Online Resource 1 — Supplementary Material

Two estimation methods agree on fit but disagree up to tenfold on the binding constant: practical non-identifiability and its prediction consequences in a simulated denosumab-like quasi-steady-state target-mediated drug disposition model

Jyotheeshwar Akshay Ravikumar, B. Anbarasan

Faculty of Pharmacy, Sri Ramachandra Institute of Higher Education and Research (Deemed to be University), Porur, Chennai 600116, Tamil Nadu, India

Corresponding author: Jyotheeshwar Akshay Ravikumar (jyotheeshwarakshay@gmail.com)

Submitted to *Journal of Pharmacokinetics and Pharmacodynamics* as Online Resource 1 (ESM_1.pdf)

Contents

S1. Complete model specification **2**
S1.1. Model equations, in full . . . 2
S1.2. Units audit, including one inconsistency in the source table . . . 3
S1.3. Complete parameter table . . . 4
S1.3.1 Free-parameter count . . . 4
S1.4. Stated departures from the published model . . . 5
S1.5. Numerical verification . . . 5
S1.5.1 Mass conservation . . . 5
S1.5.2 Independent implementation . . . 5
S1.6. Prespecified analysis criteria . . . 5
S1.6.1 Metric definitions . . . 6
S1.6.2 Recovery criteria . . . 6
S1.6.3 Operational convergence definition . . . 6
S1.6.4 Profile likelihood . . . 7
S1.6.5 Objective function comparability . . . 7
S1.7. Software environment . . . 8

S2. Analyses not shown in the main text **9**
S2.1 SAEM iteration budgets (Table S1) . . . 9
S2.2 Seed sensitivity — not a profile likelihood . . . 9
S2.3 Initialisation pilot: four starts, four terminal objective values . . . 9
S2.4 Random-effects structure . . . 10
S2.5 Censoring in this design . . . 10
S2.6 Fisher information: an independent ranking, and the sample size the design would need . . . 10
S2.7 Reproducibility of the estimation step, and a refuted hypothesis . . . 11
S2.8 Design arms: late sampling against a static target . . . 12
S2.9 Uncertainty calibration, and a reporting convention that cannot be compared across parameters . . . 14
S2.10 What the Kss error costs: little on the drug, much on the target . . . 15
S2.11 Fitting at N = 80: the extrapolation was right, and so the earlier argument was wrong . . . 16
S2.12 Convergence check and the FOCEi re-estimation . . . 17

S3. Profile likelihood in full **18**

S4. Complete list of limitations **20**

S5. Correction history **21**

S6. Software environment and reproduction **22**

S7. Supplementary data files **22**
S7.1 Reference verification 23

This supplement carries the material supporting the main text, in the order the main text refers to it. Section S5 is the complete correction history of the project, including every claim withdrawn during the work; it is supplied because the main text's workflow note refers to it and because a reader checking any number in the paper should be able to see what was wrong before it was right.

S1. Complete model specification

Written to close the specification, accounting, units and definitional items of the model of main-text Section 2.1. **S2.4 is prespecified: it was written and saved before the start-sensitivity and profile-likelihood results existed.**

S1.1. Model equations, in full

States (`depot` is an amount; all others are concentrations):

State	Meaning	Units
`depot`	subcutaneous depot drug amount	nmol
`Ctot`	total (free + bound) drug in central compartment	nmol/L
`Cp`	drug in peripheral compartment	nmol/L
`Rtot`	total (free + bound) target, RANKL	nmol/L

Quasi-steady-state algebra, giving free drug `C` as the positive root of the QSS quadratic, and complex `RC` by mass balance:

```
disc  = (Ctot - Rtot - Kss)^2 + 4 * Kss * Ctot
C     = 0.5 * ( (Ctot - Rtot - Kss) + sqrt(disc) )
Cfree = C
RC    = Ctot - Cfree
```

`disc` is a sum of a square and a non-negative term, so it is never negative, and `C` is the positive root, so it is never negative. The earlier implementation wrapped both in `max(·, 0)`; those branches are mathematically unreachable and were removed, because a non-differentiable branch inside a compiled ODE right-hand side can be resolved differently between runs by an adaptive solver.

Differential equations:

```
d/dt(depot) = -ka * depot
d/dt(Ctot)  = (ka * depot)/Vc - (CL/Vc)*Cfree - (Q/Vc)*Cfree + (Q/Vc)*Cp - kint*RC
d/dt(Cp)    = (Q/Vp)*Cfree - (Q/Vp)*Cp
d/dt(Rtot)  = ksyn - kdeg*(Rtot - RC) - kint*RC
Rtot(0)     = R0 ;  depot(0) = Cp(0) = Ctot(0) = 0
kdeg        = ksyn / R0            (derived, not estimated)
```

Observation model (combined error, matching the source paper's Eq. 12 and nlmixr2's `add() + prop()` parameterisation — identical forms, so simulation and estimation cannot diverge here):

```
DV = Ctot * (1 + eps_prop) + eps_add,    eps_prop ~ N(0, sigma_prop^2),
                                         eps_add  ~ N(0, sigma_add^2)
```

Dosing: 60 mg subcutaneously at 0, 6 and 12 months (0, 4383.4 and 8766.7 h, using 30.44 d/month), administered into `depot`.

Dose conversion, shown explicitly:

```
60 mg = 60e-3 g ;  MW = 147 kDa = 147,000 g/mol
60e-3 / 147,000 = 4.0816e-7 mol = 408.1633 nmol
```

This is the value in the `AMT` column of every simulated dataset.

S1.2. Units audit, including one inconsistency in the source table

Quantity	Units	Note
ka	h^{-1}	
Vc/F	L	apparent
Vp/F	L	apparent
CL/F	L h^{-1}	apparent
Q/F	L h^{-1}	apparent, **fixed**
kint	h^{-1}	
Kss	nmol/L	
ksyn	**nmol L^{-1} h^{-1}**	**see below — the source table labels this h^{-1}**
R0	nmol/L	
kdeg	h^{-1}	derived as ksyn/R0
sigma_add	nmol/L	
sigma_prop	dimensionless	

The ksyn inconsistency. Table 3 of the source paper labels ksyn as "(1/h)" with a value of 0.01. That label cannot be right, and this can be shown two independent ways:

1. *From the source's own Eq. 7*: `ksyn = R0 × kdeg`. With R0 in nmol/L and kdeg in h^{-1}, ksyn must be nmol L^{-1} h^{-1}.
2. *From the target ODE*: in `d/dt(Rtot) = ksyn − kdeg·(Rtot − RC) − kint·RC`, every other term is nmol L^{-1} h^{-1}, so ksyn must be too.

The value 0.01 is therefore read as 0.01 nmol L^{-1} h^{-1}, which is what the implementation uses. The published numbers are used unchanged; only the unit label is corrected.

A consequence worth stating, which follows from the published numbers and not from any choice made here. With ksyn = 0.01 nmol L^{-1} h^{-1} and R0 = 15.23 nmol/L:

```
kdeg = ksyn / R0 = 0.01 / 15.23 = 6.566e-4 h^-1
RANKL half-life  = ln(2) / kdeg = 1,055.7 h = 44.0 days
```

A 44-day half-life for a soluble cytokine is implausibly long; reported sRANKL turnover is far faster. Taken at face value, the published parameter set therefore describes a target that is **nearly static over a 6-month dosing interval**. This matters directly for the question at hand: if target concentration barely moves, the target dynamics that would inform Kss are largely absent, which is a candidate mechanism for Kss being the badly-determined parameter rather than a mere observation that it is. It is tested explicitly in the fast-turnover arm (kdeg × 100).

R0 is both an estimated parameter and the initial condition, and kdeg is derived from it. R0 therefore enters the likelihood through three routes simultaneously — as `Rtot(0)`, through `kdeg = ksyn/R0`, and through the QSS algebra — which is relevant to how its identifiability should be interpreted.

S1.3. Complete parameter table

Ground truth is Table 3 of Choi et al. (2025), postmenopausal-osteoporosis rows (ka_PMO, R0_PMO, Q/F_PMO and CL/F in Caucasian). IIV is reported there as %CV and converted as $\omega^2 = \log(1 + CV^2)$.

#	Parameter	Units	True value	Estimated?	IIV estimated?	ω^2 (true)	Source
1	ka	h^{-1}	0.0078	yes	yes	0.27764	ka_PMO
2	Vc/F	L	1.58	yes	yes	0.32249	VC/F
3	Vp/F	L	6.06	yes	yes	0.02389	VP/F
4	CL/F	L h^{-1}	0.006	yes	yes	0.06733	CL/F Caucasian
5	kint	h^{-1}	0.022	yes	yes	0.00619	kint
6	Kss	nmol/L	1.56	yes	yes	0.29102	KSS
7	ksyn	nmol L^{-1} h^{-1}	0.01	yes	yes	0.04921	ksyn
8	R0	nmol/L	15.23	yes	yes	1.25444	R0_PMO
9	**Q/F**	L h^{-1}	**0.20**	**no — fixed**	**no**	—	Q/F_PMO
—	kdeg	h^{-1}	6.566e-4	derived	—	—	= ksyn/R0
—	sigma_add	nmol/L	0.72	yes	—	—	eps_add
—	sigma_prop	—	0.07	yes	—	—	eps_prop

So: the model contains nine structural quantities; eight are estimated and Q is fixed at 0.20 L h^{-1} with no between-subject variability. The phrase "eight structural parameters" always refers to the eight estimated ones.

S1.3.1 Free-parameter count

Component	Full model	Reduced model
Fixed effects (θ)	8	8
IIV variances (ω^2, diagonal)	8	5
IIV correlations	0	0
Residual error (σ_add, σ_prop)	2	2
Total free	**18**	**15**

Fixed constants not counted: Q (1). Derived quantities not counted: kdeg (1). The reduced model drops IIV on Vp, kint and ksyn — the three lowest-%CV parameters — giving the 3-degree-of-freedom difference used in the likelihood ratio comparison.

S1.4. Stated departures from the published model

The ground truth is **not** the published denosumab model. Three deliberate simplifications:

1. The reported CL–Vp random-effect correlation (0.43) is not implemented; all random effects are independent.
2. The body-weight covariate model on Vc, Vp and CL is not implemented.
3. Between-subject variability on Q (295.99 %CV) is removed, and Q is a fixed constant in both the simulation and the estimation model.

The correct description of the system studied is therefore **a simplified denosumab-like QSS-TMDD model derived from the published model**, and no result here should be read as characterising the published clinical model.

S1.5. Numerical verification

S1.5.1 Mass conservation

Total drug = `depot + Vc·Ctot + Vp·Cp + cumulative eliminated`, where elimination accumulates as `CL·Cfree + Vc·kint·RC`. Summing the ODEs gives `d(total)/dt = −CL·Cfree − Vc·kint·RC`, i.e. the only losses are linear clearance and target-mediated internalisation. Integrating over the full three-dose, ~18-month course at tolerances rtol 1e-11 / atol 1e-13:

Case	Max absolute error	As % of 1224.5 nmol dosed
Typical subject	2.27e-12 nmol	1.9e-13 %
Low Vc (−2 SD)	4.32e-12 nmol	3.5e-13 %
High R0 (+2 SD)	3.41e-12 nmol	2.8e-13 %
High kint (+2 SD)	9.32e-12 nmol	7.6e-13 %

Mass is conserved to solver tolerance, including at population extremes. This replaces the earlier bare assertion that the equations had been "re-derived to confirm mass conservation".

S1.5.2 Independent implementation

The equations were implemented a second time, independently, in Python/SciPy (LSODA) and used for the mass-balance test above and for the reconstruction that localised the original pipeline defect. The two implementations agree: the Python reconstruction of a dataset generated by the R/rxode2 pipeline gave standardised residuals with mean −0.025 and SD 1.020 against the generating parameter vector, which is the expected result if and only if both implementations describe the same system.

S1.6. Prespecified analysis criteria

Declared before the start-sensitivity and profile-likelihood results existed. The earlier ±30 % band was a reporting threshold chosen after seeing results; it is withdrawn as a criterion and retained only as a descriptive aid.

S1.6.1 Metric definitions

For parameter θ with true value θ_true and estimates $\hat{\theta}_1 \ldots \hat{\theta}_n$ over n replicates:

```
Relative bias (%)   = 100 * ( median(θ̂) - θ_true ) / θ_true
RSD (%)             = 100 * SD(θ̂) / mean(θ̂)              [empirical, not model-based]
Relative RMSE (%)   = 100 * sqrt( mean( (θ̂- θ_true)^2 ) ) / θ_true
MCSE of median (%)  ≈ 100 * 1.253 * SD(θ̂) / ( sqrt(n) * θ_true )
```

RSD is the empirical across-replicate coefficient of variation of the estimates, **not** a model-based relative standard error. Where a model-based %RSE is reported it is labelled as such.

Uncertainty-calibration metrics, added for the S2.9 analysis:

```
Model-based rel. SE (%) = 100 * SE(log θ̂)              [per fit; scale-free]
Calibration ratio       = median(model-based rel. SE) / RSD
CI coverage             = # replicates with θ_true in [CI Lower, CI Upper] / n
η-shrinkage (%)         = f$shrink["sd shrinkage (%)", "eta_θ"]
```

A calibration ratio below 1 means the covariance step reports more precision than the estimates actually have. Confidence bounds are the software's back-transformed 95 % bounds (`CI Lower`, `CI Upper` of `parFixedDf`), which are symmetric on the estimation scale.

S1.6.2 Recovery criteria

A parameter is declared **adequately recovered** under a given scenario when both hold:

- |relative bias| ≤ 20 %, and
- RSD ≤ 30 %.

These are operational thresholds declared in advance for this analysis. They are not taken from a published standard and no claim is made that they constitute a community criterion; they exist so that "fails" has a fixed meaning that was not chosen after inspecting the numbers.

S1.6.3 Operational convergence definition

A fit counts as **converged** only if all of the following hold, and each is recorded per fit:

1. the estimation call returns without error;
2. the covariance step succeeds and all reported %RSE values are finite;
3. no between-subject variance lies at or near either boundary of the interval the estimator constrains it to. The interval is estimator-specific and is read from the control object actually used, not assumed: `foceiControl()` bounds each diagonal ω^2 multiplicatively about its **initial** value — `diagOmegaBoundLower = 100` and `diagOmegaBoundUpper = 5` give $[\omega^2_0/100, 5\cdot\omega^2_0]$, which for the standard start on K_{ss} ($\omega^2_0 = 0.2910$) is [0.00291, 1.45511] — while `saemControl()` imposes no such bounds, so the only attainable boundary there is $\omega^2 \to 0$. A component is flagged when it lies within a factor of **1.1** of either end of its interval, that is $\omega^2 \le 1.1\cdot L$ or $\omega^2 \ge U/1.1$, and, at an end where the estimator sets no finite bound, by the absolute floor $\omega^2 <$ 1e-6. Each flag is recorded per component and per end together with the bound value, so a variance pinned at a bound is distinguishable from one that is merely small;
4. maximum ETA shrinkage < 90 %;
5. the objective function is finite.

Fits failing any criterion are reported, not silently dropped. "All fits converged" on its own is not treated as evidence; the earlier analysis's use of the software's own flag is insufficient.

Three defects in this definition were found after it was declared. All are recorded here rather than corrected retrospectively.

Criterion 4 was not evaluated in 36 fits. The worker extracted shrinkage as `f$shrink[, "Shrink(SD)%"]`. In nlmixr2est 7.1.0 `f$shrink` is a 9 × 9 data frame whose **rows** are metrics (`mean`, `var`, `sd`, `skewness`, `kurtosis`, `var shrinkage (%)`, `sd shrinkage (%)`, `t statistic`, `p-value`) and whose **columns** are the η names plus `IWRES`. Column-indexing by `"Shrink(SD)%"` therefore returned `NA` silently, and criterion 4 was never applied to any fit run through `10_worker.R` — 14 profile fits, 4 initialisation fits and 18 design-arm fits. The correct access is `f$shrink["sd shrinkage (%)", setdiff(colnames(sh), "IWRES")]`. (The name `Shrink(SD)%` does exist, but as a column of `f$parFixedDf`, which is a different table.) The worker is corrected and shrinkage is reported only for fits run after the correction. The general remedy, which the project has now needed twice, is that a criterion which cannot be evaluated must fail loudly rather than return a missing value.

Criterion 2 rests on a quantity that is not comparable across parameters. For a log-transformed parameter the `%RSE` column reports 100 · SE(log θ̂)/|log θ̂|. The denominator depends on the units in which θ is expressed, so the column is large for any parameter whose log estimate is near zero and small for one whose log estimate is far from it, at identical actual precision. Verified against the fits' own output for all eight parameters: at median estimates, its median across the deposited fit sets is 26–71 % for Kss (log ≈ 0.16) while CL (log ≈ −5.1) reports 1.0 %. The scale-free quantity is SE(log θ̂) itself, which *is* the relative standard error of θ on its natural scale, and that is what should have been prespecified. The finiteness test criterion 2 applies is still meaningful; any magnitude threshold on that column would not have been.

Criterion 3 was one-sided and estimator-blind. It tested only ω^2 < 1e-6, a floor `foceiControl()` cannot reach — its defaults `diagOmegaBoundLower = 100` and `diagOmegaBoundUpper = 5` confine each diagonal variance to $[\omega^2_0/100, 5\cdot\omega^2_0]$, so the criterion could not fire for any FOCEi fit even while two of them sat on the upper bound — and it was therefore revised on 24 October 2026 to the two-sided, estimator-aware form stated above.

S1.6.4 Profile likelihood

The Kss profile fixes the **population typical value** of Kss on a log-spaced grid (±1.5 log units about the truth, 7 points) and re-estimates every remaining free parameter at each grid point, **including ω^2_Kss**. An earlier implementation in this project dropped the profiled parameter's random effect, which profiles a structurally different model at every grid point and makes the resulting ΔOFV curve uninterpretable; that behaviour is now off by default and controlled by an explicit `drop_profiled_eta` argument.

ΔOFV = 3.84 is used **only as a conventional asymptotic reference**, not as a confidence statement. The regularity conditions for a χ^2_1 likelihood-ratio calibration are not established here: variance components are estimated, the likelihood is an approximation (Gaussian quadrature), observations are censored, and boundaries exist. The profile shape is reported as a shape.

S1.6.5 Objective function comparability

Objectives come from two methods and are never mixed within a comparison.

SAEM-reported objectives. Every objective quoted for a fitted SAEM model is SAEM's −2 log-likelihood by Gaussian quadrature (nnodes = 3, nsd = 1.6). These are compared only with each other and only within one analysis, because the iteration budget differs between analyses (Table S1).

Fixed-parameter objectives (the common-objective comparison). Comparing a SAEM estimate with a FOCEi estimate requires one method applied to both, so neither estimator's own reported value is used. Each parameter vector is re-evaluated with every population parameter declared `fixed()` — all eight typical values, all eight variance components and both residual terms — and the objective taken from `est = "focei"` with `foceiControl(maxOuterIterations = 0, optExpression = FALSE)`. With no population parameter free, the estimator reduces to empirical-Bayes estimation of the individual random

effects at those fixed values and returns the objective there. The same call is used for all three vectors of a replicate (SAEM, FOCEi, truth), so the comparison is internally consistent.

Two limitations of that comparison are declared here rather than discovered later. First, the fitted variance matrix was not written to the result file for every fit, so the variance components are held at the programme's common values for all three vectors; the comparison is therefore between fixed-effect vectors under one random-effects model, which is the right construction for asking whether the likelihood is flat in the typical value of Kss, but is not a comparison of the fits' own reported objectives. Second, the resulting numbers are consequently not equal to, and must not be tabulated alongside, the SAEM-reported objectives above.

S1.7. Software environment

Component	Version
R	4.5.3
rxode2	5.1.7
nlmixr2est	7.1.0
nlmixr2	7.0.1
Platform	aarch64-apple-darwin, conda toolchain

Build note: conda-forge carries no `r-rxode2`/`r-nlmixr2est` for osx-arm64, so the stack is built from CRAN source; `cmake` and `symengine` are required, and `CPPFLAGS += -fopenmp-version=45` is needed because the conda OpenMP header uses `begin declare variant` syntax the bundled clang rejects.

One function is unusable in this build and is therefore avoided rather than worked around: `setOfv(fit, "focei")` segfaults. FOCEi itself segfaults under the default control during the symbolic pass that removes duplicate subexpressions; `foceiControl(optExpression = FALSE)` compiles and completes, and `optExpression` is a performance option that changes runtime rather than the objective or the estimates. All FOCEi fits and all fixed-parameter objective evaluations reported here were run with that setting. Neither is evidence about any other environment.

Reproduction: `00_truth.R` holds the single parameter definition and generates both the simulation and the estimation model; `assert_aligned()` fails if they diverge. `10_worker.R` plus `jobs.csv` reproduce every fit reported, one job per row, each writing its own output file.

Package pinning. `renv.lock` in the repository pins the exact version of every one of the 179 packages in the library under R 4.5.3. The lockfile has no package hashes, because the library was built directly from CRAN rather than under `renv`; `renv::restore()` will therefore reinstall by version rather than verify by checksum.

Library path. The nlmixr2 stack is a project-local CRAN build in `./Rlib` (conda-forge publishes no `r-rxode2` or `r-nlmixr2est` for osx-arm64). During this work a batch of 12 queued fits failed within seconds because the run was launched without the corresponding `R_LIBS` export. `10_worker.R` now resolves the path itself:

```
if (!requireNamespace("nlmixr2", quietly=TRUE) && dir.exists("Rlib"))
  .libPaths(c(normalizePath("Rlib"), .libPaths()))
```

Configuration that the code depends on is configuration the code should set. This is the same principle as §S2's single parameter definition, applied to the environment instead of the model.

S2. Analyses not shown in the main text

S2.1 SAEM iteration budgets (Table S1)

Table S1. SAEM iteration budget by analysis.

Analysis	Section	Burn-in	EM	Fits
Replicate recovery (24 datasets)	main-text Section 3.1	150	200	24
Seed sensitivity (1 dataset, 5 seeds)	S2.2	200	300	5
Profile likelihood (Kss, R0)	S3	100	150	14
Initialisation pilot (4 starts)	S2.3	100	150	4
Random-effects structure	S2.4	200	300	4
Reproducibility experiment	S2.7	100	150	8
Design arms	S2.8	100	150	18

One consistency check is available across analyses: the reproducibility experiment (S2.7) and the true-start fit of the initialisation pilot (S2.3) use the same dataset, budget, seed and starting values but were run by two separately written scripts. They agree on the objective and on Kss to every digit either script recorded — 2176.0429 and 2.0457 in the pilot, 2176.04294602 and 2.04569831 in the reproducibility run. This is an end-to-end agreement between two independent drivers of the same model, and is the check that the previous version of this work did not have.

S2.2 Seed sensitivity — not a profile likelihood

Within a single dataset, five estimation seeds returned objective values of 2170.54, 2176.84, 2177.95, 2179.94 and 2181.66 (spread 11.1) while placing Kss at 1.514, 2.531, 0.519, 0.862 and 2.592 nmol/L. The correct description is that **the estimation procedure produced substantially different Kss estimates across independent seeds with modest differences in objective value.** This is evidence of optimisation instability or near-equivalent solutions. It is *not* a profile likelihood, and an earlier draft wrongly inferred from it that the likelihood is flat in Kss. That inference is withdrawn. S3 attempted to replace it with a profile likelihood and reports why the attempt did not succeed.

S2.3 Initialisation pilot: four starts, four terminal objective values

Table S2. One dataset, four start configurations (100 burn-in, 150 EM).

Start	Seed	Objective	Kss	Kss vs true	max %RSE
true values	42	2176.04	2.046	+31 %	21
perturbed	42	**2173.94**	**3.283**	**+110 %**	21
perturbed	43	2197.67	1.253	−20 %	39
perturbed	44	2181.67	1.246	−20 %	124

Four starts produced four distinct terminal objective values spanning 23.7 units, and **the lowest objective was reached from a perturbed start, not the true start** — at a Kss 110 % above truth. Two

consequences follow. First, the true-start results in main-text Section 3.1 are an upper bound on achievable recovery rather than an estimate of it, and must be read as a sensitivity analysis in the sense both reviewers intend. Second, the best-fitting solution in this dataset is further from the truth in Kss than a worse-fitting one, so objective value does not order solutions by proximity to truth for this parameter.

The fit from seed 44 met every prespecified convergence criterion while carrying a 124 % model-based %RSE. Criterion 2 of S1.6.3 of the specification ("all reported %RSE finite") is therefore too permissive; a numerical threshold is needed. We record this as a post hoc observation about a prespecified criterion rather than silently revising it.

S2.4 Random-effects structure

The full structure (18 free parameters) reached 2170.54 and a reduced five-variance structure (15 parameters) 2313.27: ΔOFV +142.7 for three fewer parameters, against a conventional $\chi^2(3)$ reference of 7.81. The data support all eight variance components, so the Kss difficulty is not a symptom of over-parameterisation. Reducing the structure did not stabilise Kss (0.770 and 2.175 at two seeds).

Two seeds per structure is weak, and the number that shows how weak belongs here: the two reduced-structure seeds returned 2354.71 and 2313.27, a spread of 41.4 objective units within one structure, against 6.3 units between the two full-structure seeds (2170.54 and 2176.84). The ΔOFV quoted above is best-to-best. It is six times the within-structure seed spread, so the direction is not in doubt, but a $\chi^2(3)$ reference is no more licensed here than the χ^2_1 reference of S3 — for the same five reasons — and is quoted only as a conventional landmark.

S2.5 Censoring in this design

Across the 24 datasets a median of 1 observation per 720 fell below the quantification limit (range 0–8), against 26.49 % of post-dose samples in the clinical dataset the parameters came from. Two implications. The M1/M3 choice is nearly immaterial *in this design* — a comparison here would test the design, not the method; the informative comparison is at a raised LLOQ where censoring is material, and the behaviour of the two methods as the censored fraction grows is documented in Ahn et al. (2008), Beal (2001) and Jaber et al. (2021). And the simulated design is not like-for-like with the clinical study, whose schedule, follow-up, population and assay all differ, so the 26.49 % figure is qualitative context rather than evidence about external validity.

S2.6 Fisher information: an independent ranking, and the sample size the design would need

Profile likelihood is the appropriate instrument for practical identifiability, but S3 shows that a profile adequate for this model costs far more fits than were available. To obtain a ranking for all eight at once, and to answer the reviewer question "would N = 80 fix this?" without 80-subject refits, we computed the sensitivity matrix of the predicted total concentration with respect to the log of each structural parameter at the typical subject (central differences, $h = 10^{-4}$, 24 sampling times), weighted each residual by the prespecified error model, and formed the per-subject information matrix. The population information is approximated as N times the single-subject matrix; this is a fixed-effects approximation that ignores the variance components and is therefore a lower bound on the uncertainty, which is the direction that makes the conclusion below conservative.

Table S3. Predicted relative standard error from the Fisher information (fixed-effects approximation) against the observed across-replicate RSD of main-text Section 3.1.

Parameter	Predicted %RSE, N = 30	N = 80	N = 500	Observed RSD, N = 30
Vp	2.6	1.6	0.6	2.8
CL	2.7	1.7	0.7	5.8

Parameter	Predicted %RSE, N = 30	N = 80	N = 500	Observed RSD, N = 30
ka	2.9	1.8	0.7	7.9
ksyn	10.2	6.3	2.5	11.6
R0	10.5	6.4	2.6	20.8
Vc	10.9	6.7	2.7	11.9
kint	13.3	8.2	3.3	7.1
Kss	**128.8**	**78.8**	**31.5**	**35.1**

Three results.

The ranking is reproduced by a method that shares no code with the estimation. Kss is last by a factor of 9.7 over the next worst parameter, and the rank correlation between predicted precision and observed RSD across the eight parameters is $\rho = 0.74$ ($p = 0.037$). Vp, CL and Kss occupy the same rank positions under both. This is independent corroboration of main-text Section 3.1's ranking that does not depend on SAEM, on the starting values, or on the replicate count.

The weakest direction of the information matrix is essentially Kss itself. The matrix has condition number 9.4×10^4, and the eigenvector belonging to its smallest eigenvalue loads +0.99 on log Kss with no other loading above 0.1 in magnitude. In this model and at this design, the poorly-determined direction in parameter space is not a composite of several parameters — it is one parameter. That is a sharper statement than "ill-conditioned model", and it is what makes a per-parameter rather than a whole-model answer meaningful here.

The N^−1/2 scaling is sound; the level it is anchored to is not — and S2.11 settles which anchor is right by fitting at N = 80. Information scales linearly in N, so a standard error scales as N^−1/2 whatever its starting value. The starting value, however, is exactly what the three available instruments disagree about. At N = 30 the relative standard error of Kss is 128.8 % by the FIM, 15.0 % by the fits' own covariance step (S2.9), and 35.1 % empirically across the 24 replicates. The FIM overstates the empirical figure 3.7-fold — expected, since it ignores the eight variance components that absorb part of the between-subject variability — and the covariance step understates it 2.3-fold. Extrapolating each by N^−1/2 therefore gives incompatible answers to the reviewer's question:

Anchor at N = 30	Kss at N = 80	N for Kss below 30 %	N to match Vc at N = 30
FIM, 128.8 %	78.8 %	≈ 550	≈ 4,200
empirical RSD, 35.1 %	21.5 %	≈ 41	≈ 260
observed (S2.11)	**21.8 %**	—	—

An order of magnitude separates them. S2.11 resolves it: the empirical anchor is right to within 0.3 percentage points and the Fisher information is wrong by a factor of 3.6. The FIM's ranking and eigenstructure are corroborated elsewhere and are used below; its absolute precision level is not usable for this model and is not used.

Two caveats on the FIM itself. It can mislead for practical identifiability (Chis et al., 2011; Kreutz, 2018), and here it disagrees with the profile in two places: the signs of the Kss–ksyn and Kss–CL couplings are opposite between the two instruments (agreement on the other five). And, as above, its absolute level is miscalibrated for this model. We use it for the ranking and the eigenstructure — both corroborated by independent evidence — and not for the mechanism or the level.

S2.7 Reproducibility of the estimation step, and a refuted hypothesis

S2.2 and S2.3 both show substantial estimate-to-estimate movement under changes of seed or starting value. An obvious candidate explanation is that the fits are not reproducible at all — that non-associative

floating-point reduction across OpenMP threads makes even a fixed-seed fit irreproducible, in which case none of S2.2, S2.3 or S3 would mean anything. We tested it directly: one dataset, two seeds, two repeats each, at one thread and at four threads (8 fits).

Table S4. Reproducibility of fixed-seed fits.

Threads	Seed 42 objective	Seed 42 Kss	Seed 123 objective	Seed 123 Kss
1	2176.04294602 (×2)	2.04569831	2178.42285613 (×2)	1.91639231
4	2176.04294602 (×2)	2.04569831	2178.42285613 (×2)	1.91639231

Every value is bit-identical across repeats and across thread counts (maximum absolute difference 0.00×10^{0}). **The threading hypothesis is refuted.** A fixed-seed, fixed-start fit in this build is exactly reproducible, and thread count changes only the wall time (26.0 min at one thread, 9.4 min at four). The movement documented in S2.2 and S2.3 is therefore a property of the objective surface and of SAEM's stochastic trajectory through it, not floating-point noise.

S2.8 Design arms: late sampling against a static target

Main-text Section 3.4 names two competing mechanisms for poor Kss recovery — the sampling schedule stops at 2160 h of a 4383 h interval, and the published parameters imply a RANKL half-life of 44 days so the target barely moves. These make different predictions, and the predictions are separable by simulation. Three arms were run on six paired datasets each (same subject-level random effects across arms, so the comparison is within-subject): the reference design; a *late* design adding samples at 3600 and 4380 h after each dose; and a *fast-turnover* arm in which ksyn is scaled 100-fold so that kdeg = ksyn/R0 rises by the same factor and the RANKL half-life falls from 44 days to 10.6 h, leaving R0 unchanged.

All 18 fits returned without error.

One correction before the result, because it changes what the reference arm is evidence of. An earlier version of this section described the reference arm as reproducing main-text Section 3.1 "on six datasets it did not share with it". That is wrong: `06_sse_worker.R` and `10_worker.R` both seed the simulation with `set.seed(10000 + replicate)`, so design-arm replicates 1–6 *are* replicates 1–6 of the 24. The BLQ counts match on all six and the Kss estimates correlate 0.934. The reference arm is therefore a re-analysis of the same data by a second, independently written driver at a different iteration budget (100/150 against 150/200), not an independent replication.

Read correctly it says something less comfortable and more useful. On those same six datasets the main-text Section 3.1 analysis gives a median Kss bias of −35.8 % (RSD 38.4 %); the reference arm gives −24.7 % (35.6 %). The only differences are the driver and the iteration budget, so **the iteration budget alone moves the median Kss bias by 11 percentage points on fixed data** — a further instance of the S2.3 basin sensitivity, and a reason to treat any single reported Kss bias, including −24.6 %, as budget-conditional. It does not affect the arm comparisons below, which are paired within replicate and run at one budget.

Table S5. Kss recovery by arm. Six paired datasets per arm — the same subject-level random effects are used in all three arms of each replicate, so every comparison is within-subject. True Kss 1.56 nmol/L.

Arm	Median Kss	Rel. bias	RSD	Range	Criteria	Paired vs reference
reference	1.17	−24.7 %	35.6 %	0.81–2.05	fail (both)	—

Arm	Median Kss	Rel. bias	RSD	Range	Criteria	Paired vs reference
+ late samples (3600, 4380 h)	1.74	+11.7 %	22.5 %	1.45–2.52	pass	3/6 closer, p = 0.66
100× faster target turnover	1.43	**−8.1 %**	**16.8 %**	1.11–1.83	pass	**6/6 closer, p = 0.016**

The paired column is an exact sign test on the within-dataset change in |Kss − truth|/truth; the Wilcoxon signed-rank test on the same differences gives 0.28 for the late arm and 0.016 for the turnover arm, so the two tests agree on the turnover arm and differ on the late arm, where neither reaches significance. Two arms were compared against one reference without multiplicity adjustment; the turnover arm survives a Bonferroni correction (0.031) and the late arm does not approach it. At n = 6 the smallest attainable sign-test p-value is 0.0156, so the reported 0.016 is that floor: it establishes a consistent direction, not a large effect. No other parameter's recovery degraded: against each arm's own truth, the seven remaining structural parameters have median relative biases between −6.0 % and +5.1 % in both intervention arms (`design_arms_summary.csv`).

The static-target mechanism is supported; the late-sampling mechanism is not, at this power. Accelerating target turnover moved Kss from failing both prespecified criteria to passing both, and improved the estimate in every one of six paired datasets. Adding two late samples per dose window improved the median absolute bias and the RSD but helped in only three of six datasets and changed the sign of the bias. So of the two mechanisms named in main-text Section 3.4, the one the units audit uncovered — a target with a 44-day half-life that barely moves over a dosing interval — accounts for more of the Kss difficulty than the one the earlier draft asserted.

Three cautions, and the first is the important one.

The turnover arm is not an actionable design change. It alters the *data-generating system*, not the experiment: an analyst cannot make a patient's RANKL turn over 100× faster. Its value is diagnostic — it identifies what the information deficit is made of. The actionable reading is the contrapositive: when the target is slow relative to the dosing interval, drug-concentration data alone will not determine Kss well however the samples are placed, and the interventions worth costing are the ones that observe the target directly rather than the ones that add PK samples.

The turnover arm changes three things, not one. Scaling ksyn 100-fold raises kdeg = ksyn/R0 by the same factor, which is the intended intervention on the target's *timescale*. It also leaves R0 unchanged while multiplying the baseline synthesis flux by 100, so far more target passes through the system per unit time, more drug is cleared by target-mediated internalisation, and the target-mediated signature in the concentration profile becomes substantially stronger. And it raises the censoring fraction (below). This experiment cannot separate "Kss becomes identifiable because the target moves faster" from "Kss becomes identifiable because the target-mediated signal is larger" — the two are varied together by construction. The conclusion that survives is the negative one: it is *not* the sampling schedule. The positive attribution to the 44-day half-life specifically is weaker than a single arm can support, and separating timescale from flux requires an arm that scales ksyn and R0 together with kint compensated, which was not run.

Both arms buy their improvement partly with censored observations. Median observations below the quantification limit per dataset: 0.5 of 720 in the reference arm, 54 of 900 in the late arm, and 107 of 720 in the turnover arm. The added late samples sit near trough and most of them are discarded under M1; the turnover arm has far more target, so more drug is bound and internalised and the measured total concentration falls below the assay limit more often. The turnover arm therefore changes two things at once — target dynamics and censoring fraction — and its effect cannot be attributed to target dynamics alone from this experiment. Repeating it under M3 would separate them and has not been done.

Model-based precision deteriorates even where recovery improves. The largest reported %RSE in a fit ranged 21–358 in the reference arm, 17–69 in the late arm and 77–8,645 in the turnover arm. The turnover arm recovers Kss better across replicates while reporting far worse model-based uncertainty within replicates, which is a second instance of the S2.3 finding that the covariance step's output and the across-replicate truth are not interchangeable diagnostics. S2.9 takes this up directly.

A defect found while analysing these fits is recorded here rather than silently fixed. The worker extracted η-shrinkage as `f$shrink[, "Shrink(SD)%"]`, but in this software version `f$shrink` is a 9 × 9 data frame with *metrics as rows* and η names as columns, so the expression returned `NA` for every fit and prespecified convergence criterion 4 (shrinkage < 90 %) was never evaluated in any fit run through this worker — 36 fits across S3, S2.3 and S2.8. The correct access is by row name (`f$shrink["sd shrinkage (%)", ]`, excluding the IWRES column); the worker is corrected and S2.9 reports shrinkage from fits run after the correction. This is the same failure mode as main-text Section 4.4 at smaller scale: a quantity assumed to be checked, never checked, and silent about it. The remedy applied in both cases is the same — a criterion that cannot be evaluated must fail loudly rather than return a missing value.

S2.9 Uncertainty calibration, and a reporting convention that cannot be compared across parameters

Six further reference-design fits were run with per-parameter uncertainty captured. These six are retained here because they are the true-value-start counterpart of main-text Table 5, which reports the same quantities on the 24 multistart fits; the analysis below is therefore the start-conditional version and the factor it reports, 2.4, is superseded by the 3.7 of Table 5. The six fits captured: the model-based standard error on the estimation (log) scale, the back-transformed 95 % confidence bounds, and η-shrinkage — the last of which was unavailable in every earlier fit for the reason given in S2.8.

Table S6. Model-based against empirical uncertainty, six datasets, reference design, true-value starts. The model-based relative standard error of θ on its natural scale is SE(log θ̂); the calibration ratio is that quantity divided by the empirical across-replicate RSD, so a value below 1 means the covariance step reports more precision than the estimates actually have. Coverage is of the true value by the back-transformed 95 % interval, out of six.

Parameter	Model-based rel. SE	Empirical RSD	Calibration ratio	Coverage	η-shrinkage
Vp	2.7 %	1.8 %	1.53	6/6	9 %
R0	23.2 %	20.5 %	1.13	6/6	5 %
Vc	12.3 %	11.1 %	1.11	6/6	6 %
CL	5.1 %	4.8 %	1.07	6/6	2 %
ka	8.5 %	10.9 %	0.79	6/6	−3 %
kint	3.1 %	6.0 %	0.52	5/6	66 %
ksyn	6.2 %	13.3 %	0.47	4/6	48 %
Kss	**15.0 %**	**35.6 %**	**0.42**	**3/6**	**83 %**

The covariance step is anti-conservative exactly where the model is least identifiable. The three parameters whose model-based uncertainty most understates the empirical spread are the three target parameters, in the same order as every other ranking in this paper; the five disposition parameters have ratios between 0.79 and 1.53. Coverage follows: 6/6 for every disposition parameter, 3/6 for Kss. With six replicates a coverage figure carries an uncertainty of roughly ±20 percentage points and is reported as a direction, not an estimate — but the direction is consistent across three parameters and matches the calibration ratios. The practical consequence is that an analyst who read the standard error from a single fit of this model would conclude Kss was determined to about 15 %, when the across-dataset reality across these six fits is 36 %. On the 24 multistart fits of main-text Table 5 the same comparison is 16 % against 59 %.

η-shrinkage reproduces the ranking, and now meets the criterion. Kss has the highest η-shrinkage of any parameter at 83 %, followed by kint at 66 % and ksyn at 48 %; the disposition parameters are all below 10 %. All six fits pass prespecified criterion 4 (maximum shrinkage across η's 81.0–87.2 %, below the 90 % threshold), but only just — the threshold was set before these numbers existed and is, in hindsight, close enough to the observed values that it would not have discriminated much. Notably R0's η-shrinkage is only 5 %, which does not contradict the high shrinkage reported in the source analysis: here R0 carries a 158 % between-subject CV, so its random effect is well-informed even though its typical value is not precisely determined.

The software's `%RSE` column is not comparable across parameters, and one of our own criteria was built on it. For a log-transformed parameter the reported quantity is $100 \cdot \mathrm{SE}(\log \theta)/|\log \theta|$, whose denominator depends on the units in which θ is expressed. We verified this identity against the fits' output for all eight parameters. Because log Kss ≈ 0.16 at these values, the column's median across the deposited fit sets is 26–71 % for Kss, while for CL (log ≈ −5.1) it reports 1.0 % for the *same* relative precision. This explains the 124 % figure in S2.3 and the 8,645 % figure in S2.8: both are small standard errors divided by a log estimate close to zero, not failures of the covariance step. Prespecified convergence criterion 2 was written in terms of this column, and is therefore a weaker criterion than it appeared — the finiteness test it applies is meaningful, the magnitudes it would have thresholded are not. The scale-free quantity, $\mathrm{SE}(\log \theta)$, is what Table S6 uses and what should have been prespecified. We record this rather than revise the criterion retrospectively.

S2.10 What the Kss error costs: little on the drug, much on the target

Every section so far asks how well Kss can be estimated. None asks whether it matters — which is the question a decision-maker asks first, and the one that determines whether "Kss is practically non-identifiable" is a warning or a footnote. It needs no new fits. The typical-subject profile was solved for each of the 24 fitted parameter vectors of main-text Section 3.1 and compared with the true-parameter profile over 18 months on a 1,501-point grid, for two outputs: total drug concentration, which was observed, and free RANKL, which was not. Kss was then isolated two ways — the true vector with only Kss replaced by each fitted value, and each fitted vector with only Kss replaced by the truth (`prediction_impact.csv`, main-text Fig. 1).

Table S7. Median absolute prediction error over 18 months, across the 24 fitted vectors.

Parameter vector	Total drug (observed)	Free RANKL (not observed)
fitted, as estimated	11.5 %	30.8 %
fitted, with the true Kss substituted	11.3 %	12.9 %
truth, with only Kss set to the fitted value	**1.5 %**	**26.4 %**

Three results, and together they reframe the paper.

Kss is almost irrelevant to the observed output. Corrupting the true parameter vector by nothing but the fitted Kss error — a median −24.6 %, range −65 % to +38 % — moves the predicted drug concentration by a median 1.5 % (range 0.35–4.0 %). Not one of the 24 replicates exceeds 5 %. This is the mechanism behind every other result in this paper stated as a magnitude: Kss cannot be estimated from these data because it barely changes what these data measure. Practical non-identifiability here is not a numerical pathology, a design accident or an algorithmic failure; it is the response of an estimator to a parameter that the drug-concentration data constrain only weakly.

On the output that matters, the same error is 18 times as expensive. The identical Kss perturbation moves predicted free RANKL by a median 26.4 % (range 6.3–60 %), and 17 of 24 replicates exceed 20 %. Substituting the true Kss back into each fitted vector removes 58 % of the free-RANKL prediction

error and 0.2 percentage points of the drug-concentration error. The fitted models agree with each other and with the truth about the drug, and disagree about the target (main-text Fig. 1a).

The transfer is almost one for one. Regressing the error in predicted maximum RANKL suppression on the error in Kss across the 24 replicates gives a slope of 1.14 (SE 0.10), r = 0.92, $p = 1 \times 10^{-10}$. A 25 % error in Kss is a 25 % error in the predicted depth of target suppression. There is no averaging-out and no compensation by the other seven parameters, which is consistent with S2.6's finding that the poorly determined direction of the information matrix is log Kss alone rather than a composite.

So the unidentifiability is consequential, not benign. A sparse PK-only design of this kind yields a model whose *Kss* error costs far more on the pharmacodynamic consequence of exposure than on exposure itself — a median 1.5 % against 26.4 % in isolation, while the whole fitted vector carries 11.5 % on total drug and 30.8 % on free RANKL, so the model does not predict exposure to within 1.5 % and no such claim is made — while reporting a model-based standard error on Kss of 16 % (main-text Table 5) that understates the across-dataset spread by a factor of 3.7 and gives no indication of the problem. On the six true-value-start fits of S2.9 the same factor is 2.4. An analyst with one such fit has no internal signal that the target prediction is unreliable.

This also closes the loop with S2.8 from the other direction. The design arm that repaired Kss did so by making the target's dynamics visible in the drug data; S2.10 says the reason it is worth repairing is that Kss sets the target prediction. Both point at the same intervention — measure the target — and neither points at more subjects or later PK samples.

One bound on this analysis: it is a typical-subject calculation with the random effects set to zero, so it characterises the error in the structural prediction rather than in a population prediction with its variability, and it compares each fitted vector against the single truth rather than propagating estimation uncertainty. A full version is a VPC against the true model and is not run.

S2.11 Fitting at N = 80: the extrapolation was right, and so the earlier argument was wrong

S2.6 left the reviewer's question — would 80 subjects fix this? — bracketed between two incompatible extrapolations, and S2.10 leaves open whether a larger cohort is an alternative to measuring the target. Six datasets of 80 subjects were simulated and fitted under the reference design at the same iteration budget as the six 30-subject reference fits of S2.8, which makes the comparison like-for-like in everything but sample size (`n80.csv`).

Table S8. Kss recovery at two sample sizes, six datasets each, identical design, starts and iteration budget.

	Median Kss	Rel. bias	RSD	Range	Criteria
N = 30	1.17	−24.7 %	35.6 %	0.81–2.05	fail (both)
N = 80	1.63	**+4.8 %**	**21.8 %**	1.18–2.22	**pass (both)**

The empirical extrapolation was accurate and the Fisher information was not. The observed RSD at N = 80 is 21.8 %. Scaling the N = 30 RSD by N^−1/2 predicted 21.5 % from the 24-replicate value and 21.8 % from these same six datasets; the observed ratio is 0.611 against a predicted $\sqrt{(30/80)} = 0.612$. The Fisher information predicted 78.8 % and is wrong by a factor of 3.6. So the functional form was never in doubt and the FIM's absolute level was the whole error, which is the expected direction for a fixed-effects approximation that omits eight variance components and is worth stating because the FIM is often the only instrument available before any data exist.

The bias is finite-sample, not structural, and an earlier version of this argument was wrong on that point. S2.6 previously rested its conclusion on the claim that increasing N shrinks a variance but not a bias, so that no cohort size could deliver a *correct* Kss. The median bias moves from −24.7 % to +4.8 %. The difference is 29.5 percentage points against a pooled Monte Carlo standard error of 18.6, i.e. 1.6 MCSE units, with a Mann–Whitney p = 0.047 across the twelve estimates — enough to refute "the bias

does not move", not enough to quantify where it settles. At N = 80, Kss passes both prespecified criteria. The earlier claim is withdrawn.

These are the first fits in this work for which all five operational convergence criteria were both recorded and met. All six returned without error with a finite objective, the covariance step succeeded on all six, the largest reported %RSE was 16–49 %, maximum η-shrinkage was 71–82 % (below the 90 % threshold), and no variance component sat at a boundary. The other seven structural parameters also improved and none degraded: RSD roughly halves across the board (Vc 11.1 → 3.9 %, ka 10.9 → 5.9 %, ksyn 13.3 → 5.6 %, kint 6.0 → 3.9 %, CL 4.8 → 2.1 %, Vp 1.8 → 0.9 %, R0 20.5 → 16.6 %), and R0 retains the largest residual bias at +8.1 %.

What this does and does not change. It changes the recommendation from one route to two: at these parameter values a cohort of roughly 80 brings Kss inside the criteria, so the deficit at N = 30 is a power problem as well as an information-geometry problem. It does not touch S2.10 — a Kss error still costs 18 times more on the target prediction than on the drug, so the *cost* of residual error remains concentrated where it cannot be seen, and a 21.8 % RSD on Kss is still roughly a 22 % error on predicted target suppression. Measuring the target remains the cheaper and more robust intervention, because it attacks the 18-fold asymmetry itself rather than paying N^−1/2 to shrink one side of it. And the claim is bounded by six replicates at true-value starts under one design: the start-value sensitivity of S2.3 was not re-examined at N = 80, and a larger cohort is not evidence that a realistic analyst starting from literature values would land near the global optimum.

Table S14. Per-parameter recovery across the 24 replicate datasets at true-value starts, the comparison set for main-text Table 1. Relative bias and RSD as defined in main-text Section 2.5. Because the true values are not a maximum-likelihood starting point, this set bounds achievable recovery rather than estimating it.

Parameter	True	Median	Rel. bias (%)	RSD (%)	Range	Criteria
V_p/F (L)	6.06	6.09	+0.6	2.8	5.82–6.39	pass
σ_{prop}	0.07	0.0709	+1.3	4.5	0.0621–0.0774	pass
CL/F (L h^{-1})	0.006	0.0061	+1.0	5.8	0.0053–0.0069	pass
k_{int} (h^{-1})	0.022	0.0209	-5.0	7.1	0.0190–0.0254	pass
k_a (h^{-1})	0.0078	0.0078	+0.5	7.9	0.0066–0.0092	pass
k_{syn} (nmol L^{-1} h^{-1})	0.01	0.0096	-4.4	11.6	0.0075–0.0121	pass
σ_{add} (nmol L^{-1})	0.72	0.73	+2.0	11.9	0.56–0.91	pass
V_c/F (L)	1.58	1.59	+0.8	11.9	1.25–1.98	pass
R_0 (nmol L^{-1})	15.23	14.85	-2.5	20.8	11.12–22.40	pass
K_{ss} (nmol L^{-1})	1.56	1.18	-24.6	35.1	0.54–2.16	**fail (both)**

S2.12 Convergence check and the FOCEi re-estimation

Pre-specification. The convergence check was specified in full, including its decision rule and the replicates to be used, before any fit was run; the specification and two subsequent amendments are deposited as `CONVERGENCE_RULE.md`. No criterion, threshold or replicate was changed after a result was seen; where an amendment changed the plan, the reason and the date are recorded in that file.

Arm A — chain length. Eight replicates, named in advance (1, 8, 11, 13, 18, 19, 20, 23), were refit by SAEM at nBurn 1000 / nEm 1000 from their own phase-3 estimates, with the between-subject variances started at the programme's standard values so that chain length is the only difference. All eight returned without error with a successful covariance step. Kss changed by more than 5 % in all eight — +23.3, +279.6, +12.7, −42.3, +17.4, −31.2, −15.9 and −65.7 % — with a median absolute change of 27.2 %.

Arm B — a second estimator. The same eight replicates were fitted by FOCEi from the same starting values. Agreement with the long-chain SAEM estimate to within 10 % was reached in one of eight.

What the traces show. The per-iteration traces (`parts4/trace40N.csv`) are the substantive result. During burn-in Kss wanders over an order of magnitude; from the first EM iteration the trace is flat and does not move again, because SAEM's second phase averages the stochastic approximation and so retains wherever burn-in left the chain. Trace flatness over the second half of a run is therefore not evidence of convergence in this model, and a longer chain lengthens the frozen segment rather than extending the search. The long chain also drove ω^2(η_Kss) to a median 0.0068 against the true 0.2910 — about 2 % — where the fits reported in the main text hold it at 64–65 % of truth.

Consequence. Because a longer chain cannot establish convergence here, the remedy was changed from rerunning all 24 replicates at the longer chain to the common-objective comparison of S1.6.5, which answers the question the chain length was meant to answer.

FOCEi on all 24 replicates. All 24 returned without error; the covariance step succeeded on 24 of 24. Twenty of 24 meet all five operational convergence criteria of S1.6.3. The four exceptions are reported rather than excluded: replicates 11 and 23 return ω^2(η_Kss) at the upper bound `foceiControl()` imposes on a diagonal variance (five times its initial value, 1.4551), and replicates 10 and 16 exceed the 90 % η-shrinkage ceiling at 91.5 % and 94.8 %. Two covariance steps are uninformative rather than optimistic: the relative standard error on Kss is 1832 % in replicate 4 and 173 % in replicate 24, both finite and therefore passing a finiteness test.

Table S13. Per-replicate comparison of the two estimators under the common objective of S1.6.5 is tabulated in the main text as Table 2; the full record, including the objective at the true vector and every recorded variance component, standard error and shrinkage value, is in `focei24_fits.csv`, `focei24_objeval.csv` and `phase4_all.csv`.

S3. Profile likelihood in full

Profiles were computed for Kss and for R0 as a comparator, each on a 7-point grid spanning ±1.5 log units about the true value, fixing the population typical value and re-estimating every remaining free parameter **including that parameter's own variance component** (one dataset).

Table S9. ΔOFV from each curve's own minimum, **one start per grid point** (`profile_likelihood.csv`).

Profiled / true	0.22	0.37	0.61	1.00	1.65	2.72	4.48	Span
Kss	22.8	17.6	**0.0**	4.3	4.4	16.7	9.8	**22.8**
R0	71.0	44.5	20.5	12.0	**0.0**	3.1	27.7	**71.0**

Read at face value this says Kss is constrained about three times more weakly than R0. An earlier version of this manuscript reported that as the headline evidence for the Kss ranking, and noted as a caveat that the Kss curve is non-monotone while the R0 curve is smooth. Both reviews asked for multistart at every grid point. We ran it for both parameters — three extra log-normally perturbed starts (SD 0.3) at each of the fourteen grid points, 42 fits (`profile_multistart_full.csv`, `profile_multistart_R0.csv`, main-text Fig. 2).

Table S10. Four starts per grid point. Recovery is the single-start objective minus the best of four; spread is the range of the four objective values at that fixed value.

Fixed at / true	Kss: best of 4	recovered	spread	R0: best of 4	recovered	spread
0.22	2180.66	26.1	26.1	2237.39	15.9	19.4
0.37	2180.79	20.8	20.8	2198.31	28.4	29.3
0.61	2175.71	8.3	9.1	2177.64	25.1	27.6
1.00	2179.03	9.3	24.5	**2157.05**	37.2	37.2
1.65	**2167.98**	20.4	22.6	2157.08	25.1	25.1
2.72	2173.36	27.3	27.3	2180.16	5.1	29.4
4.48	2184.58	9.2	10.3	2171.00	38.9	38.9
span of the 4-start curve			**16.6**			**80.3**

Four findings.

Single-start profiling failed for both parameters, so the earlier caveat mis-attributed the problem. All fourteen grid points were above their conditional optimum — recovery 8.3–27.3 units for Kss (median 20.4) and 5.1–38.9 for R0 (median 25.1). The smooth single-start R0 curve was not evidence that one start sufficed for R0; it was luck. Any claim resting on a single-start profile of this model class should be regarded as unsupported, whatever the curve looks like.

With four starts, the comparator claim is restored on a valid footing and is stronger than before. Over the identical ±1.5 log-unit span the Kss curve rises 16.6 objective units and the R0 curve 80.3: **R0 is constrained 4.8 times more strongly than Kss**, against the 3.1-fold factor the invalid curves gave. The R0 four-start minimum also lands on the true value (the adjacent grid point is 0.03 units above it, so the two are tied in practice), where its single-start minimum had been displaced to 1.65× truth; the Kss minimum remains displaced, at 1.65× truth with ΔOFV 11.1 at the true value.

A between-start diagnostic separates the two cases, and it needs no knowledge of the true optima. Divide the median spread of the four objective values at one fixed value by the span of the whole four-start curve. For R0 the ratio is 29.3/80.3 = **0.36** — the curve carries about three times more signal than the scatter that threatens it. For Kss it is 22.6/16.6 = **1.36** — the scatter at a single fixed Kss is larger than the entire range of the curve across a 20-fold change in Kss. A profile is a lower envelope over starts, so more starts can only lower it; when the ratio exceeds one, the envelope at that start count is not a stable estimate and no interval or flatness statement can be read from it. R0's profile is usable at four starts and Kss's is not, which is a quantitative version of the same conclusion the rest of the paper reaches by other means.

The Kss ranking no longer depends on this section, but is now also supported by it. The independent lines are main-text Section 3.1 (24-replicate recovery under two estimators; the design-arm reference arm of S2.8 is a separate six-dataset experiment, not a reproduction of it), S2.6 (Fisher information and eigenstructure, sharing no code with the estimation), S2.8 (design-arm response), S2.9 (uncertainty calibration and η-shrinkage) and S2.10 (prediction consequence). The profile adds a fifth: at matched four starts, 4.8-fold weaker constraint than the comparator.

The practical recommendation for anyone profiling a QSS-TMDD model with SAEM follows directly: use multiple starts at every grid point, and report the between-start spread alongside the profile. Without the spread a smooth curve carries no information about whether its points are optima, and with it the usability of the profile can be judged from the figure.

The observations about which parameters move along the Kss curve are descriptive and are taken from the single-start series: the re-estimated R0 and kint rise with the forced Kss ($r = +0.97$ and $+0.97$, R0 from 17.3 to 24.1 and kint from 0.0218 to 0.0258) and Vc falls ($r = -0.87$). S2.6 reaches a different

ordering from the Fisher information, where Kss pairs most strongly with Vc (−0.94) ahead of kint (+0.63) and R0 (+0.61). We report both and claim neither as the mechanism.

On the ΔOFV = 3.84 reference. It is a conventional asymptotic value and nothing more; no confidence interval is read from these curves. A χ^2_1 calibration would require regularity conditions that are not established here: variance components are estimated, the likelihood is an approximation (Gaussian quadrature, nnodes 3), the data are left-censored, parameters have boundaries, and S2.3 shows the objective surface is multimodal.

The closest adjacent work is a preprint deriving exact total-QSS formulas for competitive TMDD (Kim et al., 2026), which independently reports a false objective-function minimum in a reduced TMDD model — a different route to the same warning that an objective value in this model class can be reached without being the conditional optimum.

These 42 fits are also the first in this project for which η-shrinkage was recorded (S2.8 explains why), and the first at which a prespecified convergence criterion excludes a fit. Maximum shrinkage across η's ranged 68.1–90.2 % over the 21 Kss fits and 74.1–88.6 % over the 21 R0 fits; **one** Kss fit — a perturbed start at 0.22 × truth — reached 90.2 % and therefore fails criterion 4. It is reported rather than dropped, and it does not affect Table S10: it was not the best start at its grid point, so excluding it leaves every best-of-four entry unchanged to the last decimal. The covariance step succeeded on all 42 and no variance component sat at a boundary.

S4. Complete list of limitations

1. The ground truth is a simplified model: no CL–Vp correlation, no body-weight covariates, Q fixed without variability.
2. Twenty-four replicates support a pilot, not confident statements about bias, precision, coverage or parameter ranking.
3. The primary analysis starts at the true values. This is no longer only a pilot limitation: all 24 replicates were re-estimated from three perturbed starts each plus a matched-budget true-value start, and the winning start was refit at the budget of main-text Table 1 (Table 2). The winning start was selected at the lower budget and only then refit, which is an approximation that has not been tested directly.
4. A single primary sampling design.
5. M1 censoring only, in a design where censoring is negligible.
6. The Kss profile is not usable for interval estimation at four starts per grid point: its between-start spread at one fixed value exceeds its own span (ratio 1.36, against 0.36 for R0), and because a profile is a lower envelope over starts, the four-start curve is an upper bound of unknown tightness (S3). The comparator claim is reported as a span ratio at matched start count, not as an interval. No kint profile was computed, and both profiles are on one dataset.
7. Uncertainty calibration now rests on the 24 multistart fits (main-text Table 5), so the coverage figures carry roughly ±10 percentage points of uncertainty and remain directional. The six-fit true-value-start version is retained as Table S6.
8. No population predictive check was run. The prediction-impact analysis of S2.10 is a typical-subject calculation against a single truth with no propagated estimation uncertainty, so it bounds the structural prediction error rather than giving a population prediction interval.
9. Multistart was not applied within the reduced-structure comparison.
10. The Fisher-information analysis of S2.6 is a fixed-effects approximation at the typical subject: it ignores the eight variance components and the censoring, and its N-scaling assumes the same design per subject. It is used for a ranking and an order-of-magnitude sample-size statement, not for interval estimation, and it disagrees with the profile on the sign of two couplings.
11. The design arms of S2.8 use six paired datasets per arm with true-value starts and a single seed, which is enough to exclude a large effect and not enough to estimate a small one.

12. SAEM iteration budgets were not harmonised across analyses (Table S1), so objective values are comparable only within an analysis.
13. The computational environment required a patched build in which two nlmixr2 functions are unusable; results were obtained with the functions that work and no claim is made about other environments.
14. S2.10 is a typical-subject calculation with random effects set to zero, compares each fitted vector against a single truth, and does not propagate estimation uncertainty; it bounds the structural prediction error, not a population prediction interval. No VPC was run.
15. All five operational convergence criteria are recorded for the 24 replicates, and criterion 3 was revised after it was declared (S1.6.3); the recovery criteria remain post hoc with respect to the true-value-start comparison set (main-text Section 2.5).
16. The design arm that repairs Kss varies target timescale, target flux and censoring fraction together, so its positive mechanistic attribution is not separable (S2.8).
17. The median Kss bias is conditional on the SAEM iteration budget: the same six datasets give −35.8 % at 150/200 iterations and −24.7 % at 100/150 (S2.8).
18. No external validation against an independent implementation of the *estimation* step, although the simulation was validated against a second, independent implementation and the two drivers of the estimation step agree exactly (S1.5.2).

Table S11. Recovery of the eight between-subject variance components across the 24 replicates, true-value starts. Criteria as for the typical values.

Variance component	True	Median	Rel. bias (%)	RSD (%)	Criteria
ω^2(ka)	0.27764	0.27240	-1.9	29.7	pass
ω^2(Vc)	0.32249	0.35980	+11.6	36.6	**fail**
ω^2(Vp)	0.02389	0.02220	-7.1	29.4	pass
ω^2(CL)	0.06733	0.06520	-3.2	35.0	**fail**
ω^2(kint)	0.00619	0.00910	+47.0	92.3	**fail**
ω^2(Kss)	0.29102	0.18990	-34.7	61.3	**fail**
ω^2(ksyn)	0.04921	0.02890	-41.3	97.2	**fail**
ω^2(R0)	1.25444	1.41340	+12.7	45.8	**fail**

S5. Correction history

Every defect found during this work is recorded in `defect_register.csv` (39 entries: 4 fatal, 15 major, 20 minor), with the evidence that established it and the fix applied. The dispositions of two external reviews of an earlier draft are in `critique_audit_FINAL.csv` (38 items, each with a machine-checked evidence probe). Four claims were withdrawn during the work and are listed here because three of them appeared in earlier drafts:

1. A random effect on Q present in the simulation and absent from every estimation script, which invalidated an entire earlier result set. Diagnosed by replaying the random-number stream and resolving the system in a second implementation; the objective-function cost of the discrepancy was localised to +1250 units, 621 of them in one subject.
2. “Kss is constrained about three times more weakly than R_0”, from single-start profiles. Withdrawn; see S3.
3. “Increasing N shrinks a variance but not a bias, so no cohort size delivers a correct K_ss.” Refuted by the N = 80 fits (S2.11).

4. "The terminal sampling region carries the K_ss information." Demoted to a hypothesis, then found to be the weaker of two candidate mechanisms (S2.8).

The full narrative account is in `DIAGNOSIS_AND_REDESIGN.md`.

S6. Software environment and reproduction

R 4.5.3, rxode2 5.1.7, nlmixr2est 7.1.0, nlmixr2 7.0.1, aarch64-apple-darwin. `renv.lock` pins all 179 packages; it carries no package hashes because the library was built directly from CRAN rather than under `renv`, so `renv::restore()` reinstalls by version rather than verifying by checksum. `setOfv(fit, "focei")` segfaults in this build and is avoided. FOCEi runs correctly with `foceiControl(optExpression = FALSE)`, which is how every FOCEi fit and every fixed-parameter objective evaluation reported here was obtained; no claim is made about other environments.

Every analysis runs through `10_worker.R` against a row of `jobs.csv`. The worker records all eight typical values, all eight variance components, both residual terms, per-parameter standard errors and back-transformed confidence bounds, per-η shrinkage, covariance-step status, boundary flags, observation and below-limit counts, and runtime — one row per fit.

S7. Supplementary data files

Table S12. Every fit in this work, counted from the files.

File	Fits	Supports
`sse_replicates.csv`	24	24 replicate SSE (main-text Table 1)
`multiseed_aligned.csv`	5	seed sensitivity (S2.2)
`reduced_etas.csv`	4	random-effects structure (S2.4)
`profile_likelihood.csv`	14	single-start profiles, Kss and R0 (S3)
`multistart_basins.csv`	4	initialisation pilot (S2.3)
`determinism.csv`	8	reproducibility (S2.7)
`design_arms.csv`	18	design arms (Table S9)
`uncertainty_calibration.csv`	6	uncertainty calibration, true-value starts (Table S6)
`multistart_recovery.csv`	120	multistart recovery: 72 perturbed starts, 24 matched-budget true-value starts, 24 refits of the winning start at Table 1's budget (main-text Table 2, Table 5)
`profile_multistart.csv`	6	multistart profile, first 2 grid points (S3)
`profile_multistart_full.csv`	15	multistart profile, remaining 5 grid points (S3)
`profile_multistart_R0.csv`	21	multistart profile, R0 comparator (S3)

File	Fits	Supports
n80.csv	6	sample size N=80 (Table S9)
sse24_full_convergence.csv	24	re-analysis of the 24 replicates, all five convergence criteria and all eight variance components (Table S1, Table S11)
phase4_all.csv	16	convergence check (S2.12): eight SAEM fits at nBurn 1000 / nEm 1000 (arm A) and eight FOCEi fits from the same starts (arm B). The Kss_150_200 column of that file is each replicate's Table 2 multistart winner, joined in for comparison and already counted among the 120; it is not an additional fit.
focei24_fits.csv	24	FOCEi on all 24 replicates, every variance component, standard error and shrinkage value (main-text Tables 1–3)
TOTAL	**315**	

Non-fit tables:

File	Rows	Supports
references_verified.csv	25	Reference list provenance
defect_register.csv	39	S5
critique_audit_FINAL.csv	38	S5

S7.1 Reference verification

Every journal reference above was resolved through the Crossref REST API from its bibliographic string, and the journal, year, volume and DOI printed here are the values Crossref returned (references_verified.csv, which also records the matched title and author list for each). This replaced an earlier list in which seven entries had been verified only from search-engine metadata, and the check found one error that had survived in the group believed verified: Jaber et al. (2021) had been cited as 10(5), 418–424 and is 10(4), 275–278. Kim et al. (2026) is a preprint, is marked as such, and was read in full; it is the closest adjacent work to Section 3.3.