exact marginal OFV, computed by adaptive Gauss-Hermite quadrature [7,34] in a separate implementation that shares no code with PKPy2 (Online Resource 1, Section S8). The comparison used R 4.5.3 [39], nlmixr2 7.0.1 (nlmixr2est 7.1.0,

rxode2 5.1.7), and saemix 3.5 on the computer described below, with one thread per fit and five concurrent processes.

Evaluation of the event-record interface

Predictions were compared with rxode2 [10] in 24 scenarios covering infusions, steady state with bolus, infusion, and constant-infusion dosing, additional doses, lag time and bioavailability, modeled rates and durations, resets, time-varying covariates, three-compartment, transit, effect-compartment, and parent-metabolite models, and the Michaelis-Menten, indirect-response, and target-mediated disposition models. Marginal OFVs of four models combining block Ω, interoccasion variability, time-varying and categorical covariates, M3 censoring, a log-normal residual model, bioavailability, and two outputs were compared at fixed parameters with an independent adaptive Gauss-Hermite quadrature implementation [7,34]. The theophylline and warfarin analyses and ten primary simulation datasets were refitted through the event-record interface for comparison with the compact interface. Diagnostics of the theophylline model were compared with nlmixr2 and the npde package at identical parameter values.

Estimation was compared with nlmixr2 FOCEi in four simulated datasets: a two-compartment infusion model at steady state with block Ω and effects of body weight and sex (60 subjects); a one-compartment oral model with 102 of 450 concentrations below the quantification limit (50 subjects); a one-compartment model with log-normal residual error; and Michaelis-Menten elimination (40 subjects each). The warfarin turnover model of the nlmixr2 examples, with transit absorption, inhibition of the production of prothrombin complex activity (PCA), and eight random effects, was fitted to the concentrations and PCA of 32 subjects [5,40] with PKPy2 and with nlmixr2 FOCEi and SAEM. The exact marginal OFV of each program's estimates was evaluated with PKPy2 importance sampling. Parameter recovery was assessed in 20 datasets each of a two-compartment steady-state infusion model with block Ω, interoccasion variability, time-varying weight, sex, and BLQ data (40 subjects) and of an indirect-response PK/PD model with two outputs (40 subjects), and the calibration of NPDE and VPC in 50 and 30 datasets of the same designs evaluated at the true parameters. The interval methods were compared on the theophylline model, and stepwise covariate selection was applied to 20 simulated datasets with two true and five null covariate effects (Online Resource 1, Section S12).

Computational evaluation

Prediction-call times were measured after just-in-time compilation for direct summation and recurrence-based evaluation. Method order alternated across seven batches, and concentration agreement was checked before timing. Measurements used 1, 10, 100, and 1,000 dose events and report the concentration-prediction component. Numerical experiments were conducted with Python 3.13.5, NumPy 2.2.6, SciPy 1.16.1 [33], and Numba 0.65.1 [41] on Windows 11 with an AMD Ryzen 5 5600 processor. Fits and uncertainty calculations with the compact interface used no more than two workers, and event-record fits used three or four threads.

Results

Numerical validation

Analytical predictions were evaluated under 48 conditions spanning the four closed-form model structures. The maximum scaled discrepancy from the independent calculations was 4.10×10^{-10} for concentration predictions and 3.41×10^{-8} for parameter sensitivities. The four comparisons with the independent Gaussian-quadrature implementation also showed close agreement. The maximum objective function difference at identical parameter values was 4.34×10^{-6}, the maximum difference between optimized coordinates was 8.88×10^{-6}, and the maximum relative difference in coordinate standard errors was 0.0009% (Table 2).

Table 1 summarizes the main changes from PKPy to PKPy2. PKPy2 replaces summaries of separately fitted individual parameters with joint marginal-likelihood estimation and adds explicit fixed-parameter and bound handling, repeated-dose calculations, NONMEM-format event records, nonlinear and PK/PD models, model diagnostics, and population-level uncertainty estimates.

Recovery of population parameters and interindividual variability

PKPy2 converged for all 100 rich-sampling and 100 sparse-sampling datasets, and every primary fit converged without the SAEM stage. PKPy returned successful individual fits for every subject in all 200 datasets. The Gaussian two-stage control was accepted in 98 rich-sampling datasets and all 100 sparse-sampling datasets.

Under sparse sampling, the relative RMSEs for CL and V were 6.18% and 4.55% with PKPy and 5.98% and 4.20% with PKPy2. Larger differences were observed for the interindividual variance components. The relative RMSE of ω^2(CL) decreased from 34.71% with PKPy to 27.95% with PKPy2, and the relative RMSE of ω^2(V) decreased from 51.66% to 32.49%. The relative bias of ω^2(V) was 40.50% with PKPy and -7.75% with PKPy2. The paired difference in the relative RMSE of ω^2(V), calculated as PKPy2 minus PKPy, was -19.17 percentage points, with a bootstrap 95% interval of -25.52 to -12.23 percentage points (Fig. 2).

For the Gaussian two-stage control, the sparse-sampling relative RMSEs were 7.43% for CL, 4.88% for V, 37.20% for ω^2(CL), and 28.52% for ω^2(V). The paired PKPy2-minus-control difference for ω^2(V) was 3.97 percentage points, with a bootstrap 95% interval of -0.84 to 8.67 percentage points.

Confidence interval coverage and additional model structures

Local 95% confidence intervals were obtained for all 200 primary fits. Empirical coverage ranged from 90% to 100% across parameters and sampling designs. Under sparse sampling, coverage was 90% for CL, 94% for V, 93% for ω^2(CL), 100% for ω^2(V), and 95% for the proportional residual standard deviation. Interindividual variance intervals were wider than the CL and V intervals, particularly under sparse sampling (Fig. 3).

All ten one-compartment oral datasets and all ten two-compartment intravenous datasets converged. Nine of the ten two-compartment oral datasets converged. Among these nine fits, the relative RMSEs were 19.81% for V1, 20.87% for Ka, and 76.01% for ω^2(V1). Structural parameter RMSEs were at most 7.67% in the one-compartment oral model and at most 4.79% in the two-compartment intravenous model.

Real-world clinical data: comparison with NONMEM, nlmixr2, and saemix

In the theophylline proportional-error model, PKPy2 estimated CL70/F as 2.7851 L/h, V70/F as 32.5444 L, and Ka as 1.5082 h^{-1}. The three structural parameter estimates and three interindividual variance estimates differed by no more than 3.37% from the expert NONMEM estimates. Structural-parameter RSEs were also similar, and larger RSE differences were observed for some variance components. The RSE of ω^2(V), for example, was 60.2% with PKPy2 and 42.2% with NONMEM (Table 3).

For warfarin, the five freely estimated parameters differed by no more than 4.56% from the expert NONMEM estimates. The largest difference was observed for Ka. Point estimates agreed more closely than RSEs, with the largest RSE differences observed for ω^2(CL) and Ka (Table 4 and Fig. 4).

The same clinical models were fitted with nlmixr2 and saemix (Table 5 and Fig. 4). For theophylline, all 24 point estimates from the four open-source programs were within 6.5% of the expert NONMEM estimates, and the exact marginal OFVs at the five sets of estimates, including NONMEM, were within 0.09 units of one another (PKPy2, 352.79; NONMEM, 352.86). For warfarin, the nlmixr2 FOCEi estimates were within 4.2% of the expert estimates, and the exact OFV at these estimates was 0.16 units higher than at the PKPy2 estimates. nlmixr2 SAEM estimated Ka and absorption lag time 17.1% and 28.9% lower than the expert estimates, and the exact OFV at its estimates was 16.4 units higher than at the PKPy2 estimates, indicating that the SAEM run stopped at an inferior solution of the same model. Standard errors differed more than point estimates across programs (Online Resource 1, Table S13).

For tobramycin, PKPy2 fits of the published model from the documented initial values and from the expert estimates converged to the same maximum-likelihood estimate (OFV 560.06), which differed from the expert solution: the body-weight exponent of V1 was -5.6, and V2 was 188 L (Online Resource 1, Table S15). The exact OFV at the expert estimates was 45.0 units higher. An independent quadrature calculation reproduced the PKPy2 OFV, and nlmixr2 FOCEi and saemix fits of the same model also gave negative weight exponents (Online Resource 1, Table S17). The data carry little information on distribution: 205 of 236 concentrations were measured 2.5 h after a dose, body weight ranged only from 50 to 75 kg and was correlated with creatinine clearance ($r = 0.44$ on the log scale), and the profile OFV varied by 18.6 units across V2 values from 5 to 600 L (Online Resource 1, Table S16). Within the parameter bounds of the expert NONMEM analysis, the best solution lay on the bounds (weight exponent 0, V2 100 L), and the fit started from the expert estimates remained near them with an OFV 18.2 units higher; the nlmixr2 FOCEi objective ranked the two solutions in the same order.

With the expert-judgment constraints, PKPy2 converged from both starting points to the same estimate, with no constraint active, and reproduced the expert estimates within 6.0% (Table 5 and Fig. 4c). The OFV of this estimate was 44.4 units above the unconstrained maximum. nlmixr2 FOCEi with the same constraints gave estimates within 12.8% of the expert estimates, and the exact OFV at its estimates was 0.72 units higher than at the PKPy2 estimates. nlmixr2 SAEM, which does not apply the bounds, estimated CL and Q 41% and 259% higher than the expert estimates, with an exact OFV 68.6 units higher than at the PKPy2 estimates.

Comparison with established software in the primary simulation

nlmixr2 FOCEi, nlmixr2 SAEM, and saemix returned estimates for all 200 primary datasets, and parameter recovery was similar across programs (Table 6 and Fig. 5). Under rich sampling, relative RMSEs differed from PKPy2 by at most 0.25 percentage points for every parameter; the only paired difference whose bootstrap interval excluded zero was for $\omega^2(V)$ against nlmixr2 FOCEi (PKPy2 minus FOCEi, 0.20 percentage points; 95% interval 0.01 to 0.41). Under sparse sampling, nlmixr2 FOCEi overestimated CL by 2.70% on average, compared with 0.70% for PKPy2, and its relative RMSEs for CL and V were higher by 0.58 and 0.24 percentage points (95% intervals for PKPy2 minus FOCEi, -0.92 to -0.21 and -0.40 to -0.07). The sparse-sampling relative RMSE of $\omega^2(V)$ was 32.49% with PKPy2, 32.31% with nlmixr2 FOCEi, 33.66% with nlmixr2 SAEM, and 37.78% with saemix; the paired differences from nlmixr2 SAEM and saemix were -1.17 (-2.03 to -0.33) and -5.29 (-8.14 to -2.53) percentage points. All paired differences are listed in Online Resource 1, Table S12.

Coverage of nominal 95% intervals for CL and V was 90% to 96% for PKPy2, 86% to 96% for nlmixr2 FOCEi, 90% to 96% for nlmixr2 SAEM, and 89% to 96% for saemix. Across the two sampling designs, median wall-clock times per fit were 1.2-1.8 s for saemix, 2.1-2.9 s for nlmixr2 FOCEi, 5.2-9.6 s for nlmixr2 SAEM, and 7.5-9.9 s for PKPy2, which computed intervals in a separate step (medians 6.0-8.4 s).

Event-record interface

Predictions agreed with rxode2 in all 24 scenarios, with maximum relative differences of 2.3×10^{-11} for linear systems and 1.1×10^{-6} for ODE models (Fig. 6a). The four marginal OFVs differed from the independent quadrature values by at most 1.3×10^{-4}. All twelve refits through the event-record interface converged, with OFVs within 0.0012 and estimates within 0.76% of the compact-interface fits (Fig. 6b). At identical parameters, PRED agreed with nlmixr2 to 6.0×10^{-14}, and IPRED, IWRES, and CWRES agreed to within 0.0017; NPDE computed from the same simulated replicates agreed with the npde package to 4.9×10^{-15} (Fig. 6c).

PKPy2 converged for the four simulated comparison datasets and the warfarin PK/PD model. The exact OFV at the nlmixr2 FOCEi estimates was 0.01 to 2.59 units higher than at the PKPy2 estimates in the simulated datasets, and for warfarin PK/PD the OFV was 334.34 units higher at the FOCEi estimates and 17.94 units higher at the SAEM estimates (Fig. 6d). Estimates of all programs are listed in Online Resource 1, Tables S20 and S21. For warfarin PK/PD, PKPy2

estimated IMAX at 0.9999 (nlmixr2 SAEM, 0.962; FOCEi, 0.879), CL at 0.134 L/h, and V at 7.73 L; because IMAX approached 1, the information matrix was not positive definite and local intervals were not reported. In the visual predictive checks, 32 of 42 observed concentration percentiles and 19 of 24 PCA percentiles lay within their simulated 95% intervals; the simulated 5th-95th percentile range was wider than the observed range at several times, and observed PCA was capped at 100% at baseline (Fig. 7). The fit took 1,139 s with PKPy2 (four threads), 1,052 s with nlmixr2 FOCEi, and 608 s with nlmixr2 SAEM (one thread each).

In the recovery simulations, all fits converged in both designs, with median fit times of 486 s for the infusion design and 1,201 s for the PK/PD design. Relative bias was within ±3.1% for typical values, -12.3% to -2.4% for the covariate coefficients, and -12.3% to +2.8% for variance components; the additive residual SD of the infusion model, which the concentrations informed little, had a relative bias of -32% (Online Resource 1, Table S22). At the true parameters of the two designs, the mean NPDE was -0.008 and 0.004, the NPDE variance 1.02 and 1.01, and 95% and 94% of observed VPC percentiles lay within their 95% intervals. For theophylline, the Wald, sandwich, profile-likelihood, bootstrap, and SIR intervals were similar for the typical values and differed more for the variance components (Fig. 7). Stepwise covariate selection retained both true effects in 19 of 20 datasets, each true effect was selected in every dataset, and a null effect was selected in 1 of 100 null-candidate evaluations.

Repeated-dose prediction efficiency

Prediction-call times after compilation were similar for direct summation and recurrence with a single dose. With 10 dose events, recurrence was 1.91-2.13 times faster than direct summation. The speedup increased to 20.04-26.31 times with 100 doses and 440.05-906.47 times with 1,000 doses (Online Resource 1, Table S10). The largest ratio was observed for the two-compartment intravenous model with 1,000 dose events.

Discussion

PKPy2 advances PKPy by replacing population summaries of individual fits with joint marginal-likelihood inference. The advantage was most evident for interindividual variability under sparse sampling. The relative RMSE of $\omega^2(V)$ decreased from 51.66% with PKPy to 32.49% with PKPy2, and relative bias changed from 40.50% to -7.75%. CL and V were already recovered with low error by both workflows, so differences in their typical values were smaller. Joint estimation uses all subjects simultaneously and estimates residual error together with population variability, reducing the contribution of uncertainty from separate individual fits to the reported variance components.

The Gaussian-likelihood two-stage control provided an informative reference because it used the generating residual standard deviation. Its sparse-sampling RMSE for $\omega^2(V)$ was 28.52%, compared with 32.49% for PKPy2, and the paired bootstrap interval for the difference included zero. PKPy2 estimated residual error jointly with the population parameters and improved both

interindividual variance estimates relative to the original PKPy implementation. The comparison shows how population estimation and residual-error specification jointly shape variance recovery.

Independent calculations verified the numerical accuracy of the implementation. All 48 prediction conditions agreed with the independent matrix-exponential and numerical-integration calculations, and the independent quadrature comparisons closely reproduced objective values, optimized coordinates, and coordinate standard errors. The maximum same-point OFV difference was 4.34×10^{-6}, and the maximum relative difference in coordinate standard errors was 0.0009%. All 200 primary fits converged and produced local confidence intervals, demonstrating stable marginal-likelihood estimation and information-matrix calculation across both sampling designs.

Empirical coverage of the nominal 95% intervals ranged from 90% to 100%. Sparse-sampling coverage was 90% for CL and 100% for $\omega^2(V)$, showing parameter-specific differences in finite-sample calibration. Variance-component intervals were wider than structural-parameter intervals, particularly under sparse sampling. The combination of coverage and interval width provides a direct view of the precision available for each parameter and sampling design.

Performance extended beyond the primary one-compartment intravenous model. 29 of the 30 additional-structure datasets converged. Structural parameter RMSEs were at most 7.67% in the one-compartment oral model and at most 4.79% in the two-compartment intravenous model. The two-compartment oral model was more demanding, with relative RMSEs of 19.81% for V1, 20.87% for Ka, and 76.01% for $\omega^2(V1)$. These results identify absorption and distribution models as the next priority for precision improvements.

The clinical analyses showed close agreement with established NONMEM results. In the theophylline analysis, all six structural and interindividual variance estimates were within 3.37% of the expert NONMEM estimates. In the warfarin analysis, five freely estimated parameters differed by no more than 4.56% from the expert estimates. RSEs varied more across implementations, consistent with their distinct likelihood approximations and covariance calculations. The close point-estimate agreement across both datasets supports the use of PKPy2 for declared oral pharmacokinetic models.

With identical datasets, models, and starting values, PKPy2, nlmixr2, and saemix recovered structural parameters and variance components with similar accuracy under rich sampling. Under sparse sampling, the FOCEi approximation overestimated clearance by 2.70% on average, and the SAEM-based programs estimated the variance of volume less precisely than PKPy2. In the clinical datasets the programs agreed closely, except for the nlmixr2 SAEM fits of warfarin and tobramycin, whose exact objective function values were well above those at the PKPy2 estimates. PKPy2 reports convergence after projected marginal scores are small in two independent integration banks, a criterion designed to detect such non-stationary solutions. nlmixr2 and saemix additionally support non-Gaussian observation models such as categorical and time-to-event data, and saemix and nlmixr2 FOCEi had shorter median fit times.

The tobramycin analysis shows how PKPy2 responds to pharmacological judgment. The published model had a maximum-likelihood solution with a negative weight exponent and a large peripheral volume, because concentrations were measured mostly at a single time after dose and body weight varied little and was correlated with creatinine clearance; nlmixr2 and saemix reached similar solutions. The expert solution reflects additional knowledge of how tobramycin distributes. Entered as fixed values and bounds, this knowledge led PKPy2 to the best solution within the constraints from either starting point, and that solution reproduced the expert estimates. Declared constraints therefore let a clinical pharmacologist steer an analysis toward pharmacologically plausible solutions while the likelihood remains the criterion within them, and the exact OFV shows how much fit each constraint costs.

The analytical recurrence for repeated dosing provided a substantial computational advantage for long dose histories. With 1,000 dose events, prediction calls were 440- to 906-fold faster than direct summation across the three accelerated models. Population estimation evaluates concentrations repeatedly across subjects, integration samples, and parameter values, so this reduction directly strengthens the computational core of the estimation workflow.

The event-record interface extends the same estimation and audit to the data structures and models of routine population PK/PD analysis. Its predictions reproduced rxode2 to near machine precision for linear systems and to the integration tolerance for ODE models, its likelihoods reproduced an independent quadrature, and its diagnostics matched nlmixr2 and the npde package. Its estimates reached exact OFVs at least as low as those of nlmixr2 in all comparisons. These checks cover the numerical components that users cannot inspect directly: event handling, likelihood integration, and residual definitions.

The primary evaluation used 30-subject datasets, diagonal random effects, specified linear compartmental models, and ten replicates for each additional structure, and the event-record interface was evaluated with 20 replicates per design, datasets of up to 60 subjects, and one PK/PD application. The NONMEM comparisons used the expert estimates as reported, and nlmixr2 and saemix were run with default settings. The tobramycin constraints were formulated with knowledge of the expert solution; the data alone supported a different solution. Larger simulation studies across population sizes, sparse PK/PD designs, and stiff systems will further characterize precision and computing time for complex models.

Overall, PKPy2 provides a unified Python workflow for explicit PopPK and PK/PD model specification with fixed values and bounds, NONMEM-format event records, joint estimation, numerical convergence assessment, model diagnostics, and uncertainty quantification. Its improved recovery of interindividual variability under sparse sampling, close agreement with independent calculations, expert NONMEM estimates, rxode2, and nlmixr2, and efficient handling of long dose histories demonstrate the practical value of nonlinear mixed-effects analysis in a standalone Python environment.

Conclusions

PKPy2 extends PKPy from separate individual fits to joint nonlinear mixed-effects population estimation. In the primary simulation, all 200 fits converged, and the sparse-sampling relative RMSE of the interindividual variance in volume decreased from 51.66% with PKPy to 32.49% with PKPy2. Independent calculations closely matched PKPy2 predictions, likelihoods, optimized parameters, and standard errors. On the same simulated and clinical datasets, PKPy2 performed comparably to nlmixr2 and saemix. Theophylline and warfarin estimates agreed closely with published expert NONMEM estimates, tobramycin estimates reproduced the expert estimates once the expert's pharmacological judgment was declared as fixed values and bounds, and analytical dose-state propagation accelerated prediction for long dose histories. An event-record interface extends estimation, diagnostics, and uncertainty assessment to multi-compartment, nonlinear, and PK/PD models with correlated and interoccasion random effects and censored data, and its calculations agreed with rxode2, an independent quadrature implementation, nlmixr2, and the npde package.

PKPy2 provides an integrated Python environment for explicit PopPK and PK/PD model specification, joint estimation of structural parameters and variability, numerical convergence assessment, model diagnostics, and uncertainty quantification. The framework supports reproducible population analysis with a transparent model interface and a standalone estimation engine.

Declarations

Funding No funding was received for conducting this study.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable. The study used simulated data and publicly available, de-identified pharmacokinetic datasets.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability The theophylline data are available in the R datasets package [36], the warfarin and tobramycin data in the dataset directory of the PKGPT repository (https://github.com/Gumgo91/PKGPT) [37], and the warfarin PK/PD data in the nlmixr2data R package [5,40]. The analysis datasets, per-dataset estimates of all programs, numerical checks, and timings underlying all tables and figures are provided in Online Resources 2 and 3, with variable definitions in Online Resource 4. The nlmixr2 and saemix comparison scripts, the tobramycin analyses, and the saved fit records are available in the PKPy2 repository (https://github.com/Gumgo91/PKPy2).

Code availability PKPy2, installation instructions, runnable examples, and the numerical validation script are available at https://github.com/Gumgo91/PKPy2 under the MIT License.

Author contributions Conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, and writing – original draft: Hyunseung Kong. Supervision and writing – review and editing: Inyoung Kim. All authors read and approved the final manuscript.

References

1. Mould DR, Upton RN (2012) Basic concepts in population modeling, simulation, and model-based drug development. CPT Pharmacometrics Syst Pharmacol 1:e6. https://doi.org/10.1038/psp.2012.4
2. Bauer RJ (2019) NONMEM tutorial part II: estimation methods and advanced examples. CPT Pharmacometrics Syst Pharmacol 8:538−556. https://doi.org/10.1002/psp4.12422
3. ICON plc (2026) NONMEM. https://www.iconplc.com/solutions/technologies/nonmem. Accessed 16 September 2026
4. Comets E, Lavenu A, Lavielle M (2017) Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. J Stat Softw 80(3):1−41. https://doi.org/10.18637/jss.v080.i03
5. Fidler M, Wilkins JJ, Hooijmaijers R, Post TM, Schoemaker R, Trame MN, Xiong Y, Wang W (2019) Nonlinear mixed-effects model development and simulation using nlmixr and related R open-source packages. CPT Pharmacometrics Syst Pharmacol 8:621−633. https://doi.org/10.1002/psp4.12445
6. Kong H, Kim I, Zhang B-T (2025) PKPy: a Python-based framework for automated population pharmacokinetic analysis. PeerJ 13:e20258. https://doi.org/10.7717/peerj.20258
7. Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4:12−35. https://doi.org/10.1080/10618600.1995.10474663
8. Delyon B, Lavielle M, Moulines E (1999) Convergence of a stochastic approximation version of the EM algorithm. Ann Stat 27:94−128. https://doi.org/10.1214/aos/1018031103
9. Beal SL, Boeckmann AJ, Sheiner LB (1992) NONMEM users guide, part VI: PREDPP guide. NONMEM Project Group, University of California, San Francisco
10. Fidler ML, Wang W, Hallow M (2026) rxode2 additional model types. rxode2 documentation. https://nlmixr2.github.io/rxode2/articles/rxode2-model-types.html. Accessed 16 September 2026
11. Owen AB (1998) Scrambling Sobol' and Niederreiter−Xing points. J Complex 14:466−489. https://doi.org/10.1006/jcom.1998.0487
12. Beal SL, Sheiner LB (1988) NONMEM users guide, part II: users supplemental guide. NONMEM Project Group, University of California, San Francisco
13. Oehlert GW (1992) A note on the delta method. Am Stat 46:27−29. https://doi.org/10.1080/00031305.1992.10475842
14. Savic RM, Jonker DM, Kerbusch T, Karlsson MO (2007) Implementation of a transit compartment model for describing drug absorption in pharmacokinetic studies. J Pharmacokinet Pharmacodyn 34:711−726. https://doi.org/10.1007/s10928-007-9066-0
15. Mager DE, Jusko WJ (2001) General pharmacokinetic model for drugs exhibiting target-mediated drug disposition. J Pharmacokinet Pharmacodyn 28:507−532. https://doi.org/10.1023/A:1014414520282

16. Gibiansky L, Gibiansky E, Kakkar T, Ma P (2008) Approximations of the target-mediated drug disposition model and identifiability of model parameters. J Pharmacokinet Pharmacodyn 35:573–591. https://doi.org/10.1007/s10928-008-9102-8
17. Dayneka NL, Garg V, Jusko WJ (1993) Comparison of four basic models of indirect pharmacodynamic responses. J Pharmacokinet Biopharm 21:457–478. https://doi.org/10.1007/BF01061691
18. Moler C, Van Loan C (2003) Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev 45:3–49. https://doi.org/10.1137/S00361445024180
19. Dormand JR, Prince PJ (1980) A family of embedded Runge-Kutta formulae. J Comput Appl Math 6:19–26. https://doi.org/10.1016/0771-050X(80)90013-3
20. Shampine LF, Reichelt MW (1997) The MATLAB ODE suite. SIAM J Sci Comput 18:1–22. https://doi.org/10.1137/S1064827594276424
21. Karlsson MO, Sheiner LB (1993) The importance of modeling interoccasion variability in population pharmacokinetic analyses. J Pharmacokinet Biopharm 21:735–750. https://doi.org/10.1007/BF01113502
22. Beal SL (2001) Ways to fit a PK model with some data below the quantification limit. J Pharmacokinet Pharmacodyn 28:481–504. https://doi.org/10.1023/A:1012299115260
23. Hooker AC, Staatz CE, Karlsson MO (2007) Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. Pharm Res 24:2187–2197. https://doi.org/10.1007/s11095-007-9361-x
24. Brendel K, Comets E, Laffont C, Laveille C, Mentré F (2006) Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. Pharm Res 23:2036–2049. https://doi.org/10.1007/s11095-006-9067-5
25. Comets E, Brendel K, Mentré F (2008) Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. Comput Methods Programs Biomed 90:154–166. https://doi.org/10.1016/j.cmpb.2007.12.002
26. Savic RM, Karlsson MO (2009) Importance of shrinkage in empirical Bayes estimates for diagnostics: problems and solutions. AAPS J 11:558–569. https://doi.org/10.1208/s12248-009-9133-0
27. Bergstrand M, Hooker AC, Wallin JE, Karlsson MO (2011) Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. AAPS J 13:143–151. https://doi.org/10.1208/s12248-011-9255-z
28. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26. https://doi.org/10.1214/aos/1176344552
29. Dosne AG, Bergstrand M, Harling K, Karlsson MO (2016) Improving the estimation of parameter uncertainty distributions in nonlinear mixed effects models using sampling importance resampling. J Pharmacokinet Pharmacodyn 43:583–596. https://doi.org/10.1007/s10928-016-9487-8

30. Dosne AG, Bergstrand M, Karlsson MO (2017) An automated sampling importance resampling procedure for estimating parameter uncertainty. J Pharmacokinet Pharmacodyn 44:509–520. https://doi.org/10.1007/s10928-017-9542-0
31. Jonsson EN, Karlsson MO (1998) Automated covariate model building within NONMEM. Pharm Res 15:1463–1468. https://doi.org/10.1023/A:1011970125687
32. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55
33. Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2
34. Golub GH, Welsch JH (1969) Calculation of Gauss quadrature rules. Math Comput 23:221–230. https://doi.org/10.1090/S0025-5718-69-99647-1
35. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209–212. https://doi.org/10.1080/01621459.1927.10502953
36. R Core Team (2026) Theoph: pharmacokinetics of theophylline. R datasets package documentation. https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/Theoph.html. Accessed 16 September 2026
37. Kwack H, Kong H, Lim J, Zhang B-T, Hahn J, Chang MJ (2026) PKGPT: expert-orchestrated recursive LLM agent for automated NONMEM PopPK modeling with human benchmarking. Pharmaceutics 18:501. https://doi.org/10.3390/pharmaceutics18040501
38. Schoemaker R, Fidler M, Laveille C, Wilkins JJ, Hooijmaijers R, Post TM, Trame MN, Xiong Y, Wang W (2019) Performance of the SAEM and FOCEI algorithms in the open-source, nonlinear mixed effect modeling tool nlmixr. CPT Pharmacometrics Syst Pharmacol 8:923–930. https://doi.org/10.1002/psp4.12471
39. R Core Team (2026) R: a language and environment for statistical computing, version 4.5.3. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
40. O'Reilly RA, Aggeler PM (1968) Studies on coumarin anticoagulant drugs: initiation of warfarin therapy without a loading dose. Circulation 38:169–177. https://doi.org/10.1161/01.CIR.38.1.169
41. Lam SK, Pitrou A, Seibert S (2015) Numba: a LLVM-based Python JIT compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. ACM, New York, Article 7. https://doi.org/10.1145/2833157.2833162

744 **Table 1** Development from the evaluated PKPy components to PKPy2

Component	PKPy	PKPy2
Structural models	One- and two-compartment models with intravenous administration or first-order absorption	The same four closed-form structures with oral lag; through event records, one- to three-compartment, zero-order, transit, parent-metabolite, Michaelis-Menten, target-mediated, effect-compartment and indirect-response models and user-defined ODEs
Population inference	Separate individual fits followed by population summaries	Joint marginal-likelihood estimation
Random effects and residual error	Covariance of log-transformed individual estimates	Diagonal or block Ω and interoccasion variability estimated jointly with additive, proportional, combined, or log-normal residual error for each output
Covariate models	Regression of individual estimates on covariates with forward selection	Power, exponential, linear, and categorical effects, including time-varying covariates, within the population model; stepwise selection by likelihood-ratio tests
Diagnostics	Goodness-of-fit plots of the individual fits	PRED, IPRED, CWRES, NPDE, shrinkage, visual predictive checks, and diagnostic plots
Fixed components and bounds	Specified within individual model fitting	Explicit fixed or estimated θ, Ω, σ, and covariate coefficients, with optional lower and upper bounds
Uncertainty evaluated here	Individual fitting and population summaries	Full local population information; sandwich, profile-likelihood, bootstrap, and SIR intervals
Dosing and observations	Identified as an area for extension	Explicit dose histories and analytical superposition; NONMEM event records with infusions, steady state, additional doses, resets, several outputs, and censored observations
Numerical reporting	Individual optimizer status	Independent numerical audit, termination reason, and saved estimates

Table 2 Independent quadrature comparisons in one-random-effect intravenous models

Sampling	WT effect on CL	Node-refinement OFV gap	Same-point OFV gap	Maximum coordinate gap	Maximum SE difference (%)
Rich	Absent	4.97×10^{-13}	7.44×10^{-7}	6.47×10^{-7}	1.42×10^{-4}
Sparse	Absent	2.74×10^{-13}	9.53×10^{-7}	7.65×10^{-6}	8.67×10^{-4}
Rich	Estimated	4.26×10^{-13}	8.46×10^{-9}	9.78×10^{-7}	2.85×10^{-4}
Sparse	Estimated	6.79×10^{-13}	4.34×10^{-6}	8.88×10^{-6}	8.78×10^{-4}

Coordinates are logarithms of positive parameters and untransformed covariate coefficients; interindividual variability uses log-SD coordinates. SE differences refer to these coordinates. Node refinement increased quadrature from 4,097 to 8,193 nodes.

Table 3 Theophylline: comparison with the expert NONMEM analysis

Parameter	NONMEM estimate	PKPy2 estimate	Relative difference (%)	NONMEM RSE (%)	PKPy2 RSE (%)
CL/F (L/h)	2.8280	2.7851	−1.52	8.0	7.6
V/F (L)	32.5500	32.5444	−0.02	4.3	4.4
Ka (h^{-1})	1.4600	1.5082	3.30	21.4	20.7
ω^2(CL)	0.0646	0.0653	1.16	NR	45.3
ω^2(V)	0.0145	0.0148	2.05	42.2	60.2
ω^2(Ka)	0.4450	0.4600	3.37	NR	48.1
σ_prop	NR	0.1577	NR	NR	7.8

NONMEM estimates and RSEs are from the expert analysis [37]; CL and V were reported per kg and are shown for 70 kg. NR, not reported.

Table 4 Warfarin: comparison with the expert NONMEM analysis

Parameter	NONMEM estimate	PKPy2 estimate	Relative difference (%)	NONMEM RSE (%)	PKPy2 RSE (%)
CL/F (L/h)	0.1350	0.13406	−0.70	4.8	4.6
V/F (L)	7.860	7.84580	−0.18	3.0	1.7
Ka (h^{-1})	1.150	1.20245	4.56	30.5	24.1
ALAG (h)	0.8250	0.82728	0.28	8.6	4.8
ω^2(CL)	0.0643	0.06451	0.33	37.3	27.0
σ_prop	NR	0.13660	NR	NR	6.4

NONMEM estimates and RSEs are from the expert analysis [37]. CL and V are standardized to 70 kg. Ka interindividual variance (0.5), additive variance (0.117), and WT exponents on CL and V (0.75 and 1) were fixed. NR, not reported.

Table 5 Clinical datasets: estimates from NONMEM, PKPy2, nlmixr2 and saemix

Parameter	NONMEM	PKPy2	nlmixr2 FOCEi	nlmixr2 SAEM	saemix
Theophylline (12 subjects)					
CL/F (L/h)	2.828	2.785 (−1.5)	2.834 (0.2)	2.780 (−1.7)	2.781 (−1.7)
V/F (L)	32.550	32.544 (0.0)	32.542 (0.0)	32.265 (−0.9)	32.530 (−0.1)
Ka (h^{-1})	1.460	1.508 (3.3)	1.460 (0.0)	1.483 (1.6)	1.510 (3.4)
ω^2(CL)	0.0646	0.0653 (1.2)	0.0648 (0.3)	0.0644 (−0.3)	0.0658 (1.9)
ω^2(V)	0.0145	0.0148 (2.0)	0.0150 (3.3)	0.0136 (−6.5)	0.0153 (5.3)
ω^2(Ka)	0.4450	0.4600 (3.4)	0.4359 (−2.1)	0.4534 (1.9)	0.4539 (2.0)
Marginal OFV	352.86	352.79	352.87	352.85	352.80
Warfarin (32 subjects)					
CL/F (L/h)	0.1350	0.1341 (−0.7)	0.1348 (−0.2)	0.1341 (−0.7)	—
V/F (L)	7.860	7.846 (−0.2)	7.851 (−0.1)	7.884 (0.3)	—
Ka (h^{-1})	1.150	1.202 (4.6)	1.102 (−4.2)	0.953 (−17.1)	—
ALAG (h)	0.8250	0.8273 (0.3)	0.8208 (−0.5)	0.5870 (−28.9)	—
ω^2(CL)	0.0643	0.0645 (0.3)	0.0627 (−2.5)	0.0650 (1.2)	—
Marginal OFV	755.54	755.47	755.63	771.90	—
Tobramycin (97 subjects)					
CL (L/h)	2.950	2.931 (−0.7)	3.030 (2.7)	4.174 (41.5)	—
V1 (L)	4.590	4.863 (5.9)	5.177 (12.8)	8.600 (87.4)	—
Q (L/h)	6.850	7.099 (3.6)	7.418 (8.3)	24.610 (259.3)	—
V2 (L)	13.200	13.281 (0.6)	13.259 (0.4)	16.661 (26.2)	—
CLCR exponent	0.2360	0.2418 (2.5)	0.2377 (0.7)	0.6755 (186.2)	—
ω^2(CL)	0.0280	0.0283 (0.9)	0.0297 (6.1)	0.1038 (270.7)	—
Marginal OFV	605.06	604.44	605.17	673.07	—

Estimates with the relative difference from NONMEM (%) in parentheses. NONMEM, expert analyses [37]. CL and V are standardized to 70 kg. Marginal OFV: −2 log L of the declared model at each set of estimates, evaluated by independent adaptive Gauss-Hermite quadrature (Online Resource 1, Section S8); for NONMEM, σ_prop (not reported) was set to its maximum-likelihood value. Tobramycin was fitted with the expert-judgment constraints (WT exponent of V1 fixed at 1, V1 ≤ 10 L, V2 ≤ 30 L; Methods); nlmixr2 SAEM does not apply parameter bounds. saemix was not applied to warfarin or tobramycin. —, not available. Standard errors are given in Online Resource 1, Table S13.

Table 6 Primary simulation: PKPy2 compared with established open-source NLME software

Parameter	Quantity	PKPy2	nlmixr2 FOCEi	nlmixr2 SAEM	saemix
Rich sampling (100 datasets)					
Estimates returned		100	100	100	100
CL	Bias (RMSE), %	0.45 (5.64)	1.32 (5.79)	0.46 (5.64)	0.57 (5.65)
V	Bias (RMSE), %	0.17 (3.81)	0.48 (3.84)	0.17 (3.81)	0.26 (3.83)
ω²(CL)	Bias (RMSE), %	−4.27 (27.67)	−5.46 (27.42)	−4.17 (27.66)	−4.08 (27.84)
ω²(V)	Bias (RMSE), %	−4.09 (25.71)	−3.75 (25.51)	−4.31 (25.67)	−4.35 (25.76)
σ_prop	Bias (RMSE), %	−0.31 (4.97)	−0.18 (4.99)	−0.29 (4.96)	−0.28 (4.98)
CL	95% CI coverage, %	94 (n = 100)	89 (n = 100)	94 (n = 100)	94 (n = 100)
V	95% CI coverage, %	96 (n = 100)	96 (n = 100)	96 (n = 100)	96 (n = 100)
Median fit time, s		9.9	2.9	9.6	1.8
Sparse sampling (100 datasets)					
Estimates returned		100	100	100	100
CL	Bias (RMSE), %	0.70 (5.98)	2.70 (6.55)	0.72 (5.99)	1.04 (6.06)
V	Bias (RMSE), %	0.48 (4.20)	1.36 (4.44)	0.59 (4.20)	0.75 (4.27)
ω²(CL)	Bias (RMSE), %	−4.91 (27.95)	−6.95 (27.53)	−4.59 (28.13)	−4.96 (28.14)
ω²(V)	Bias (RMSE), %	−7.75 (32.49)	−7.09 (32.31)	−9.38 (33.66)	−13.28 (37.78)
σ_prop	Bias (RMSE), %	−0.57 (12.17)	−0.98 (12.10)	0.46 (12.56)	1.26 (13.80)
CL	95% CI coverage, %	90 (n = 100)	86 (n = 100)	90 (n = 100)	89 (n = 100)
V	95% CI coverage, %	94 (n = 100)	93 (n = 100)	91 (n = 100)	93 (n = 100)
Median fit time, s		7.5	2.1	5.2	1.2

Relative bias and RMSE (%) against the generating values CL = 4 L/h, V = 40 L, ω²(CL) = 0.09, ω²(V) = 0.04 and σ_prop = 0.15. Coverage refers to 95% Wald intervals on the log scale; n, datasets with an interval. PKPy2 fit times exclude the uncertainty calculation. Paired-bootstrap RMSE differences are given in Online Resource 1, Table S12.

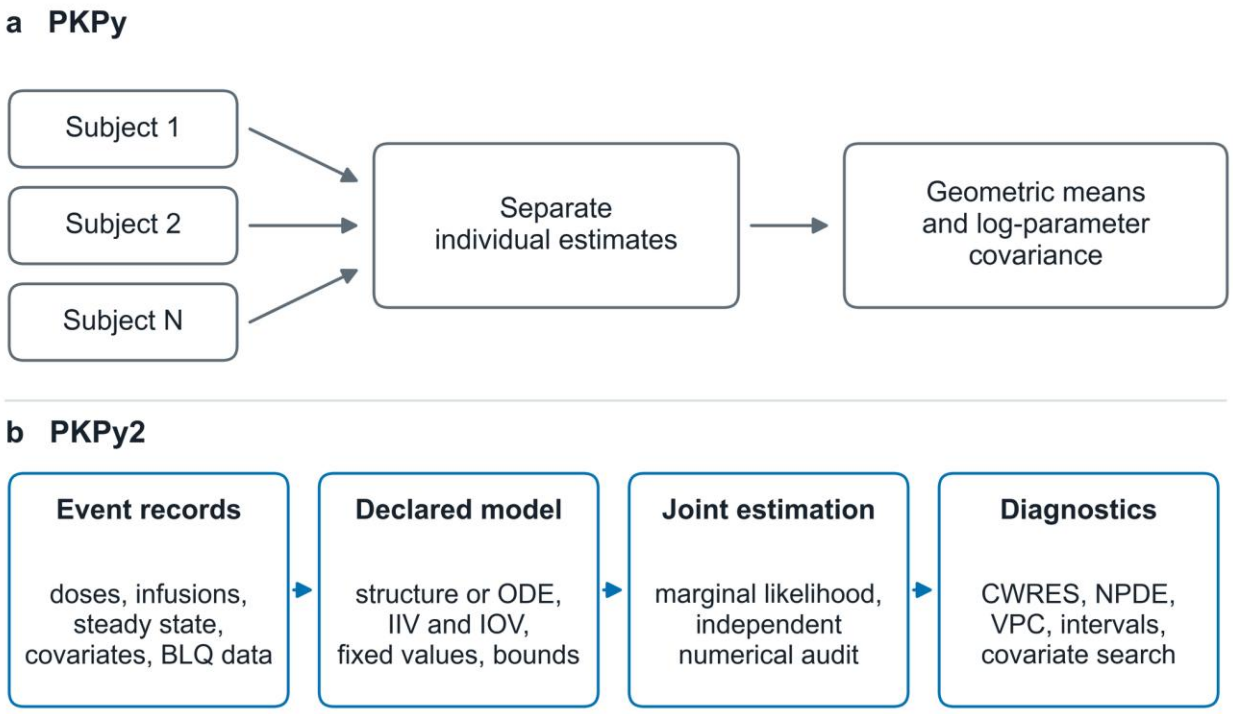


Fig. 1 Population inference in PKPy and PKPy2. (a) PKPy fits each subject separately and summarizes the individual log parameters. (b) PKPy2 reads event records, fits the declared population model jointly by marginal likelihood with an independent numerical audit, and provides diagnostics and interval estimates

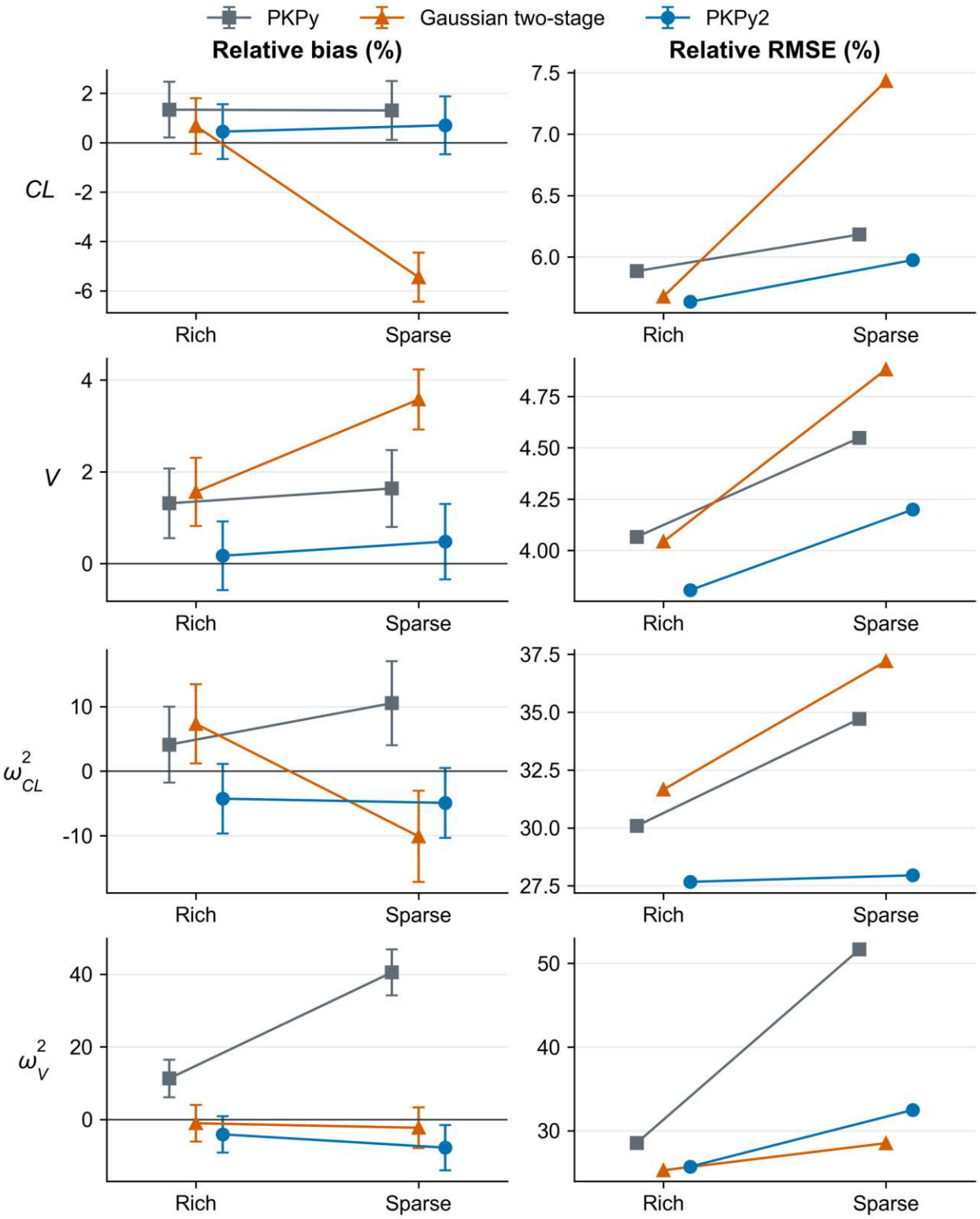


Fig. 2 Parameter recovery in the primary simulation by PKPy, the Gaussian two-stage control, and PKPy2. Left, relative bias with ±1.96 Monte Carlo standard errors; right, relative root mean squared error (RMSE). Each point summarizes 100 datasets (98 for the rich-sampling Gaussian control)

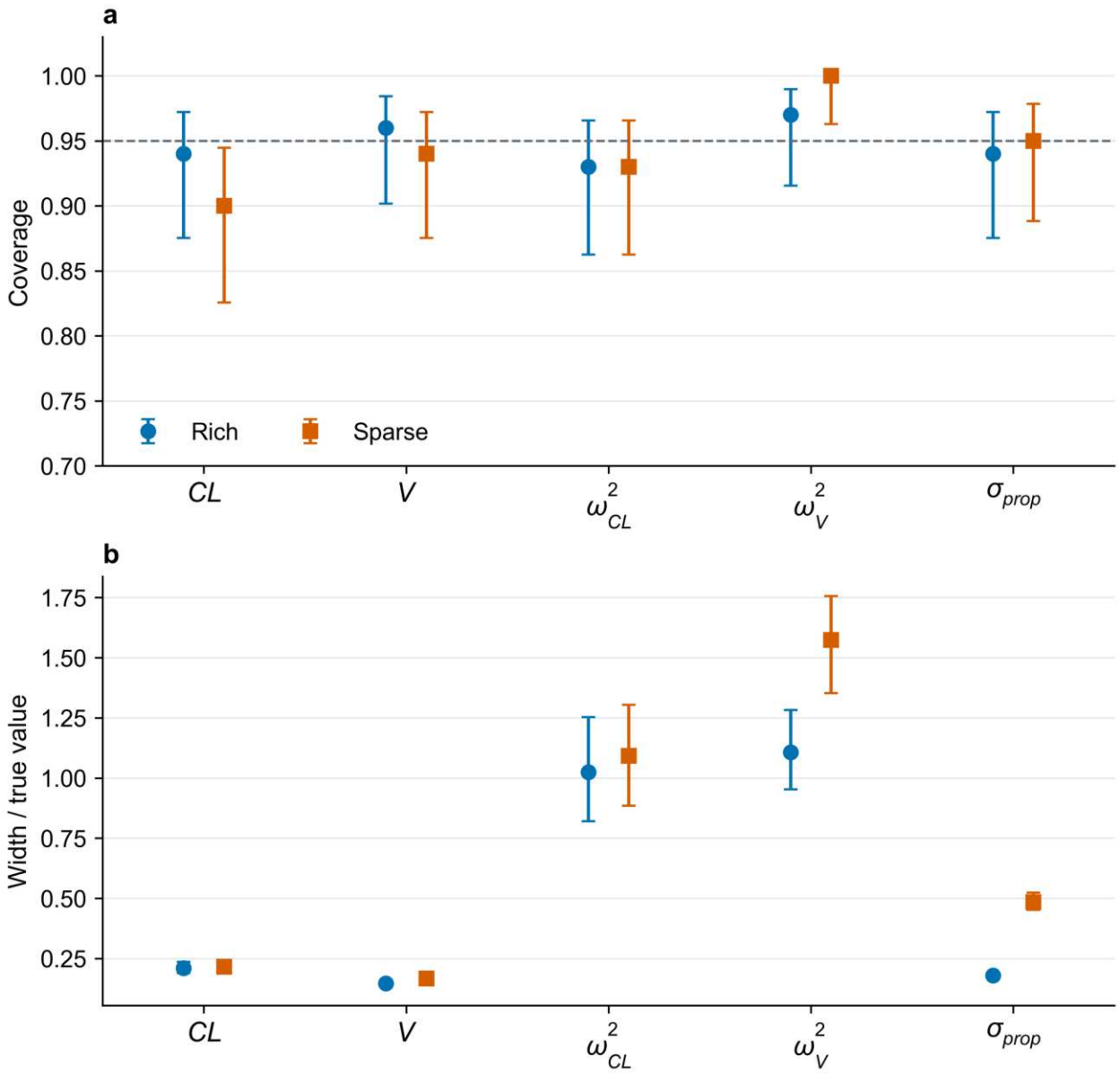


Fig. 3 Coverage (a) and width (b) of PKPy2 95% confidence intervals in the primary simulation. Error bars in (a) are Wilson 95% intervals, and the dashed line marks 95%. Widths are medians relative to the true value, with interquartile ranges

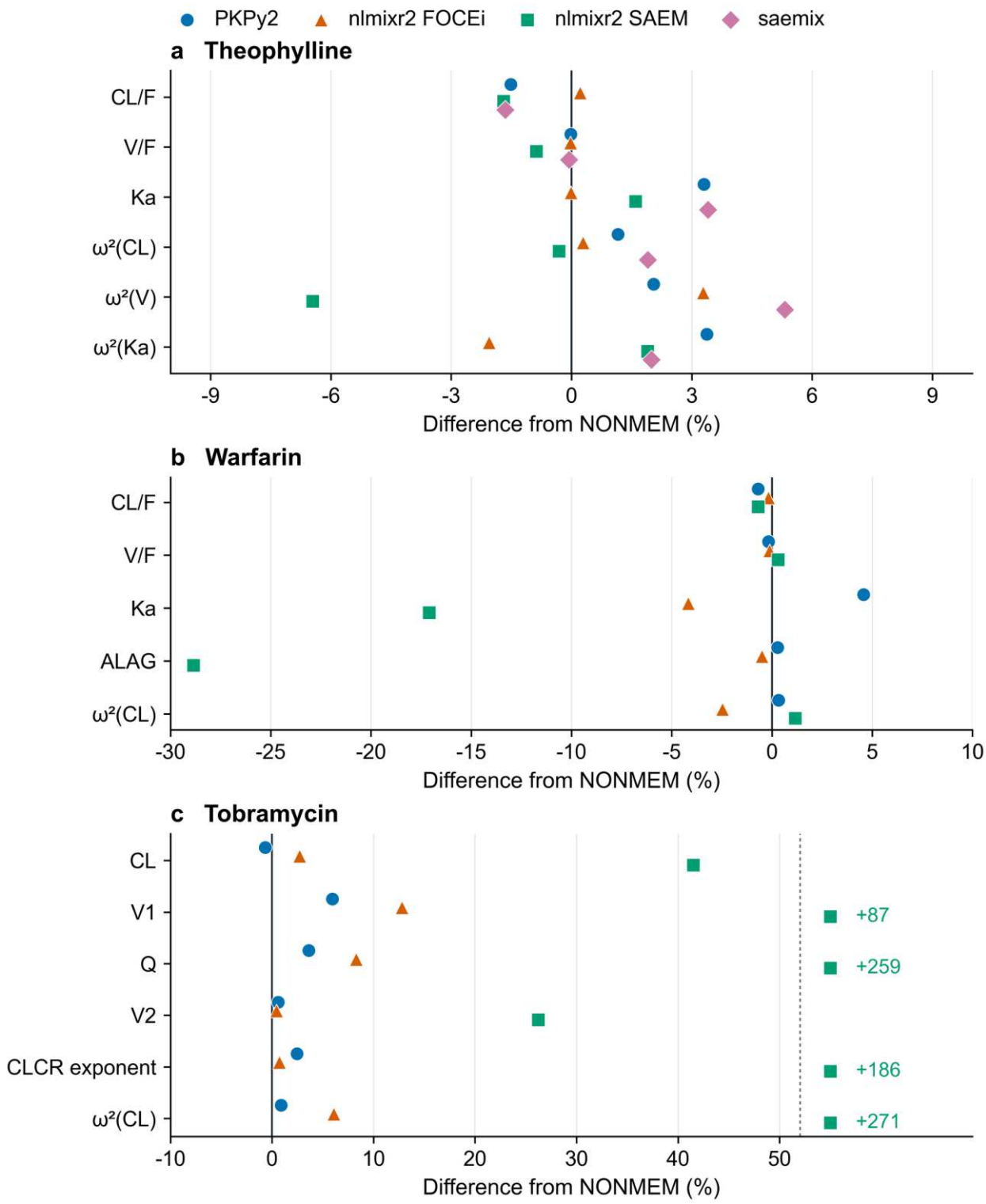


Fig. 4 Differences of the theophylline (a), warfarin (b), and tobramycin (c) estimates from the published expert NONMEM estimates. Tobramycin estimates were obtained with the expert-judgment constraints. Differences larger than 50% are plotted to the right of the dotted line and labeled with their values

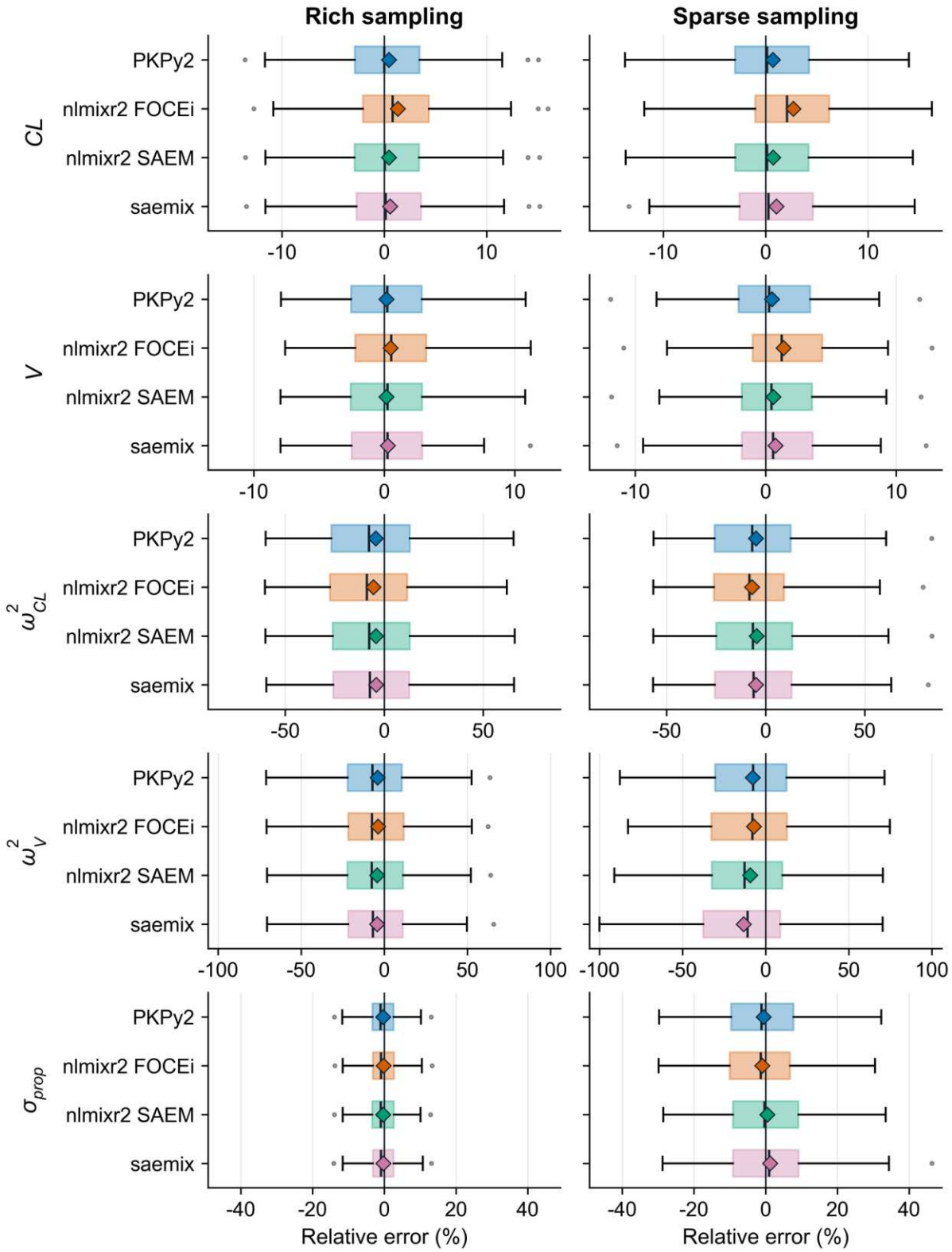


Fig. 5 Relative errors of PKPy2, nlmixr2 FOCEi, nlmixr2 SAEM, and saemix estimates in the 200 primary simulation datasets. Boxes show medians and interquartile ranges, whiskers extend to 1.5 times the interquartile range, and diamonds mark means

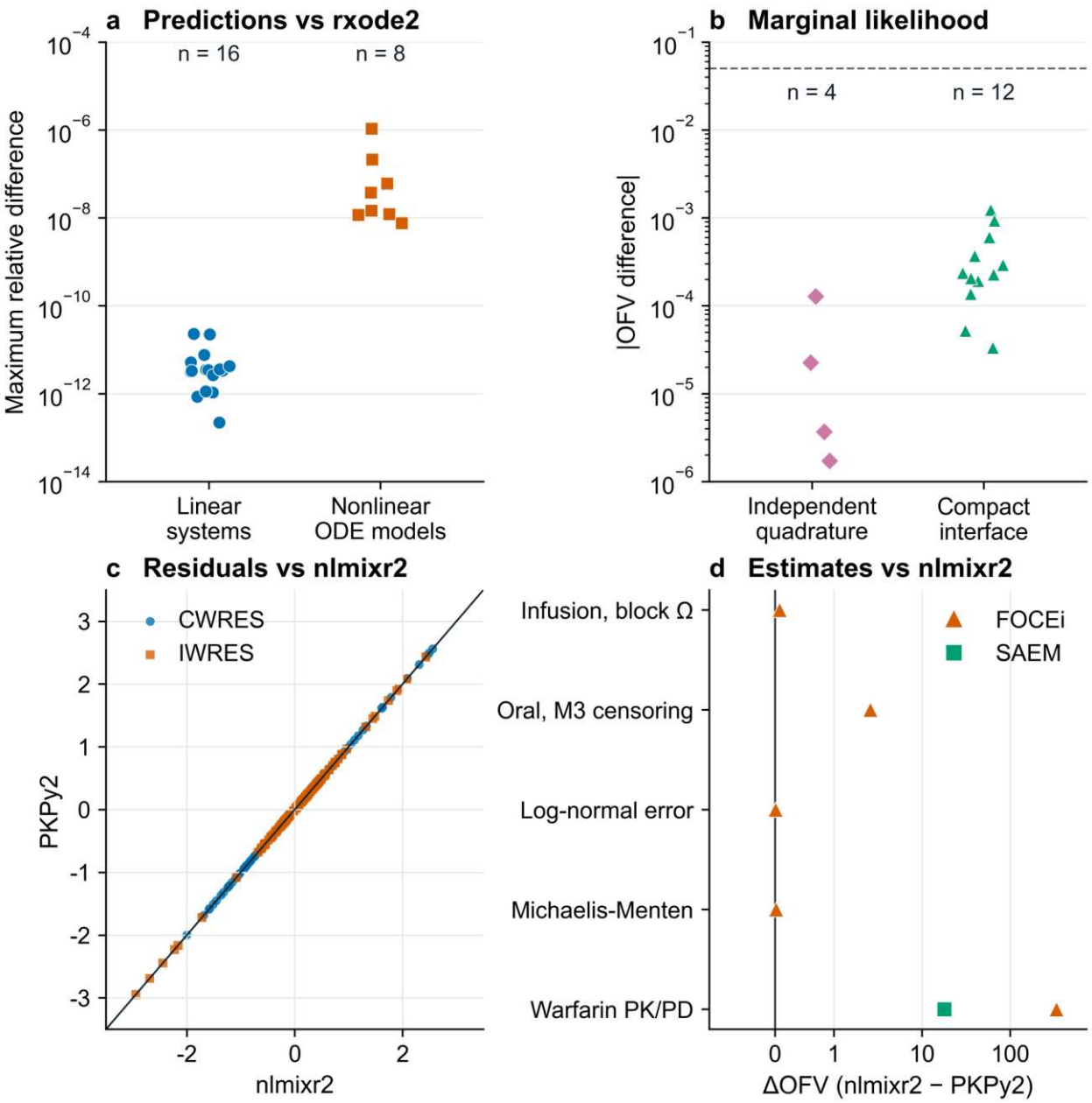


Fig. 6 Verification of the event-record interface. (a) Maximum relative differences of PKPy2 predictions from rxode2 in linear and nonlinear (ODE) scenarios. (b) Absolute OFV differences from an independent quadrature implementation and from the compact interface; the dashed line marks the convergence tolerance of 0.05. (c) CWRES and IWRES of the theophylline model from PKPy2 and nlmixr2 at identical parameters. (d) Exact OFV at the nlmixr2 FOCEi and SAEM estimates minus that at the PKPy2 estimates

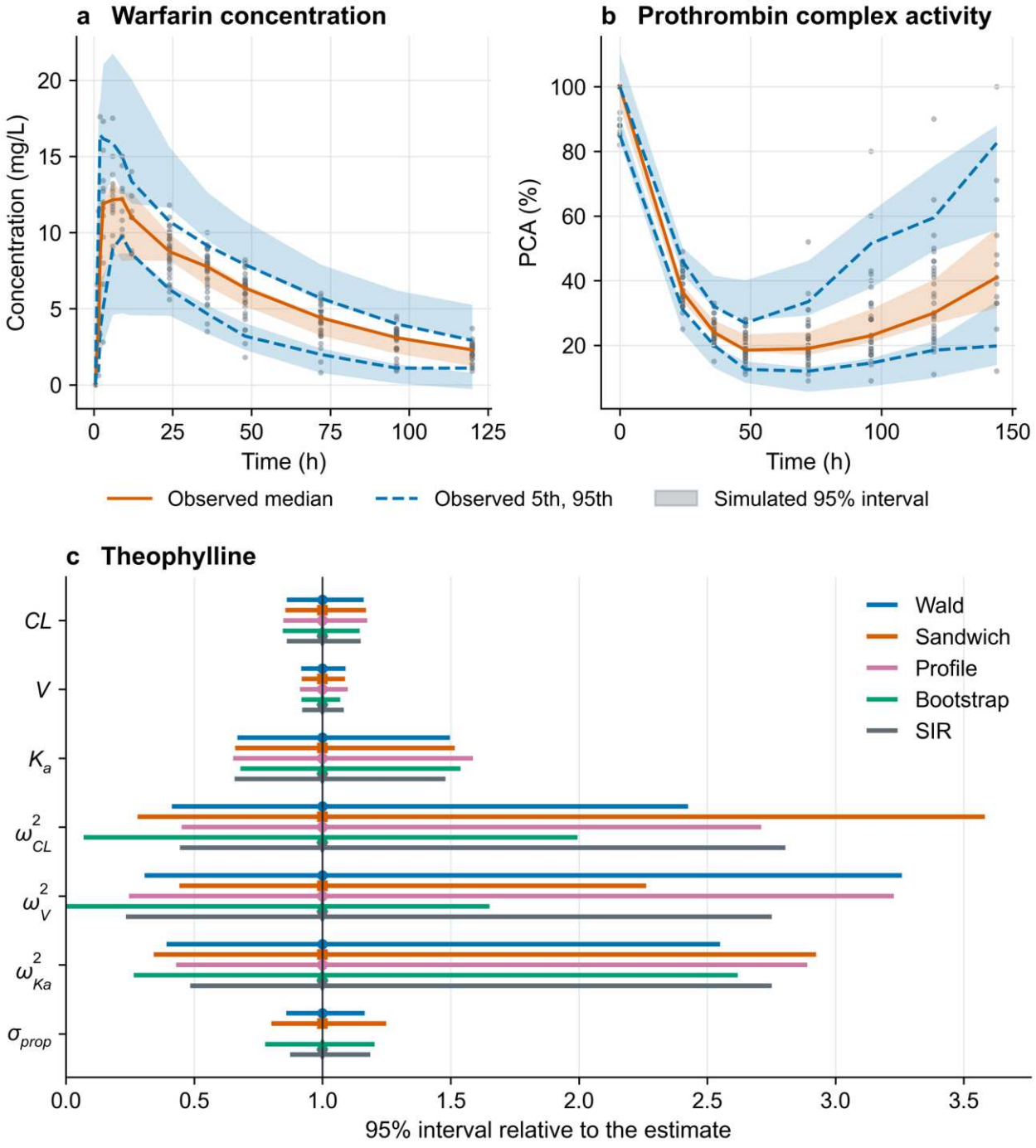


Fig. 7 Warfarin PK/PD application and interval methods. (a, b) Visual predictive checks of plasma concentration and prothrombin complex activity (PCA): observed 5th, 50th, and 95th percentiles (lines), 95% intervals of the simulated percentiles (bands), and observations (points). (c) 95% intervals for the theophylline model from the Wald, sandwich, profile-likelihood, bootstrap, and SIR methods, relative to the estimate

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.xlsx
- ESM3.zip
- ESM4.xlsx