A Structure-Constrained Neural Simulator for Population PK under Regimen Shift (ConstrainNODE-PK)

Ali Issa[1] ali.issa.21@ucl.ac.uk
Tarjinder Sahota[2] tarjinder.z.sahota@gsk.com
Núria Buil-Bruna[2] nuria.x.builbruna@gsk.com
Joseph F. Standing[1] j.standing@ucl.ac.uk
Frank Kloprogge[1] f.kloprogge@ucl.ac.uk
[1] *University College London, London, United Kingdom*
[2] *GSK, Stevenage, United Kingdom*

Abstract

Prospective dose selection requires simulating concentration distributions under regimens for which no concentrations have yet been observed. In this setting, success is not defined by low individual prediction error alone but by calibrated population simulation under regimen shift: a model can fit held-out individual observations yet still miscalibrate exposure distributions and underestimate high-exposure risk when the dosing schedule changes. We present ConstrainNODE-PK, a structure-constrained neural population pharmacokinetic simulator for intravenous bolus dosing under linear kinetics. The model expresses concentration as an exact superposition of dose events applied to a learned unit-dose impulse response, so transfer to a new regimen changes only the event sequence. The response is parameterized as a positive exponential mixture with a monotone residual neural ordinary differential equation, enforcing non-negativity, stability and monotone post-bolus decay; together with exact superposition, this gives exact dose proportionality while remaining flexible enough to capture one- to three-compartment-like behavior without per-compound compartment selection. We evaluate on six simulated settings and two public cohorts, using held-out subjects throughout and assessing both individualized prediction and population-level simulation calibration. The primary stress test is strict transfer from once-daily to three-times-daily dosing, with no target-regimen concentrations available at inference time. In this setting, ConstrainNODE-PK maintains near-nominal 95% prediction-interval coverage (95.6%), whereas a matched unconstrained dose-aware neural ordinary differential equation falls to 45.8%. Across settings, the model remains competitive with matched nonlinear mixed-effects references on individualized prediction and preserves near-nominal prior-predictive population calibration. Strict regimen-transfer evidence comes from controlled simulated intravenous-bolus settings; the African Research on Kidney Disease iohexol and Tobramycin cohorts provide complementary held-out within-cohort population-calibration checks rather than strict real-world transfer validation. More broadly, when a healthcare model is used to simulate a future intervention, the relevant validation target is calibration under the induced intervention shift, not only in-regimen fit.

1. Introduction

Dosing is one of the most consequential and adjustable interventions in clinical care and drug development. Clinicians and trialists routinely compare alternative dose *amounts* and *schedules* to balance efficacy, toxicity risk and feasibility. In these settings, the key requirement is prospective simulation rather than retrospective fit: before concentrations are observed under a candidate regimen, one needs credible predictions of the distribution of exposures that regimen would induce in a target population.

Clinical motivation and intended use. The direct use of ConstrainNODE-PK is population pharmacokinetic simulation for prospective regimen comparison and study design in linear intravenous-bolus settings, not stand-alone bedside dose adjustment. In early-phase and special-population pharmacokinetic studies, candidate schedules often cannot all be sampled fully, especially in pediatric and neonatal studies, pregnancy, critical illness and other vulnerable cohorts. A calibrated simulator can compare candidate intravenous-bolus schedules, estimate exposure quantiles, assess high-exposure risk and plan sampling windows before target-regimen concentrations are available. Treatment decisions would still require covariates, pharmacodynamic response, safety outcomes and clinical-risk models; the present work addresses the narrower pharmacokinetic simulation component.

Terminology. Here, intravenous bolus means an immediate intravenous dose with no absorption phase modeled; QD and TID mean once-daily and three-times-daily dosing; AUC_{0-t} is exposure over the evaluated time window; $C_{\max}$ is the largest concentration in that window; PI95 is empirical coverage of nominal 95% prediction intervals; NPDE and VPC are population-calibration checks; IPRED is individualized prediction after observed concentrations; and POP is population simulation before target-regimen concentrations are available.

Population pharmacokinetic (popPK) models are a standard tool for this purpose. They map a dosing history to a concentration–time trajectory, quantify between-subject variability and support simulation of trial-level quantities such as exposure quantiles, probability of target attainment and risk of exceeding safety thresholds (Mould and Upton, 2012, 2013). In practice, these models are judged not only by how well they fit observed individuals, but also by whether simulated populations resemble the observed cohort. For that reason, popPK analyses emphasize population-level diagnostics such as visual predictive checks (VPC) and normalized prediction distribution errors (NPDE), which ask whether the observed data look like a plausible draw from the modeled population, including in the tails of the exposure distribution (Bergstrand et al., 2011; Comets et al., 2008; Nguyen et al., 2017).

Use cases: IPRED versus POP. We distinguish two related settings. In individual prediction (IPRED), the model conditions on a patient's observed concentrations to infer an individualized PK profile, as in therapeutic drug monitoring (Wicha et al., 2021; Stankeviciute et al., 2023). In prior-predictive population simulation (POP), no concentrations are yet available under the target regimen and the question is what distribution of concentrations a population would show if that regimen were used. Trial simulation and prospective regimen comparison are canonical POP use cases, because target-regimen concentration data are not yet available and the quantity of interest is the population exposure

distribution. The same prior-predictive framing also arises at treatment initiation, before patient-specific concentration data are available. A model can look strong on IPRED yet still be unreliable as a simulator if, under a new schedule, it places too little probability mass in the high-exposure tail—that is, it underestimates the upper percentiles of the population exposure distribution.

Existing ML-based PK studies have examined cross-regimen generalization, but often under conditional forecasting: early measurements from the target regimen are available and the task is to predict later measurements for the same patient (Lu et al., 2021b; Tang, 2023). That is a clinically relevant conditional-forecasting problem, especially for therapeutic drug monitoring and model-informed precision dosing. Here we study the stricter simulation problem: before any concentrations are observed under the target regimen, can the model generate a calibrated population-level concentration distribution for that regimen?

Why regimen shift is hard. Because nonlinear mixed-effects (NLME) modeling—the traditional pharmacometric framework for population PK—can require substantial manual structural work, researchers have explored machine-learning alternatives for PK (Sibieude et al., 2022; Stankeviciute et al., 2023; Lu et al., 2021a; Tang, 2023). Yet regimen shift remains difficult because unconstrained neural dynamics models do not automatically respect basic PK structure under changes in schedule, dose or prediction horizon. When evaluated outside the training regimens, such models are not guaranteed to preserve non-negativity, monotone post-bolus decay or stable late-time behavior. These failures are not cosmetic: they distort uncertainty propagation and can mislead regimen comparisons. From an ML perspective, regimen change is a structured intervention on the dose-event sequence rather than passive covariate drift: it changes both local post-dose behavior and drug accumulation under repeated dosing. That makes this a particularly demanding form of distribution shift, because small violations of PK structure can compound under transfer even when in-regimen interpolation looks acceptable.

For IV bolus dosing under linear kinetics, the underlying system obeys strong structural properties: (i) **superposition** (exact dose linearity), so concentrations under any regimen are sums of shifted unit-dose impulse responses; (ii) **positivity and stability**, so concentrations remain non-negative and decay to zero; and (iii) **monotone post-bolus decay** for the impulse response. Classical compartmental models enforce these properties by construction (Mould and Upton, 2013). Flexible neural dynamics models, including unconstrained neural ODEs (Chen et al., 2018; Lu et al., 2021b; Bräm et al., 2024), do not enforce these properties by construction and must instead learn them from finite data. Under regimen shift, even small violations can degrade calibration despite acceptable in-regimen interpolation.

Model and experimental setup. We develop ConstrainNODE-PK, a constrained neural impulse-response popPK model for linear IV bolus data. For subject-specific latent z, concentration under dosing history $\{(t_k, D_k)\}$ is

$$C(t \mid z) = \sum_k D_k \, g(t - t_k \mid z) \, \mathbb{I}[t \geq t_k], \qquad g(\Delta t \mid z) = 0 \text{ for } \Delta t < 0.$$

Changing the regimen therefore changes only the dose-event sequence, not the structural model. This localizes the key modeling requirement to the unit-dose impulse response,

$g(\cdot \mid z)$: if g respects the IV-bolus PK constraints, then the superposed trajectory inherits the corresponding non-negativity, between-dose decay and stability properties under any dosing history. This impulse response is parameterized as a positive exponential mixture multiplied by a monotone residual Neural ODE term, which preserves non-negativity and stable post-dose decay while allowing controlled departures from a pure multi-exponential form. Combined with exact dose superposition, this yields exact dose linearity. Architecturally, ConstrainNODE-PK is a variational autoencoder (VAE)-style latent-variable simulator: observed histories are encoded into subject latents and a regimen-conditioned Gaussian prior regularizes that latent space so that unseen subjects can be simulated by prior sampling for POP, while observed subjects can be individualized by posterior inference for IPRED.

NLME is the natural reference because it encodes the IV-bolus structure mechanistically and remains the standard tool for regimen simulation. Our question is narrower and practical: when compound-specific structural search is costly, can a single reusable constrained neural model remain competitive with that reference while preserving calibration-sensitive behavior under regimen change?

We evaluate on six simulated IV-bolus settings with known data-generating mechanisms and two public cohorts. The simulated settings span one- to three-compartment dynamics under single and multiple dosing. The simulation suite is deliberately broad: because ConstrainNODE-PK uses one reusable impulse-response family rather than per-dataset compartment selection, it must be tested across the range of compartmental behaviors that a mechanistic workflow would ordinarily specify manually. In addition to individualized prediction, we evaluate prior-predictive population calibration, strict once-daily (QD)→three-times-daily (TID) regimen transfer without using target-regimen concentrations for inference and extrapolation beyond the observation window. Because ConstrainNODE-PK is exactly dose linear by construction, we report dose proportionality under global scaling as a supporting structural check in Appendix A.3. These experiments are designed to mirror both the intended use case and the structural flexibility claimed by the model.

Contributions. We formulate prior-predictive population simulation under regimen shift as a distinct evaluation target for dosing models, separate from conditional individualized forecasting. We introduce ConstrainNODE-PK, a constrained neural popPK simulator matched to linear IV-bolus transfer through exact dose-event superposition and a constrained learned impulse response. We then evaluate it with calibration-sensitive diagnostics on strict QD→TID transfer, extrapolation, six simulated cohorts with known ground truth and two public cohorts.

Generalizable Insights about Machine Learning in the Context of Healthcare

- When an ML model will be used to compare candidate clinical interventions before outcomes under those interventions are observed, validation should target calibration under the induced intervention shift, not only conditional forecast accuracy.
- When the deployment shift is governed by known scientific structure, encoding those invariances in the model class can improve robustness under shift.

- Evaluations limited to held-out subjects under interventions already seen during development or that condition on observations from the target intervention can substantially overstate readiness for prospective regimen design.

2. Related Work

Classical popPK is built from compartmental differential-equation models fit with nonlinear mixed-effects (NLME) methods (Mould and Upton, 2013). In linear IV-bolus settings, these models encode the key transfer structure by construction: non-negativity, stable decay and exact superposition across dose events. They are also routinely assessed with population-level diagnostics such as visual predictive checks (VPC) and normalized prediction distribution errors (NPDE) (Bergstrand et al., 2011; Comets et al., 2008), which makes NLME the natural reference for any proposed alternative.

A parallel line of work aims to reduce the manual burden of NLME workflows through automated covariate screening, structural assistance, initialization and improved tooling (Sibieude et al., 2022; Fidler et al., 2019; Huang et al., 2025; Richardson et al., 2025). These efforts improve efficiency and reproducibility but typically remain within the compartmental modeling paradigm.

Machine-learning approaches to PK and PK/PD have been explored for interpolation from sparse data, automation and covariate discovery (Tang, 2023; Uno et al., 2024; Huang et al., 2024). Neural ODE-based PK models have also shown cross-regimen forecasting ability and can be regularized with PK-motivated structure (Lu et al., 2021b,a; Bräm et al., 2024; Giacometti et al., 2025). However, some cross-regimen evaluations remain conditional on early target-regimen data. In Lu et al. (2021b), for example, first-cycle PK observations from the target regimen are used to forecast later concentrations for the same subject. That setting is clinically useful for individualized later-cycle forecasting, but it is less stringent than the prior-predictive population-simulation problem studied here, because early target-regimen concentrations are already available at inference time.

We therefore study a reusable constrained neural simulator that preserves the IV-bolus transfer structure by construction and evaluate it with pharmacometric population-calibration diagnostics rather than pointwise error alone. To our knowledge, this specific evaluation setting remains rare in the ML-PK literature.

3. Methods

3.1. Problem setting

We consider linear IV-bolus PK. For subject i, dose events occur at times $\{t_{ik}\}_{k=1}^{K_i}$ with amounts $\{a_{ik}\}_{k=1}^{K_i}$; concentrations are observed at times $\{u_{ij}\}_{j=1}^{n_i}$. The latent PK simulator operates on the concentration scale, while observations are encoded and scored in standardized log concentration. At each observation, the encoder input consists of $\log(1 +$ time since first dose$)$, the standardized log concentration and $\log(1 +$ time since last dose$)$; Appendix B.1 gives the exact definitions.

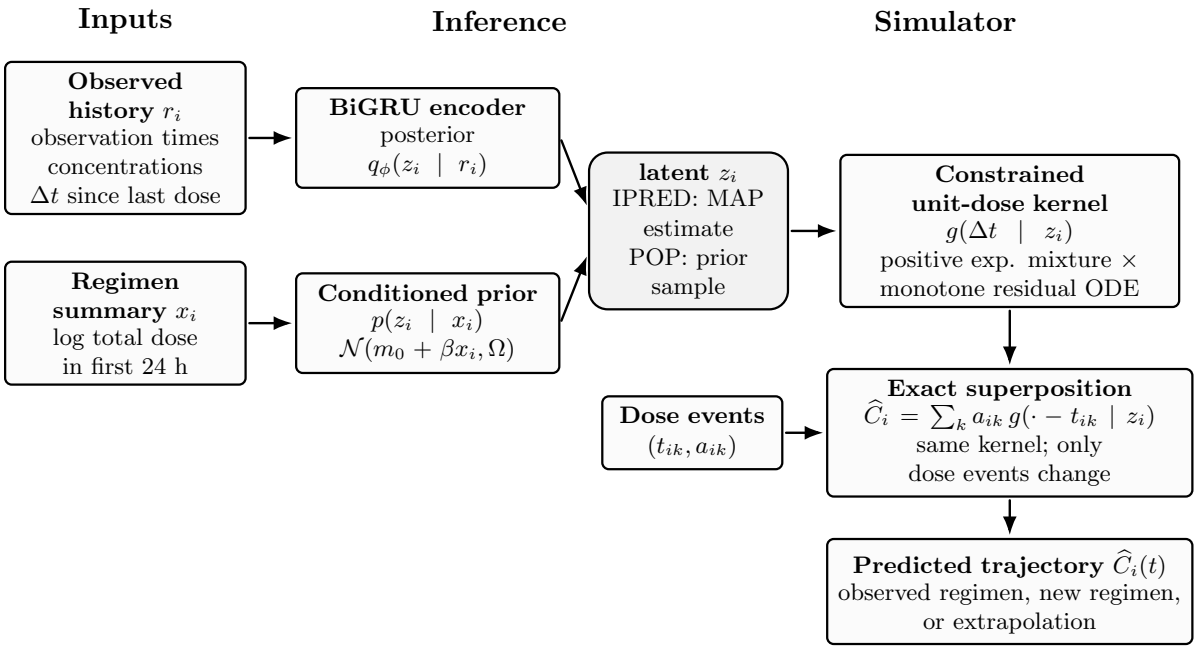


Figure 1: Architecture of ConstrainNODE-PK. A BiGRU encoder maps observed concentrations to a subject latent z_i, regularized by a regimen-conditioned Gaussian prior. For individualized prediction (IPRED), z_i is represented by its maximum a posteriori (MAP) estimate; for prior-predictive population simulation (POP), z_i is sampled from the prior. The simulator constructs a constrained unit-dose kernel and combines it with dose events by exact superposition to produce $\widehat{C}_i(t)$. Under regimen transfer, the subject-specific kernel stays fixed and only the dose-event table changes.

3.2. Model

Figure 1 shows the architecture of ConstrainNODE-PK and highlights the separation between subject-level inference and regimen-level simulation.

For subject i, ConstrainNODE-PK predicts concentration by exact superposition of a learned unit-dose impulse response:

$$\widehat{C}_i(t) = \sum_{k=1}^{K_i} a_{ik}\, g(t - t_{ik} \mid z_i)\, \mathbb{I}[t \geq t_{ik}],$$

where z_i is a subject-level latent variable. Architecturally, ConstrainNODE-PK is a variational autoencoder (VAE)-style latent-variable model whose decoder is the constrained PK simulator. A bidirectional GRU encoder maps the observed history to an approximate Gaussian posterior over z_i (Cho et al., 2014). For individualized prediction (IPRED), we represent z_i by its maximum a posteriori (MAP) estimate given that subject's observations; for prior-predictive population simulation (POP), we generate unseen subjects by sampling

z_i from the regimen-conditioned prior. The regimen-conditioned prior is

$$p(z_i \mid x_i) = \mathcal{N}(m_0 + \beta x_i, \Omega),$$

where x_i is the standardized log total dose in the first 24 hours. Thus, the "prior sample" path in Fig. 1 samples z_i from the regimen-conditioned Gaussian prior rather than from a fixed standard Normal prior.

The base impulse response is parameterized as a positive exponential mixture,

$$g_{\text{base}}(t \mid z) = \frac{1}{V(z)} \sum_{j=1}^{J} w_j(z) \exp(-\lambda_j(z) t),$$

with $V(z) > 0$, $w_j(z) \geq 0$, $\sum_j w_j(z) = 1$ and $\lambda_j(z) > 0$, where $V(z)$ is a positive volume-like scale parameter, $w_j(z)$ are mixture weights, $\lambda_j(z)$ are decay rates and J is a fixed maximum number of exponential components. This lets a single reusable kernel express one- to three-compartment-like decay patterns without choosing a compartment count for each compound; unused components can shrink toward zero. We then apply a monotone residual factor

$$g(t \mid z) = g_{\text{base}}(t \mid z) \exp(-s(t \mid z)), \qquad s(0) = 0, \quad \frac{ds}{dt} \geq 0,$$

where s is generated by a small neural ODE (Chen et al., 2018). This kernel construction preserves non-negativity and monotone post-dose decay; combined with exact superposition, the overall model preserves exact dose linearity.

3.3. Training, calibration and inference

Training minimizes a weighted Gaussian negative log-likelihood in standardized log space, plus Kullback–Leibler (KL) regularization toward the regimen-conditioned prior, an auxiliary population-fit term evaluated at the prior mean $z = m(x_i) = m_0 + \beta x_i$ and a small penalty on the residual ODE state. After network training, we apply a train+validation-only empirical-Bayes calibration step. Concretely, with global model parameters fixed, we re-estimate each subject latent by maximum a posteriori (MAP) optimization, i.e., the latent value that best fits that subject's observed concentrations while remaining probable under the regimen-conditioned prior. We then re-estimate the observation-noise scale, optionally allow extra noise immediately after dosing and refit the mean and covariance of the latent Gaussian prior to the collection of MAP-refined train/validation latents. The same post-training calibration is applied to ConstrainNODE-PK and both neural baselines, so differences in transfer performance are not an artifact of one method receiving extra uncertainty tuning.

The observation model is Gaussian in standardized log concentration, with a shared residual scale and optional early-post-dose inflation calibrated using train+validation only. This lets POP uncertainty reflect both between-subject variability and residual noise without giving any method access to test-set information.

Table 1: Implementation details used for all neural experiments.

Component	Specification
Event input	`dataset,id,time,dv,evid,amt`; dose history and concentrations only
Encoder	BiGRU, hidden size 128, inputs from Appendix B.1
Latent prior	$D = 6$; $p(z_i \mid x_i) = \mathcal{N}(m_0 + \beta x_i, \Omega)$
Kernel	$J = 8$ positive exponential components with monotone residual ODE
Residual ODE	hidden size 256; `dopri5`; rtol 10^{-6}; atol 10^{-8}
Training	AdamW; LR 5×10^{-4}; weight decay 10^{-6}; global clipping 1.0
Stopping	up to 150 epochs; ReduceLROnPlateau patience 5; early stopping patience 20
Calibration	empirical Bayes using train+validation only; no test data and no target-regimen concentrations
Splits	fixed subject-level train/validation/test splits shared by all methods

3.4. Baselines, metrics and ablations

Across the full eight-dataset sweep, we compare ConstrainNODE-PK against matched nonlinear mixed-effects references and two matched dose-aware neural baselines. DoseJump-NODE is a latent Neural ODE that evolves continuously between dose times and receives additive jumps at dose events, without enforcing exact superposition or monotone post-bolus decay. Kernel-MLP retains exact dose superposition but replaces the constrained exponential-mixture kernel with a generic positive MLP kernel. Both neural baselines share the same encoder, regimen-conditioned prior, loss structure and empirical-Bayes calibration pipeline as ConstrainNODE-PK, so differences isolate the effect of structural constraints rather than auxiliary training choices. The primary tables report the nonlinear mixed-effects reference and representative neural-baseline stress tests; the full neural-baseline sweep is summarized in Appendix A.4. For all comparisons, matching means the same event-table inputs, the same subject-level train/validation/test splits and no additional demographic or laboratory covariates. In the simulated settings with known data-generating mechanisms, the nonlinear mixed-effects reference is evaluated within the correct mechanistic family; on the public datasets, its structural specification is fixed using train/validation data only. Reported evaluations include IPRED logRMSE, median absolute percentage error for AUC_{0-t} and $C_{\max}$, 95% prediction-interval coverage, NPDE mean and standard deviation for POP, strict QD→TID regimen transfer, forecast logRMSE and targeted ablations. Full encoder features, objective definitions, optimization details, calibration steps, baseline specifications, nonlinear mixed-effects reference specifications, metric definitions and ablation settings are provided in Appendix B.

4. Cohort and Data Sources

We evaluate on eight datasets: six simulated linear IV-bolus cohorts and two public cohorts. Unless otherwise noted, each simulated cohort contains $N = 120$ subjects. The simulated datasets cover one-, two- and three-compartment dynamics under single-dose and multiple-dose regimens. The public cohorts are the final filtered African Research on Kidney Disease (ARK) iohexol analysis cohort (`ARKinput_final.csv`, $N = 2578$) (Kalyesubula et al., 2020; Fabian et al., 2022) and the Tobramycin population PK dataset ($N = 97$)

(Aarons et al., 1989). All experiments use shared subject-level train/validation/test splits targeted at approximately 60/20/20, with integer rounding at the subject level, across ConstrainNODE-PK, NLME and neural baselines. This yields 72/24/24 subjects for each simulated cohort ($N = 120$), 1547/516/515 for ARK and 58/19/20 for Tobramycin.

All datasets are represented in a common event-table format with subject identifier, time, observation value, dosing indicator and dose amount. Model inputs are dosing history and concentration observations only. No demographic or laboratory covariates are used by ConstrainNODE-PK in the experiments reported here. In the simulated cohorts, body weight is used only to generate subject-specific parameters and weight-based dosing; it is not provided to any model as a covariate during training, inference, or subject-level re-estimation. We made this choice to isolate whether structural constraints alone improve calibration under regimen shift. When an observation and a dose share the same nominal timestamp, the observation is placed immediately before the dose to preserve trough semantics. No below-the-limit-of-quantification (BLQ) observations were present in the final analysis datasets, so no BLQ censoring or imputation was required.

The simulated datasets were generated from linear compartment models with log-normal inter-individual variability, weight-based dosing, realistic sampling jitter and log-normal residual error. ARK provides a large sparse single-dose setting, whereas Tobramycin provides sparse heterogeneous repeated dosing, with most concentrations collected 2–6 h post-dose. Full cohort curation, parameter values, sampling templates and generation equations are provided in Appendix C.

5. Results

Main findings. In prospective simulation under regimen shift, ConstrainNODE-PK remains near-nominal on population-calibration diagnostics when matched flexible baselines do not. In the representative two-compartment multidose simulated QD→TID transfer experiment, ConstrainNODE-PK maintains near-nominal coverage on the unseen target regimen (PI95= 95.6%) while remaining close to the nonlinear mixed-effects reference on transfer accuracy, whereas a matched unconstrained dose-aware Neural ODE falls to PI95= 45.8%. Across simulated cohorts and public held-out within-cohort checks, ConstrainNODE-PK remains competitive on individualized prediction while preserving near-nominal prior-predictive population diagnostics.

5.1. Evaluation protocol and datasets

We use the eight datasets and fixed subject-level train/validation/test splits described in Section 4 and report all primary metrics on held-out test subjects. Subject-level splitting is the relevant evaluation unit for popPK use in new patients and avoids leakage across repeated measurements from the same subject. We report **IPRED** point and exposure metrics using MAP-refined subject latents and **POP** diagnostics from unconditional prior-predictive simulation.

Metrics reported were: (i) logRMSE in log concentration for IPRED; (ii) median absolute percentage error (MdAPE) for AUC_{0-t} (area under the concentration–time curve over the evaluated window) and $C_{\max}$ (maximum observed concentration), computed over the observed sampling window in ARK and Tobramycin; (iii) 95% prediction-interval coverage

(PI95) for POP; and (iv) NPDE mean and standard deviation for POP. Because the use case is prospective regimen evaluation, POP metrics are the primary test of whether the model acts as a calibrated simulator under regimen shift. The simulated cohorts with known data-generating mechanisms are essential here because they let us distinguish schedule-shift model failure from ambiguity in sparse real-world observations. They also test whether one reusable constrained kernel can cover one-, two- and three-compartment behavior without per-dataset structural engineering.

5.2. Primary stress test: strict regimen transfer without target-regimen concentrations

Our primary empirical test is *strict* regimen transfer from once-daily dosing (QD) to three-times-daily dosing (TID) on the representative two-compartment multidose simulated dataset. Total daily dose is held fixed across QD and TID, so this experiment isolates schedule change rather than dose-scale change. All fitting and calibration are completed before any TID test concentrations are seen; at evaluation time, subject latents z_i are inferred from QD data only and the full TID concentration profile is predicted without using any TID concentration measurements for inference or recalibration. We report (i) subject-specific transfer accuracy on TID, with latents inferred from QD data only and (ii) unconditional POP predictive checks on TID, together in a single table for clarity (Table 2). For the POP diagnostics, PI95 near 95% and NPDE mean/sd near 0/1 indicate well-calibrated population simulation.

Comparison to neural baselines. We additionally compare against the two matched dose-aware neural baselines introduced in Section 3.4: DoseJump-NODE and Kernel-MLP.

Table 2: Strict QD→TID regimen transfer on the representative two-compartment multidose simulated dataset, evaluated on an unseen target regimen. **Transfer IPRED:** infer z from QD data only, then predict TID with *no TID concentrations available to the model at inference time.* **POP:** unconditional prior-predictive checks on TID, also without TID concentrations used for inference or recalibration. The two neural baselines use the same encoder, regimen-conditioned prior, loss structure and train+validation-only empirical-Bayes calibration pipeline as ConstrainNODE-PK.

	Transfer IPRED on TID			**POP on TID**		
Model	logRMSE↓	MdAPE AUC_{0-t} (%)↓	MdAPE C_{max} (%)↓	PI95 (%)→ 95	NPDE mean→ 0	NPDE sd→ 1
NLME (nlmixr2)	0.166	4.53	20.22	94.9	+0.014	1.061
ConstrainNODE-PK	0.169	1.95	16.34	95.6	+0.009	0.975
Kernel-MLP (positive)	0.199	5.11	27.71	96.4	+0.012	0.768
DoseJump-NODE	0.847	88.54	180.13	45.8	-0.722	2.061

For prospective regimen simulation, the main difference in Table 2 is not only point error but uncertainty. The flexible baselines are not simply weak in-regimen models: on the representative QD test split, DoseJump-NODE and Kernel-MLP achieve QD POP PI95 of 95.0% and 95.6%, respectively (Appendix A.1), yet under strict QD→TID transfer DoseJump-NODE collapses to PI95= 45.8% and Kernel-MLP becomes over-dispersed (PI95= 96.4%,

NPDE sd = 0.768). This shows that acceptable in-regimen performance does not guarantee reliable simulation under schedule shift. In contrast, ConstrainNODE-PK preserves near-nominal TID coverage while remaining close to the nonlinear mixed-effects reference on subject-specific transfer accuracy.

The broader neural-baseline checks support the same interpretation. Across the full eight-dataset POP calibration sweep, ConstrainNODE-PK had mean $|\text{PI95} - 95| = 1.3$ percentage points, max $|\text{PI95} - 95| = 2.9$ pp and mean $|\text{NPDE sd} - 1| = 0.09$, compared with 3.8/7.5/0.18 for Kernel-MLP and 5.5/12.0/0.24 for DoseJump-NODE (Appendix A.4). In strict QD→TID transfer over all multidose simulations, ConstrainNODE-PK remained closest to nominal, with mean/worst $|\text{PI95} - 95| = 1.8/3.4$ pp, compared with 5.0/8.5 pp for Kernel-MLP and 18.0/49.2 pp for DoseJump-NODE. The pre/post empirical-Bayes check shows that ConstrainNODE-PK was already close before EB (PI95=93.8%, NPDE sd=1.04) and that EB mainly tuned uncertainty scale (post-EB PI95=95.6%, NPDE sd=0.975), while DoseJump-NODE remained poorly calibrated after the same EB step (PI95=45.8%, NPDE sd=2.061; Appendix A.5). Regimen-level exposure diagnostics for ConstrainNODE-PK gave q95 AUC error 6.5%, q95 $C_{\max}$ error 12.0%, high-exposure-risk error 4.0 pp and correct QD-vs-TID risk ranking in 3 of 3 multidose settings (Appendix A.6). These are simulator-calibration checks, not clinical outcome endpoints.

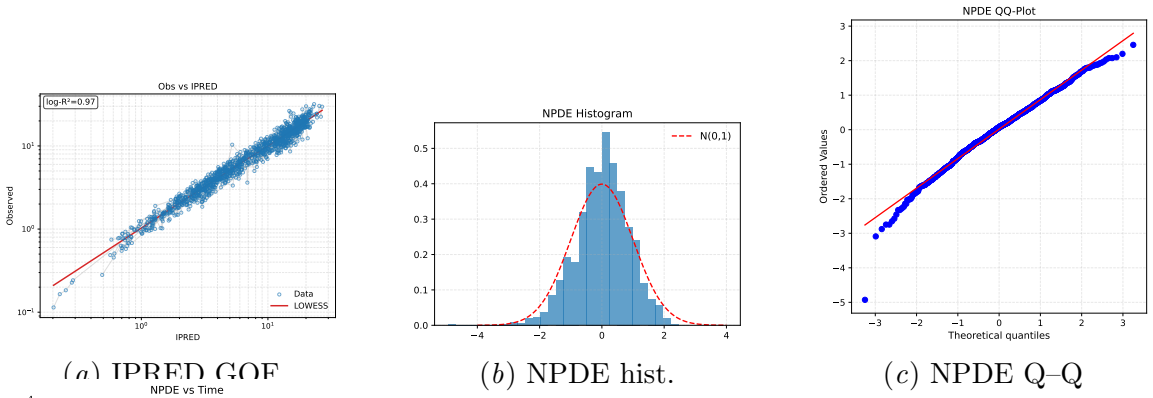
5.3. In-regimen interpolation on simulated cohorts (held-out 20% test)

Table 3 summarizes held-out test performance across the six simulated datasets. For readers less familiar with pharmacometric diagnostics, IPRED metrics assess how well the model fits an individual's trajectory after conditioning on that subject's observations. PI95 asks whether the nominal 95% predictive interval achieves the correct empirical coverage and NPDE asks whether predictive residuals are approximately centered at 0 with standard deviation near 1 under the simulated population model. In these synthetic settings, ConstrainNODE-PK is usually not the best individual fitter, but it stays close to nominal on the population diagnostics that matter for simulation under regimen shift.

Figure 2 shows representative GOF/NPDE/VPC diagnostics for the representative two-compartment multidose simulated dataset. Additional held-out individual-fit examples for the representative two-compartment multidose simulated dataset are provided in Appendix A (Fig. 4).

Table 3: Held-out test performance for the six simulated IV-bolus datasets. IPRED metrics use MAP-refined latents; POP metrics use unconditional prior simulation.

Dataset	Model	logRMSE (IPRED)↓	MdAPE AUC_{0-t} (IPRED, %)↓	MdAPE $C_{\max}$ (IPRED, %)↓	PI95 (POP, %)→ 95	NPDE μ/σ (POP)→ 0/1
1-compartment, single-dose	ConstrainNODE-PK	0.192	2.47	11.99	96.7	-0.083/1.064
	NLME (nlmixr2)	0.248	2.30	9.50	95.0	+0.049/0.977
1-compartment, multidose	ConstrainNODE-PK	0.166	1.29	15.45	94.2	-0.117/0.960
	NLME (nlmixr2)	0.240	0.90	15.90	93.9	+0.015/1.024
2-compartment, single-dose	ConstrainNODE-PK	0.169	2.50	12.67	93.8	+0.076/1.133
	NLME (nlmixr2)	0.165	2.20	6.20	94.6	+0.054/0.984
2-compartment, multidose	ConstrainNODE-PK	0.193	2.11	21.34	96.9	-0.017/0.943
	NLME (nlmixr2)	0.154	0.90	12.10	93.4	+0.019/1.054
3-compartment, single-dose	ConstrainNODE-PK	0.167	3.21	10.24	94.3	-0.093/1.092
	NLME (nlmixr2)	0.158	2.20	5.80	95.7	+0.068/0.952
3-compartment, multidose	ConstrainNODE-PK	0.185	1.61	16.92	96.0	-0.015/0.925
	NLME (nlmixr2)	0.149	0.60	12.40	93.2	+0.007/1.062



(*a*) IPRED GOF (*b*) NPDE hist. (*c*) NPDE Q–Q

(*d*) NPDE vs time (*e*) POP GOF (*f*) VPC

Figure 2: Representative diagnostics on the representative two-compartment multidose simulated dataset. Panels show observed versus IPRED, NPDE distribution, NPDE Q–Q, NPDE over time, observed versus POP prediction and VPC.

5.4. Public cohorts

Table 4 reports ConstrainNODE-PK and NLME performance on the held-out test splits of the ARK iohexol and Tobramycin cohorts. These datasets stress complementary parts of the problem: ARK is a large sparse single-dose setting, whereas Tobramycin is a smaller but clinically heterogeneous repeated-dose setting.

Table 4: Held-out test performance for the two public cohorts. IPRED metrics use MAP-refined subject latents; POP metrics use unconditional simulation. AUC_{0-t} and $C_{\max}$ are computed over each subject's observed sampling window.

Dataset	Model	logRMSE (IPRED)↓	MdAPE AUC_{0-t} (IPRED, %)↓	MdAPE $C_{\max}$ (IPRED, %)↓	PI95 (POP, %)→ 95	NPDE μ/σ (POP)→ 0/1
ARK iohexol ($N = 2578$)	ConstrainNODE-PK	0.164	5.38	15.05	94.5	-0.043/1.020
	NLME (nlmixr2)	0.171	11.34	15.19	95.2	-0.052/1.028
Tobramycin ($N = 97$)	ConstrainNODE-PK	0.180	5.15	10.80	94.7	-0.026/0.933
	NLME (nlmixr2)	0.175	7.71	5.14	96.7	+0.088/0.899

Together, ARK and Tobramycin provide complementary held-out within-cohort POP-calibration checks. They do not constitute strict real-world regimen-transfer validation because comparable source-regimen and target-regimen sampling pairs are not available.

Table 5: Ablations on the representative two-compartment multidose simulated dataset (held-out test split). Forecast logRMSE evaluates future-only prediction after fitting on the first half of each subject's samples.

Variant	logRMSE (IPRED)↓	MdAPE AUC_{0-t} (%)↓	MdAPE C_{max} (%)↓	PI95 (%)→ 95	NPDE $\mu/\sigma \to 0/1$	Forecast logRMSE↓
Full model	0.193	2.11	21.34	96.9	-0.017/0.943	0.289
Prior without regimen conditioning	0.193	1.81	21.34	92.2	+0.008/0.957	0.288
Single-exponential kernel	0.215	5.28	24.04	95.6	+0.025/0.912	0.270
Uniform likelihood weighting	0.213	4.38	16.60	95.9	-0.040/0.826	0.296
No residual ODE	0.236	6.35	26.06	97.9	+0.002/0.815	0.413

5.5. Structural stress test: extrapolation stability beyond the observation window

We evaluated stability beyond the last observation by extrapolating each held-out test subject to an extended horizon spanning multiple inferred dosing intervals after the final observation. We then checked for non-physical behavior in the extrapolation region, in particular positive upward steps in $\log C$ after extrapolation begins. Figure 3 shows a representative subject: both PRED and IPRED trajectories remain stable in the shaded extrapolation region, without spurious oscillations or late-time upturns.

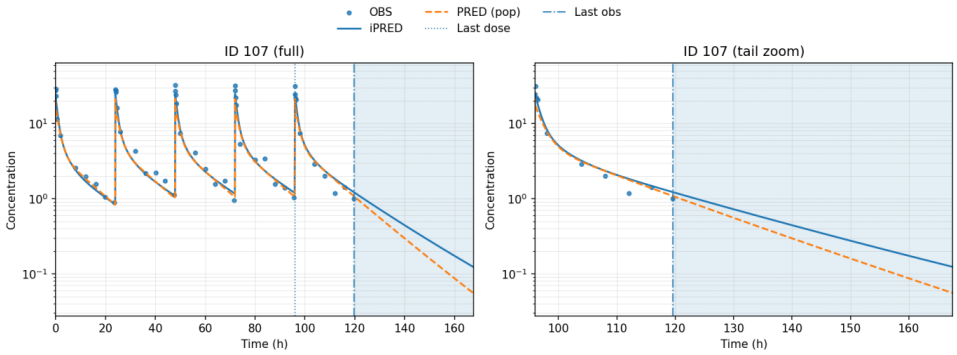


Figure 3: Time extrapolation check for a representative held-out subject. Points are observations; solid and dashed curves show IPRED and PRED. Shading marks extrapolation beyond the last observation.

Because ConstrainNODE-PK is exactly dose linear by construction, dose proportionality under global dose scaling is reported as a supporting structural check in Appendix A.3.

5.6. Ablations on the representative two-compartment multidose dataset: accuracy, calibration and forecasting

To isolate key components, we ran targeted ablations on the representative two-compartment multidose dataset. In addition to standard test metrics, we report a **forecast logRMSE** computed by fitting/refining z on the first half of each subject's observations and scoring predictions on the held-out second half (future-only).

No single ablation dominates every metric. Removing regimen conditioning from the prior leaves IPRED nearly unchanged but reduces PI95 from 96.9% to 92.2%, showing that this component primarily affects population calibration. Restricting the kernel to a single exponential slightly improves forecast logRMSE but clearly worsens IPRED fit and exposure recovery, consistent with a simpler kernel that regularizes forecasting at the cost of misspecifying multi-compartment structure. Using uniform likelihood weighting lowers $C_{\max}$ error but worsens AUC and NPDE spread, indicating a shift in where the model allocates error rather than a uniform gain. Removing the residual ODE produces the clearest failure, substantially worsening forecast logRMSE and exposure error despite superficially high PI95. Taken together, the full model offers the most balanced operating point across fit, calibration and extrapolation.

6. Discussion

Why the constraints matter. The core claim is that if a model will be used as a simulator under regimen shift, then the model class should respect the PK invariances that govern transfer. In ConstrainNODE-PK, exact superposition handles dose scaling by construction and the monotone residual ODE suppresses the unstable tails that flexible continuous-time models can develop outside the training schedule. The results suggest that these constraints matter most when the question changes from interpolation within an observed regimen to simulation under a new regimen.

Why this matters for regimen design. For prospective regimen design, the main risk is not only a modest increase in average pointwise error but overconfident population simulation under a schedule for which no concentrations are yet available. A simulator that underestimates the upper tail can make a more aggressive schedule look safer than it is. The present model is not a bedside dosing policy: it estimates pharmacokinetic exposure under candidate regimens and would need covariate, pharmacodynamic, safety and clinical-risk components before use in individual treatment decisions.

Why the evaluation distinction matters. Conditional target-regimen forecasting and unconditional regimen-transfer simulation answer different deployment questions. The first asks whether a model can predict later measurements after conditioning on observations from the target regimen. The second asks whether it can support prospective regimen design before any target-regimen concentrations are available. Our QD→TID experiment targets the second setting by inferring z from QD data only and evaluating both individualized transfer accuracy and unconditional POP calibration on TID. Importantly, a standard held-out-subject evaluation performed only within regimens already seen during development would have substantially overestimated readiness for transfer: the flexible baselines looked acceptable on the QD test split, yet their calibration deteriorated sharply under QD→TID transfer. Because real-world test sets will often consist only of held-out subjects under already-observed regimens, any method proposed for prospective regimen comparison should also be stress-tested on unseen regimens.

Position relative to NLME. These results should not be read as an argument against NLME. When compound-specific mechanistic knowledge is available and expert structural

refinement is feasible, NLME remains the most interpretable and often the most accurate approach. Our contribution is to identify a middle ground: a reusable constrained model that gives up some mechanistic specificity in exchange for less per-compound structural search, while preserving the properties that matter most for reliable simulation under schedule change.

What the ablations imply. The ablations suggest that the main components contribute differently to fit and calibration rather than yielding uniform gains across all metrics. In particular, regimen conditioning mainly affects POP calibration, while the residual ODE is most important for forecast stability and exposure recovery. We therefore view the full model not as the optimizer of every single metric, but as the best overall operating point across fit, calibration and extrapolation.

Interpretation of the public cohorts. The two public cohorts were chosen to be complementary rather than to validate strict regimen transfer. ARK provides a large sparse single-dose check dominated by elimination-phase samples, while Tobramycin provides a smaller irregular repeated-dose check with routine sampling and regimen heterogeneity. Their role is to test held-out subject prediction and POP calibration under real clinical sampling patterns; the strict source-to-target regimen-transfer claim remains supported by the controlled simulations.

Broader ML-for-health relevance. The ML contribution is to show that when the prediction target is the consequence of a *future clinical intervention*, both the model class and the evaluation protocol should be intervention-aligned. In this case, that means encoding the scientific invariances that govern regimen transfer and testing calibration under the actual shift induced by changing the dosing schedule. PK provides a clean testbed, but the principle applies more broadly to healthcare simulators used for treatment comparison, policy evaluation, or prospective planning.

Clinical scope and limitations. The direct scope is population pharmacokinetic regimen and study-design simulation for linear intravenous-bolus settings. The model is not a stand-alone bedside dosing policy and should not be used alone for individualized dose adjustment under evolving clinical status. Such use would require measured covariates, pharmacodynamic response, safety outcomes and clinical-risk models.

Real-world strict regimen-transfer validation would require comparable source-regimen and target-regimen samples while withholding target-regimen concentrations at inference. The public ARK and Tobramycin cohorts do not provide that design, so we use simulations with known data-generating mechanisms as the controlled test of regimen transfer and public cohorts as held-out within-cohort POP-calibration checks. ARK mainly tests sparse elimination-phase behavior; Tobramycin adds sparse repeated dosing and regimen heterogeneity. These sparse cohorts do not make ConstrainNODE-PK a low-sample-size substitute for nonlinear mixed-effects modeling. The method still needs enough cohort-level data to learn the latent population distribution.

Assumptions, failure modes and extensions. The current model assumes reliable dose-event records, linear intravenous-bolus kinetics, stable nonnegative disposition and monotone post-bolus decay of the impulse response. The unmodified model is not expected

to be valid for oral dosing with absorption, finite infusions without changing the input operator, nonlinear or saturable elimination, strong time-varying covariates, between-occasion pharmacokinetic changes, major dose-recording errors or new populations where the constraints and latent population distribution have not been re-derived and externally validated. For finite infusions, the disposition kernel could be retained while replacing bolus impulses with rate inputs over infusion windows. For oral dosing with linear absorption and disposition, an absorption kernel would be needed before disposition. For nonlinear pharmacokinetics, exact superposition no longer holds and a constrained state-space model with positivity, mass balance and saturable elimination would need separate validation.

Computation. We do not claim a runtime advantage over nonlinear mixed-effects modeling. Cost is dominated by neural training and MAP/EB latent refinement. After fitting, counterfactual simulation mainly changes the dose-event table. In these experiments we used AdamW, gradient clipping, learning-rate scheduling and early stopping; we did not observe instability specific to the constrained kernel.

7. Conclusion

For prospective regimen-design questions, the relevant model is not just a forecaster but a calibrated simulator under intervention shift. In linear intravenous-bolus pharmacokinetics, ConstrainNODE-PK implements this idea through exact dose-event superposition, a constrained learned unit-dose response and a population prior that supports both individualized prediction and prior-predictive simulation. The controlled transfer experiments show that these constraints preserve population calibration under schedule change better than matched flexible neural baselines, while the public cohorts provide held-out within-cohort calibration checks. ConstrainNODE-PK is best viewed as a reusable complement to expert nonlinear mixed-effects modeling when manual structural search is a bottleneck and trustworthy simulation under regimen shift is needed. The broader lesson is that healthcare simulators used for future interventions should encode the scientific invariances that govern the intervention and should be evaluated under the shift they are meant to support.

Code and data availability. Code, processed event-table inputs where redistribution is permitted, fixed subject-level splits, synthetic-data generation code, neural training and evaluation utilities, paper-facing metric tables and nonlinear mixed-effects reference templates are available at https://github.com/aliissa26/constrainnode-pk. The camera-ready release is tagged as `v1.0-mlhc2026`.

Acknowledgments

This work was conducted through a collaboration between University College London and GSK. Ali Issa was supported by a UKRI EPSRC studentship co-funded by GSK. Tarjinder Sahota and Núria Buil-Bruna are employees of GSK and hold equity in GSK. Tarjinder Sahota and Núria Buil-Bruna also hold honorary appointments at University College London. Frank Kloprogge conducted this work as part of a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant number 220587/Z/20/Z).

References

L. Aarons, S. Vozeh, M. Wenk, P. H. Weiss, and F. Follath. Population pharmacokinetics of tobramycin. *British Journal of Clinical Pharmacology*, 28:305–314, 1989. doi: 10.1111/j.1365-2125.1989.tb05431.x.

B. J. Anderson and N. H. G. Holford. Mechanism-based concepts of size and maturity in pharmacokinetics. *Annual Review of Pharmacology and Toxicology*, 48:303–332, 2008. doi: 10.1146/annurev.pharmtox.48.113006.094708.

Martin Bergstrand, Andrew C. Hooker, Johan E. Wallin, and Mats O. Karlsson. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *The AAPS Journal*, 13:143–151, 2011. doi: 10.1208/s12248-011-9255-z.

Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. Low-dimensional neural ODEs and their application in pharmacokinetics. *Journal of Pharmacokinetics and Pharmacodynamics*, 51:123–140, 2024. doi: 10.1007/s10928-023-09886-4.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, pages 6572–6583, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Emmanuelle Comets, Karl Brendel, and France Mentré. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: The npde add-on package for R. *Computer Methods and Programs in Biomedicine*, 90:154–166, 2008. doi: 10.1016/j.cmpb.2007.12.002.

June Fabian, Robert Kalyesubula, Joseph Mkandawire, Christian Holm Hansen, Dorothea Nitsch, Eustasius Musenge, Wisdom P. Nakanga, Josephine E. Prynn, Gavin Dreyer, Tracy Snyman, Billy Ssebunnya, Michele Ramsay, Liam Smeeth, Stephen Tollman, Saraladevi Naicker, Amelia Crampin, Robert Newton, Jaya A. George, and Laurie Tomlinson. Measurement of kidney function in malawi, south africa, and uganda: A multicentre cohort study. *The Lancet Global Health*, 10:e1159–e1169, 2022. doi: 10.1016/S2214-109X(22)00239-X. On behalf of the African Research on Kidney Disease Consortium.

Matthew Fidler, Justin J. Wilkins, Richard Hooijmaijers, Teun M. Post, Rik Schoemaker, Mirjam N. Trame, Yuan Xiong, and Wenping Wang. Nonlinear mixed-effects model development and simulation using nlmixr and related R open-source packages. *CPT: Pharmacometrics & Systems Pharmacology*, 8:621–633, 2019. doi: 10.1002/psp4.12445.

Tommaso Giacometti, Ettore Rocchi, Pier Giorgio Cojutti, Federico Magnani, Daniel Remondini, Federico Pea, and Gastone Castellani. Leveraging neural ODEs for population

pharmacokinetics of dalbavancin in sparse clinical data. *Entropy*, 27(6):602, 2025. doi: 10.3390/e27060602.

Zhonghui Huang, Paolo Denti, Hitesh Mistry, and Frank Kloprogge. Machine learning and artificial intelligence in PK-PD modeling: Fad, friend, or foe? *Clinical Pharmacology & Therapeutics*, 115(4):652–654, 2024. doi: 10.1002/cpt.3165.

Zhonghui Huang, Matthew Fidler, Minshi Lan, Iek Leng Cheng, Frank Kloprogge, and Joseph F. Standing. An automated pipeline to generate initial estimates for population pharmacokinetic base models. *Journal of Pharmacokinetics and Pharmacodynamics*, 52 (6):60, 2025. doi: 10.1007/s10928-025-10000-z.

Robert Kalyesubula, June Fabian, Wisdom Nakanga, Robert Newton, Billy Ssebunnya, Josephine Prynn, Jaya A. George, Alisha N. Wade, Janet Seeley, Dorothea Nitsch, Christian Hansen, Moffat Nyirenda, Liam Smeeth, Saraladevi Naicker, Amelia C. Crampin, and Laurie A. Tomlinson. How to estimate glomerular filtration rate in sub-saharan africa: Design and methods of the african research into kidney diseases (ARK) study. *BMC Nephrology*, 21:1–12, 2020. doi: 10.1186/s12882-020-1688-0.

James Lu, Brendan Bender, Jin Y. Jin, and Yuanfang Guan. Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling. *Nature Machine Intelligence*, 3:696–704, 2021a. doi: 10.1038/s42256-021-00357-4.

James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. *iScience*, 24:102804, 2021b. doi: 10.1016/j.isci.2021.102804.

D. R. Mould and R. N. Upton. Basic concepts in population modeling, simulation, and model-based drug development. *CPT: Pharmacometrics & Systems Pharmacology*, 1:e6, 2012. doi: 10.1038/psp.2012.4.

D. R. Mould and R. N. Upton. Basic concepts in population modeling, simulation, and model-based drug development—part 2: Introduction to pharmacokinetic modeling methods. *CPT: Pharmacometrics & Systems Pharmacology*, 2:e38, 2013. doi: 10.1038/psp.2013.14.

Thi-Huyen-Tram Nguyen, Mohamad-Samer Mouksassi, Nick Holford, Nidal Al-Huniti, Immanuel Freedman, Andrew C. Hooker, Jyothy John, Mats O. Karlsson, Diane R. Mould, Juan Jose Perez-Ruixo, Elodie L. Plan, Rada Savic, Johan G. C. van Hasselt, Benjamin Weber, Chao Zhou, Emmanuel Comets, and France Mentre. Model evaluation of continuous data pharmacometric models: Metrics and graphics. *CPT: Pharmacometrics & Systems Pharmacology*, 6(2):87–109, 2017. doi: 10.1002/psp4.12161.

Sam Richardson, Itziar Irurzun Arana, Andrzej Nowojewski, Diansong Zhou, Jacob Leander, Weifeng Tang, Richard Dearden, and Megan Gibbs. A machine learning approach to population pharmacokinetic modelling automation. *Communications Medicine*, 5:327, 2025. doi: 10.1038/s43856-025-01054-8.

Rik Schoemaker, Matthew Fidler, Christian Laveille, Justin J. Wilkins, Richard Hooijmaijers, Teun M. Post, Mirjam N. Trame, Yuan Xiong, and Wenping Wang. Performance of the SAEM and FOCEI algorithms in the open-source, nonlinear mixed effect modeling tool nlmixr. *CPT: Pharmacometrics & Systems Pharmacology*, 8:923–930, 2019. doi: 10.1002/psp4.12471.

Emeric Sibieude, Akash Khandelwal, Pascal Girard, Jan S. Hesthaven, and Nadia Terranova. Population pharmacokinetic model selection assisted by machine learning. *Journal of Pharmacokinetics and Pharmacodynamics*, 49:257–270, 2022. doi: 10.1007/s10928-021-09793-6.

Kamile Stankeviciute, Jean-Baptiste Woillard, Richard W. Peck, Pierre Marquet, and Mihaela van der Schaar. Bridging the worlds of pharmacometrics and machine learning. *Clinical Pharmacokinetics*, 62:1551–1565, 2023. doi: 10.1007/s40262-023-01310-x.

Albert Tang. Machine learning for pharmacokinetic/pharmacodynamic modeling. *Journal of Pharmaceutical Sciences*, 112:1460–1475, 2023. doi: 10.1016/j.xphs.2023.01.010.

Mizuki Uno, Yuta Nakamaru, and Fumiyoshi Yamashita. Application of machine learning techniques in population pharmacokinetics/pharmacodynamics modeling. *Drug Metabolism and Pharmacokinetics*, 56:101004, 2024. doi: 10.1016/j.dmpk.2024.101004.

Sebastian G. Wicha, Anne-Grete Märtson, Elisabet I. Nielsen, Birgit C. P. Koch, Lena E. Friberg, Jan-Willem Alffenaar, and Iris K. Minichmayr. From therapeutic drug monitoring to model-informed precision dosing for antibiotics. *Clinical Pharmacology & Therapeutics*, 109:928–941, 2021. doi: 10.1002/cpt.2202.

Appendix A. Supplementary figures and tables

A.1 Additional neural-baseline results (in-regimen QD test)

Table 6: In-regimen performance on the QD test split for the representative two-compartment multidose simulated dataset: IPRED and unconditional POP diagnostics.

	IPRED on QD			POP on QD		
Model	logRMSE↓	MdAPE AUC_{0-t} (%)↓	MdAPE C_{max} (%)↓	PI95 (%)→ 95	NPDE mean→ 0	NPDE sd→ 1
ConstrainNODE-PK	0.193	2.11	21.34	96.9	-0.017	0.943
DoseJump-NODE	0.220	6.34	14.94	95.0	+0.004	0.870
Kernel-MLP (positive)	0.237	4.77	22.87	95.6	-0.087	0.941

A.2 Additional held-out individual fits for the representative two-compartment multidose simulated dataset

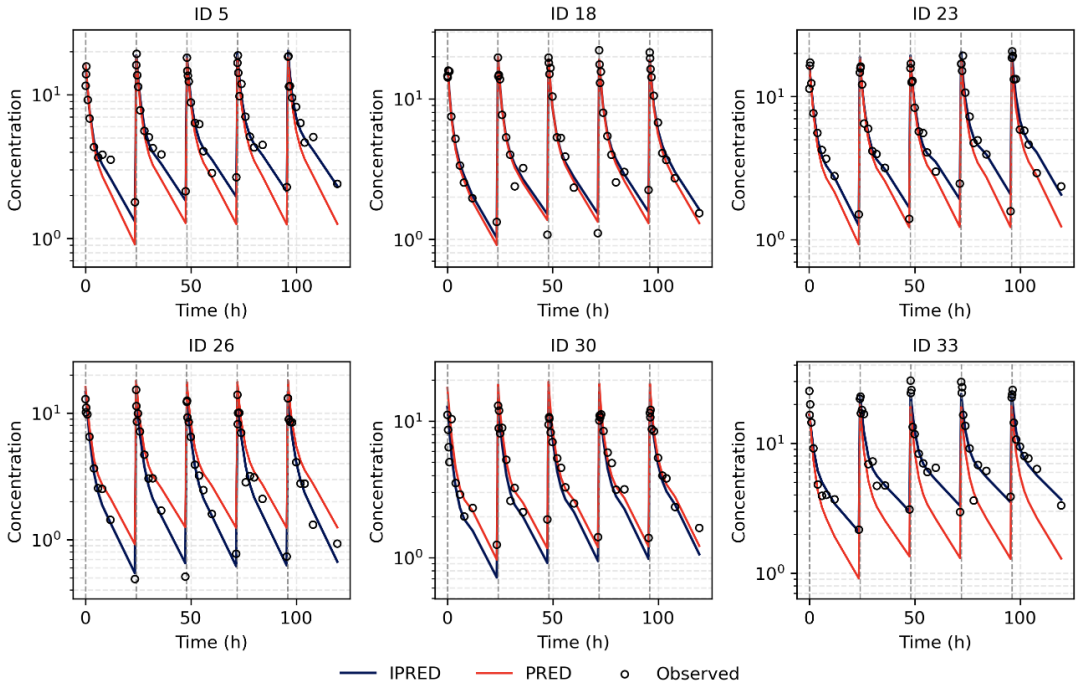


Figure 4: Held-out individual-fit examples for the representative two-compartment multidose simulated dataset. Shown are six held-out test subjects. Open circles denote observed concentrations, blue lines denote IPRED (MAP-refined), red lines denote population predictions (PRED) and vertical dashed lines mark dose times. These qualitative examples complement the aggregate simulated-data results in Table 3, the representative diagnostics in Fig. 2 and the ablation results in Table 5.

A.3 Dose proportionality under global dose scaling

For linear IV bolus kinetics, scaling all doses by a factor f should scale concentrations and exposure endpoints by the same factor. Because ConstrainNODE-PK enforces exact dose linearity by construction, this is a supporting structural verification rather than a primary empirical stress test. Figure 5 shows an example subject under global dose scaling $f \in \{0.5, 1, 2\}$; the resulting curves preserve the expected proportional spacing on the log scale.

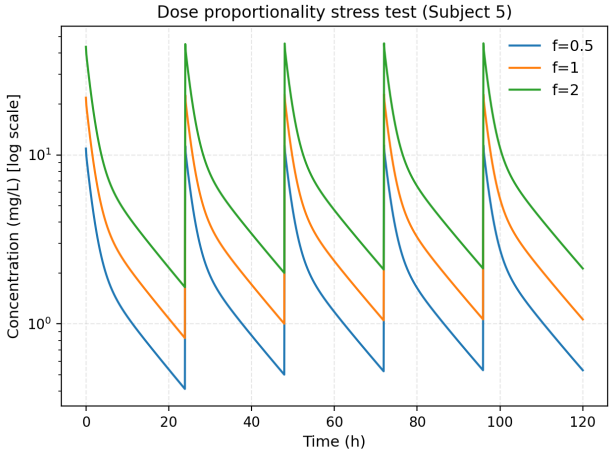


Figure 5: Dose proportionality under global dose scaling for an example subject, with $f \in \{0.5, 1, 2\}$.

A.4 Full neural-baseline calibration sweep

Table 7: Summary of the full eight-dataset neural-baseline population-calibration sweep. All neural models use the same event-table inputs, subject-level splits and train+validation-only empirical-Bayes calibration.

Model	Mean $\lvert PI95 - 95\rvert$ (pp)	Max $\lvert PI95 - 95\rvert$ (pp)	Mean $\lvert NPDE\ sd - 1\rvert$
ConstrainNODE-PK	1.3	2.9	0.09
Kernel-MLP	3.8	7.5	0.18
DoseJump-NODE	5.5	12.0	0.24

Table 8: Strict QD→TID transfer summary over all multidose simulations. Values are absolute deviations from nominal PI95.

Model	Mean $\lvert PI95 - 95\rvert$ (pp)	Worst $\lvert PI95 - 95\rvert$ (pp)
ConstrainNODE-PK	1.8	3.4
Kernel-MLP	5.0	8.5
DoseJump-NODE	18.0	49.2

The released repository contains the paper-facing metric tables used to produce these summaries.

A.5 Pre/post empirical-Bayes calibration ablation

Table 9: Pre/post empirical-Bayes calibration check on the strict QD→TID stress test. EB uses train+validation data only and is applied consistently across neural models.

Model	Calibration	PI95 (%)	NPDE mean	NPDE sd
ConstrainNODE-PK	pre-EB	93.8	–	1.040
ConstrainNODE-PK	post-EB	95.6	+0.009	0.975
Kernel-MLP	post-EB	96.4	+0.012	0.768
DoseJump-NODE	post-EB	45.8	-0.722	2.061

A.6 Regimen-level exposure diagnostics

Table 10: Upper-tail exposure diagnostics on strict QD→TID transfer simulations. q95 denotes the 95th percentile of the exposure summary. These are simulator-relevant exposure checks, not clinical outcome endpoints.

Model	q95 AUC error (%)	q95 $C_{\max}$ error (%)	High-exposure-risk error (pp)	Correct risk ranking
ConstrainNODE-PK	6.5	12.0	4.0	3 of 3

Appendix B. Full model, training, calibration, baselines and metrics

B.1 Encoder inputs and latent posterior

For observation j of subject i, let $t_{0,i}$ be the time of the first dose and define

$$\Delta_{ij}^{\text{last}} = u_{ij} - \max\{t_{ik} : t_{ik} \le u_{ij}\}, \qquad \Delta_{ij}^{\text{last}} \leftarrow \max(\Delta_{ij}^{\text{last}}, 0).$$

The encoder input is

$$r_{ij} = \begin{bmatrix} \log(1 + (u_{ij} - t_{0,i})) \\ \tilde{y}_{ij} \\ \log(1 + \Delta_{ij}^{\text{last}}) \end{bmatrix}.$$

A bidirectional GRU maps the observation history to a diagonal Gaussian posterior

$$q_\phi(z_i \mid r_i) = \mathcal{N}(\mu_\phi(r_i),\ \mathrm{diag}\left(\exp(\log v_\phi(r_i))\right)).$$

B.2 Regimen-conditioned prior

The regimen-conditioned prior is

$$p(z_i \mid x_i) = \mathcal{N}(m(x_i), \Omega), \qquad m(x_i) = m_0 + \beta x_i, \qquad \Omega = LL^\top.$$

We use a deliberately minimal regimen summary based on total dose in the first 24 hours:

$$\mathrm{DailyDose}_i = \sum_{k:\ t_{ik} \in [t_{0,i},\, t_{0,i}+24\mathrm{h})} a_{ik}, \qquad x_i = \frac{\log(1+\mathrm{DailyDose}_i) - \mu_{\mathrm{dd}}}{\sigma_{\mathrm{dd}}}.$$

This captures dose scale without encoding the target schedule itself.

B.3 Decoder and observation model

The base unit-dose impulse response is

$$g_{\mathrm{base}}(t \mid z) = \frac{1}{V(z)} \sum_{j=1}^{J} w_j(z) \exp(-\lambda_j(z) t), \qquad t \geq 0,$$

with $V(z) > 0$, $w_j(z) \geq 0$, $\sum_j w_j(z) = 1$ and $\lambda_j(z) > 0$. A monotone residual factor is then applied:

$$g(t \mid z) = g_{\mathrm{base}}(t \mid z) \exp(-s(t \mid z)), \qquad s(0) = 0, \qquad \frac{ds}{dt} \geq 0.$$

In implementation,

$$\phi(t) = [\log(1+t),\ (1+t)^{-1}], \tag{1}$$

$$\frac{ds}{dt} = \mathrm{softplus}(\mathrm{drive}_\psi([s, \phi(t), z])) \exp(-\gamma(z) t), \tag{2}$$

with $\gamma(z) > 0$. The ODE is solved using `torchdiffeq.odeint` with Dormand–Prince (`dopri5`), rtol $= 10^{-6}$ and atol $= 10^{-8}$.

Predictions are made in standardized log space via

$$\widehat{\mu}_{ij} = \frac{\log(\widehat{C}_i(u_{ij})) - \mu_{\log}}{\sigma_{\log}}, \qquad \tilde{y}_{ij} \sim \mathcal{N}(\widehat{\mu}_{ij}, \widehat{\sigma}_{ij}^2).$$

The default observation model uses constant Gaussian standard deviation in standardized log space,

$$\widehat{\sigma}_{ij} = \sigma_{\mathrm{base}}, \qquad \sigma_{\mathrm{base}} = \exp(\texttt{log\_sigma}),$$

with an optional early-post-dose inflation term

$$\widehat{\sigma}_{ij} = \sigma_{\mathrm{base}} \left(1 + a \exp\left(-\frac{\Delta_{ij}^{\mathrm{last}}}{\tau_{\mathrm{peak}}}\right)\right).$$

B.4 Training objective and implementation

The pointwise Gaussian negative log-likelihood is

$$\ell_{ij} = \frac{1}{2}\left(\frac{(\tilde{y}_{ij} - \hat{\mu}_{ij})^2}{\hat{\sigma}_{ij}^2} + 2\log\hat{\sigma}_{ij} + \log(2\pi)\right).$$

To emphasize early post-dose kinetics, we use weights

$$w_{ij} = 1 + w_{\text{amp}} \exp\left(-\frac{\Delta_{ij}^{\text{last}}}{w_\tau}\right),$$

with $w_{\text{amp}} = 4$ and $w_\tau = 0.10$ h. The weighted individual-fit term is

$$\mathcal{L}_{\text{NLL,ind}}^{\text{w}} = \frac{\sum_{i,j} w_{ij}\ell_{ij}}{\sum_{i,j} w_{ij}}.$$

Training also includes KL regularization to the regimen-conditioned prior,

$$\mathcal{L}_{\text{KL}} = \sum_i \text{KL}(q_\phi(z_i \mid r_i) \parallel p(z_i \mid x_i)),$$

a weighted population-anchor term evaluated at $z_i = m(x_i)$,

$$\mathcal{L}_{\text{NLL,pop}}^{\text{w}} = \frac{\sum_{i,j} w_{ij}\ell_{ij}^{\text{pop}}}{\sum_{i,j} w_{ij}},$$

where ℓ_{ij}^{pop} is the Gaussian negative log-likelihood evaluated at $z_i = m(x_i)$. We also include a residual ODE penalty

$$\mathcal{L}_{\text{res}} = \mathbb{E}_{t\in\mathcal{T}_{\text{lag}}}[s(t)^2].$$

The final loss is

$$\mathcal{L} = \mathcal{L}_{\text{NLL,ind}}^{\text{w}} + \lambda_{\text{pop}}\mathcal{L}_{\text{NLL,pop}}^{\text{w}} + \lambda_{\text{KL}}\frac{1}{B}\mathcal{L}_{\text{KL}} + \lambda_{\text{res}}\mathcal{L}_{\text{res}}.$$

We train with AdamW (learning rate 5×10^{-4}, weight decay 10^{-6}), gradient clipping at global norm ≤ 1.0 and ReduceLROnPlateau (patience 5, factor 0.5). Models are trained for up to 150 epochs with early stopping (patience 20). Typical hyperparameters are latent dimension $D = 6$, bidirectional GRU hidden size 128, number of exponential components $J = 8$ and residual ODE hidden size 256.

B.5 Post-training empirical Bayes calibration

After training, we calibrate using train+validation only. MAP subject latents are computed by minimizing

$$\mathcal{J}(z_i) = \mathcal{L}_{\text{NLL}}(\tilde{y}_i \mid z_i) + \lambda_{\text{prior}}\frac{1}{2}(z_i - m(x_i))^\top\Omega^{-1}(z_i - m(x_i)).$$

Residual standard deviation is then updated from MAP residuals

$$e_{ij} = \tilde{y}_{ij} - \hat{\mu}_{ij}(z_i^{\mathrm{MAP}}), \qquad \sigma_{\mathrm{base}} \leftarrow \sqrt{\frac{1}{N_{\mathrm{obs}}} \sum_{i,j} e_{ij}^2}$$

When sufficient early points exist, the optional peak-noise parameter a is calibrated so that the upper quantile of normalized absolute residuals matches the standard Normal reference. Finally, (m_0, β) are refit by least-squares regression of z_i^{MAP} on x_i and Ω is refit from the residual latent covariance with diagonal jitter as needed.

B.6 Baselines and strict transfer protocol

We compare against a matched nonlinear mixed-effects reference and two matched dose-aware neural baselines. DoseJump-NODE evolves a latent state with a neural ODE between dose times and applies additive jumps at dose events. Kernel-MLP uses the same dose-superposition form as ConstrainNODE-PK but parameterizes the unit-dose kernel with a generic positive MLP. To ensure a fair comparison, both neural baselines use the same variational encoder, regimen-conditioned prior, weighted likelihood, population-anchor loss and train+validation-only empirical Bayes calibration pipeline as ConstrainNODE-PK.

For strict QD→TID transfer, total daily dose is held fixed between source and target regimens so that the transfer isolates schedule change rather than dose-scale change. All model fitting and calibration are completed before any TID test concentrations are seen. At evaluation time, subject latents are inferred from QD data only and the full TID profile is predicted without using any TID concentrations for inference or recalibration.

B.7 NLME reference specification

The NLME references were fit with `nlmixr2` using the same event-table inputs and the same subject-level train/validation/test splits as ConstrainNODE-PK, without additional demographic or laboratory covariates (Fidler et al., 2019; Schoemaker et al., 2019). For the simulated cohorts with known data-generating mechanisms, the reference model was fit within the correct mechanistic family by construction: one-compartment for Drug1C, two-compartment for Drug2C and three-compartment for Drug3C, each under linear IV-bolus kinetics. Random effects were placed on the principal structural parameters of the chosen model and the residual error model was proportional on the concentration scale.

For the public datasets, a small candidate set of plausible linear IV-bolus compartmental models was fit using training data, with structural selection performed using validation IPRED logRMSE only; the held-out test split was not used for structural selection or population-parameter estimation. Estimation used `nlmixr2` with `FOCEI`. For held-out IPRED evaluation, subject-specific random effects were obtained by MAP optimization with the fitted population parameters held fixed. For POP, unconditional simulations were generated by sampling random effects from the fitted Ω distribution and residual noise from the fitted observation model and PI95/NPDE were computed from these simulations.

B.8 Metrics and ablations

IPRED point accuracy is measured with pooled logRMSE,

$$\text{logRMSE} = \sqrt{\frac{1}{N_{\text{obs}}} \sum_{i,j} \left(\log C_{ij} - \widehat{\log C}_{ij} \right)^2}.$$

where N_{obs} is the total number of evaluated observations.

Exposure summaries are median absolute percentage error (MdAPE) for AUC_{0-t} and $C_{\max}$. POP calibration is measured by PI95 and NPDE, using rank-based probability integral transforms from unconditional prior-predictive simulations. Forecast logRMSE is computed by fitting/refining z on the first half of each subject's observations and evaluating on the second half.

The ablations are: Full model; Prior without regimen conditioning ($\beta = 0$); Single-exponential kernel ($J = 1$); Uniform likelihood weighting ($w_{ij} \equiv 1$); and No residual ODE (residual multiplier fixed to 1).

Appendix C. Detailed cohorts and synthetic generation

C.1 Shared data representation and splits

All experiments use subject-level train/validation/test splits targeted at approximately 60/20/20, with integer rounding at the subject level, shared across ConstrainNODE-PK, NLME and neural baselines. All datasets are represented in a common long event-table format with columns

$$\{\texttt{dataset}, \texttt{id}, \texttt{time}, \texttt{dv}, \texttt{evid}, \texttt{amt}\}.$$

Dose events are encoded by $\texttt{evid} = 1$ with $\texttt{amt} > 0$ and $\texttt{dv} = 0$, while concentration observations are encoded by $\texttt{evid} = 0$ with $\texttt{amt} = 0$ and $\texttt{dv} > 0$. When an observation and a dose share the same nominal timestamp, the observation is placed immediately before the dose to preserve trough semantics.

C.2 Synthetic generation scheme

Synthetic datasets were generated under linear compartment PK models with log-normal inter-individual variability, weight-based dosing, realistic sampling jitter and log-normal residual error.

Body weight was sampled and clipped to a physiologically plausible range:

$$WT_i^* \sim \mathcal{N}(70,\ 15^2)\ \text{kg}, \tag{3}$$

$$WT_i = \text{clip}(WT_i^*;\ 40,\ 100)\ \text{kg}. \tag{4}$$

Subject-specific PK parameters followed

$$\theta_i = \theta_{\text{typ}}(WT_i) \exp(\eta_i), \qquad \eta_i \sim \mathcal{N}(0, \Omega),$$

with coefficient-of-variation parameters converted to log-scale standard deviations by

$$\omega = \sqrt{\log(1 + \text{CV}^2)}.$$

Clearances were allometrically scaled with exponent 0.75 and volumes with exponent 1.0 (Anderson and Holford, 2008):

$$CL_i = CL_{70}\left(\frac{WT_i}{70}\right)^{0.75}\exp(\eta_{CL,i}), \qquad V_i = V_{70}\left(\frac{WT_i}{70}\right)^{1.0}\exp(\eta_{V,i}).$$

Single-dose cohorts received one dose at $t = 0$. Multiple-dose cohorts used

$$t_{ik} = k\tau, \qquad k = 0, \ldots, n_{\text{doses}} - 1, \qquad a_{ik} = \text{dose}_{\text{mg/kg}} \cdot WT_i.$$

Observation times were generated from fixed templates with interval-level shift, pointwise Gaussian jitter, sorting and a minimum-gap rule to avoid overlap with dose times. Observation noise was multiplicative log-normal:

$$C_i^{\text{obs}}(t) = C_i(t)\exp(\epsilon), \qquad \epsilon \sim \mathcal{N}(0, \sigma_{\log}^2), \qquad \sigma_{\log} = \sqrt{\log(1 + \text{CV}_{\text{prop}}^2)}.$$

No BLQ observations were present in the final analysis datasets, so no BLQ handling was required.

C.3 Drug1C

The one-compartment cohort used typical parameters at 70 kg of $CL_{70} = 4.0\,\text{L/h}$ and $V_{70} = 20.0\,\text{L}$, with correlated IIV on (CL, V):

$$\text{CV}_{CL} = 0.30, \qquad \text{CV}_V = 0.25, \qquad \rho_{CL,V} = 0.30.$$

For subject i,

$$k_{e,i} = CL_i/V_i, \qquad g_i(t) = \frac{1}{V_i}\exp(-k_{e,i}t), \qquad t \geq 0.$$

Multiple-dose simulations used $\tau = 24\,\text{h}$, $n_{\text{doses}} = 5$ and $\text{dose}_{\text{mg/kg}} = 5.0$. Nominal single-dose sampling times (hours) were

$$(0.083,\ 0.25,\ 0.5,\ 1,\ 2,\ 4,\ 6,\ 8,\ 12,\ 24),$$

and nominal within-interval multiple-dose sampling times were

$$(0.083,\ 0.25,\ 0.5,\ 1,\ 2,\ 4,\ 6,\ 8,\ 12,\ 23.5).$$

Residual variability used $\text{CV}_{\text{prop}} = 0.15$.

C.4 Drug2C

The two-compartment cohort used typical parameters at 70 kg of

$$CL_{70} = 4.0, \quad V_{c,70} = 20.0, \quad Q_{70} = 6.0, \quad V_{p,70} = 30.0,$$

with correlated blocks (CL, V_c) and (Q, V_p), each with correlation 0.30 and CVs of 0.30 for clearances and 0.25 for volumes.

Micro-rate constants were

$$k_{10} = \frac{CL_i}{V_{c,i}}, \qquad k_{12} = \frac{Q_i}{V_{c,i}}, \qquad k_{21} = \frac{Q_i}{V_{p,i}}.$$

The macro-rate constants (α_i, β_i) were

$$\alpha_i, \beta_i = \frac{1}{2}\left(k_{10} + k_{12} + k_{21} \pm \sqrt{(k_{10} + k_{12} + k_{21})^2 - 4k_{10}k_{21}}\right),$$

and the central concentration response to bolus dose D_i was

$$C_i(t) = \frac{D_i}{V_{c,i}}\left(A_i e^{-\alpha_i t} + B_i e^{-\beta_i t}\right), \qquad t \geq 0,$$

with

$$A_i = \frac{\alpha_i - k_{21}}{\alpha_i - \beta_i}, \qquad B_i = \frac{\beta_i - k_{21}}{\beta_i - \alpha_i}.$$

Multiple-dose profiles were generated by superposition. Residual variability used $\mathrm{CV}_{\mathrm{prop}} = 0.15$.

C.5 Drug3C

The three-compartment cohort used typical parameters at 70 kg of

$$CL_{70} = 4.0, \quad V_{c,70} = 20.0, \quad Q_{1,70} = 6.0, \; V_{p1,70} = 30.0, \quad Q_{2,70} = 4.0, \; V_{p2,70} = 25.0,$$

with correlated blocks (CL, V_c), (Q_1, V_{p1}) and (Q_2, V_{p2}), each with correlation 0.30 and CVs of 0.30 for clearances and 0.25 for volumes.

With state $A(t) = [A_c(t), A_{p1}(t), A_{p2}(t)]^\top$, the dynamics were

$$\frac{dA}{dt} = K_i A, \qquad K_i = \begin{bmatrix} -(k_{10} + k_{12} + k_{13}) & k_{21} & k_{31} \\ k_{12} & -k_{21} & 0 \\ k_{13} & 0 & -k_{31} \end{bmatrix},$$

where

$$k_{10} = \frac{CL_i}{V_{c,i}}, \quad k_{12} = \frac{Q_{1,i}}{V_{c,i}}, \quad k_{21} = \frac{Q_{1,i}}{V_{p1,i}}, \quad k_{13} = \frac{Q_{2,i}}{V_{c,i}}, \quad k_{31} = \frac{Q_{2,i}}{V_{p2,i}}.$$

For a bolus dose D_i into the central compartment,

$$C_i(t) = \frac{1}{V_{c,i}}\, e_1^\top \exp(K_i t)\, e_1\, D_i, \qquad e_1 = [1, 0, 0]^\top,$$

which is a real sum of exponentials. In practice, the impulse response was evaluated via eigendecomposition and used within the linear superposition formula for repeated dosing.

C.6 ARK iohexol cohort

For evaluation on a public cohort, we used the African Research on Kidney Disease (ARK) Consortium iohexol study (Kalyesubula et al., 2020; Fabian et al., 2022). We used the final filtered public `ARKinput_final.csv` analysis cohort ($N = 2578$), after the ARK repository-defined inclusion/exclusion filters were applied. Using the same approximate 60/20/20 subject-level split rule as above yielded 1547/516/515 train/validation/test participants. The available data are dominated by sparse elimination-phase samples, nominally around 120, 180 and 240 minutes. Dose and concentration records were converted into the shared event-table format without additional BLQ censoring or imputation.

C.7 Tobramycin cohort

For a second public-cohort evaluation, we used the Tobramycin population PK dataset originally reported by Aarons and colleagues (Aarons et al., 1989). The public version contains 97 patients and 322 concentration measurements, with 1–9 measurements per patient and available covariates including age, weight, sex and creatinine clearance. Patients received intermittent intravenous Tobramycin under multiple dosing regimens; doses range from 20 to 140 mg and the event table is dominated by roughly q8h dosing but also includes regimen changes and irregular timings. Most concentration measurements occur 2–6 h post-dose. Using the same approximate 60/20/20 subject-level split rule as above yielded 58/19/20 train/validation/test subjects. Original dosing and observation columns were mapped to the shared event-table format, with dose administrations represented as event times and predose observations placed immediately before a same-time dose event when needed.