

Fitness functions for population pharmacokinetic modeling

Zhonghui Huang

zhonghui.huang.20@ucl.ac.uk

University College London

Matthew Fidler

Novartis Pharmaceuticals Corporation

Joseph F Standing

University College London

Frank Kloprogge

University College London

Research Article

Keywords: population pharmacokinetics, automated modeling, fitness function, search algorithms

Posted Date: July 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10418026/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. Matthew Fidler is an employee of Novartis. All other authors declared no competing interests for this work.

Fitness functions for population pharmacokinetic modeling

Zhonghui Huang¹, Matthew Fidler², Joseph F Standing^{1,3}, Frank Klopprogge⁴

¹Great Ormond Street Institute of Child Health, University College London, London, UK

²Novartis Pharmaceuticals Corporation, Fort Worth, Texas, USA

³Great Ormond Street Hospital for Children, London, UK

⁴Institute for Global Health, University College London, London, UK.

Corresponding author: Zhonghui Huang

Affiliation: Great Ormond Street Institute of Child Health, University College London, London, UK

Address: 30 Guilford St, London WC1N 1EH

E-mail: zhonghui.huang.20@ucl.ac.uk

Keywords population pharmacokinetics · automated modeling · fitness function · search algorithms

Abstract

Automated pharmacokinetic model building requires predefined criteria for model evaluation to rank candidate models and guide model selection. Evaluation criteria that rely on a single metric or are excessively stringent may bias the model toward overfitting or toward an oversimplified structure. Therefore, an appropriate fitness function is needed to enable comprehensive evaluation across multiple dimensions while balancing potential trade-offs among these criteria. This study aimed to systematically assess the model evaluation metrics commonly used in population pharmacokinetic (PopPK) modeling and to identify a fitness function for algorithm-driven model development. Binary and step penalty designs were compared for model selection performance. To this end, 48 datasets were simulated based on optimal design, and an exhaustive search approach was used to test all candidate models in the predefined model space. This allowed the original model structure captured in the simulated datasets to serve as the reference, and enabled assessment of the function's ability to recover that structure within a complete search space. The results indicate that the relative standard error (RSE) of parameter estimates, eta-shrinkage, and the omega values of inter-individual variability are key evaluation metrics for accurately identifying the true model and the step penalty was more robust than the binary penalty in selecting model structure. Based on these findings, this study proposes a fitness function combining BIC with threshold-based step penalties. The fitness function could play a central role in automated PopPK model building and provide insights for conventional model selection.

37 Introduction

38 Nonlinear mixed-effect (NLME) models have become a predominant approach for population pharmacokinetic
39 (PopPK) analysis to characterize typical drug behavior and quantify inter-individual variability [1, 2].
40 Conventional model development often requires extensive manual intervention and expert adjustment, resulting
41 in a time-consuming process. Consequently, automated NLME modelling has received increasing attention and is
42 gradually being adopted in practice [3–10]. In general, these algorithms formulate model development as a search
43 process over a predefined model space. Candidate models are constructed from predefined model features,
44 assessed for their quality, and then iteratively updated according to a search strategy to identify the model that
45 best describes the observed data.

46 A central issue in this search process is how model quality is evaluated to choose the “best” model. In manual
47 modelling, model evaluation entails a comprehensive assessment including goodness-of-fit diagnostics,
48 simulation-based diagnostics, assessment of parameter estimate precision and consideration of physiological
49 plausibility [1, 2, 11]. Empirical thresholds are often used to assess parameter precision, with the relative standard
50 errors (RSEs) for typical parameters generally expected to be below 20-30% [12, 13] and inter-individual
51 variability (IIV) shrinkage below 20-30% [14]. In practice, modelers may accept relatively high RSE values under
52 certain circumstances. For example, an RSE of up to 60% was ever accepted [15] in light of the overall model
53 performance. However, automated modelling cannot rely on experience and flexible judgement without a human-
54 in-the-loop. Algorithm-driven processes need a clear and rigid core optimization objective, namely a fitness
55 function. The fitness function assesses the quality of each candidate solution or model by mapping it to a
56 comparable numerical value, which enables comparison and ranking, and guides selection and updates [16].

57 For fitness function design, one common approach is to base model selection solely on information criteria such
58 as the Akaike information criterion (AIC) [5] or the Bayesian information criterion (BIC) [9]. Although these
59 criteria explicitly trade off goodness of fit against model complexity, they do not account for parameter precision
60 or plausibility of parameter estimates themselves. Alternatively, the fitness value can be derived from the objective
61 function with additional penalty terms to incorporate constraint violations [17]. Penalty terms such as convergence
62 failure, unsuccessful covariance step, IIV correlations exceeding 0.95, and shrinkage exceeding 70% have been
63 incorporated into previous fitness functions [3, 6]. Penalties were imposed in a binary fashion, determined solely
64 by whether a given threshold was crossed. Strict binary penalties tend to bias model selection toward conservative
65 outcomes [6]. Moreover, it has been argued that clear guidance on selecting an appropriate fitness function to
66 balance model accuracy and complexity is lacking [18].

67 To address this, a penalty function with step-like characteristics was proposed [4]. Instead of using binary
68 thresholds, it applied increasing penalties in discrete ranges. However, the thresholds and penalty magnitudes
69 were first guided by domain expertise and then optimized using four clinical datasets. No research to date has
70 systematically assessed the design of the fitness function itself and its performance in model selection or employed
71 simulated optimal-design datasets, to provide ground truth benchmarks. Without exhaustive searches of the model
72 space, one cannot guarantee that the model selected by a given fitness function is truly optimal, since unvisited
73 regions of the space may contain better-fitting models that were never considered.

The aim of this study was to develop a robust and interpretable fitness function framework to support automated PopPK model development. This framework has been incorporated into nlmixr2auto [19], a new open-source R package for automated PopPK model building and selection, where it serves as a model-evaluation component. The core objectives were to explore how each penalty term independently contributes to model selection, how different penalty magnitudes influence selection outcomes, and how binary versus step-penalty designs compare in their ability to identify the data-generating models, as well as to derive a practically implementable fitness function for model selection.

Method

Study workflow

The overall framework for the design and evaluation of fitness functions is illustrated in **Fig. 1**. Simulated datasets were generated from 48 distinct nonlinear mixed-effect models using optimal sampling design. For each dataset, parameter estimation was performed exhaustively for all candidate models within the predefined search space, including 711 models for intravenous (IV) administration or 1422 models for oral routes. For fitness function design, candidate fitness functions were initially constructed under a binary penalty strategy by varying three information criteria (AIC, BIC, and objective function value (OFV)), eight penalty-term sets, eight penalty magnitudes. The step penalty designs were subsequently evaluated, and threshold sensitivity analyses were conducted to refine the choice of cutoff values. The model with the lowest fitness value under a given fitness function was identified as the best-fitted model. The best-fitted model was then compared to the original model, and then the retrieval rate of each model component was calculated across all 48 datasets.

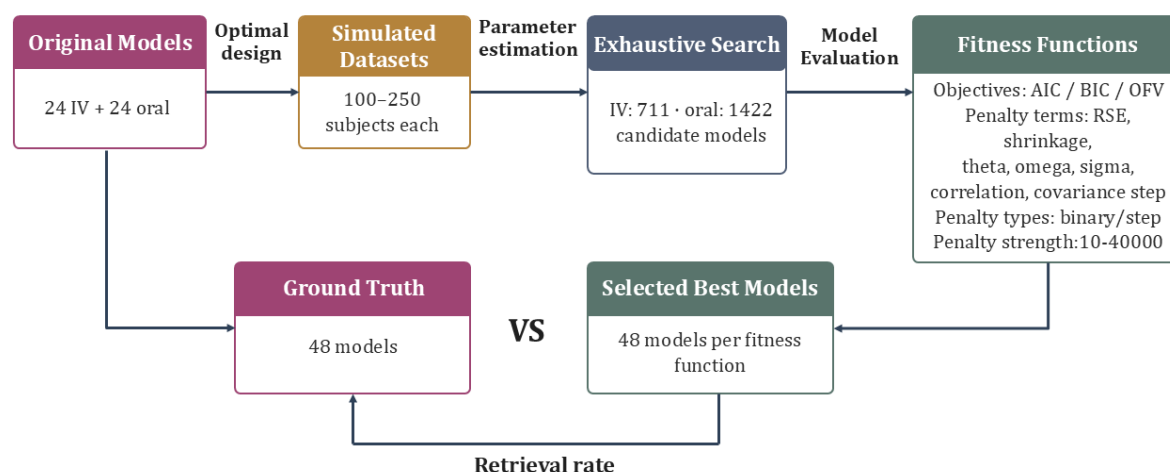


Fig. 1 Framework for the design and evaluation of fitness functions in automated PopPK model selection. RSE, relative standard error of theta (fixed-effect) parameters; omega, inter-individual variability (IIV) expressed as coefficient of variation (%CV); shrinkage, shrinkage of IIV estimates; sigma, residual unexplained variability; correlation, coefficient between IIV parameters

Simulated datasets

The 48 PK models comprised 24 intravenous and 24 oral models. Within each administration route, eight models were specified for each of the additive, proportional, and combined residual error models. Each set of eight models

consisted of two one-compartment models, three two-compartment models, and three three-compartment models, which differed in their IIV configurations. For the intravenous one-compartment models, IIV was assigned either to clearance (CL) alone or jointly to CL and central volume of distribution (V_c). For the intravenous two- and three-compartment models, the three IIV configurations included IIV on CL and V_c , on CL and peripheral volume of distribution (V_p), or jointly on CL, V_c , V_p , and Q . The oral models followed the same overall design, except that the CL and V_c IIV configuration in one of the one-compartment models was extended to include absorption rate constant (K_a).

All models share a set of parameter values. The CL and steady-state volume of distribution (V_{ss}) were derived from the median values reported across more than 1,000 drug compounds in a reference study [20], with CL set to 18.9 L/h and V_{ss} set to 63 L. V_{ss} was used as the total volume of distribution to determine the individual compartment volumes, with V_{ss} partitioned into V_c and V_p for the two-compartment model and into V_c , V_{p1} and V_{p2} for the three-compartment model. For the two-compartment model, the ratio $V_c:V_p$ was 1:2, whereas for the three-compartment model, the ratio $V_c:V_{p1}:V_{p2}$ was 1:1:3. IIV variance was set at 30% on the CV scale for all parameters. The Q was set at 10 L/h. Additive error was set as 2ng/ml and proportional error was set at 20%. A single dose of 100 mg was administered across all models.

An optimal design was applied to generate the sampling schemes. Sampling time points were determined with PopED [21] based on the D-optimality criterion, maximizing the determinant of the Fisher information matrix. The adequacy of the design was verified through re-estimation for parameters, requiring relative standard errors (RSE) below 20% for key parameters and shrinkage below 30%. The number of sampling points was increased stepwise from the minimum corresponding to the number of typical parameters, while the sample size started from 100 individuals. Expansion of sampling points or sample size was terminated once no further improvement in precision or shrinkage was observed. The characteristics of the simulated optimal design datasets are summarized in **Supplemental Material 1** and the simulated datasets are provided in **Supplemental Material 2**.

Exhaustive search

To evaluate the selection performance of the fitness functions, an exhaustive search was performed over the model space for each dataset. The model space was defined by the following factors:

- Number of compartments (no. compartment): one-, two- or three-compartment models;
- IIV specifications: configurations of IIV on V_c , V_p , V_{p2} , Q , and Q_2 ;
- IIV covariance structure: independent (diagonal IIV covariance matrix), assuming independence across IIV terms or unstructured, allowing correlation among all IIV terms;
- Elimination type: first-order elimination or Michaelis–Menten kinetics;
- Residual error models: additive, proportional, or combined.

All candidate models included IIV on CL or maximum elimination rate (V_{max}) by default to ensure at least one IIV component per model. Under this specification, the search space comprised 711 intravenous and 1422 oral candidate models.

Fitness function design

The fitness functions used in this study follow a penalty-based formulation [17], combining a baseline objective criterion with penalty terms that reflect violations of predefined diagnostic criteria. Each penalty term is constructed as the product of a penalty weight and a function indicating whether, or to what extent, a diagnostic threshold is violated. The overall fitness value is obtained by summing the baseline criterion and all associated penalties.

$$Fitness(M) = Base(M) + \sum_{m=1}^M \lambda_m p_m(M).$$

Here, M denotes a candidate PopPK model evaluated. $Base(M)$ denotes the model evaluation criterion (OFV, AIC, or BIC). λ_m is the penalty weight for the m_{th} diagnostic criterion. $p_m(M)$ is the corresponding non-negative penalty function, which maps diagnostic violations to a penalty magnitude.

Threshold-based constraints are imposed on three aspects:

- Parameter boundaries: including fixed-effect parameters (THETA, population-level effects), inter-individual variability parameters (OMEGA, variances of random effects expressed as coefficient of variation (CV%)), residual error parameters (SIGMA, including the standard deviation of the additive error σ_{add} and the proportional error σ_{prop}) as well as correlations between IIV parameters.
- Parameter quality diagnostic metrics: RSE % of fixed-effect parameters and shrinkage of IIV estimates.
- Model run diagnostics: the successful completion of the covariance step.

This study designed two schemes (Fig. 2) for the penalty term, a binary scheme that maps diagnostic violations to penalty magnitudes and a step-penalty scheme that assigns increasing penalty levels according to severity of the violation, as described below.

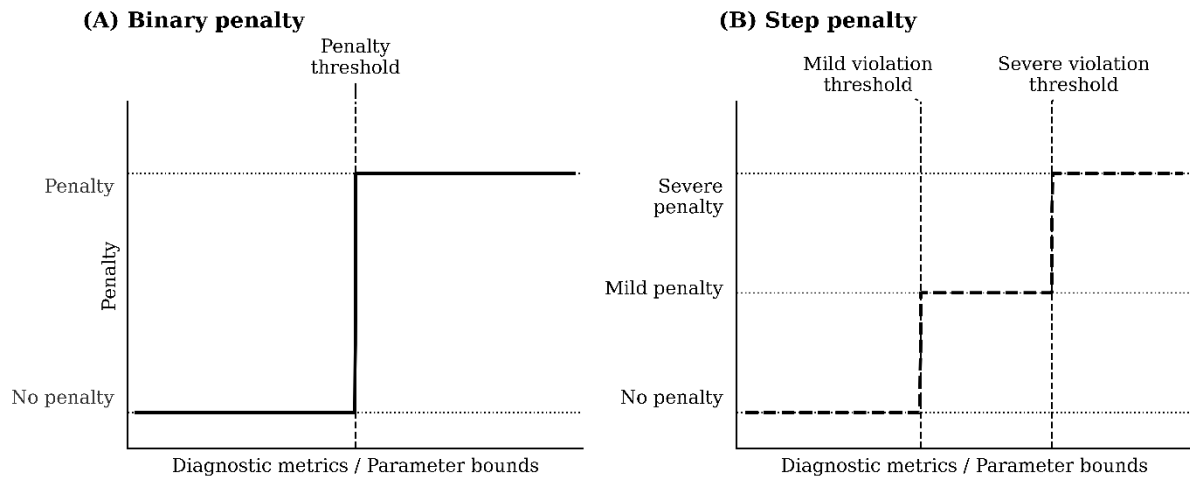


Fig. 2 Comparison of binary and step-penalty frameworks for diagnostic metrics. The binary penalty (A) applies a fixed penalty when diagnostic thresholds are crossed, based on a binary (0/1) indicator. The step penalty (B) introduces graded cutoffs, distinguishing between no penalty, mild violation, and severe violation, thereby reflecting different degrees of deviation in diagnostic metrics

Binary penalty

A single threshold is specified for each diagnostic criterion, and a fixed penalty is imposed once the threshold is crossed. The formulation is given as follows:

$$\begin{aligned}
 Fitness = Base &+ \sum_{i=1}^{n_{\theta}} \lambda_{RSE} \cdot 1_{\{RSE_i > 20\%\}} + \sum_{j=1}^{n_{\omega}} \lambda_{Shrinkage} \cdot 1_{\{Shrinkage_j > 30\%\}} \\
 &+ \sum_{i=1}^{n_{\theta}} \lambda_{\theta} \cdot 1_{\{\theta_i > \theta_i^{bound}\}} + \sum_{j=1}^{n_{\omega}} \lambda_{BSV_j} \cdot 1_{\{BSV_j < 10\%\}} \\
 &+ \lambda_{\sigma_{add}} \cdot 1_{\{\sigma_{add} < 0.001 \cdot C_{min}\}} + \lambda_{\sigma_{prop}} \cdot 1_{\{\sigma_{prop} < 0.001\}} \\
 &+ \sum_{k=1}^{n_{corr}} \lambda_{\rho} \cdot 1_{\{\rho_k < 0.1 \vee \rho_k > 0.8\}} + \lambda_{covStep} \cdot 1_{\{covariance\ step\ failed\}}
 \end{aligned}$$

Here, *Base* denotes the unpenalized objective function that represents model goodness-of-fit (OFV, AIC, or BIC). $1_{\{\cdot\}}$ refers to the indicator function that takes the value of 1 when a diagnostic condition is satisfied and 0 otherwise. Each λ represents the corresponding penalty score. θ_i denotes the i_{th} fixed-effect parameter, and n_{θ} is the number of θ_i . RSE_i is the relative standard error of the i_{th} θ estimate. n_{ω} denotes the number of IIV parameters. $Shrinkage_j$ is the eta-shrinkage for the j_{th} IIV parameter. BSV_j denotes the between-subject variability (BSV) on the CV% scale, computed as $BSV_j = 100\sqrt{e^{\omega_j} - 1}$. θ_i^{bound} denotes the predefined bound for the θ_i . C_{min} refers to the average minimum observed concentration across subjects in the dataset. n_{corr} is the number of unique correlation pairs between IIV parameters, and ρ_k denotes the k_{th} correlation estimate between two IIV parameters. *Covariance step failed* indicates that standard errors could not be computed because the variance-covariance matrix of the parameter estimates was not successfully obtained.

Diagnostic thresholds were defined with reference to the following considerations. The RSE of a typical parameter is considered insufficiently precise when it exceeds 25% [12]. A shrinkage of 30% or higher indicates substantial uncertainty in model estimation [14, 22]. An IIV of 10% CV is generally regarded as indicative of low variability [23], and has even been considered abnormally low variability [24]. The threshold for residual error parameters is set at $\sigma_{add} < 0.001$ times C_{min} or $\sigma_{prop} < 0.001$, given that the linear dynamic range of LC-MS/MS is typically about three orders of magnitude [25]. At this level, the residual error is close to the lower limit of quantification (LLOQ), where an additive error component at the LLOQ scale is not reliably supported by the data [26]. In addition, correlations between IIV parameters with coefficients less than 0.1 can generally be considered negligible or weak, following statistical interpretations [27]. In contrast, correlations exceeding 0.8 suggest strong associations among random effects, indicating that they cannot be independently and reliably estimated in PopPK models [28]. For θ parameters, all estimates were required to remain finite. In addition, only K_m (the Michaelis-Menten constant, representing the substrate concentration at which the metabolic rate reaches half of its maximum) was subject to a lower-bound constraint, as in practice [29, 30] estimates exceeding the observed concentration range suggest that saturation is not supported by the data.

Under the binary penalty framework, fitness functions were constructed either by applying all penalty terms simultaneously (full penalty) or by activating a single penalty term in isolation (e.g., RSE only, shrinkage only) while holding all other terms at zero. These single-term configurations were designed to assess the individual contribution of each penalty term to model retrieval with penalty scores ranging from 10 to 400,000.

Step penalty

Building upon the binary penalty scheme, this study introduced a step penalty framework to distinguish different degrees of deviation in diagnostic criteria. Each diagnostic metric is divided into three regions: no penalty, mild violation, and severe violation, with the initial cutoff thresholds summarized in **Table 1**. Taking RSE as an example, the mild cutoff was set at 20% and the severe cutoff at 40%. The two cutoffs specified that RSE values at or below 20% incurred no penalty, values above 20% but not exceeding 40% were treated as mild violations, and values above 40% were treated as severe violations. The mild violation default thresholds were adopted from the binary penalty scheme. Severe violation thresholds were optimized through local sensitivity analysis, where each candidate cutoff was adjusted individually while all others were kept fixed at their default value. The severe cutoffs for RSE and shrinkage were varied from 40 to 100% and from 50 to 100%, respectively, both in 10% increments. For omega values, the severe cutoff was explored across the range of 0-10%. In the residual error model, severe cutoffs were varied between 0 and 0.001 for σ_{prop} , and between 0 and 0.001 times C_{min} for σ_{add} . For IIV correlations, severe cutoffs were evaluated across the lower bound ranging from 0.01-0.05 and the upper bound ranging from 0.8-1.0. The strength of the mild penalty was also assessed, with values tested from 10 to 100 in increments of 10. In addition, failure of the covariance step was directly assigned a severe penalty.

Table 1 Initial cutoff thresholds for mild and severe violations

Metric	Mild cutoff	Severe cutoff
THETA -- RSE, %	20%	40%
OMEGA -- value (CV%)	10%	0
OMEGA -- η -shrinkage, %	30%	50%
SIGMA -- σ_{add} (rel. C_{min})	0.001	0
SIGMA -- σ_{prop}	0.001	0
Correlation -- lower (ρ)	0.1	0.05
Correlation -- upper (ρ)	0.8	0.95

Notes: RSE: relative standard error of THETA (fixed-effect) parameters; OMEGA value: inter-individual variability (IIV expressed as coefficient of variation (CV%)); η -shrinkage: eta-shrinkage of IIV estimates; σ_{add} : additive residual error (scaled relative to C_{min}); σ_{prop} : proportional residual error; ρ : correlation coefficient between IIV parameters.

Fitness function evaluation

The performance of each fitness function was evaluated based on its ability to recover the original models across 48 simulated datasets. Five evaluation metrics were defined, including structural model, IIV, residual error model, correlation model, and overall concordance across all dimensions. The IIV retrieval rate was calculated based only on datasets for which the structural model was correctly identified. For each fitness function, the selected models from 48 datasets were compared with the corresponding original models and retrieval rates defined as the number of correct identifications divided by the total were then calculated for each evaluation criterion. Stratified analyses were further performed to evaluate the retrieval rates across specific model components, including number of compartments (one-, two- or three-compartment), the presence or absence of M-M elimination, residual error model type (additive, proportional, or combined), and IIV groups (CL only; CL and V_c ; CL, V_c and K_a ; CL and V_p ; CL, V_c , V_p and Q).

Software

All analyses were performed in R [31], and model parameter estimation was carried out using the *nlmixr2* package [32], with parameter initialization facilitated by the *nlmixr2autoinit* [33].

Results

Penalty term contributions under binary penalty design

Fig. 3 presents the retrieval performance of the BIC-based fitness function for the full-penalty design and seven single-term designs, stratified by data type. In the single-term analyses, RSE, shrinkage, and omega values were the primary contributors, and the effects were component-specific. RSE improved structural identification in oral datasets, increasing overall retrieval from 88% at baseline to a peak of 92%, but then declining to 83% at higher penalties. For intravenous datasets, structural model retrieval remained consistently high, reaching 96%. Shrinkage enhanced both IIV and residual error model identification. The IIV retrieval rate increased from 59% in the absence of a penalty to 78% at the highest penalty, with a larger increase observed in intravenous datasets than in oral datasets, from 74% to 95% and 43% to 56%, respectively. Residual error model identification was also improved, with the overall retrieval rate increasing from 75% to 83%. Meanwhile, structural model identification decreased from 92% to 83%. Omega values also contributed to IIV identification, raising the overall retrieval rate to 80%, with rates of 86% and 72% for the intravenous and oral datasets. Correlation penalties provided only a slight improvement in correlation model identification, increasing retrieval from 96% to 98%. Little impact on other model components was observed when the remaining metrics were used alone.

The full-penalty design demonstrated the best overall performance. For IIV component identification, retrieval performance further improved to 82%, reaching 91% in intravenous datasets and 71% in oral datasets. In terms of the residual error model, 79% of datasets were correctly identified. For correlation model identification, performance remained stable at 95% overall, with intravenous data at 100% and oral data at 90%. However, the decline in structural identification associated with RSE, shrinkage, and omega value penalties was also retained. Accordingly, structural model retrieval decreased from 92% without penalty to 89% at a score of 10 and further to 81% as the penalty increased. This is largely driven by oral datasets, in which the retrieval rate declined from 88% to 78%. For penalty strength, the percentage of correctly identified models remains stable at penalty scores of 10000 or higher.

The corresponding results for AIC-based and OFV-based fitness functions are provided in **Supplemental Material 3, Figures 1 and 2**. Both followed trends similar to those observed with BIC-based fitness functions, but the corresponding retrieval rates were lower. At a penalty of 10000 with the full-penalty design, the rates across all data were 86%, 76%, 75%, and 82% for the structural model, IIV model, residual error model, and correlation model, whereas the OFV-based fitness function achieved 81%, 79%, 79% and 85%. In addition to RSE, shrinkage, and omega value penalties, correlation and theta penalties also contributed to model identification. The benefit of the RSE penalty for structural model identification was more evident with OFV-based fitness functions, increasing retrieval from 60% without penalty to 75% with penalty. Correlation penalties mainly improved correlation model identification, increasing from 23% to 65% for the OFV-based fitness function and 73% to 85%. With the introduction of the theta-value penalty, the structural model identification rate was 7 %

260 higher under both AIC and OFV. Consistent with BIC-based results, shrinkage and omega mainly enhanced IIV
261 and residual error identification, whereas structure identification declined by about 8%.

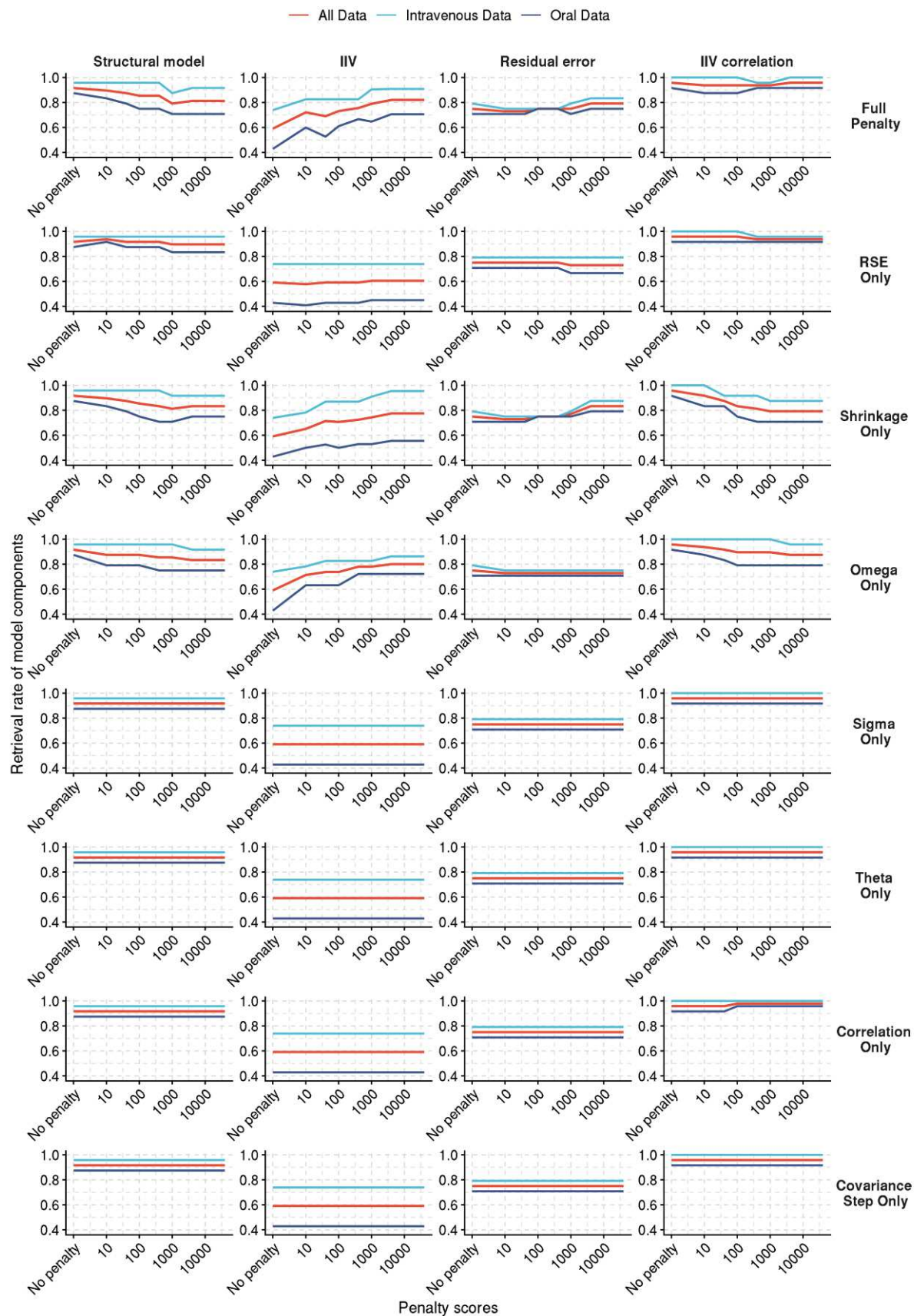


Fig. 3 Retrieval performance under different binary penalty configurations with BIC. Rows show penalty settings, columns show model components (structural model, IIV, residual error, IIV correlations). Configurations include single-term penalties (THETA, OMEGA, SIGMA, correlations, shrinkage, RSE, covariance step), the *full penalty*, and *no penalty*. The *full penalty* configuration considered all penalty terms jointly, while the *no penalty* configuration applied none. IIV, inter-individual variability; RSE, relative standard error; THETA, fixed-effect parameters; OMEGA, inter-individual variability expressed as coefficient of variation; SIGMA, residual error, including additive (σ_{add}) and proportional (σ_{prop}) components.

Step penalty and cutoff analysis

Sensitivity analysis results for the cutoff settings of RSE, shrinkage, and omega values are presented in **Fig. 4**, with those for sigma components and correlation values in **Supplemental Material 3, Figure 3**. Regarding RSE, the retrieval rate for structural model increased from 88 to 90% when a severe penalty was applied to RSE values exceeding 40% and remained stable thereafter. For shrinkage, retrieval for the structural model and IIV model peaked at 50%, leading to the selection of 50% as the severe cutoff. In terms of omega values, when the cutoff was set at 8% or higher, the retrieval rate for structural model increased to 90% and remained stable, while correlation performance also improved to about 94%. By contrast, the cutoffs for additive and proportional errors showed little sensitivity, with retrieval remaining essentially unchanged across settings, and both were therefore set to 0. The cutoffs for correlation were only weakly sensitive. When the lower cutoff was adjusted from 0.01 to 0.1, only a minor effect was observed at 0.1, with no substantive changes otherwise and the original setting (lower cutoff of 0.05) was retained. Changes to the upper cutoff produced no discernible impact, and the original value (0.95) was therefore preserved.

Comparison of binary and step penalty

Fig. 5 presents the final comparison between the binary and step penalty designs across different severe penalty scores after all penalty terms were incorporated. The primary differences between two penalty designs were observed in the structural model retrieval. Overall, structural model retrieval under the step penalty remained stable at 90% when the penalty score reached 1,000 or higher, in contrast to the declining trend observed under the binary penalty. Intravenous results remained consistent at 96% using the AIC-, BIC-, and OFV-based fitness functions, while oral results showed no decline, remaining at 83% with AIC and BIC, and 92% with OFV.

For IIV, residual error, and correlation model identification, binary and step penalty designs yielded essentially similar results when the penalty score reached 10,000. Using the BIC-based fitness functions, IIV and residual error retrieval differed by no more than 2-3 percentage points, with IIV retrieval of 82% and 81%, residual error retrieval of 79% and 77%, and correlation retrieval of 96% and 94%, respectively. Similar patterns were found under AIC and OFV, with differences remaining within 4-6%, respectively. Consistent with these findings, for full model retrieval, where correct identification of structural model, IIV, residual error, and correlation was required simultaneously, binary and step penalties also showed similar performance. The BIC-based fitness function under the step penalty achieved the highest full model retrieval rate (56%), outperforming the AIC- and OFV-based fitness functions (both 46%). This superiority was consistently observed in both the intravenous and oral datasets, with corresponding retrieval rates of 71% vs. 58% and 42% vs. 33%, respectively.

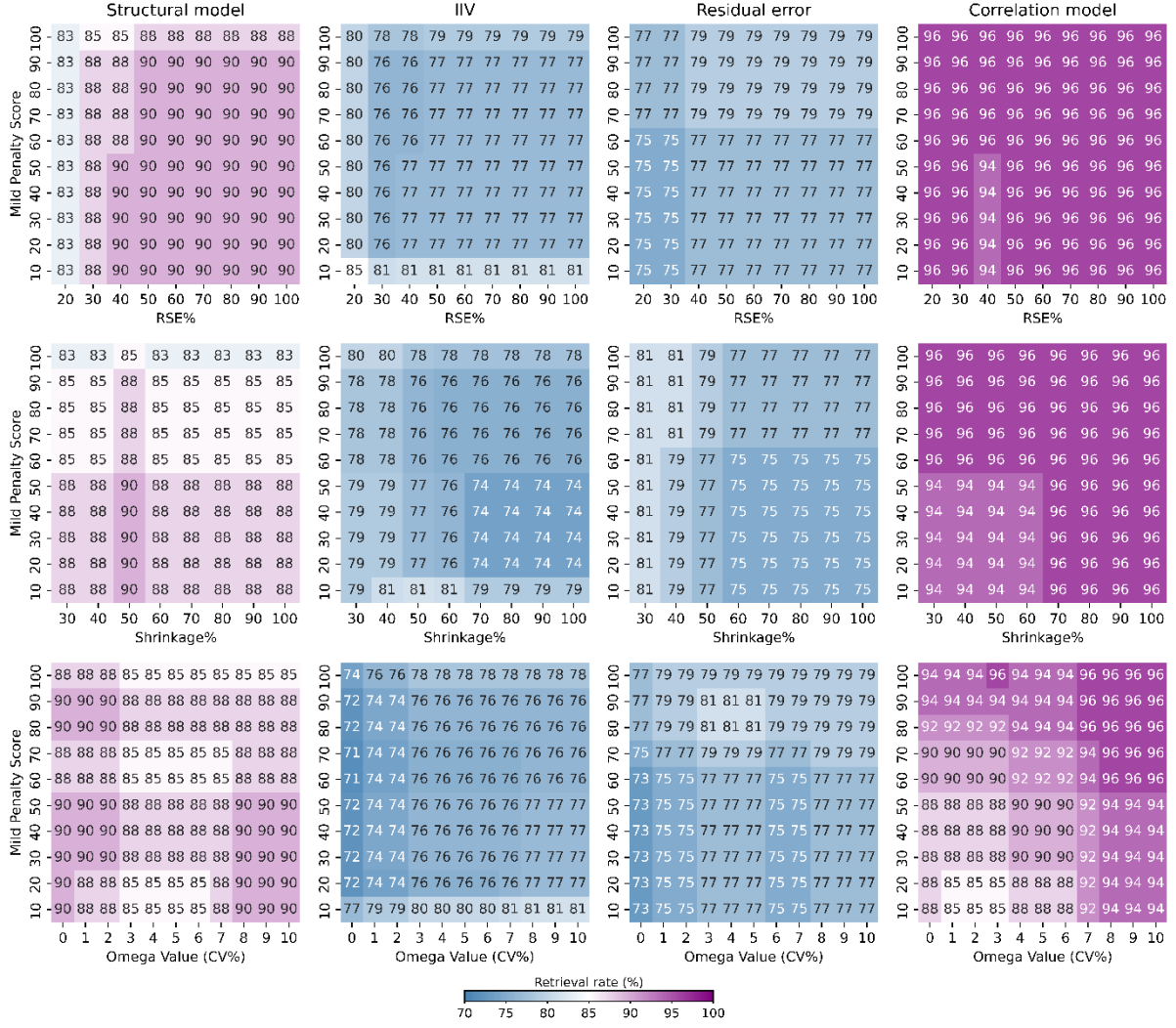


Fig. 4 Sensitivity analysis of cutoff settings for penalty function. Heatmaps show retrieval rates (in %) of correct model identification across varying cutoff thresholds and mild penalty scores. Rows correspond to cutoff criteria applied to relative standard error (RSE%), shrinkage (%), omega value (CV%). Columns represent evaluation dimensions: structural model, IIV model, residual error model and correlation model

Final fitness function and performance

The final fitness function was determined as follows:

$$\begin{aligned}
 \text{Fitness} = & BIC + \sum_{i=1}^{n_{\theta}} \lambda_{RSE} \cdot 1_{\{RSE_i > c_{RSE}\}} + \sum_{j=1}^{n_{\omega}} \lambda_{Shrinkage} \cdot 1_{\{Shrinkage_j > c_{Shrinkage}\}} \\
 & + \sum_{i=1}^{n_{\theta}} \lambda_{\theta} \cdot 1_{\{\theta_i > \theta_i^{bound}\}} + \sum_{j=1}^{n_{\omega}} \lambda_{BSV} \cdot 1_{\{BSV_j < c_{BSV}\}} \\
 & + \lambda_{\sigma_{add}} \cdot 1_{\{\sigma_{add} < c_{\sigma_{add}}\}} + \lambda_{\sigma_{prop}} \cdot 1_{\{\sigma_{prop} < c_{\sigma_{prop}}\}} \\
 & + \sum_{k=1}^{n_{corr}} \lambda_{\rho} \cdot 1_{\{\rho_k < c_{\rho}^{low} \vee \rho_k > c_{\rho}^{high}\}} + \lambda_{covStep} \cdot 1_{\{\text{covariance step failed}\}}
 \end{aligned}$$

where c denotes the cutoff threshold for the corresponding metric, and λ denotes the penalty weight associated with the metric. The specific cutoff thresholds and penalty weights are summarized in the following Table 2.

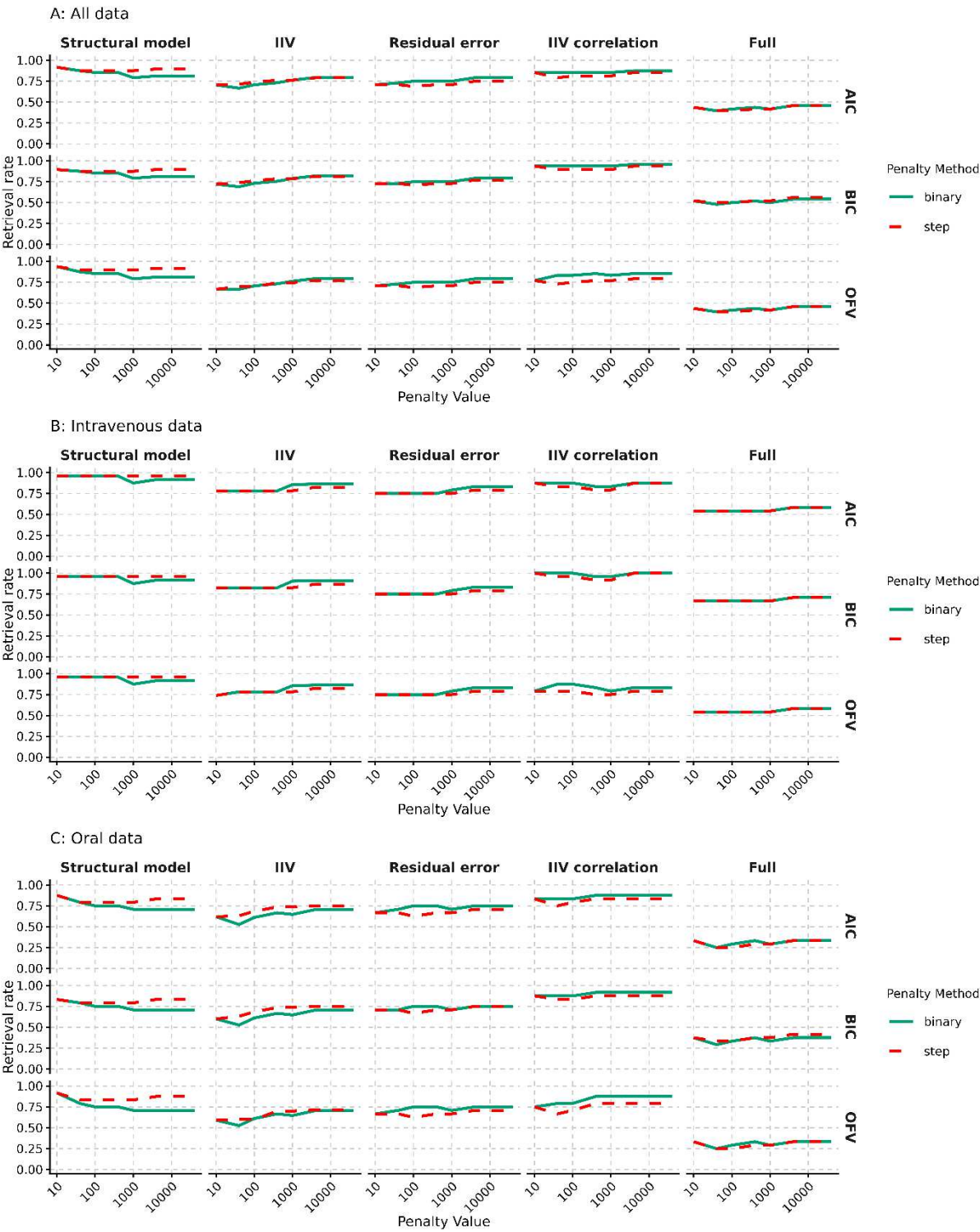


Fig. 5 Comparison of binary and step penalty methods under all penalties. Retrieval performance across five dimensions (structural model, inter-individual variability, residual error, correlation model, and complete model

retrieval (“Full”)) under binary versus step penalty methods, using AIC, BIC, and OFV as objective functions. Panels show results for (A) all data, (B) intravenous data, and (C) oral data

Table 2 Penalty scheme with mild and severe cutoffs for each metric

Metric	Symbol	Mild value	Severe value
THETA parameters - RSE	c_{RSE}	20%	40%
	λ_{RSE}	10	10000
OMEGA parameters - shrinkage	$c_{Shrinkage}$	30%	50%
	$\lambda_{Shrinkage}$	10	10000
OMEGA parameters - value (BSV, CV%)	c_{BSV}	10%	10%
	λ_{BSV}	10	10000
Residual error - additive (σ_{add}, %C_{min})	$c_{\sigma_{add}}$	0.001	0
	$\lambda_{\sigma_{add}}$	10	10000
Residual error - proportional (σ_{prop})	$c_{\sigma_{prop}}$	0.001	0
	$\lambda_{\sigma_{prop}}$	10	10000
Correlation - lower bound (ρ)	$c_{\rho_{low}}$	0.1	0.05
	λ_{ρ}	10	10000
Correlation - upper bound (ρ)	$c_{\rho_{high}}$	0.8	0.95
	λ_{ρ}	10	10000

Notes: c_{RSE} , $c_{Shrinkage}$, c_{BSV} , $c_{\sigma_{add}}$, $c_{\sigma_{prop}}$, $c_{\rho_{low}}$, $c_{\rho_{high}}$ denote cutoff thresholds for each metric. λ terms denote the corresponding penalty weights. RSE: relative standard error of THETA parameters; BSV: between-subject variability (expressed as coefficient of variation, CV%); σ_{add} : additive residual error (expressed relative to C_{min}); σ_{prop} : proportional residual error; ρ : correlation coefficient between parameters.

Table 3 presents the identification results and corresponding accuracy rates for each structural component using the proposed fitness function. Regarding the number of compartments, the one-compartment model was correctly identified in all 12 cases (12/12, 100%), and the two-compartment model was misclassified in only one case (17/18, 94%) as a three-compartment model. The three-compartment model was correctly identified in 14 out of 18 cases (78%), with four cases misclassified as two-compartment models. In selecting elimination models, all cases were successfully identified as linear elimination. In the screening of IIV structures, models with fewer IIV terms were more accurately identified. The IIV settings of CL, CL with V_c , and CL with V_p were nearly perfectly identified, with only one case of CL with V_p failing to detect V_p . For the more complex settings, the model with IIV on CL, V_c , and K_a was correctly identified in 7 of 9 cases, with the remaining two cases fitted as having IIV on CL and K_a . The model with IIV on CL, V_c , V_p , and Q was correctly identified in 7 of 12 cases, with five cases fitted as alternative IIV structures. In identifying residual variability models, the correct rates were 94% for the proportional model and 100% for the combined model, whereas the additive model had a lower success rate of 38%, with most misclassifications as the combined model. In identifying correlations, only three cases were misclassified as correlated.

Table 3 Comparison of model component identification results using the final fitness function

Number of compartments					
Fitted\original	1	2	3		
1	12	0	0		
2	0	17	4		
3	0	1	14		
Residual error					
Fitted\original	Additive	Combined	Proportional		
Additive	6	0	0		
Combined	10	16	1		
Proportional	0	0	15		
Inter-individual Variability					
Fitted\original	CL	CL+V _c	CL+V _c +K _a	CL+V _c +V _p +Q	CL+V _p
CL	6	0	0	0	0
CL+V _c	0	9	0	1	0
CL+V _c +K _a	0	0	7	0	0
CL+V _c +V _p +Q	0	0	0	7	0
CL+V _p	1	0	0	0	11
CL+K _a	0	0	2	0	0
CL+V _c +Q	0	0	0	2	0
CL+V _c +V _p +Q+K _a	0	0	0	1	0
CL+V _p +Q+K _a	0	0	0	1	0
Correlation model					
Fitted\original	IIV correlated	No correlation			
No correlation	3	45			
Elimination (M-M model)					
Fitted\original	First-order elimination				
First-order elimination	48				

Note: Rows correspond to the original models, and columns correspond to the fitted models. Diagonal entries indicate correct identification, while off-diagonal entries correspond to misclassifications.

Discussion

This study constructed a systematic framework for the design and evaluation of fitness functions in automated NLME modeling and ultimately proposed a BIC-based fitness function incorporating step-penalty strategies, ensuring objectivity and interpretability without expert intervention. Evaluation results showed that this fitness function effectively identified model features, accurately recovering the true data-generating characteristics represented in the test datasets.

The results from **Fig. 3** indicated that among the various candidate penalty terms, RSE, shrinkage, and omega values are the three most influential factors. Under BIC, the effect of RSE penalty was mainly observed in oral

346 datasets rather than the intravenous datasets. The latter already achieved a near-saturated structure retrieval rate
347 before the RSE penalty was applied, with the true model structure correctly identified in 96% of cases and stable
348 performance across the full penalty range. In contrast, the retrieval rate increased by 4% for the AIC-based fitness
349 function and by 16% for the OFV-based fitness function after the RSE penalty was introduced. This was likely
350 because the penalty term of BIC increases with sample size, leading it to favor more parsimonious models
351 compared with AIC [34]. Thus, because RSE similarly penalizes unstable or poorly estimated models, it provided
352 limited additional improvement. For oral data, however, the larger candidate model space (1422 vs 711), greater
353 structural complexity, and lower baseline retrieval rate (88%) allowed the RSE penalty to exert a greater influence
354 on model selection.

355 Regarding shrinkage, both binary and step penalty designs improved IIV retrieval, suggesting that shrinkage-
356 based penalization is useful for reducing unreliable IIV selection. Further results from **Fig. 4** show that when the
357 severe cutoff in the step penalty was relaxed above 60%, the retrieval rates of structure, IIV, and residual error all
358 declined by about 2%. Although thresholds as high as 70% have been reported under the first-order conditional
359 estimation with interaction (FOCEI) algorithm [6], shrinkage exceeding approximately 30 % or 50% renders
360 individual empirical Bayes estimates (EBEs) unreliable [14, 22]. The present evidence therefore suggests that
361 shrinkage should not be fixed at a high cutoff but rather determined systematically through cutoff analysis to
362 ensure reliable IIV detection without compromising structural accuracy. In contrast, omega values below 10%
363 required stronger penalization to avoid retaining implausibly small IIV terms. This was particularly evident in the
364 binary penalty design, where applying a strong penalty to omega values below 10% increased IIV retrieval from
365 74% to 95% in intravenous datasets and from 43% to 56% in oral datasets. However, under the step penalty design,
366 lowering the severe cutoff to 1-7% did not lead to a comparable improvement, suggesting that the penalty strength
367 was insufficient to eliminate very small, but unsupported IIV estimates. This finding is consistent with recent
368 automated modeling approaches that also penalize omega values below 10% to avoid spurious IIV retention [4].

369 This study found that overly stringent shrinkage penalties led to lower structural identification rates under binary
370 designs as shown in **Fig. 5**. One reason lies in the nature of the data. Although the data were generated from an
371 optimal design, such design tools mainly optimize the precision of parameter estimates via the Fisher information
372 matrix and do not explicitly target shrinkage [21]. As a result, some complex models may still exhibit shrinkage
373 above 30%. Consequently, the graded design of the step penalty imposes only mild restrictions when shrinkage
374 slightly exceeds 30%, reducing the risk of discarding multi-compartment models due to modest shrinkage
375 elevations, while applying stronger penalties only above the severe cutoff of 50%. This approach preserved
376 structural identification stability while simultaneously enhancing IIV retrieval, especially when the IIV value was
377 not set high in this study, as a previous study reported it may increase shrinkage [35].

378 Component-wise evaluation showed good retrieval performance for the structural model, IIV structure, and
379 correlation model. Residual error model retrieval was limited, mainly because additive error models were
380 frequently misclassified as combined models: only 6 of 16 additive cases were correctly identified, with most of
381 the rest misclassified as combined. Combined residual models were almost always correctly identified. This
382 pattern may be related to the observation in this study that adding a very small proportional residual error
383 component to an otherwise additive error model can sometimes improve estimation stability compared with using
384 a purely additive residual error model under SAEM estimation. There was a previous suggestion that combined

proportional and additive residual error models may describe PK data better than single proportional or additive residual error models [36]. Notably, these additive-to-combined misclassifications therefore contributed to the lower complete-model retrieval rate, where all model components were evaluated simultaneously.

Nonetheless, this study has certain limitations. Future research should evaluate a broader and more diverse range of parameter settings to further verify the robustness and applicability of the proposed method. Although simulated data allow controlled evaluation of penalty functions, real clinical data are typically more complex, with issues such as missing observations, measurement errors, and heterogeneous patient characteristics. Therefore, further validation on diverse clinical datasets is necessary to establish the generalizability and robustness of the penalty-based selection framework. In addition, the model search space was relatively limited, with the tested PK scenarios focusing primarily on simple linear absorption processes and not considering more complex structures such as lag time, transit absorption, or flip-flop absorption. Future work should extend validation to more complex pharmacokinetic contexts to further assess the applicability of the proposed approach. Although nonlinear elimination and IIV correlations were included in the candidate search space, their evaluation in the present study mainly assessed whether the fitness function could avoid selecting unsupported nonlinear or correlated structures, rather than its ability to recover true nonlinear elimination or true IIV correlation structures.

Through systematic evaluation of multiple candidate penalty terms, this study ultimately proposed a BIC-based fitness function combined with a step penalty. The method demonstrated high robustness and accuracy in simulated datasets, effectively recovering the original underlying models. These results highlight its potential utility in automated PopPK modeling. It provides a reference for a more systematic consideration of parameter estimation quality in model selection and offers a promising tool to support automated model selection in the nlmixr2 platform.

References

1. Committee for Medicinal Products for Human Use (CHMP) (2007) Guideline on Reporting the Results of Population Pharmacokinetic Analyses. London, UK
2. U.S. Food and Drug Administration (2022) Population Pharmacokinetics: Guidance for Industry. Silver Spring, MD
3. Bies RR, Muldoon MF, Pollock BG, et al (2006) A Genetic Algorithm-Based, Hybrid Machine Learning Approach to Model Selection. *J Pharmacokinet Pharmacodyn* 33:195–221. <https://doi.org/10.1007/s10928-006-9004-6>
4. Richardson S, Irurzun Arana I, Nowojewski A, et al (2025) A machine learning approach to population pharmacokinetic modelling automation. *Commun Med* 5:327. <https://doi.org/10.1038/s43856-025-01054-8>
5. Ismail M, Sale M, Yu Y, et al (2022) Development of a genetic algorithm and NONMEM workbench for automating and improving population pharmacokinetic/pharmacodynamic model selection. *J Pharmacokinet Pharmacodyn* 49:243–256. <https://doi.org/10.1007/s10928-021-09782-9>
6. Sibieude E, Khandelwal A, Girard P, et al (2022) Population pharmacokinetic model selection assisted by machine learning. *J Pharmacokinet Pharmacodyn* 49:257–270. <https://doi.org/10.1007/s10928-021-09793-6>
7. Li X, Sale M, Nieforth K, et al (2024) pyDarwin machine learning algorithms application and comparison in nonlinear mixed-effect model selection and optimization. *J Pharmacokinet Pharmacodyn* 51:785–796. <https://doi.org/10.1007/s10928-024-09932-9>

- 424 8. Sale M, Sherer EA (2015) A genetic algorithm based global search strategy for population
425 pharmacokinetic/pharmacodynamic model selection. *British Journal of Clinical Pharmacology* 79:28–39.
426 <https://doi.org/10.1111/bcp.12179>
- 427 9. Chen X, Nordgren R, Belin S, et al (2024) A fully automatic tool for development of population
428 pharmacokinetic models. *CPT: Pharmacometrics & Systems Pharmacology* 13:1784–1797.
429 <https://doi.org/10.1002/psp4.13222>
- 430 10. Uno M, Nakamaru Y, Yamashita F (2024) Application of machine learning techniques in population
431 pharmacokinetics/pharmacodynamics modeling. *Drug Metabolism and Pharmacokinetics* 56:101004.
432 <https://doi.org/10.1016/j.dmpk.2024.101004>
- 433 11. Byon W, Smith M, Chan P, et al (2013) Establishing Best Practices and Guidance in Population
434 Modeling: An Experience With an Internal Population Pharmacokinetic Analysis Guidance. *CPT:*
435 *Pharmacometrics & Systems Pharmacology* 2:51. <https://doi.org/10.1038/psp.2013.26>
- 436 12. Owen JS, Fiedler-Kelly J (2014) Model Building: Typical Process. In: *Introduction to Population*
437 *Pharmacokinetic/Pharmacodynamic Analysis with Nonlinear Mixed Effects Models*. Wiley, Hoboken, NJ, pp 90–
438 177
- 439 13. Williams PJ, Kim YH, Ette EI (2007) *The Epistemology of Pharmacometrics*. John Wiley & Sons, Ltd,
440 pp 223–244
- 441 14. Savic RM, Karlsson MO (2009) Importance of Shrinkage in Empirical Bayes Estimates for Diagnostics:
442 Problems and Solutions. *AAPS J* 11:558–569. <https://doi.org/10.1208/s12248-009-9133-0>
- 443 15. van Esdonk MJ, Lindeman I, Okkerse P, et al (2018) Population Pharmacokinetic/Pharmacodynamic
444 Analysis of Nociceptive Pain Models Following an Oral Pregabalin Dose Administration to Healthy Subjects.
445 *CPT Pharmacometrics Syst Pharmacol* 7:573–580. <https://doi.org/10.1002/psp4.12318>
- 446 16. Harman M, Jones BF (2001) Search-based software engineering. *Information and Software Technology*
447 43:833–839. [https://doi.org/10.1016/S0950-5849\(01\)00189-6](https://doi.org/10.1016/S0950-5849(01)00189-6)
- 448 17. Coello Coello CA (2002) Theoretical and numerical constraint-handling techniques used with
449 evolutionary algorithms: a survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering*
450 191:1245–1287. [https://doi.org/10.1016/S0045-7825\(01\)00323-1](https://doi.org/10.1016/S0045-7825(01)00323-1)
- 451 18. Janssen A, Bennis FC, Mathôt RAA (2022) Adoption of Machine Learning in Pharmacometrics: An
452 Overview of Recent Implementations and Their Considerations. *Pharmaceutics* 14:..
453 <https://doi.org/10.3390/pharmaceutics14091814>
- 454 19. Huang Z, Fidler M, Standing JF, Klopogge F (2026) Automated Model Selection in Population
455 Pharmacokinetics Using Metaheuristic Algorithms. *Journal of Pharmacokinetics and Pharmacodynamics*
- 456 20. Lombardo F, Berellini G, Obach RS (2018) Trend Analysis of a Database of Intravenous
457 Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug Metabolism and Disposition* 46:1466–
458 1477. <https://doi.org/10.1124/dmd.118.082966>
- 459 21. Nyberg J, Ueckert S, Strömberg EA, et al (2012) PopED: An extended, parallelized, nonlinear mixed
460 effects models optimal design tool. *Computer Methods and Programs in Biomedicine* 108:789–805.
461 <https://doi.org/10.1016/j.cmpb.2012.05.005>
- 462 22. Karlsson MO, Savic RM (2007) Diagnosing Model Diagnostics. *Clinical Pharmacology & Therapeutics*
463 82:17–20. <https://doi.org/10.1038/sj.clpt.6100241>
- 464 23. Derendorf H, Schmidt S (2020) Chapter 12: Variability. In: Rowland and Tozer's *Clinical*
465 *Pharmacokinetics and Pharmacodynamics: Concepts and Applications*, 5th ed. Wolters Kluwer, Philadelphia
- 466 24. Al-Sallami HS, Cheah SL, Han SY, et al (2014) Between-subject variability: should high be the new
467 normal? *Eur J Clin Pharmacol* 70:1403–1404. <https://doi.org/10.1007/s00228-014-1740-8>

25. Yuan L, Zhang D, Jemal M, Aubry A-F (2012) Systematic evaluation of the root cause of non-linearity in liquid chromatography/tandem mass spectrometry bioanalytical assays and strategy to predict and extend the linear standard curve range. *Rapid Communications in Mass Spectrometry* 26:1465–1474. <https://doi.org/10.1002/rcm.6252>
26. Keizer RJ, Jansen RS, Rosing H, et al (2015) Incorporation of concentration data below the limit of quantification in population pharmacokinetic analyses. *Pharmacology Research & Perspectives* 3:e00131. <https://doi.org/10.1002/prp2.131>
27. Schober P, Boer C, Schwarte LA (2018) Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 126:1763. <https://doi.org/10.1213/ANE.0000000000002864>
28. Mould D, Upton R (2013) Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development—Part 2: Introduction to Pharmacokinetic Modeling Methods. *CPT: Pharmacometrics & Systems Pharmacology* 2:38. <https://doi.org/10.1038/psp.2013.14>
29. Mathis A, Sale M, Cornpropst M, et al (2022) Population pharmacokinetic modeling and simulations of berotralstat for prophylactic treatment of attacks of hereditary angioedema. *Clinical and Translational Science* 15:1027–1035. <https://doi.org/10.1111/cts.13233>
30. Dong Q, Chen C, Taubert M, et al (2024) Understanding adefovir pharmacokinetics as a component of a transporter phenotyping cocktail. *Eur J Clin Pharmacol* 80:1069–1078. <https://doi.org/10.1007/s00228-024-03673-x>
31. R Core Team (2026) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria
32. Fidler M, Wilkins JJ, Hooijmaijers R, et al (2019) Nonlinear Mixed-Effects Model Development and Simulation Using nlmixr and Related R Open-Source Packages. *CPT: Pharmacometrics & Systems Pharmacology* 8:621–633. <https://doi.org/10.1002/psp4.12445>
33. Huang Z, Fidler M, Lan M, et al (2025) An automated pipeline to generate initial estimates for population Pharmacokinetic base models. *J Pharmacokinet Pharmacodyn* 52:60. <https://doi.org/10.1007/s10928-025-10000-z>
34. Burnham KP, Anderson DR (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* 33:261–304. <https://doi.org/10.1177/0049124104268644>
35. Xu XS, Yuan M, Karlsson MO, et al (2012) Shrinkage in Nonlinear Mixed-Effects Population Models: Quantification, Influencing Factors, and Impact. *AAPS J* 14:927–936. <https://doi.org/10.1208/s12248-012-9407-9>
36. Proost JH (2017) Combined proportional and additive residual error models in population pharmacokinetic modelling. *European Journal of Pharmaceutical Sciences* 109:S78–S82. <https://doi.org/10.1016/j.ejps.2017.05.021>

Supplementary Files

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

- [ESM1.csv](#)
- [ESM2.csv](#)
- [ESM3.docx](#)