

Online Resource 1

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

Journal: Journal of Pharmacokinetics and Pharmacodynamics

Authors: Hyunseung Kong, Inyoung Kim

Corresponding author: Inyoung Kim, Department of Defense Science, Korea National Defense University, Nonsan, Republic of Korea; inyoungkim@korea.kr

Numerical settings, complete simulation summaries, clinical reference details, the comparison with nlmixr2 and saemix, the tobramycin analyses, and the numerical methods and evaluation of the event-record interface. Notation follows the main article, and reference numbers refer to its reference list.

S1. Protocols and numerical thresholds

The primary experiment contains 200 datasets and the additional-model experiment 30 datasets; their seeds are recorded in the protocols, which were fixed before execution.

Each fit used at most two workers. Estimation began with a deterministic Laplace exploration: the Laplace objective (twice the joint negative log density at the individual modes plus the log determinant of the individual Hessians, normal constants included) was minimized with L-BFGS-B and central-difference gradients from the starting values and from up to seven scrambled-Sobol perturbations of the typical values and covariate coefficients, alternately within $\pm \log 3$ and $\pm \log 10$, within 120 CPU seconds. Declared bounds were respected throughout. Exploration stopped early when the first four starts reached the same optimum within 0.1 OFV units; optima differing by more than this were recorded as distinct. The best Laplace point was then refined with the importance-sampled marginal likelihood: initial refinement within 20 CPU seconds, SAEM when required within 60 CPU seconds, and marginal refinement within 300 CPU seconds and at most 90 stages. Final acceptance required maximum absolute projected score ≤ 0.2 in both independent banks, their OFV difference ≤ 0.05 , and every subject ESS ≥ 100 . Reaching a budget limit was recorded as a separate termination reason.

The integration proposal was an equally weighted mixture of the population Gaussian and a subject-specific Gaussian, with importance weights computed using the full mixture density. Pilot moments used at least 2^{14} samples per subject, increasing to the subject's current integration power when higher. Pilot ESS below 50 or maximum normalized weight above 0.1 triggered verification of a conditional mode and its curvature; the proposal covariance was $2H_{\text{joint}}^{-1} + 0.01\Omega$, where H_{joint} is the Hessian of the conditional negative log density. Marginal refinement began with 2^{12} samples per subject, used L-BFGS stages of at most 40 iterations, and increased integration power for subjects showing integration disagreement, up to power 16. A failed stationarity audit increased all subjects to power 16. Final acceptance used two independent banks, each with $2^{16} = 65,536$ samples per subject, so integration accuracy was assessed with samples independent of those used for optimization.

When required, SAEM used four chains and at most 400 exploration and 400 smoothing iterations within its CPU budget. The stochastic approximation weight was one during exploration and $(m + 2)^{-0.7}$ at smoothing index m beginning at zero. Phase decisions were assessed every 25 iterations with a stability tolerance of 0.01. Eligible log-parameter means used regression updates and diagonal variances used conditional moments; the remaining coordinates, including residual terms, used a generalized L-BFGS-B M-step with at most six iterations, accepted when it improved the complete-data objective. SAEM was always followed by marginal refinement and the independent convergence assessment. The estimation paths are reported in Table S11.

Intervals used two independent banks of power 16, finite-difference steps 0.002 and 0.001, and the same saved estimate. Acceptance required positive information, relative Hessian changes across banks and steps ≤ 0.1 , relative coordinate SE change ≤ 0.2 , difference from the saved objective ≤ 0.05 , ESS ≥ 100 , and relative Hessian asymmetry ≤ 0.1 . RSEs above 50% were flagged as large. Numerical floors were 1e-10 for prediction, 1e-12 for residual variance, and 1e-8 for interindividual variance.

The primary true values were CL = 4 L/h, V = 40 L, $\omega^2(\text{CL}) = 0.09$, $\omega^2(\text{V}) = 0.04$, and $\sigma_{\text{prop}} = 0.15$. Rich times were 0.25, 0.5, 1, 2, 4, 6, 8, 12, 18, and 24 h; sparse times were 1, 8, and 24 h. Each subject received 100 mg. Structural starting values were 80% of truth, initial estimated interindividual variances were 0.08, and initial proportional SD was 0.2. The Gaussian two-stage controls used the known generating residual SD.

S2. Complete primary simulation summaries

Table S1. Returned and numerically accepted population estimates.

Sampling	Method	Parameter	Set	n	Bias (%)	RMSE (%)
Rich	PKPy2	CL	All	100	0.45	5.64
Rich	PKPy2	CL	Accepted	100	0.45	5.64
Rich	PKPy2	V	All	100	0.17	3.81
Rich	PKPy2	V	Accepted	100	0.17	3.81
Rich	PKPy2	$\omega^2(\text{CL})$	All	100	-4.27	27.67
Rich	PKPy2	$\omega^2(\text{CL})$	Accepted	100	-4.27	27.67
Rich	PKPy2	$\omega^2(\text{V})$	All	100	-4.09	25.71
Rich	PKPy2	$\omega^2(\text{V})$	Accepted	100	-4.09	25.71
Rich	PKPy2	σ_{prop}	All	100	-0.31	4.97
Rich	PKPy2	σ_{prop}	Accepted	100	-0.31	4.97
Rich	PKPy	CL	All	100	1.34	5.89
Rich	PKPy	CL	Accepted	100	1.34	5.89
Rich	PKPy	V	All	100	1.31	4.07
Rich	PKPy	V	Accepted	100	1.31	4.07
Rich	PKPy	$\omega^2(\text{CL})$	All	100	4.11	30.09
Rich	PKPy	$\omega^2(\text{CL})$	Accepted	100	4.11	30.09
Rich	PKPy	$\omega^2(\text{V})$	All	100	11.29	28.54
Rich	PKPy	$\omega^2(\text{V})$	Accepted	100	11.29	28.54
Rich	Gaussian TST	CL	All	100	0.76	5.69
Rich	Gaussian TST	CL	Accepted	98	0.67	5.68
Rich	Gaussian TST	V	All	100	1.57	4.02
Rich	Gaussian TST	V	Accepted	98	1.56	4.04
Rich	Gaussian TST	$\omega^2(\text{CL})$	All	100	7.16	31.36
Rich	Gaussian TST	$\omega^2(\text{CL})$	Accepted	98	7.33	31.66
Rich	Gaussian TST	$\omega^2(\text{V})$	All	100	-0.21	26.19
Rich	Gaussian TST	$\omega^2(\text{V})$	Accepted	98	-1.03	25.29
Sparse	PKPy2	CL	All	100	0.70	5.98
Sparse	PKPy2	CL	Accepted	100	0.70	5.98

Sparse	PKPy2	V	All	100	0.48	4.20
Sparse	PKPy2	V	Accepted	100	0.48	4.20
Sparse	PKPy2	$\omega^2(\text{CL})$	All	100	-4.91	27.95
Sparse	PKPy2	$\omega^2(\text{CL})$	Accepted	100	-4.91	27.95
Sparse	PKPy2	$\omega^2(\text{V})$	All	100	-7.75	32.49
Sparse	PKPy2	$\omega^2(\text{V})$	Accepted	100	-7.75	32.49
Sparse	PKPy2	σ_{prop}	All	100	-0.57	12.17
Sparse	PKPy2	σ_{prop}	Accepted	100	-0.57	12.17
Sparse	PKPy	CL	All	100	1.30	6.18
Sparse	PKPy	CL	Accepted	100	1.30	6.18
Sparse	PKPy	V	All	100	1.64	4.55
Sparse	PKPy	V	Accepted	100	1.64	4.55
Sparse	PKPy	$\omega^2(\text{CL})$	All	100	10.53	34.71
Sparse	PKPy	$\omega^2(\text{CL})$	Accepted	100	10.53	34.71
Sparse	PKPy	$\omega^2(\text{V})$	All	100	40.50	51.66
Sparse	PKPy	$\omega^2(\text{V})$	Accepted	100	40.50	51.66
Sparse	Gaussian TST	CL	All	100	-5.45	7.43
Sparse	Gaussian TST	CL	Accepted	100	-5.45	7.43
Sparse	Gaussian TST	V	All	100	3.58	4.88
Sparse	Gaussian TST	V	Accepted	100	3.58	4.88
Sparse	Gaussian TST	$\omega^2(\text{CL})$	All	100	-10.09	37.20
Sparse	Gaussian TST	$\omega^2(\text{CL})$	Accepted	100	-10.09	37.20
Sparse	Gaussian TST	$\omega^2(\text{V})$	All	100	-2.28	28.52
Sparse	Gaussian TST	$\omega^2(\text{V})$	Accepted	100	-2.28	28.52

All, every returned estimate, including unsuccessful optimizer states; Accepted, estimates meeting the method-specific numerical criteria. The Gaussian two-stage control (TST) used the true residual SD.

Table S2. PKPy2 interval coverage and width.

Sampling	Parameter	Available / planned	Covered / available	Wilson 95% interval	Median relative width
Rich	CL	100/100	94/100	87.5-97.2%	0.210
Rich	V	100/100	96/100	90.2-98.4%	0.147
Rich	$\omega^2(\text{CL})$	100/100	93/100	86.3-96.6%	1.025
Rich	$\omega^2(\text{V})$	100/100	97/100	91.5-99.0%	1.107
Rich	σ_{prop}	100/100	94/100	87.5-97.2%	0.180
Sparse	CL	100/100	90/100	82.6-94.5%	0.216
Sparse	V	100/100	94/100	87.5-97.2%	0.167
Sparse	$\omega^2(\text{CL})$	100/100	93/100	86.3-96.6%	1.092
Sparse	$\omega^2(\text{V})$	100/100	100/100	96.3-100.0%	1.574
Sparse	σ_{prop}	100/100	95/100	88.8-97.8%	0.484

Width is divided by the true parameter value.

Table S3. Paired-bootstrap differences in relative RMSE.

Sampling	Comparator	Parameter	Paired n	Difference (pp)	Bootstrap 95% interval (pp)
Rich	PKPy	CL	100	-0.25	-0.42 to -0.07
Rich	Gaussian TST	CL	98	-0.03	-0.21 to 0.16
Rich	PKPy	V	100	-0.26	-0.47 to -0.03
Rich	Gaussian TST	V	98	-0.20	-0.55 to 0.14
Rich	PKPy	$\omega^2(\text{CL})$	100	-2.42	-4.66 to -0.11
Rich	Gaussian TST	$\omega^2(\text{CL})$	98	-3.82	-6.80 to -0.82
Rich	PKPy	$\omega^2(\text{V})$	100	-2.83	-5.60 to 0.10
Rich	Gaussian TST	$\omega^2(\text{V})$	98	-0.21	-2.78 to 2.43
Sparse	PKPy	CL	100	-0.21	-0.35 to -0.06
Sparse	Gaussian TST	CL	100	-1.46	-2.46 to -0.40
Sparse	PKPy	V	100	-0.35	-0.56 to -0.12
Sparse	Gaussian TST	V	100	-0.68	-1.29 to -0.05
Sparse	PKPy	$\omega^2(\text{CL})$	100	-6.76	-12.49 to -1.11
Sparse	Gaussian TST	$\omega^2(\text{CL})$	100	-9.25	-12.22 to -6.27
Sparse	PKPy	$\omega^2(\text{V})$	100	-19.17	-25.52 to -12.23
Sparse	Gaussian TST	$\omega^2(\text{V})$	100	3.97	-0.84 to 8.67

PKPy2 minus comparator in percentage points (pp); negative values favor PKPy2. Intervals are from 10,000 paired bootstrap resamples of datasets accepted by both methods (seed 261001) and are unadjusted for multiplicity.

S3. Additional model structures

Table S4. Additional-model generating values and starting values.

Model	Structural truth	IIV variance truth	Structural start
One-compartment oral	CL = 4; V = 40; Ka = 1.2	CL = 0.09; V = 0.04; Ka = 0.16	CL = 3.2; V = 32; Ka = 0.96
Two-compartment IV	CL = 4; V1 = 25; Q = 8; V2 = 50	CL = 0.09; V1 = 0.04	CL = 3.2; V1 = 20; Q = 6.4; V2 = 40
Two-compartment oral	CL = 4; V1 = 25; Q = 8; V2 = 50; Ka = 1.2	CL = 0.09; V1 = 0.04; Ka = 0.16	CL = 3.2; V1 = 20; Q = 6.4; V2 = 40; Ka = 0.96

Each structure used ten independently generated datasets with 30 subjects, a single 100 mg dose per subject, and samples at 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8, 12, 18, 24, 36, and 48 h. Clearance and intercompartmental clearance are in L/h, volumes in L, and Ka in h⁻¹. IIV entries are diagonal log-parameter variances. Generating proportional SD was 0.15, starting proportional SD was 0.2, and every estimated IIV variance started at 0.08. Additive SD and oral lag were fixed at zero, with no covariates and F = 1. Structural starts were 80% of truth.

Table S5. Additional-model parameter recovery among accepted fits.

Model	Accepted / planned	Parameter	n	Bias (%)	RMSE (%)
1-comp oral	10/10	CL	10	-0.29	3.45
1-comp oral	10/10	V	10	-0.12	3.03
1-comp oral	10/10	Ka	10	-2.22	7.67

1-comp oral	10/10	$\omega^2(\text{CL})$	10	5.27	27.01
1-comp oral	10/10	$\omega^2(\text{V})$	10	0.96	39.21
1-comp oral	10/10	$\omega^2(\text{Ka})$	10	-5.23	30.52
1-comp oral	10/10	σ_{prop}	10	-1.83	3.53
2-comp IV	10/10	CL	10	0.21	4.79
2-comp IV	10/10	V1	10	0.22	2.96
2-comp IV	10/10	Q	10	-0.74	3.03
2-comp IV	10/10	V2	10	-0.49	2.49
2-comp IV	10/10	$\omega^2(\text{CL})$	10	-3.66	19.11
2-comp IV	10/10	$\omega^2(\text{V1})$	10	4.49	31.37
2-comp IV	10/10	σ_{prop}	10	-0.07	2.97
2-comp oral	9/10	CL	9	0.08	5.30
2-comp oral	9/10	V1	9	-9.69	19.81
2-comp oral	9/10	Q	9	5.50	11.97
2-comp oral	9/10	V2	9	5.30	7.60
2-comp oral	9/10	Ka	9	-11.49	20.87
2-comp oral	9/10	$\omega^2(\text{CL})$	9	-2.69	25.14
2-comp oral	9/10	$\omega^2(\text{V1})$	9	13.06	76.01
2-comp oral	9/10	$\omega^2(\text{Ka})$	9	-20.97	29.14
2-comp oral	9/10	σ_{prop}	9	-2.76	6.73

Two-compartment oral replicate 8 did not pass the convergence assessment (TimeoutError: CPU budget exhausted); it is reported as unresolved.

S4. Clinical datasets and expert reference analyses

The expert NONMEM analyses are from the PKGPT study [37]; their published estimates and RSEs are used as reference values.

Table S6. Models of the expert analyses and of PKPy2.

Dataset	Expert NONMEM analysis [37]	PKPy2 analysis
Theophylline	One-compartment first-order absorption; FOCE-I; proportional error; IIV on CL, V and Ka; CL and V per kg; 12 subjects, 120 observations	Same model; CL and V scaled by WT/70
Warfarin	One-compartment first-order absorption with lag; FOCE-I; combined error with fixed additive variance 0.117; IIV on CL, fixed $\omega^2(\text{Ka}) = 0.5$; fixed WT exponents 0.75 (CL) and 1 (V); 32 subjects, 251 observations	Same model
Tobramycin	Two-compartment IV bolus; FOCE-I; combined error with fixed additive variance 1.85×10^{-6} ; IIV on CL; power effects of CLCR on CL (reference 58 mL/min) and of WT on V1 (reference 62 kg); 97 subjects, 236 observations, 858 doses	Same model; fitted without constraints and with the constraints described in S10

The expert theophylline model estimates CL and V per kg using mg/kg doses. For identical η values, multiplying CL and V by individual WT while multiplying dose by the same WT preserves concentrations exactly; PKPy2 implements this with CL70 and V70 and fixed $(\text{WT}/70)^1$ factors. The warfarin data retained all four zero observations. Initial values for CL, V, Ka and ALAG were (0.2, 10, 0.8, 0.5) for the primary warfarin fit and (0.1, 6, 2, 1) for a sensitivity fit; estimated CL variance started at 0.1 and proportional SD at 0.2. The first start was designated as primary before fitting.

Table S7. Estimation and uncertainty in the expert analyses and in PKPy2.

Component	Expert NONMEM [37]	PKPy2
Estimation	FOCE-I	Importance-integrated marginal likelihood
Uncertainty	NONMEM covariance step (RSEs as published)	Inverse observed marginal information, $2H_{\text{OFV}}^{-1}$
Reported quantities	Estimates; RSEs where published	Estimates, SEs, RSEs and 95% intervals (Table S8)

Table S8. PKPy2 clinical standard errors and 95% intervals.

Dataset	Parameter	Estimate	SE	RSE (%)	95% interval
Theophylline	CL/F (L/h)	2.7851	0.213	7.6	2.3976 to 3.2353
Theophylline	V/F (L)	32.544	1.44	4.4	29.845 to 35.488
Theophylline	Ka (h ⁻¹)	1.5082	0.312	20.7	1.0056 to 2.262
Theophylline	$\omega^2(\text{CL})$	0.065348	0.0296	45.3	0.026892 to 0.1588
Theophylline	$\omega^2(\text{V})$	0.014797	0.00891	60.2	0.0045438 to 0.048188
Theophylline	$\omega^2(\text{Ka})$	0.46002	0.221	48.1	0.17938 to 1.1797
Theophylline	σ_{prop}	0.15769	0.0123	7.8	0.13539 to 0.18366
Warfarin	CL/F (L/h)	0.13406	0.00623	4.6	0.12239 to 0.14685
Warfarin	V/F (L)	7.8458	0.136	1.7	7.5829 to 8.1178
Warfarin	Ka (h ⁻¹)	1.2024	0.29	24.1	0.74977 to 1.9284
Warfarin	ALAG (h)	0.82728	0.04	4.8	0.75254 to 0.90945
Warfarin	$\omega^2(\text{CL})$	0.064513	0.0174	27.0	0.038035 to 0.10942
Warfarin	σ_{prop}	0.1366	0.00868	6.4	0.12061 to 0.15471
Tobramycin	CL (L/h)	2.9305	0.403	13.8	2.2376 to 3.838
Tobramycin	V1 (L)	4.8626	3.28	67.5	1.2946 to 18.264
Tobramycin	Q (L/h)	7.0989	2.94	41.4	3.1529 to 15.983
Tobramycin	V2 (L)	13.281	1.77	13.3	10.231 to 17.239
Tobramycin	$\omega^2(\text{CL})$	0.028254	0.0094	33.3	0.014724 to 0.054216
Tobramycin	σ_{prop}	0.21873	0.0129	5.9	0.19485 to 0.24553
Tobramycin	CLCR exponent	0.24182	0.0792	NR	0.086591 to 0.39706

Intervals are log-coordinate Wald intervals transformed to the reported scale. Interindividual effects are reported as variances and residual components as SDs.

Table S9. Theophylline combined-error sensitivity analysis.

Parameter	Estimate	RSE (%)	95% interval
CL/F (L/h)	2.8036	8.4	2.3786 to 3.3045
V/F (L)	31.986	4.6	29.213 to 35.022
Ka (h ⁻¹)	1.5434	19.9	1.0442 to 2.2812
$\omega^2(\text{CL})$	0.071457	48.4	0.027655 to 0.18464
$\omega^2(\text{V})$	0.017094	56.3	0.0056675 to 0.051559
$\omega^2(\text{Ka})$	0.42717	47.8	0.16732 to 1.0905
σ_{prop}	0.074734	40.9	0.033509 to 0.16668
σ_{add} (mg/L)	0.58899	20.2	0.39654 to 0.87484

The combined-error model is a sensitivity analysis. Warfarin fits took 30.85 s and 26.58 s from the two starts, and the primary uncertainty assessment took 59.51 s; estimates from the two starts differed by at most 0.014% and their OFVs by 0.000106. Tobramycin estimates in Table S8 are those obtained with the expert-judgment constraints (S10).

S5. Independent numerical calculations and prediction timing

The one-random-effect quadrature comparisons optimized with 4,097 nodes and were checked with 8,193 nodes. The independent code has its own prediction, likelihood, and finite-difference Hessian calculations and includes the Gaussian density constants.

Table S10. Warm prediction-call microbenchmark.

Model	Doses	Observations	Direct (μs)	Recurrence (μs)	Ratio
1cmt_iv	1	64	1.26	1.30	0.97
1cmt_iv	10	64	2.85	1.33	2.13
1cmt_iv	100	300	82.93	3.36	24.66
1cmt_iv	1000	3000	10267.80	23.33	440.05
1cmt_oral	1	64	1.58	1.63	0.97
1cmt_oral	10	64	4.82	2.53	1.91
1cmt_oral	100	300	199.23	9.94	20.04
1cmt_oral	1000	3000	42097.37	85.40	492.94
2cmt_iv	1	64	2.59	2.64	0.98
2cmt_iv	10	64	3.77	1.95	1.94
2cmt_iv	100	300	147.53	5.61	26.31
2cmt_iv	1000	3000	34506.20	38.07	906.47

Medians of seven alternating-order batches after compilation. With a single dose, recurrence was slightly slower than direct summation.

S6. Computational environment and estimation paths

PKPy was run with its numerical model, objective, and optimizer unchanged; an unused plotting import was stubbed because its optional dependency was absent. The comparison covers PKPy's individual fitting and population summaries.

The calculation host was Windows 11, AMD Ryzen 5 5600 (6 cores, 12 threads), Python 3.13.5, NumPy 2.2.6, SciPy 1.16.1, Numba 0.65.1, and threadpoolctl 3.6.0. Each saved result records data fingerprints, numerical settings, and seeds, and can be restored only with matching input data.

Table S11. Estimation paths in the simulation experiments.

Experiment	Condition	Path after Laplace exploration	Final status	n
Additional	One-compartment oral	Initial refinement	Converged	10
Additional	Two-compartment IV	Initial refinement	Converged	10
Additional	Two-compartment oral	SAEM then refinement	Converged	9
Additional	Two-compartment oral	SAEM then refinement	Partial	1
Primary	Rich	Initial refinement	Converged	100
Primary	Sparse	Initial refinement	Converged	100

Counts refer to the 200 primary and 30 additional datasets. The Laplace exploration reached more

than one distinct optimum in 36 of these 230 fits.

S7. Executable use of the two interfaces

Compact interface. The accompanying examples/pkpy2_explicit_model.py creates synthetic observations for 30 subjects. After installing PKPy2 with Python 3.13, run the example with an unused output directory. It declares fixed volume and a fixed weight exponent for illustration, estimates clearance, clearance variability, and proportional residual SD, inspects the fit and uncertainty states, and restores the saved result with the same data. The model and result-handling core is shown in Listing 1; data generation and assertions are included in the complete executable file.

Listing 1. Model declaration and numerical result handling (excerpt).

```
from pkpy2 import ModelSpec, Parameter as P, Covariate
from pkpy2 import fit, load_fit

specification = ModelSpec(
    'lcmt_iv',
    theta={'CL': P(3.), 'V': P(40., fixed=True)},
    omega={'CL': P(.08)},
    sigma_prop=P(.2), sigma_add=P(0., fixed=True),
    covariates=(Covariate('CL', 'WT', 70.,
                          P(.75, fixed=True))),
)
result = fit(subjects, specification, seed=20260915, workers=2,
             saem_options={'cpu_budget_seconds': 60.},
             refinement_options={'cpu_budget_seconds': 300.,
                                'max_stages': 90, 'analytic_non_eta': True})
print(result.status, result.audit.get('passed'))
if result.converged:
    report = result.uncertainty(seed=20260916, workers=2,
                               power=16, step=.002)
    print(report['status'])
result.save(destination)
restored = load_fit(destination, subjects)
```

Here, subjects holds the generated Subject records and destination is a new JSON path. The complete example checks that fixed terms are preserved and that estimates, diagnostics, and intervals are restored.

Event-record interface. The accompanying examples/theophylline_diagnostics.py reads the NONMEM-format theophylline data of the main article (paper/data/theophylline.csv in the PKPy2 repository), fits the theophylline model through the event-record interface, and draws diagnostic plots with the plotting module (Listing 2 and Fig. S1). Its estimates differed from those of the compact-interface analysis by at most 0.32% and its OFV by 0.0002.

Listing 2. Event-record analysis with diagnostics and plots (excerpt).

```
import pkpy2
from pkpy2 import Model, Residual, Covariate, Parameter as P, structures as S, plots

data = pkpy2.read_nonmem("paper/data/theophylline.csv", covariates=["WT"])
model = Model(
    S.pk(1, "first_order"), # depot (CMT 1), central (CMT 2)
    theta={'CL': P(3.), "V": P(30.), "Ka": P(1.)}, # at 70 kg
    omega={'CL': P(.1), "V": P(.03), "Ka": P(.3)},
    covariates=(Covariate("CL", "WT", 70., P(1., fixed=True)),
                Covariate("V", "WT", 70., P(1., fixed=True))),
    residual=Residual(proportional=P(.15)),
)
result = pkpy2.fit(data, model, seed=20260930)
table, summary = pkpy2.diagnostics(result) # PRED, IPRED, CWRES, NPDE
plots.gof(table).savefig("gof.png")
plots.individual_fits(table).savefig("individual_fits.png")
plots.vpc(pkpy2.vpc(result)).savefig("vpc.png")
```

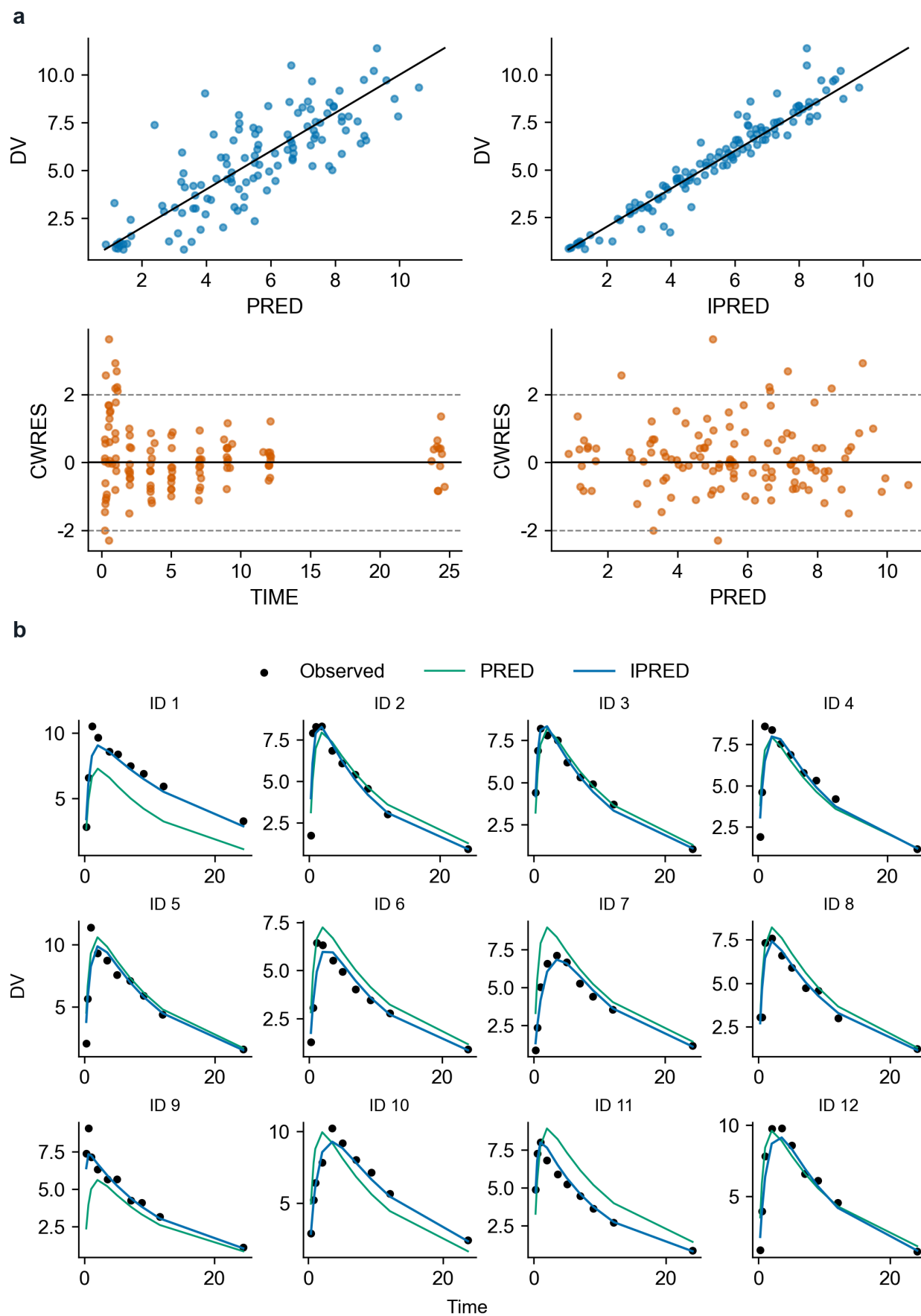


Fig. S1. Diagnostic plots of the theophylline analysis drawn by the plotting module (Listing 2). (a) Observations against population (PRED) and individual (IPRED) predictions, and CWRES against time and PRED. (b) Observations, PRED, and IPRED of the 12 subjects.

S8. Comparison with nlmixr2 and saemix

This section describes the comparison with nlmixr2 and saemix (Tables 5 and 6, Figs. 4 and 5).

Data. The observation records analysed by PKPy2 were exported unchanged to NONMEM-style event files (ID, TIME, AMT, DV, EVID, MDV, CMT, and covariates) for the 200 primary simulation datasets and the clinical datasets. Oral doses were placed in the depot compartment. A manifest of SHA-256 hashes links each exported file to its PKPy2 input.

Models and starting values. All programs used the models, fixed terms, and starting values of the PKPy2 analyses (S1 and S4). The residual variance was $\sigma_{\text{prop}}^2 f^2 + \sigma_{\text{add}}^2$ in every program (nlmixr2 addProp = "combined2").

Program settings. nlmixr2 7.0.1 (nlmixr2est 7.1.0, rxode2 5.1.7) was run with FOCEi and SAEM using default estimation and covariance settings (FOCEi: sandwich "r,s"; SAEM: stochastic approximation "sa", seed 260923). The simulation and theophylline models used analytical linear-compartment solutions, and the warfarin model used ordinary differential equations for the absorption lag. saemix 3.5 was run with default iterations, seed 260923, and standard errors from the linearized Fisher information; it was not applied to warfarin, whose model fixes nonzero variance components. For tobramycin, nlmixr2 used the expert-judgment constraints of S10; nlmixr2 SAEM does not apply parameter bounds, and saemix, which does not support bounds, was not applied. Tobramycin fits without constraints are reported in S10. All fits used R version 4.5.3 (2026-03-11 ucrt) on the computer described in S6, with one rxode2 thread per fit and five concurrent processes. All 610 external fits completed without errors.

Metrics. Relative bias, relative RMSE, Monte Carlo standard errors, Wilson intervals, and paired-bootstrap differences (10,000 resamples of datasets) were computed as in the primary analysis. Wald 95% intervals for CL and V were formed on the log scale from the reported standard errors (for saemix, SE divided by the estimate).

Exact marginal likelihood at the clinical estimates. Because the programs report objective functions under different approximations and constants, each set of clinical estimates was evaluated with the exact marginal OFV (normal constants included), computed by adaptive Gauss-Hermite quadrature in a separate Python implementation that shares no code with PKPy2. For each subject, the conditional mode was found by BFGS, the Hessian by central finite differences, and a product Gauss-Hermite rule was centred and scaled by the mode and the Cholesky factor of the inverse Hessian. For the published NONMEM estimates, which do not include σ_{prop} , σ_{prop} was set to its maximum-likelihood value. At the PKPy2 estimates, this calculation reproduced the OFV reported by PKPy2 (352.785 vs 352.783 for theophylline; 755.471 vs 755.471 for warfarin).

Table S12. Paired-bootstrap differences in relative RMSE between PKPy2 and nlmixr2 or saemix.

Sampling	Parameter	Comparator	Pairs	PKPy2 – comparator (pp)	Bootstrap 95% interval
Rich	CL	nlmixr2 FOCEi	100	-0.16	-0.32 to 0.01
Rich	V	nlmixr2 FOCEi	100	-0.03	-0.09 to 0.03
Rich	$\omega^2(\text{CL})$	nlmixr2 FOCEi	100	0.25	-0.07 to 0.58
Rich	$\omega^2(\text{V})$	nlmixr2 FOCEi	100	0.20	0.01 to 0.41
Rich	σ_{prop}	nlmixr2 FOCEi	100	-0.02	-0.05 to 0.01
Rich	CL	nlmixr2 SAEM	100	-0.01	-0.02 to 0.00
Rich	V	nlmixr2 SAEM	100	0.00	-0.01 to 0.01
Rich	$\omega^2(\text{CL})$	nlmixr2 SAEM	100	0.01	-0.08 to 0.11
Rich	$\omega^2(\text{V})$	nlmixr2 SAEM	100	0.05	-0.10 to 0.19

Rich	σ_{prop}	nlmixr2 SAEM	100	0.02	-0.01 to 0.04
Rich	CL	saemix	100	-0.02	-0.04 to 0.01
Rich	V	saemix	100	-0.02	-0.05 to 0.01
Rich	$\omega^2(\text{CL})$	saemix	100	-0.17	-0.37 to 0.02
Rich	$\omega^2(\text{V})$	saemix	100	-0.05	-0.36 to 0.26
Rich	σ_{prop}	saemix	100	0.00	-0.05 to 0.04
Sparse	CL	nlmixr2 FOCEi	100	-0.58	-0.92 to -0.21
Sparse	V	nlmixr2 FOCEi	100	-0.24	-0.40 to -0.07
Sparse	$\omega^2(\text{CL})$	nlmixr2 FOCEi	100	0.42	-0.10 to 0.95
Sparse	$\omega^2(\text{V})$	nlmixr2 FOCEi	100	0.18	-0.29 to 0.65
Sparse	σ_{prop}	nlmixr2 FOCEi	100	0.06	-0.10 to 0.23
Sparse	CL	nlmixr2 SAEM	100	-0.02	-0.04 to 0.01
Sparse	V	nlmixr2 SAEM	100	0.00	-0.04 to 0.05
Sparse	$\omega^2(\text{CL})$	nlmixr2 SAEM	100	-0.17	-0.35 to 0.00
Sparse	$\omega^2(\text{V})$	nlmixr2 SAEM	100	-1.17	-2.03 to -0.33
Sparse	σ_{prop}	nlmixr2 SAEM	100	-0.40	-0.70 to -0.09
Sparse	CL	saemix	100	-0.08	-0.16 to 0.01
Sparse	V	saemix	100	-0.08	-0.22 to 0.06
Sparse	$\omega^2(\text{CL})$	saemix	100	-0.19	-0.55 to 0.16
Sparse	$\omega^2(\text{V})$	saemix	100	-5.29	-8.14 to -2.53
Sparse	σ_{prop}	saemix	100	-1.63	-2.76 to -0.57

Negative values favor PKPy2. pp, percentage points.

Table S13. Relative standard errors (%) of the clinical estimates.

Dataset	Parameter	NONMEM	PKPy2	nlmixr2 FOCEi	nlmixr2 SAEM	saemix
Theophylline	CL/F (L/h)	8.0	7.6	8.2	7.5	7.7
Theophylline	V/F (L)	4.3	4.4	4.3	3.9	4.4
Theophylline	Ka (h ⁻¹)	21.4	20.7	1.3	20.1	20.5
Theophylline	$\omega^2(\text{CL})$	NR	45.3	NR	NR	NR
Theophylline	$\omega^2(\text{V})$	42.2	60.2	NR	NR	NR
Theophylline	$\omega^2(\text{Ka})$	NR	48.1	NR	NR	NR
Theophylline	σ_{prop}	NR	7.8	11.6	7.7	7.7
Warfarin	CL/F (L/h)	4.8	4.6	4.8	4.6	n/a
Warfarin	V/F (L)	3.0	1.7	3.1	2.2	n/a
Warfarin	Ka (h ⁻¹)	30.5	24.1	1.4	17.9	n/a
Warfarin	ALAG (h)	8.6	4.8	3.2	13.7	n/a
Warfarin	$\omega^2(\text{CL})$	37.3	27.0	NR	NR	n/a
Warfarin	σ_{prop}	NR	6.4	11.8	7.1	n/a
Tobramycin	CL (L/h)	NR	13.8	5.3	12.9	NR
Tobramycin	V1 (L)	NR	67.5	13.8	229.5	NR
Tobramycin	Q (L/h)	NR	41.4	13.5	261.6	NR

Tobramycin	V2 (L)	NR	13.3	19.7	104.3	NR
Tobramycin	CLCR exponent	NR	32.8	29.2	19.8	NR
Tobramycin	$\omega^2(\text{CL})$	NR	33.3	NR	NR	NR
Tobramycin	σ_{prop}	NR	5.9	11.8	6.0	NR

NR, not reported; n/a, not applied. nlmixr2 RSEs of structural parameters are SEs of log-scale parameters.

Table S14. Exact marginal OFV at the clinical estimates of each program.

Dataset	Estimates from	OFV	$\Delta\text{OFV vs PKPy2}$	Node-refinement change
Theophylline	PKPy2	352.785	0.000	8.6e-03
Theophylline	saemix	352.802	0.017	8.3e-03
Theophylline	nlmixr2 SAEM	352.848	0.063	8.1e-03
Theophylline	NONMEM (σ_{prop} profiled)	352.855	0.070	8.0e-03
Theophylline	nlmixr2 FOCEi	352.873	0.088	7.9e-03
Tobramycin	PKPy2	604.444	0.000	1.2e-08
Tobramycin	NONMEM (σ_{prop} profiled)	605.062	0.618	1.2e-08
Tobramycin	nlmixr2 FOCEi	605.168	0.724	1.4e-08
Tobramycin	nlmixr2 SAEM	673.074	68.630	7.7e-07
Warfarin	PKPy2	755.471	0.000	1.0e-02
Warfarin	NONMEM (σ_{prop} profiled)	755.538	0.067	6.2e-03
Warfarin	nlmixr2 FOCEi	755.632	0.161	1.9e-03
Warfarin	nlmixr2 SAEM	771.901	16.430	8.0e-03

Lower OFV indicates higher marginal likelihood.

S9. Raw data files and codebook

The raw-data workbook (Online Resource 2) contains a README sheet and 19 data sheets, also provided as CSV files (Online Resource 3). S1 and S3 hold the simulated concentration-time records of the 200 primary and 30 additional-structure datasets with the true individual random effects. S2 and S4 hold per-dataset estimates, 95% intervals, coverage indicators, and fit times for every program. S5 and S6 hold the clinical analysis datasets and the clinical estimates of all programs. S7 to S9 hold the prediction checks, quadrature comparisons, and per-batch prediction timings. S10 to S18 hold the evaluation of the event-record interface (S12), including the simulated comparison datasets and the warfarin PK/PD dataset in NONMEM event format, and S19 holds the diagnostics plotted in Fig. S1. The codebook (Online Resource 4) defines every variable, unit, and categorical code, including the numerically coded EVID, MDV, CMT, and 0/1 indicators.

S10. Tobramycin: fits without and with expert-judgment constraints

Data and model. The tobramycin dataset of the PKGPT repository [37] contains 97 patients, 858 intravenous bolus doses, and 236 concentrations (2 to 6 per patient). Of these concentrations, 205 (87%) were measured 2.5 h after the preceding dose. Body weight ranged from 50 to 75 kg and was correlated with creatinine clearance (Pearson correlation of the logarithms, 0.44). The expert model (Table S6) is a two-compartment model with $\text{CL} = \theta_1(\text{CLCR}/58)^{\theta_2} \exp(\eta)$, $\text{V1} = \theta_3(\text{WT}/62)^{\theta_4}$, $\text{Q} = \theta_5$ and $\text{V2} = \theta_6$. The documented initial values of the expert control stream were CL 3.4 L/h, V1 20.3 L, Q 5 L/h, V2 20 L, and both exponents 1, with $\omega^2(\text{CL})$ 0.1; σ_{prop} started at 0.2. Each setting was fitted from these values (with V1 = 8 L when V1 was restricted to 10 L) and from the expert estimates.

Settings. (1) The published model without bounds. (2) The \$THETA bounds of the expert control stream: CL (0, 20) L/h, CLCR exponent (0, 2), V1 (0, 100) L, WT exponent (0, 2), Q (0, 50) L/h, V2 (0, 100) L. (3) The WT exponent fixed at 1 with $V2 \leq 30$ L and the other expert bounds. (4) The expert-judgment constraints used in the main comparison: V1 proportional to body weight (WT 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 other expert bounds.

Table S15. PKPy2 tobramycin estimates under each setting.

Setting	Constraints	Start	OFV	CL	V1	Q	V2	CLCR exp.	WT exp.	$\omega^2(\text{CL})$	Laplace optima
1	None	Documented	560.06	2.33	18.08	2.07	188.0	0.431	-5.60	0.0218	6
		Expert estimates	560.06	2.33	18.08	2.07	188.0	0.431	-5.60	0.0218	6
2	Expert NONMEM bounds	Documented	586.21	2.94	19.51	1.64	100.0	0.363	0.00	0.0481	3
		Expert estimates	604.44	2.91	4.73	6.97	13.3	0.243	1.11	0.0280	4
3	WT exp. = 1; $V2 \leq 30$ L	Documented	603.81	3.49	19.29	0.93	30.0	0.272	1.00	0.0500	1
		Expert estimates	604.44	2.93	4.87	7.10	13.3	0.242	1.00	0.0283	3
4	Expert judgment	Documented	604.44	2.93	4.86	7.10	13.3	0.242	1.00	0.0283	1
		Expert estimates	604.44	2.93	4.86	7.10	13.3	0.242	1.00	0.0283	3
	Expert NONMEM estimates [37]		605.06	2.95	4.59	6.85	13.2	0.236	1.07	0.0280	

CL and Q in L/h, volumes in L, at CLCR 58 mL/min and WT 62 kg. OFV, exact marginal $-2 \log L$ reported by PKPy2. For the expert NONMEM estimates, the OFV was computed by the independent quadrature of S8 with σ_{prop} at its maximum-likelihood value. Laplace optima, number of distinct optima found by the Laplace exploration. All fits met the convergence criteria.

Findings. Without constraints, both starts converged to the same maximum-likelihood estimate (OFV 560.06), with a WT exponent of -5.6 and V2 of 188 L. The independent 60-node Gauss-Hermite calculation used for the profile in Table S16 gave 560.061 at this estimate (PKPy2, 560.061). Across V2 from 5 to 600 L, the profile OFV varied by 18.6 units, and every profile point had a negative WT exponent. With the expert bounds, the best solution lay on the bounds (WT exponent 0.00, V2 100 L), and the fit started from the expert estimates stayed near them with an OFV 18.2 units higher. The nlmixr2 FOCEi objective confirmed this ordering: 155.6 at the bounded optimum and 173.8 at the solution near the expert estimates (published NONMEM OFV, 173.11); an nlmixr2 FOCEi fit with the expert bounds from the documented initial values stopped at a third solution (V1 10.3 L, V2 10.7 L, WT exponent 0.00; objective 172.9), 17.3 units above the FOCEi objective at the bounded PKPy2 optimum. Fixing the WT exponent and limiting V2 (setting 3) was not sufficient from the documented initial values, which reached V1 19.3 L with V2 on its bound. With the expert-judgment constraints (setting 4), both starts reached the same estimate, close to the expert estimates, with no constraint active; its local 95% intervals are given in Table S8.

Table S16. Profile likelihood in V2 for the published model without constraints.

V2 (L)	OFV	CL	V1	Q	CLCR exp.	WT exp.	$\omega^2(\text{CL})$
5	574.63	3.36	17.04	3.83	0.318	-7.11	0.0201
7	573.04	3.26	16.22	5.26	0.340	-7.34	0.0193
10	575.44	3.19	14.33	8.17	0.374	-7.77	0.0207
13.2	578.65	3.29	19.56	1.97	0.335	-5.78	0.0210
20	577.68	3.32	19.14	1.38	0.334	-5.73	0.0223
30	574.43	3.24	18.87	1.27	0.337	-5.74	0.0222
50	569.06	3.07	18.60	1.39	0.348	-5.70	0.0213
100	562.41	2.73	18.29	1.70	0.378	-5.63	0.0205
188	560.06	2.33	18.08	2.07	0.431	-5.60	0.0218
300	561.24	1.96	17.96	2.37	0.497	-5.63	0.0259
600	566.17	1.30	17.82	2.94	0.698	-5.73	0.0460

Independent calculation: 60-node Gauss-Hermite quadrature for the CL random effect, closed-form two-compartment predictions, and Nelder-Mead optimization of the other parameters from three starts at each V2.

Table S17. nlmixr2 and saemix tobramycin estimates.

Model	Program	CL	V1	Q	V2	CLCR exp.	WT exp.	$\omega^2(\text{CL})$
Published model, no bounds	nlmixr2 FOCEi	3.30	16.19	5.18	7.0	0.336	-7.37	0.0186
Published model, no bounds	nlmixr2 SAEM	3.41	23.45	1.61	16.9	0.342	-0.82	0.0353
Published model, no bounds	saemix	3.19	15.91	4.73	7.4	0.315	-7.62	0.0145
Expert judgment	nlmixr2 FOCEi	3.03	5.18	7.42	13.3	0.238	1 (fixed)	0.0297
Expert judgment	nlmixr2 SAEM	4.17	8.60	24.61	16.7	0.676	1 (fixed)	0.1038

Without constraints, saemix used a proportional error model because it cannot fix the additive term. With the expert-judgment constraints, nlmixr2 SAEM applied the fixed WT exponent but not the bounds, and saemix was not applied.

S11. Event-record interface: numerical methods

Records and events. Records are ordered by time within subject (data order at equal times) after additional doses (ADDL) are expanded at $\text{TIME} + k \cdot \text{II}$. Dose records (EVID 1 and 4) are converted to events: a bolus at the record time plus the lag time, or an infusion start and end, with the amount multiplied by the bioavailability; for $\text{RATE} > 0$ the rate is kept and the duration changes (NONMEM convention), $\text{RATE} = -1$ takes the rate and $\text{RATE} = -2$ the duration from model parameters. EVID 3 resets all compartments, EVID 4 resets and doses. $\text{SS} = 1$ replaces the state by the steady state reached with the record's dose and interval before the dose is given, $\text{SS} = 2$ adds that steady state, and SS with $\text{II} = 0$ and $\text{RATE} > 0$ sets the steady state of a constant infusion, which then stops. Parameters and covariates change at their records (last observation carried forward).

Linear systems. The system matrix $A(\theta)$ is assembled from the declared transfer terms. Between events, $x(t + \Delta) = \exp(A\Delta) x(t) + J(\Delta)$, where $J(\Delta)$ is the integral of $\exp(As) r$ over $0 \leq s \leq \Delta$ for the active infusion rates r . When A has real eigenvalues separated by more than 10^{-6} in relative terms and an eigenvector basis with 1-norm condition number below 10^6 , propagation uses $A = V\Lambda V^{-1}$, with the integral $\phi(\lambda, \Delta) = (\exp(\lambda\Delta) - 1)/\lambda$ per eigenvalue; otherwise the matrix exponential is computed by (6,6) Padé approximation with scaling and squaring [18], and the infusion integral from the

exponential of an augmented block matrix. Steady states use closed forms, with $\mathbf{u}$ the unit vector of the dosed compartment: for a bolus of amount a with interval τ and lag L , $\mathbf{x} = (\mathbf{I} - \exp(\mathbf{A}\tau))^{-1} \exp(\mathbf{A}(\tau - L)) \mathbf{u} a$; for an infusion of duration D , the state at the start of an infusion is $\mathbf{z} = (\mathbf{I} - \exp(\mathbf{A}\tau))^{-1} \exp(\mathbf{A}(\tau - D)) \mathbf{J}(D)$, including infusions that extend beyond the interval; for a constant infusion at rate ρ , $\mathbf{x} = -\mathbf{A}^{-1} \mathbf{u} \rho$.

Nonlinear systems. ODEs are integrated with the Dormand-Prince 5(4) method with first-same-as-last stages and adaptive steps (relative tolerance 10^{-8} , absolute tolerance 10^{-10}) [19]; when the step count exceeds 100,000, the interval is recomputed with a linearly implicit Rosenbrock method of order 2(3) and a finite-difference Jacobian [20]. Steady states are obtained by repeating the dosing interval until the relative change of the state is below 10^{-9} (at most 1,000 intervals). Linear PK parts of ODE models (indirect-response, Michaelis-Menten and TMDD models) use the same transfer terms as the linear library.

Statistical model. For parameter k of subject i at record r , $\psi_{ik}(r) = g^{-1}(\theta_k + \Sigma \text{ effects} + \eta_{ik} + \kappa_{ik, \text{occ}}(r))$, where g is the log, logit or identity function and $\theta_k = g(\theta_k)$. Covariate effects are $\beta \log(z/c)$ (power), $\beta(z - c)$ (exponential), $\log(1 + \beta(z - c))$ (linear) and $\beta \cdot 1[z = \text{level}]$ (categorical). Ω is parameterized by the Cholesky factor of each declared block (log diagonal) and interoccasion variances by log SDs. The combined residual variance is $\sigma_{\text{prop}}^2 f^2 + \sigma_{\text{add}}^2$; for log-normal outputs $\log y \sim N(\log f, \sigma^2)$, and the OFV is evaluated on the log scale, as in log-transform-both-sides analyses. A censored observation contributes $\Phi((\text{LLOQ} - f)/\text{SD})$ (on the log scale for log-normal outputs); $\text{CENS} = -1$ gives the upper tail and LIMIT the other end of an interval.

Estimation. Importance particles are drawn in the space of the transformed individual parameters, $\phi_i = \mu_i + \eta_i$, so that typical values and time-invariant covariate effects of parameters with interindividual variability, variances, and residual SDs change only the importance weights; their score contributions are analytic. For the remaining coordinates (typical values without random effects and time-varying covariate effects), scores are central differences (step 10^{-4}) of the objective with fixed particles; a time-varying effect on a parameter with random effects enters the typical value at the subject's mean covariate and the predictions through the within-subject deviation. The Laplace stage finds conditional modes by Newton steps with Fisher-type curvature $\mathbf{H} = \mathbf{J}^T \mathbf{W} \mathbf{J} + \Omega^{-1}$ ($\mathbf{J}$, prediction derivatives by central differences) and minimizes $2\mathbf{J}(\mathbf{b}^*) + \log \det \mathbf{H} - d \log 2\pi$ with L-BFGS-B from the starting values and up to three scrambled-Sobol perturbations of the typical values and covariate coefficients ($\pm \log 3$ and $\pm \log 10$), stopping when three starts agree within 0.1. Importance banks use 2^p scrambled Sobol points per subject, half from the population distribution and half from a subject-specific Gaussian adapted twice to the weighted mean and 1.5 times the weighted covariance of a pilot sample of 2^{10} points. Refinement begins at $p = 10$ and takes quasi-Newton steps within a trust box, with curvature from central differences of the bank gradient updated by damped BFGS; each step is validated on an independent bank, and p increases up to 16 when the banks disagree or have fewer than 100 effective particles. The final assessment uses two independent banks with $p = 14$ (or the working p when higher). Acceptance requires their OFVs to agree within 0.05 and every subject $\text{ESS} \geq 100$, as in S1, and each projected score to be at most 0.2 or, for a coordinate whose standard error from the quasi-Newton curvature, $\text{SE} = \sqrt{2[\mathbf{H}^{-1}]_{jj}}$, is below one, at most $0.2/\text{SE}$; such coordinates then lie within about 0.1 standard errors of the stationary point.

Diagnostics and tools. PRED uses $\eta = 0$; IPRED and IWRES use the conditional modes. WRES and CWRES use first-order linearizations at $\eta = 0$ and at the modes [23], decorrelated with the Cholesky factor of their covariance. NPDE use 1,000 simulated replicates of each subject's design, decorrelated with the Cholesky factor of the simulated covariance [24,25]. Shrinkage is $1 - \text{SD}(\text{EBE})/\omega$ for η and $1 - \text{SD}(\text{IWRES})$ for ε [26]. VPC bins observations by nominal time (one bin per distinct time when there are at most twice as many distinct times as requested bins, otherwise by time quantiles) and compares the 5th, 50th and 95th observed percentiles with the 95% intervals of the same percentiles in 500 simulated replicates, optionally after prediction correction [27]; with a quantification limit,

observed and simulated values below it are set to the limit. The sandwich covariance is $I^{-1} S I^{-1}$ with $I = H/2$ and S the sum of outer products of the per-subject scores. Profile intervals are where the refitted OFV exceeds its minimum by 3.84, with linear interpolation on the log scale. The bootstrap resamples subjects with replacement and refits from the original estimates. SIR draws 1,000 vectors per iteration from a multivariate normal proposal centred at the estimate, starting from the local covariance and then using the covariance of the previous resample, and resamples 500 vectors per iteration [29,30]. Stepwise covariate selection adds the effect with the largest OFV decrease while $p < 0.05$ and removes effects while $p \geq 0.01$ (likelihood-ratio tests, one degree of freedom) [31].

S12. Evaluation of the event-record interface

Table S18. PKPy2 predictions compared with rxode2.

Scenario	Model	Values	Max. relative difference
One-compartment infusions	Linear	20	3.2e-12
Two-compartment infusion, steady state	Linear	11	8.5e-13
Oral, steady state, lag, F	Linear	14	3.3e-12
Oral, additional doses, lag	Linear	22	1.1e-12
Three-compartment bolus and infusion	Linear	22	5.1e-12
Zero-order absorption (D1)	Linear	20	3.5e-12
Modeled infusion rate (RATE = -1)	Linear	20	3.4e-12
Three transit compartments	Linear	20	2.3e-11
Resets (EVID 3 and 4)	Linear	9	2.2e-13
Time-varying weight on CL	Linear	12	3.3e-12
Effect compartment, sigmoid Emax	Linear	40	7.6e-12
Parent-metabolite	Linear	40	2.2e-11
Steady state with SS = 2	Linear	8	1.1e-12
Constant-infusion steady state	Linear	4	2.6e-12
Infusion with bioavailability	Linear	8	3.6e-12
Steady-state infusion past the interval	Linear	9	4.2e-12
Michaelis-Menten, repeated IV	ODE	20	3.8e-08
Michaelis-Menten, oral steady state	ODE	9	6.1e-08
Indirect response I	ODE	42	1.2e-08
Indirect response II	ODE	42	1.5e-08
Indirect response III	ODE	42	1.2e-08
Indirect response IV	ODE	42	7.6e-09
TMDD, full	ODE	46	1.1e-06
TMDD, quasi-steady state	ODE	46	2.1e-07

rxode2 solved every scenario as an ODE system with relative and absolute tolerances of 10^{-12} (including steady-state tolerances). Differences are divided by the rxode2 value, or by 10^{-6} times the subject's largest value when that is larger.

Table S19. Likelihood, engine and diagnostic checks.

(a) Marginal OFV at fixed parameters compared with independent adaptive Gauss-Hermite quadrature (PKPy2 banks with 2^{16} particles per subject).

Model	Subjects	Observations	Censored	PKPy2 OFV (two banks)	Quadrature OFV	Difference
Block Ω , time-varying and categorical covariates	12	96	0	208.16021 / 208.16024	208.16023	-1.7e-06
IOV, M3 censoring, log-normal residual	10	100	24	45.20350 / 45.20451	45.20388	1.3e-04
Bioavailability (logit), linear covariate	12	108	0	24.50450 / 24.50455	24.50455	-2.3e-05
Indirect response, two outputs	8	72	0	280.86987 / 280.86988	280.86988	-3.7e-06

(b) Refits through the event-record interface compared with the compact-interface fits (event-record minus compact).

Dataset	Status	OFV difference	Max. estimate difference (%)
Theophylline	converged	0.00024	0.762
Warfarin	converged	0.00003	0.394
Primary rich 000	converged	-0.00019	0.014
Primary rich 001	converged	-0.00005	0.140
Primary rich 002	converged	0.00124	0.137
Primary rich 003	converged	-0.00029	0.124
Primary rich 004	converged	0.00014	0.152
Primary sparse 000	converged	0.00060	0.638
Primary sparse 001	converged	0.00037	0.412
Primary sparse 002	converged	0.00093	0.651
Primary sparse 003	converged	0.00023	0.437
Primary sparse 004	converged	-0.00021	0.041

(c) Diagnostics of the theophylline model at identical parameters compared with nlmixr2 (CWRES standardized by its own variance and NPDE with the upper Cholesky factor, as computed by nlmixr2) and with the npde package applied to the same simulated replicates.

Quantity	n	Max. abs. difference	Correlation
PRED	120	6.0e-14	1.000000
IPRED	120	1.4e-03	1.000000
IWRES	120	1.7e-03	1.000000
CWRES	120	9.1e-04	1.000000
NPDE	120	3.2e-01	0.996019
EBE (η)	12	7.7e-04	-
NPDE vs npde package (same replicates)	120	4.9e-15	1.000000

Table S20. Simulated datasets: true values and estimates of PKPy2 and nlmixr2 FOCEi.

Dataset	Parameter	True	PKPy2	nlmixr2 FOCEi
Infusion, block Ω , WT and sex	CL	5	5.171	5.188
Infusion, block Ω , WT and sex	V1	20	20.81	20.8
Infusion, block Ω , WT and sex	Q	8	8.245	8.286
Infusion, block Ω , WT and sex	V2	40	39.54	39.56
Infusion, block Ω , WT and sex	$\beta(\text{CL} \sim \text{WT})$	0.75	0.5112	0.5101
Infusion, block Ω , WT and sex	$\beta(\text{V1} \sim \text{SEX} = 1)$	-0.2	-0.1682	-0.1586
Infusion, block Ω , WT and sex	$\omega^2(\text{CL})$	0.09	0.114	0.1142
Infusion, block Ω , WT and sex	$\omega^2(\text{V1})$	0.06	0.04586	0.04566
Infusion, block Ω , WT and sex	$\omega(\text{V1}, \text{CL})$	0.04	0.05177	0.05176
Infusion, block Ω , WT and sex	σ_{prop} (CP)	0.12	0.1268	0.127
Infusion, block Ω , WT and sex	σ_{add} (CP)	0.05	0.05147	0.0523
Infusion, block Ω , WT and sex	Exact OFV	-	2067.952	2068.033
Oral, M3 censoring	CL	3	3.184	3.189
Oral, M3 censoring	V	35	36.34	37.45
Oral, M3 censoring	Ka	1.2	1.179	1.184
Oral, M3 censoring	$\omega^2(\text{CL})$	0.09	0.1032	0.09365
Oral, M3 censoring	$\omega^2(\text{V})$	0.04	0.04002	0.04526
Oral, M3 censoring	$\omega^2(\text{Ka})$	0.2	0.3117	0.2543
Oral, M3 censoring	σ_{prop} (CP)	0.15	0.1584	0.1669
Oral, M3 censoring	Exact OFV	-	238.231	240.824
Log-normal residual	CL	4	3.764	3.78
Log-normal residual	V	40	38.61	38.62
Log-normal residual	$\omega^2(\text{CL})$	0.09	0.09981	0.0986
Log-normal residual	$\omega^2(\text{V})$	0.04	0.03138	0.03149
Log-normal residual	$\sigma_{\log}$ (CP)	0.2	0.1948	0.1947
Log-normal residual	Exact OFV	-	34.160	34.170
Michaelis-Menten	V	20	19.45	19.47
Michaelis-Menten	VMAX	40	41.01	41.19
Michaelis-Menten	KM	4	3.936	3.963
Michaelis-Menten	$\omega^2(\text{V})$	0.05	0.05485	0.05488
Michaelis-Menten	$\omega^2(\text{VMAX})$	0.09	0.08424	0.08359
Michaelis-Menten	σ_{prop} (CP)	0.1	0.1022	0.1023
Michaelis-Menten	σ_{add} (CP)	0.1	0.1023	0.1017
Michaelis-Menten	Exact OFV	-	1682.264	1682.281

Exact OFV: marginal OFV at each program's estimates evaluated by PKPy2 with two independent banks of 2^{14} particles per subject. Residual components fixed at zero are not shown.

Table S21. Warfarin PK/PD: estimates of each program and exact OFV at these estimates.

(a) Estimates of each program.

Parameter	PKPy2	nlmixr2 FOCEi	nlmixr2 SAEM
KTR	1.154	1.026	1.032
Ka	1.153	1.022	1.287
CL	0.1343	0.1062	0.1338
V	7.729	9.182	7.707
R0	96.53	96.6	96.52
KOUT	0.05207	0.06131	0.05532
IMAX	0.9999	0.8786	0.9615
IC50	1.152	0.509	0.9721
$\omega^2(\text{KTR})$	0.399	0.8958	0.3523
$\omega^2(\text{Ka})$	0.4085	0.8979	0.498
$\omega^2(\text{CL})$	0.07865	0.3223	0.0777
$\omega^2(\text{V})$	0.05115	0.2312	0.05061
$\omega^2(\text{R0})$	0.002762	0.1489	0.002469
$\omega^2(\text{KOUT})$	0.007888	0.2629	0.008712
$\omega^2(\text{IMAX})$	0.3093	0.3619	0.007739
$\omega^2(\text{IC50})$	0.1838	0.3996	0.2322
$\sigma_{\text{prop}}(\text{CP})$	0.09881	0.1233	0.124
$\sigma_{\text{add}}(\text{CP})$	0.5265	0.1683	0.4073
$\sigma_{\text{add}}(\text{R})$	3.704	6.086	3.942

(b) Exact marginal OFV at the estimates of each program, evaluated by PKPy2 as in Table S20.

Estimates	Exact OFV	$\Delta\text{OFV vs PKPy2}$	Objective reported by the program
PKPy2	2223.144	0.000	2223.144
nlmixr2 FOCEi	2557.483	334.339	1800.553
nlmixr2 SAEM	2241.082	17.938	1457.960

PKPy2 local intervals are not reported: IMAX approached 1 and the marginal Hessian was not positive definite. The objectives reported by the programs use different approximations and constants. The model is the nlmixr2 example turnover model: transit (KTR) and first-order (Ka) absorption, CL and V, and an indirect response with inhibition of production by $\text{IMAX} \cdot \text{C} / (\text{IC50} + \text{C})$, baseline R0 and KOUT; IMAX is logit-transformed and all eight parameters have interindividual variability.

Table S22. Parameter recovery with the event-record interface.

Design	Parameter	True	n	Relative bias (%)	Relative RMSE (%)
Two-compartment infusion	CL	5	20	0.2	6.5
Two-compartment infusion	V1	20	20	1.5	4.1
Two-compartment infusion	Q	8	20	0.9	3.8
Two-compartment infusion	V2	40	20	-0.1	1.9
Two-compartment infusion	$\omega^2(\text{CL})$	0.09	20	-10.4	19.1
Two-compartment infusion	$\omega^2(\text{V1})$	0.06	20	-12.0	18.7

Two-compartment infusion	$\omega(V1, CL)$	0.04	20	-12.3	24.2
Two-compartment infusion	IOV(CL)	0.03	20	2.1	30.1
Two-compartment infusion	$\beta(CL \sim WT)$	0.75	20	-2.4	27.1
Two-compartment infusion	$\beta(V1 \sim SEX = 1)$	-0.2	20	-12.3	31.4
Two-compartment infusion	σ_{prop} (CP)	0.12	20	-0.5	4.0
Two-compartment infusion	σ_{add} (CP)	0.05	20	-32.1	62.0
Indirect-response PK/PD	CL	2	20	0.4	4.1
Indirect-response PK/PD	V	25	20	1.4	2.9
Indirect-response PK/PD	Ka	1.3	20	-0.3	5.3
Indirect-response PK/PD	R0	80	20	0.0	0.5
Indirect-response PK/PD	KOUT	0.15	20	3.1	9.7
Indirect-response PK/PD	EMAX	1.5	20	-0.7	4.3
Indirect-response PK/PD	EC50	1	20	1.2	14.0
Indirect-response PK/PD	$\omega^2(CL)$	0.09	20	-2.5	23.9
Indirect-response PK/PD	$\omega^2(V)$	0.04	20	-4.3	25.2
Indirect-response PK/PD	$\omega^2(KOUT)$	0.09	20	2.2	25.0
Indirect-response PK/PD	$\omega^2(EC50)$	0.16	20	2.8	26.2
Indirect-response PK/PD	σ_{prop} (CP)	0.12	20	-1.5	7.0
Indirect-response PK/PD	σ_{add} (R)	4	20	0.0	5.2

Table S23. Calibration of NPDE and VPC at the true parameters.

Design	Replicates	NPDE mean	NPDE variance	KS test p < 0.05	VPC percentiles within 95% interval
Two-compartment infusion	50	-0.008	1.025	6%	CP 95.2%
Indirect-response PK/PD	30	0.004	1.008	10%	CP 95.1%; R 93.8%

KS, Kolmogorov-Smirnov test of the NPDE against $N(0, 1)$.

Table S24. Theophylline 95% intervals from five methods.

Parameter	Estimate	Wald	Sandwich	Profile	Bootstrap	SIR
CL	2.787	2.400 to 3.236	2.383 to 3.259	2.361 to 3.274	2.356 to 3.192	2.398 to 3.202
V	32.54	29.85 to 35.48	29.90 to 35.42	29.69 to 35.76	29.86 to 34.80	29.96 to 35.26
Ka	1.503	1.004 to 2.251	0.9919 to 2.279	0.9793 to 2.385	1.023 to 2.313	0.9887 to 2.225
$\omega^2(CL)$	0.06502	0.02680 to 0.1577	0.01815 to 0.2329	0.0293 to 0.1762	0.004463 to 0.1297	0.02884 to 0.1823
$\omega^2(V)$	0.0147	0.004511 to 0.04792	0.006499 to 0.03326	0.003608 to 0.04746	0.00004767 to 0.02428	0.003441 to 0.04045
$\omega^2(Ka)$	0.4565	0.179 to 1.164	0.1561 to 1.335	0.1961 to 1.320	0.1207 to 1.196	0.2211 to 1.256
σ_{prop} (CP)	0.1576	0.1353 to 0.1835	0.1262 to 0.1968	-	0.1224 to 0.1896	0.1377 to 0.1869

Profile likelihood was computed for the typical values and variances (-, not computed). Bootstrap: 200 of 200 replicates converged. SIR effective sample size in the final iteration: 213.

Table S25. Stepwise covariate selection in 20 simulated datasets.

Candidate effect	True effect	Selected (%)
CL~AGE	no	5
CL~CRCL	yes	100
CL~SEX=1	no	0
CL~WT	no	0
V~AGE	no	0
V~SEX=1	no	0
V~WT	yes	100

Data: 60 subjects, one-compartment oral model, eight samples per subject; true effects on CL of (CRCL/80) to the power 0.7 and on V of (WT/70) to the power 1.