

Auditing model-based antiviral recommendations: a decision-relevance diagnostic for seasonal influenza surveillance

Tian-Yu Wang

wangty0709@hotmail.com

Guangdong Peizheng College

Research Article

Keywords: identifiability, sloppiness, simulation-based inference, decision-relevant uncertainty, seasonal influenza, antiviral allocation

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11350156/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Antiviral stockpile decisions rest on transmission models calibrated to surveillance data. Seasonal influenza surveillance records reported cases rather than infections, so the reporting rate ρ and the initial infection size i_0 enter expected counts only through their product: the likelihood is exactly flat along a ridge in (ρ, i_0) . Standard calibration checks can pass while the recommended intervention remains unidentified.

Methods

We built a six-parameter SEIRS-type influenza model with negative-binomial reporting, fit neural posterior estimators to reported counts, and scored decision stability with a variance-weighted decision-relevance score (DRS), the share of the decision margin's posterior variance coming from weakly identified directions. A scale-invariant positive control checked the score. We audited a 192-cell synthetic grid and 204 real CDC influenza region-seasons, comparing the score with a purely geometric alternative.

Results

The ridge is exact: compensating i_0 across ρ 's full prior range changes the single-season log-likelihood by at most 0.261. On the synthetic grid the bounded posterior flipped the plug-in action in 27% of cells, losing 12% of the oracle's value. The geometric index failed its own control arm, whereas the variance-weighted score separated the arms (0.30 vs 0.13, Fisher/Laplace geometry; 0.25 vs 0.05, neural posterior). Parameter-level non-identifiability did not predict decision instability (Spearman + 0.107, $p = 0.14$). On real seasons the score separated sensitive from insensitive decisions by up to several hundredfold, and seven of 204 region-seasons (3.4%) had a margin below its own posterior standard deviation, six in the 2022–2023 season.

Conclusions

A calibrated posterior is not sufficient evidence that a model-based recommendation is identified, and parameter-level diagnostics do not reveal when this happens. Computed from the posterior a team already holds and needing no new data, the score flags the seasons in which the recommendation would be untrustworthy; we suggest running it as a routine audit alongside calibration checks.

Introduction

Influenza places a large and recurring burden on health systems, and dynamic transmission models are routinely calibrated to seasonal influenza surveillance data to support planning [1,2]. Antiviral stockpiles are among the few levers a health agency can pull once a season is under way. That lever is not a switch but a magnitude: a programme commits to a coverage tier, meaning how many treatment courses to release and how many people to reach. The costs are asymmetric, since over-commitment wastes drug and money while under-commitment leaves preventable illness in the population. Modellers are routinely asked to recommend a tier from weekly surveillance counts, and the recommendation is only as trustworthy as the mapping from counts to the decision.

Two literatures bound what surveillance can deliver, and both stop short of the decision. The first is about information and timing. Parag and colleagues show that under-reporting and reporting delay place a floor on the uncertainty of any epidemic quantity and use that to rank surveillance streams by information content [3]. On the same footing they show that real-time detection of resurgence in the effective reproduction number has a fundamental time limit [4], and that the asymmetry between starting and relaxing non-pharmaceutical interventions survives any improvement in the inference algorithm [5]. The second literature is geometric. Case and colleagues show that the summary statistics of an outbreak become informative only once enough of the epidemic has unfolded [6], while work on model sloppiness shows that many parameter combinations remain practically unidentifiable even when they are formally identifiable [7,8]. What neither literature settles is whether the residual uncertainty changes the action. A decision can survive a large parameter trade-off if the trade-off runs along a direction the loss does not care about, and it can be fragile even when every parameter looks tight. Parameter identifiability and decision identifiability are different properties, and only the second one matters to the agency. Whereas Parag and colleagues establish that noisy surveillance limits the timing of binary intervention decisions [3,4,5], we address a different question: conditional on a well-calibrated posterior, whether the magnitude or choice of an intervention is identifiable at all. We provide a decision-level diagnostic rather than a detection-threshold result.

This paper makes that distinction operational. We study a seasonal influenza transmission model in which only reported cases are observed, so that the reporting rate ρ and the initial infection size i_0 are linked by an exact, analytically transparent ridge. We then ask whether the optimal antiviral tier is identifiable conditional on a posterior that is well calibrated by conventional standards, and we build a diagnostic that answers the question at the level of the loss rather than the parameter vector. The diagnostic is a decision-relevance score (DRS): the share of the decision margin's posterior variance that comes from weakly identified directions. Our score sits within the family of loss-aware inference criteria that weight posterior uncertainty by its effect on a decision [9,10,11] and alongside work on decision-focused learning that moves the training objective toward the downstream task [12]. The epidemiological setting is deliberately small, so that the geometry can be checked against exact likelihood profiles before any neural approximation is introduced.

We show first that a well-calibrated neural posterior can still leave the antiviral tier substantially unidentified, with the plug-in action flipped in 27% of grid cells and the plug-in rule losing 12% of the

oracle's value. The obvious geometric index of decision-relevant sloppiness turns out to be actively misleading: it fails its own scale-invariant control arm in most cells because the decision gradient is not aligned with the flat ridge, while the variance-weighted index we propose behaves correctly under two independent covariance sources. On a decade of real CDC influenza surveillance the index separates sensitive from insensitive decisions by up to several hundredfold, and we report one natural hypothesis that our own data falsify. The score is computed from the posterior a modelling team already holds, needs no additional data, and runs on a CPU; we intend it as a routine audit alongside standard calibration checks rather than as a new inference method.

The rest of the paper is organised as follows. The Methods section specifies the transmission and observation models, the decision problem, the estimators, the grid design, and the data provenance. The Results section establishes the ridge, evaluates the two decision-relevance indices, reports the synthetic grid, the calibration results, the real-data audits, and the cost-ratio sensitivity, and states the open questions. The Discussion interprets the geometric failure, and the Conclusions section states what a modelling team should do differently.

Methods

Transmission model

We model a single influenza season with a susceptible-exposed-infectious-recovered system and waning immunity. With S , E , I , and R the population fractions,

$$\frac{dS}{dt} = -\beta SI + \omega R,$$

$$\frac{dE}{dt} = \beta SI - \sigma E,$$

$$\frac{dI}{dt} = \sigma E - \gamma I,$$

$$\frac{dR}{dt} = \gamma I - \omega R,$$

and transmission rate $\beta = R_0 \gamma \psi_a$, where $\psi_a = 1 - \varepsilon_\beta \tau_a$ is the reduction in transmission under action a . The system is integrated with a fourth-order Runge-Kutta scheme at a step of 0.02 weeks over a horizon of 20 weeks. Initial conditions are $S(0) = 1 - i_0$, $E(0) = 0.30 i_0$, $I(0) = 0.70 i_0$, and $R(0) = 0$, so i_0 is the total seeded infection fraction split between exposed and infectious states.

The six unknown parameters are R_0 , σ , γ , ω , ρ , and i_0 . Their prior is uniform on the log box with ranges R_0 in $[1.15, 1.50]$, σ in $[2.0, 7.0]$ per week, γ in $[1.0, 2.5]$ per week, ω in $[0.005, 0.040]$ per week, ρ in $[0.05,$

0.60], and i_0 in $[1.0 \times 10^{-6}, 5.0 \times 10^{-5}]$. All constants of the study design are declared, not estimated (Table 1).

Table 1
Model parameters, priors, and declared constants.

Symbol	Meaning	Range or value
R_0	basic reproduction number	uniform, log[1.15, 1.50]
sigma	rate of progression E to I	uniform, log[2.0, 7.0] per week
gamma	recovery rate	uniform, log[1.0, 2.5] per week
omega	rate of immunity waning	uniform, log[0.005, 0.040] per week
rho	case reporting rate	uniform, log[0.05, 0.60]
i_0	initial infection fraction	uniform, log[1.0×10^{-6} , 5.0×10^{-5}]
ε_β	antiviral efficacy on transmission	0.50 per unit coverage
tau	commitment levels (actions)	{0.10, 0.30, 0.60}
lambda	intervention cost ratio	9.827704×10^{-3}
phi	negative-binomial dispersion	10.0
N_{POP}	catchment population	1.0×10^6
T	burden horizon	20 weeks

Observation model

Reported counts are drawn from a negative-binomial distribution with mean $\mu_t = \rho N_{POP} \text{incidence}_t(\theta)$ and dispersion $\varphi = 10$, treated as known:

$$y_t \sim \text{NegBinomial}(\text{mean} = \mu_t, \text{dispersion} = \varphi), t = 1, \dots, W.$$

The negative binomial is the observation model of the ridge. Because incidence_t is proportional to i_0 in the epidemic's exponential phase, μ_t depends on ρ and i_0 only through the product, and the likelihood is exactly flat along the log-space direction $(1, -1)/\sqrt{2}$.

Decision problem

The three actions are antiviral commitment levels τ in {0.10, 0.30, 0.60} of the population. The per-capita loss of action a at parameters θ is

$$l(a, \theta) = \frac{1}{N_{POP}} B_a(\theta) + \lambda \tau_a,$$

where B_a is the cumulative number of infections over the 20-week horizon under action a and λ is the cost per unit coverage. The optimal action is the minimiser of l . The scale-invariant control loss replaces the absolute burden with the relative reduction in transmission,

$$l_{ctrl}(a, \theta) = 1 - \frac{B_0 - B_a}{B_0} + \lambda \tau_a,$$

which carries no i_0 dependence and is therefore the positive control for the ridge.

The cost ratio is calibrated by a declared rule that makes no reference to any downstream result: λ is the median per-capita uncontrolled burden over the prior divided by the maximum commitment level. An earlier implementation of this rule double-counted i_0 and produced a value 113,343 times too small; the consequences are described below.

Decision-relevance scores

Let $u = \log \theta$ and let $g = d/du [l(a_{hi}, \theta) - l(a_{lo}, \theta)]$ be the decision gradient at the pair of actions with the two smallest posterior-mean losses. The decision margin is $m = l(a_{hi}) - l(a_{lo})$ and its posterior variance is $g' \Sigma g$, where Σ is the posterior covariance in u -space.

The geometric score projects g onto the sloppy subspace S of the posterior Hessian $H = -d^2/du^2 \log q(u | x)$, defined by eigenvalues below a threshold γ times the largest eigenvalue:

$$DRS_{geo} = \frac{\|P_S g\|^2}{\|g\|^2}, \quad S = \text{span} \{v_i : \lambda_i \leq \gamma \lambda_{max}\}.$$

The variance-weighted score instead asks what share of the margin's posterior variance comes from directions the data failed to shrink. Writing the eigendecomposition $\Sigma = \sum_i \lambda_i v_i v_i'$, a direction is sloppy when λ_i is at least half the prior variance along that same direction, and

$$DRS_{var} = \frac{\left[\sum_{i \in S} (v_i' g)^2 \lambda_i \right]}{(g' \Sigma g)}.$$

We report the variance-weighted score as the primary index and the geometric score as the methodological negative control. The gradient is evaluated by central finite differences in u -space with step 1×10^{-5} and cross-checked against exact autograd on the Runge-Kutta map.

Neural posterior estimation

We estimate the posterior with a hand-written conditional masked autoregressive flow, so that the exact Hessian required by the geometric score is available by autograd. A gated recurrent unit encodes the length- W count series (applied to $\log(1 + y)$, hidden width 64) into a 48-dimensional summary embedding. A conditional masked autoregressive flow with five affine layers, each built on a masked autoencoder with four hidden layers of width 96, maps the embedding and the parameter vector to a density over u . Training uses the automatic posterior transformation objective of Greenberg et al. [13], a form of sequential neural posterior estimation that treats $\log q(u_j | x_i)$ as unnormalised logits and minimises the cross-entropy with the diagonal target. Optimisation is Adam with learning rate 1×10^{-3} under a cosine schedule.

For the bounded estimator the flow models an unconstrained variable w and the parameters are obtained through the per-dimension sigmoid bijector $u_k = \text{lo}_k + (h_i - \text{lo}_k) \cdot \text{sigmoid}(w_k)$, with the exact log-Jacobian added to the flow log-density. The logit argument is clamped to $[1 \times 10^{-6}, 1 - 1 \times 10^{-6}]$ so that $\log_{\text{prob}}(u, x)$ stays finite and twice differentiable in u . All computation is CPU-only.

We place our estimator in the wider simulation-based inference programme [14,15] but implement it directly rather than through a standard toolkit, because the decision-relevance score needs a twice-differentiable log-density in the parameter argument and a support constraint that the toolkit interfaces do not expose.

Simulation-based calibration

Calibration is assessed in the simulation-based calibration style. For 200 synthetic seasons per window length, we draw parameters from the prior, sample a length- W observation, draw 1,000 posterior samples, and record the z-scores of the true parameters. We report the mean z-score standard deviation across parameters, the 90% parameter coverage, and the fraction of posterior draws that fall inside the prior box. A calibrated estimator has a mean z-score standard deviation near 1 and 90% coverage near 0.9.

Grid design

The headline grid crosses W in $\{6, 10\}$, ϕ in $\{2, 10, 50\}$, and four reporting-rate levels spaced evenly in log space across the prior box $[0.05, 0.60]$, with eight true seasons per level, giving 192 cells. For each cell we draw a season, sample counts at the cell's over-dispersion, run the neural posterior with 2,000 draws, and compute the decision-theoretic indices. The same grid is run with the unbounded estimator as a negative control. Seeds are fixed throughout (season seed 7, observation seed 999, posterior seed equal to the cell index).

Fisher/Laplace geometry

The variance-weighted score is evaluated under two covariance sources. The first is expected Fisher information: the Jacobian of the expected counts with respect to u , weighted by the inverse negative-binomial variance, combined with the prior covariance as $\Sigma = (F + C_{\text{prior}} - 1) - 1$. This Laplace route needs no neural network and is used as an independent check on the neural posterior. The second source is the sample covariance of the neural posterior draws.

External-validity data

Two public surveillance sources were retrieved on 2026-09-15. WHO FluNet virological surveillance was downloaded from the FLUMART public API (xmart-api-public.who.int/FLUMART/VIW_FNT), and CDC FluView outpatient influenza-like-illness surveillance was downloaded through the CMU DELPHI Epidata API (api.delphi.cmu.edu/epidata/fluview). Both endpoints returned HTTP 200 and neither required registration or an API key. We define seasons as week 40 of year Y through week 20 of year $Y + 1$, keep Northern-Hemisphere units with at least eight observed weeks inside the window, and retain ten seasons from 2015–2016 to 2024–2025. The final panel has 25,836 rows across 85 geographic units and 795 unit-seasons, of which 204 FluView region-seasons supply a complete 6-week window. Full field definitions, screening rules, and known pitfalls are documented with the data.

Because neither the reporting rate nor the catchment population is known for real surveillance, we estimate the combined reporting multiplier in two passes: a wide log-uniform search to locate the scale, followed by a log-uniform prior centred on that estimate with a width of one, two, or three orders of magnitude. The reported statistic is how the decision-relevance score responds to the assumed uncertainty in the reporting multiplier.

Statistics

Associations between the margin signal-to-noise ratio and the peak expected weekly count are summarised with Spearman rank correlation, because the relation is monotone but not linear, with Pearson correlation reported alongside. All intervals and p-values are two-sided. Because the synthetic grid is a fixed design rather than a random sample, we report correlations as descriptive summaries of the design points. Comparisons between the failing and passing region-seasons in the real-data audit use two-sided Mann-Whitney U tests with tie correction.

Cost-ratio sensitivity

The decision-relevance indices depend on the calibrated cost ratio $\lambda = 9.827704 \times 10^{-3}$. To check that the conclusions do not depend on this calibration, we recomputed the posterior-probability-of-flip over the 192-cell grid with λ scaled by factors of 0.1, 0.3, 3, and 10, holding the posterior draws and all seeds fixed.

Self-corrections

Three implementation issues are disclosed in full. The unbounded posterior estimator places almost all of its mass outside the prior box and is retained as a negative control rather than discarded. The

intervention cost ratio was initially mis-calibrated by a double-counted i_0 , giving $\lambda = 8.670733 \times 10^{-8}$ instead of 9.827704×10^{-3} , a factor of 113,343. With the erroneous value the decision degenerates to always choosing the maximum tier and the flip fraction is identically zero; the corrected value is used in every result reported here, and the pre-fix grid is retained (results/cells_lambda_model.csv). We also report the in-box fraction and the effective sample size for every posterior, so that a decision-level analysis built on an unconstrained flow would be visible as such.

Declaration of generative AI use

During the preparation of this work the author used Kimi K3 (Moonshot AI) for code development and language editing. The author reviewed and edited all AI-generated content and takes full responsibility for the integrity and accuracy of the publication.

Results

A structural ridge links the reporting rate to the initial infection size

The model is a standard SEIRS-type influenza system with susceptible, exposed, infectious, and recovered compartments and waning immunity (Methods). Weekly new infections are converted to reported counts through the observation model

$$\mu_t = \rho N_{POP} incidence_t(\theta),$$

1

where θ collects the six unknown parameters ($R_0, \sigma, \gamma, \omega, \rho, i_0$), N_{POP} is the catchment population, and $incidence_t$ is the weekly infection fraction produced by the ODE. While the epidemic is in its exponential phase and susceptible depletion is negligible, $incidence_t$ is proportional to i_0 . Substituting that proportionality into Eq. (1) shows that μ depends on ρ and i_0 only through the product ρi_0 . In log space the two parameters therefore trade off along the direction $(1, -1)/\sqrt{2}$, and any pair of parameter vectors on that line produces identical expected counts. This is an exact multiplicative ridge, not a weak or approximate trade-off.

We verified the ridge two ways. Moving ρ across its prior range (a twelve-fold span, the full prior box) while compensating i_0 leaves the single-season log-likelihood almost unchanged: the largest absolute change over the profiled path is 0.261 log-likelihood units (results/ridge_profile.csv). Over a 300-season ensemble the median maximum change along the ridge is 0.009 and the 90th percentile is 0.081 (results/ridges.csv). The ridge is deliberately planted by the observation model, and the likelihood confirms it to numerical accuracy.

The ridge is also the source of a tension that the rest of the paper quantifies. A longer observation window adds information about the growth parameters, but it also pushes the epidemic further from the linear phase in which the ridge is exact. Table 2 records this trade-off. At a 6-week window the median expected peak weekly count is 9 and the median relative residual of the observable along the ridge is 0.9%. At 10 weeks the same quantities are 30 and 3.1%. At 20 weeks they are 622 and 47%. Longer windows buy growth-rate information at the cost of a ridge that no longer holds exactly, which is precisely the regime in which a decision-relevance diagnostic has to be verified rather than assumed.

Table 2

Observation window W	Median peak weekly count	Median ridge residual
6 weeks	9	0.9%
10 weeks	30	3.1%
20 weeks	622	47%

A geometric decision-relevance index fails; a variance-weighted one works

We first define what it means for a decision to be unidentifiable. Let g be the gradient, with respect to the log-parameters $u = \log \theta$, of the loss difference between the two actions that are closest in posterior-mean loss. The decision margin is the scalar $m = l(a_{hi}) - l(a_{lo})$, and $g = dm/du$. A parameter direction matters to the decision only through its projection onto g . A natural index projects g onto the sloppy subspace of the posterior Hessian and reports the squared share that lands there (the geometric DRS). The control arm replaces the burden loss with a scale-invariant loss that depends only on the relative reduction in transmission and carries no i_0 term; by construction this control should be insensitive to the ridge.

The geometric index fails. Table 3 reports the two arms under two independent covariance sources and two evaluation points. Under Fisher/Laplace geometry at the true parameters the main arm has a median score of 0.0013 against a control of 0.0049, and the main exceeds the control in only 20.3% of cells. Under the neural posterior the geometric score is 0.0336 against a control of 0.0967, with the main exceeding the control in 11.5% of cells. The failure is not a threshold artefact: sweeping the sloppy-subspace cut from 1×10^{-1} down to 1×10^{-8} never lifts the main arm above the control in more than 17.2% of cells, and the sloppy subspace contains five of the six dimensions even at a cut of 1×10^{-2} (results/cells.csv). Figure 1 shows the four histograms; the two geometric panels are visibly inverted, with the control distribution sitting to the right of the main distribution.

The variance-weighted index reverses this. Instead of asking whether g lies in the sloppy subspace, it asks how much of the decision margin's posterior variance $g' \Sigma g$ comes from directions the data failed to shrink, with a direction classified as sloppy when its posterior variance is at least half the prior variance along that same direction. Under Fisher/Laplace geometry at the true parameters the main arm

scores 0.2996 against a control of 0.1255, separating cleanly in 78.1% of cells. Evaluating the same geometry at the plug-in estimate θ -hat gives 0.2981 against 0.0779 (79.2%). Under the neural posterior the variance-weighted score is 0.2493 against 0.0498 (74.5%). Table 3 summarises all six comparisons and Fig. 1 plots the four distributions.

Table 3

Decision-relevance scores under two geometries and two covariance sources. Main arm uses the burden loss; control arm uses the scale-invariant loss. $\text{frac}(\text{main} > \text{control})$ is the fraction of grid cells in which the main arm exceeds the control. Values are medians over 192 cells.

Index and covariance source	Evaluation point	Main (median)	Control (median)	$\text{frac}(\text{main} > \text{control})$	Verdict
Geometric, Fisher/Laplace	true theta	0.0013	0.0049	0.203	fails
Geometric, Fisher/Laplace	plug-in theta-hat	0.0011	0.0036	0.151	fails
Geometric, neural posterior	posterior mean	0.0336	0.0967	0.115	fails
Variance-weighted, Fisher/Laplace	true theta	0.2996	0.1255	0.781	passes
Variance-weighted, Fisher/Laplace	plug-in theta-hat	0.2981	0.0779	0.792	passes
Variance-weighted, neural posterior	posterior mean	0.2493	0.0498	0.745	passes

The reason for the reversal is geometric and can be stated exactly. The uncontrolled burden is far more elastic to R_0 than to i_0 : numerically the elasticity of the burden to R_0 is of order rT , about ten to twenty over the prior (measured median 18), whereas the elasticity to i_0 is exactly 1 because the burden is proportional to the seed. The decision gradient therefore loads mainly on directions along which the data are informative, and the angle between g and the ridge direction is close to a right angle. The median absolute cosine between the decision gradient and the ridge is 0.036 for the bounded posterior, our primary estimator, and 0.023 for the unbounded negative control; either value is effectively orthogonal. A purely geometric index asks whether the decision responds to the unidentified direction; it does not, and the index reports low decision-relevance even though the posterior is wide enough to flip the action. The variance-weighted index instead asks how variable the margin itself is under the posterior, which is the quantity that matters. Figure 2 relates the two levels directly, and the answer is negative. Across the 192 grid cells the parameter-layer non-identifiability, measured as the maximum relative posterior standard deviation, and decision instability, measured as the posterior probability of flipping the plug-in action, show no detectable relationship (Spearman + 0.107, $p = 0.14$, $n = 192$). This is a null result and we report it as such: a parameter posterior can be poorly identified along several directions and still leave the decision stable, and a tightly identified one can still leave it close to flipping. Figure 3 relates the conditioning of the parameter posterior to the variance-weighted decision-relevance score. Across the

200 Laplace cells the two are positively rank-correlated (Spearman + 0.616, $p = 2.8 \times 10^{-22}$), and the association survives partialling out the peak count (partial Spearman + 0.405), but it is a rank relation (the marginal Pearson correlation is -0.012) and the conditioning measure itself rises almost one-for-one with the case count (Spearman + 0.885). Conditioning therefore informs the decision-relevance score without determining it, and it says nothing about the flip probability: the two levels are connected through an information proxy, not through the ridge.

A calibrated posterior requires bounded support, not more training

A decision-relevance score is only meaningful if the posterior it is computed from is calibrated. Our first neural posterior estimator, a neural density estimator in the simulation-based inference tradition [13,14,15] trained with the automatic posterior transformation objective [13], was badly over-dispersed. Its average z-score standard deviation across parameters was 0.26 (0.261 for $W = 6$ and 0.273 for $W = 10$) rather than the value of 1 expected from a calibrated posterior, the 90% parameter coverage was a trivial 1.000, and only 0.05% of posterior draws fell inside the declared prior box. A posterior that lives almost entirely outside the support implied by the priors cannot support any decision-theoretic conclusion; it inflates every instability index and destroys the control arm, which failed in 98.4% of unbounded cells.

The fix is structural. We replaced the unbounded base distribution with a per-dimension sigmoid (logit) bijector $u_k = l_{o_k} + (h_{i_k} - l_{o_k}) \cdot \text{sigmoid}(w_k)$, so the density is confined to the prior box by construction and the log-Jacobian is added exactly. The bijector passes three self-tests: the forward-inverse round trip has a maximum error of 3.3×10^{-13} , the normalised density integrates to 0.9949 by volume-weighted expectation, and the autograd Hessian of $\log q$ is symmetric to machine precision. With the bounded flow the average z-score standard deviation rises to 0.966 ($W = 6$) and 0.959 ($W = 10$), 90% parameter coverage is 0.944 and 0.947, and every posterior draw lies in the box. Table 4 compares the two estimators.

Table 4

Posterior calibration of the unbounded and bounded estimators. SBC-style summary over 200 synthetic seasons with 1,000 posterior draws each.

Quantity	Unbounded flow	Bounded flow (W = 6)	Bounded flow (W = 10)
mean z-score sd	0.26	0.966	0.959
90% parameter coverage	1.000	0.944	0.947
fraction of draws inside prior box	0.0005	1.000	1.000

More training does not substitute for support. Increasing the number of epochs from 70 to 200 moved the parameter-wise z-score standard deviation only from the range 0.08 to 0.46 up to 0.13 to 0.52, and left the coverage and in-box fractions essentially unchanged. The over-dispersion is a construction problem, not an optimisation problem, which is why we report the unbounded run throughout as a negative control rather than a discarded pilot (results/cells_unbounded.csv).

Decision instability on the synthetic grid

The headline experiment is a 192-cell grid that crosses observation window (W in $\{6, 10\}$), observation over-dispersion (ϕ in $\{2, 10, 50\}$), and four reporting-rate levels spanning the prior box, with eight true seasons per level. For each cell we draw a season, sample weekly reported counts from the negative-binomial observation model, run the neural posterior, and compute the decision-theoretic indices.

Table 5 reports the median indices before and after the bounded-support fix. The bounded posterior flips the plug-in action in 26.7% of cells and the plug-in rule loses 11.7% of the oracle's value. The maximum relative posterior standard deviation falls from 44.7 to 0.78, and the control-arm variance-weighted score falls from 0.8154 to 0.0498 while the main arm stays at 0.2493, so the positive control starts working exactly where it should.

Table 5
Headline grid results, 192 cells. Medians over the crossed design. The unbounded column is the negative control.

Quantity	Unbounded NPE	Bounded NPE (final)
Posterior-probability-of-flip (PPF)	0.5205	0.2665
Relative excess risk (ER_{rel})	0.2315	0.1174
Fraction with plug-in action $\neq$ true action	0.4479	0.2969
Maximum relative posterior sd	44.708	0.7764
Variance-weighted DRS (neural posterior)	0.4957	0.2493
Variance-weighted DRS, control arm	0.8154	0.0498

One entry in Table 4 is easy to misread. The bounded posterior's draws lie inside the prior box by construction, so the prior-reweighted PPF coincides with the raw PPF and the importance-sampling effective sample size is 1.000. That equivalence is a property of the construction rather than a failure to compute the correction, and we report both numbers because it is the check that the bounded flow is doing what it claims. The correction itself is not cosmetic. Under the unbounded estimator the plug-in rule drifted systematically toward the most aggressive tier, with the action distribution $\{0: 43, 1: 66, 2: 83\}$ against a truth of $\{0: 72, 1: 78, 2: 42\}$. After the fix the plug-in distribution is $\{0: 70, 1: 81, 2: 41\}$, so the systematic bias is gone and the residual instability is genuine posterior width rather than a skewed estimator.

Additional file 1: Figure S1 plots the variance-weighted score against the flip rate. Their Spearman rank correlation is only -0.105 ($p = 0.15$), which is itself informative: a high sloppy-variance share does not by itself imply a high flip rate, because the flip rate depends on the margin in relation to the posterior spread as well as on where that spread sits.

Decision margins are comparable to their own uncertainty

A decision is hard when the gap between the best and second-best action is small relative to the posterior standard deviation of that gap. We measure this with the margin signal-to-noise ratio, the ratio of the decision margin to its posterior standard deviation. Across the 192 cells the median ratio under Fisher/Laplace geometry at the true parameters is 0.983, and the ratio falls below 1 in 50.0% of cells and below 0.5 in 28.1%. The neural posterior gives a median of 0.721 with 60.9% of cells below 1, and an independent Laplace route gives 62.5% below 1. Across three different ways of estimating the same margin, then, roughly half to two-thirds of cells have a decision margin no larger than its own uncertainty. In those cells the sign of the margin is not determined by the data, and the optimal tier is not identifiable, even though the posterior passes the calibration checks of Table 4.

Real influenza seasons reproduce the pattern and invert a natural hypothesis

To test external validity we built a surveillance panel from two independent public sources: WHO FluNet virological surveillance and CDC FluView outpatient influenza-like-illness surveillance delivered through the CMU DELPHI Epidata API (Methods). The panel covers ten seasons from 2015–2016 to 2024–2025, 85 geographic units, and 25,836 source-by-region-by-season-by-week rows, of which 204 are FluView region-seasons with a complete 6-week observation window.

On these real series the variance-weighted score behaves as designed. As the prior on the reporting multiplier is widened from one to two to three orders of magnitude, the main arm rises from 0.476 to 0.665 to 0.733 while the control arm stays near zero at 0.051, 0.0008, and 0.0049. The separation between the arms is 9-fold at the narrowest prior, 787-fold in the middle, and 149-fold at the widest. The mechanism is visible in the same data: admitting more uncertainty about the reporting multiplier is exactly what pushes the decision margin's variance into the unidentified direction, and the scale-invariant control is immune to it. Table 6 collects the real-data results.

Table 6

External validity on 204 real FluView seasons. Reporting-multiplier prior width is given in orders of magnitude. Values are medians over region-seasons.

Reporting-prior width	Main DRS (var.)	Control DRS (var.)	Separation	Median margin SNR	frac(SNR < 1)
1 decade	0.476	0.051	9x	7.54	2.5%
2 decades	0.665	0.0008	787x	4.68	2.9%
3 decades	0.733	0.0049	149x	4.16	3.4%

We had expected the opposite of what the data show. A natural hypothesis is that decision identifiability should degrade as counts get smaller, because low counts carry less information and should leave the margin noisier. The measured relation runs the other way, and it is consistent across three independent synthetic routes and the real data. Against the peak expected weekly count, the margin signal-to-noise ratio has a Spearman correlation of -0.352 ($p = 5.4 \times 10^{-7}$) under Fisher/Laplace geometry and -0.219 ($p = 2.2 \times 10^{-3}$) under the neural posterior, both on 192 cells; an independent Laplace path gives -0.363 ($p = 1.2 \times 10^{-7}$). On the real FluView seasons the Spearman correlations are -0.164 , -0.189 , and -0.113 for the three prior widths. Higher counts are associated with less identifiable decisions, not more. The Pearson correlations are much weaker (-0.267 synthetic, -0.151 neural, -0.123 Laplace), so the relation is monotone in rank rather than linear. Figure 4 shows the two synthetic panels and the real-season panel, each annotated with its fitted rank correlation and the share of cells falling below unit signal-to-noise.

We can state the mechanism of this inversion no more precisely than that it exists. The three synthetic routes use different covariance approximations and agree on the sign and on the rank nature of the relation, so the finding is not an artefact of one estimator. But we have not identified the structural reason, and we do not offer a speculative one. A second discrepancy compounds the puzzle. On synthetic seasons roughly half to two-thirds of cells have a decision margin below its own uncertainty, whereas on real seasons only 2.5% to 3.4% do. We tested whether the difference is simply one of information content and rejected that explanation: the count scales and the window lengths are not far enough apart to account for a twentyfold change in the share of unidentifiable decisions. Both the inverted relation and the synthetic-to-real gap remain open questions, and we flag them here rather than smoothing them over.

Which real seasons fail the audit

The seven region-seasons whose decision margin falls below its own posterior standard deviation are not distributed randomly across the decade of FluView data. Six of the seven fall in the 2022–2023 season, where 6 of 21 region-seasons (29%) fail the audit, against 0% to 5% in every other season; the remaining failure is a single region in 2017–2018. The failing and passing groups differ in the parameters that drive transmission, not in the amount of data available. The median R_0 of the failing group is 1.22 against 1.15 for the passing group (Mann-Whitney U, $p < 0.001$), and the median variance-weighted score is 0.66 against 0.73 ($p < 0.001$), so the flagged decisions are attached to more intense transmission and to a decision-relevance score that was already lower. The reporting multiplier does not distinguish the groups (median 103.3 vs 103.9, $p = 0.93$), and neither does the mean weekly case count (median 5,045 vs 1,571, $p = 0.09$), so the failures are not explained by sparse reporting or by small counts; if anything, the failing seasons recorded more cases. The 2022–2023 season is the highest-volume season in the panel, with a median mean weekly count of 6,191 against 768 to 3,369 in every other season, and it is the first season after the COVID-19 pandemic disruption of influenza transmission. The audit flags exactly this season as the one in which a model-based antiviral recommendation would have been least trustworthy.

Sensitivity to the intervention cost ratio

The headline instability figures are computed at the neutral cost ratio $\lambda = 9.827704 \times 10^{-3}$. Because that calibration could in principle drive the flip rate, we recomputed the posterior-probability-of-flip over the 192-cell grid with λ scaled by factors of 0.1, 0.3, 3, and 10, holding the posterior draws and seeds fixed (Methods).

Within a threefold band around the calibrated value (0.3x to 3x) the conclusions are stable. The median PPF stays between 0.19 and 0.31 across the band (0.310 at 0.3x, 0.267 at 1x, 0.190 at 3x), and the plug-in rule misclassifies the true action in 25% to 34% of cells at every factor, so the central result of this paper does not depend on how the cost ratio was calibrated. At the extremes it is the decision problem that degenerates rather than the inference becoming more reliable: at 0.1x the maximum tier is optimal in 62.5% of cells and the median PPF is 0.283, while at 10x the minimum tier is optimal in 68.8% of cells, the maximum tier is never optimal, and the median PPF falls to 0.068. Even at 10x, however, the plug-in rule still flips in 10.4% of cells. One caveat is measured rather than assumed: the share of cells with a margin signal-to-noise ratio below 1 does not decrease monotonically with λ (67.2% at 0.1x, 71.9% at 0.3x, 50.0% at 1x, 40.6% at 3x, 21.9% at 10x); from the calibrated value upward the share falls, because λ adds a constant to the loss and raises the margin without touching the posterior covariance, and the calibrated value itself sits exactly at one half, so any citation of the SNR statistic carries this premise. As a validation check, the 1x column reproduces the main grid exactly: the maximum absolute difference in PPF against results/cells.csv is 0.0, and the margin SNR agrees to machine precision (Table 7).

Table 7

Cost-ratio sensitivity of the headline grid, 192 cells. Posterior draws and seeds held fixed; only the action pair, the optimal action, and the margin are recomputed at each factor.

lambda factor	lambda	Median PPF	Plug-in != true (%)	Median SNR	frac(SNR < 1) (%)
0.1x	9.828×10^{-4}	0.283	18.2	0.82	67.2
0.3x	2.948×10^{-3}	0.310	34.4	0.76	71.9
1x (calibrated)	9.828×10^{-3}	0.267	29.7	0.98	50.0
3x	2.948×10^{-2}	0.190	25.0	1.94	40.6
10x	9.828×10^{-2}	0.068	10.4	4.22	21.9

A negative control and a self-correction

Two reporting decisions matter here. We keep the unbounded posterior run as a standing negative control rather than discarding it, because it is the run in which the control arm is visibly broken and therefore the sharpest evidence that the bounded-support fix is doing real work. We also have to report an earlier calibration of the intervention cost ratio that contained an error and would have hidden the entire mechanism. The cost parameter had been computed by multiplying the seed i_0 into the burden twice, making it 113,343 times too small (8.670733×10^{-8} instead of 9.827704×10^{-3}). With that value the intervention cost is negligible for every parameter in the prior, the optimal action is always the maximum tier, and the flip fraction is identically zero. The bug and its consequences are documented in full (results/ λ _calibration_diagnosis.json), the corrected run is the one reported throughout, and the prefix grid is retained for the record (results/cells_ λ _model.csv).

Discussion

The central finding is that parameter identifiability and decision identifiability are distinct properties, and that the standard geometric way of linking them is unsafe. The two are not interchangeable: across the 192 grid cells the parameter-layer non-identifiability carries no detectable relationship to decision instability (Spearman + 0.107, $p = 0.14$). The geometric shortcut, which treats decision-relevance as the projection of the loss gradient onto the sloppy subspace, gets the sign wrong. Our model contains an exact ridge along which ρ and i_0 are interchangeable, so a parameter-based diagnostic correctly reports that the pair is unidentified. Yet the optimal antiviral tier depends on that ridge only through the product ρi_0 in the observation model, and not at all through the burden that enters the loss. The decision's gradient therefore runs nearly perpendicular to the ridge. A geometric index that asks whether the decision lives in the sloppy subspace answers no and concludes that the decision is fine, while the posterior is wide enough to flip the action in a quarter of cases. The failure is not marginal: the geometric index failed its own control in 80% to 89% of cells and could not be rescued by any threshold on the sloppy subspace.

The variance-weighted index avoids the trap because it measures the variability of the margin itself rather than the geometry of the gradient. We take seriously the need for a positive control, and we make the control arm a scale-invariant loss that is blind to the unidentified direction by construction. Under that control the variance-weighted index passes on both covariance sources, with the main arm separating from the control in 74.5% to 79.2% of cells. The consistency across Fisher/Laplace geometry and a neural posterior matters here, because the two covariance sources are built from different approximations and would be expected to disagree if the result were an artefact.

The calibration result generalises past this model. Over-dispersed posteriors from unbounded normalising flows are not a training failure and cannot be fixed by training longer. Confining the posterior to the prior support with a bounded bijector removes the problem at the level of the estimator's construction, and the improvement is large: the maximum relative posterior standard deviation fell by a factor of about 58, the flip rate halved, and the plug-in bias toward the most aggressive tier disappeared. Any decision-level analysis built on an unbounded flow should be treated with suspicion until the in-box fraction and the control arm are checked.

The external-validity arm shows that the index is not a synthetic curiosity. On 204 real FluView seasons spanning a decade, widening the prior on the reporting multiplier moves the main score from 0.476 to 0.733 while leaving the control near zero, producing separations of up to several hundredfold. This is the behaviour an operational diagnostic needs: it should respond to genuine decision uncertainty and ignore uncertainty that the decision does not feel. The audit also resolves at the level of whole seasons: of the seven region-seasons that fail it, six fall in 2022–2023, the season with the highest surveillance volume in the panel (median mean weekly count 6,191, against 768 to 3,369 in every other season) and the first after the COVID-19 pandemic disruption.

Several caveats qualify these conclusions. The inverted count-identifiability relation is one we can document but not explain: higher counts go with less identifiable decisions across three synthetic routes and the real panel, and we have no mechanism for it. The gap between the synthetic and real arms in how often the margin is smaller than its uncertainty compounds it, and it is far larger than the difference in count scale can explain. The model itself is deliberately small, with six parameters and a single observation stream, so the specific numbers will not transfer to richer models. What we expect to transfer is the diagnostic and the warning about geometric shortcuts.

Several directions follow. The decision-focused learning literature has shown how to move a training objective toward a downstream task [12], and loss-calibrated inference has shown how to weight the posterior by the decision [9,10,11]; the natural next step is to fold the variance-weighted score into training so that the posterior is shaped around the decision rather than the parameters alone. The same machinery should be tried where the decision is a sequence rather than a single tier, since the asymmetry between starting and stopping an intervention is known to be hard [5]. A different route treats the tier choice as a sequential control problem; instabilities in meta-learned adaptation are documented [16], and safe meta-reinforcement learning has been developed for constrained settings [17]. The diagnostic is also cheap enough to run as a routine audit alongside conventional calibration checks, and we think it belongs there.

Conclusions

A posterior that passes standard calibration checks can still leave a model-based antiviral recommendation unidentified, and the parameter-level diagnostics a modelling team normally runs do not reveal when this happens. The variance-weighted decision-relevance score proposed here is a practical audit for that gap. It is computed from the posterior the team already holds, requires no additional data, and runs on a CPU (the full 192-cell cost-ratio sensitivity sweep, which re-scores every cell, completes in 64 seconds on one machine). In a decade of CDC influenza surveillance it flagged 3.4% of region-seasons as untrustworthy, with six of the seven failures concentrated in the 2022–2023 season. We suggest running this audit alongside calibration checks, so that a recommendation is not passed on until both the fit and the decision have been checked.

Declarations

Ethics approval and consent to participate

Not applicable. The study uses only publicly available, aggregated surveillance data and synthetic simulations; no human participants, human data, or animals were involved.

Consent for publication

Not applicable.

Competing interests

The author declares that they have no competing interests.

Funding

The author received no specific funding for this work.

Author Contribution

T.-Y.W.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. The author read and approved the final manuscript.

Data Availability

All data used here are public and were retrieved on 2026-09-15. WHO FluNet virological surveillance is available from the FLUMART public API at https://xmart-api-public.who.int/FLUMART/VIW_FNT. CDC FluView outpatient influenza-like-illness surveillance is available through the CMU DELPHI Epidata API at <https://api.delphi.cmu.edu/epidata/fluview>. The processed analysis panel is included with the code. All analysis code required to reproduce every figure and statistic, together with trained posterior checkpoints and grid outputs, is publicly available at <https://github.com/GustaveKateb0709/decision-relevance-influenza> under the MIT License. Weekly influenza surveillance data are publicly available from WHO FluNet (<https://www.who.int/tools/flunet>) and CDC FluView (<https://www.cdc.gov/fluview/>), the latter accessed programmatically through the CMU DELPHI Epidata API (<https://cmu-delphi.github.io/delphi-epidata/>). No restricted data were used.

References

1. Goeyvaerts N, Willem L, Van Kerckhove K, Vandendijck Y, Hanquet G, Beutels P, et al. Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-

- specific influenza-like illness incidence. *Epidemics*. 2015;13:1–9.
2. Marmara V, Cook AR, Kleczkowski A. Estimation of force of infection based on different epidemiological proxies: 2009/2010 Influenza epidemic in Malta. *Epidemics*. 2014;9:52–61.
 3. Parag KV, Donnelly CA, Zarebski AE. Quantifying the information in noisy epidemic curves. *Nat Comput Sci*. 2022;2(9):584–94.
 4. Parag KV, Donnelly CA. Fundamental limits on inferring epidemic resurgence in real time using effective reproduction numbers. *PLoS Comput Biol*. 2022;18(4):e1010004.
 5. Parag KV, Lambert B, Donnelly CA, Beregi S. Asymmetric limits on timely interventions from noisy epidemic data. *Commun Phys*. 2025;8:450.
 6. Case BKM, Young J-G, Hébert-Dufresne L. Accurately summarizing an outbreak using epidemiological models takes time. *R Soc Open Sci*. 2023;10(9):230634.
 7. Chiş O-T, Villaverde AF, Banga JR, Balsa-Canto E. On the relationship between sloppiness and identifiability. *Math Biosci*. 2016;282:147–61.
 8. Adenane R, Andreu-Villarraig C, Avram F, Villanueva RJ. Calibration and comparison of SIR, SEIR/SLIR and SLAIR models for influenza dynamics: insights from the 2016–2017 season in the Valencian Community, Spain. *Math Med Biol*. 2024;41(4):277–303.
 9. Lacoste-Julien S, Huszár F, Ghahramani Z. Approximate inference for the loss-calibrated Bayesian. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR. 2011;15:416–424.
 10. Vadera MP, Ghosh S, Ng K, Marlin BM. Post-hoc loss-calibration for Bayesian neural networks. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. 2021.
 11. Gorecki M, Macke JH, Deistler M. Amortized Bayesian decision making for simulation-based models. *Trans Mach Learn Res*. 2024.
 12. Mandi J, Kotary J, Berden S, Mulamba M, Bucarey V, Guns T, et al. Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities. *J Artif Intell Res*. 2024;80:1623–701.
 13. Greenberg DS, Nonnenmacher M, Macke JH. Automatic posterior transformation for likelihood-free inference. In: *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR. 2019;97:2404–2414.
 14. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. *Proc Natl Acad Sci U S A*. 2020;117(48):30055–62.
 15. Tejero-Cantero A, Boelts J, Deistler M, Lueckmann J-M, Durkan C, Gonçalves PJ, et al. sbi: A toolkit for simulation-based inference. *J Open Source Softw*. 2020;5(52):2505.
 16. Deleu T, Bengio Y. The effects of negative adaptation in model-agnostic meta-learning. *arXiv:1812.02159 [Preprint]*. 2018.
 17. Xu S, Zhu M. Efficient Safe Meta-Reinforcement Learning: Provable Near-Optimality and Anytime Safety. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2025.

Figures

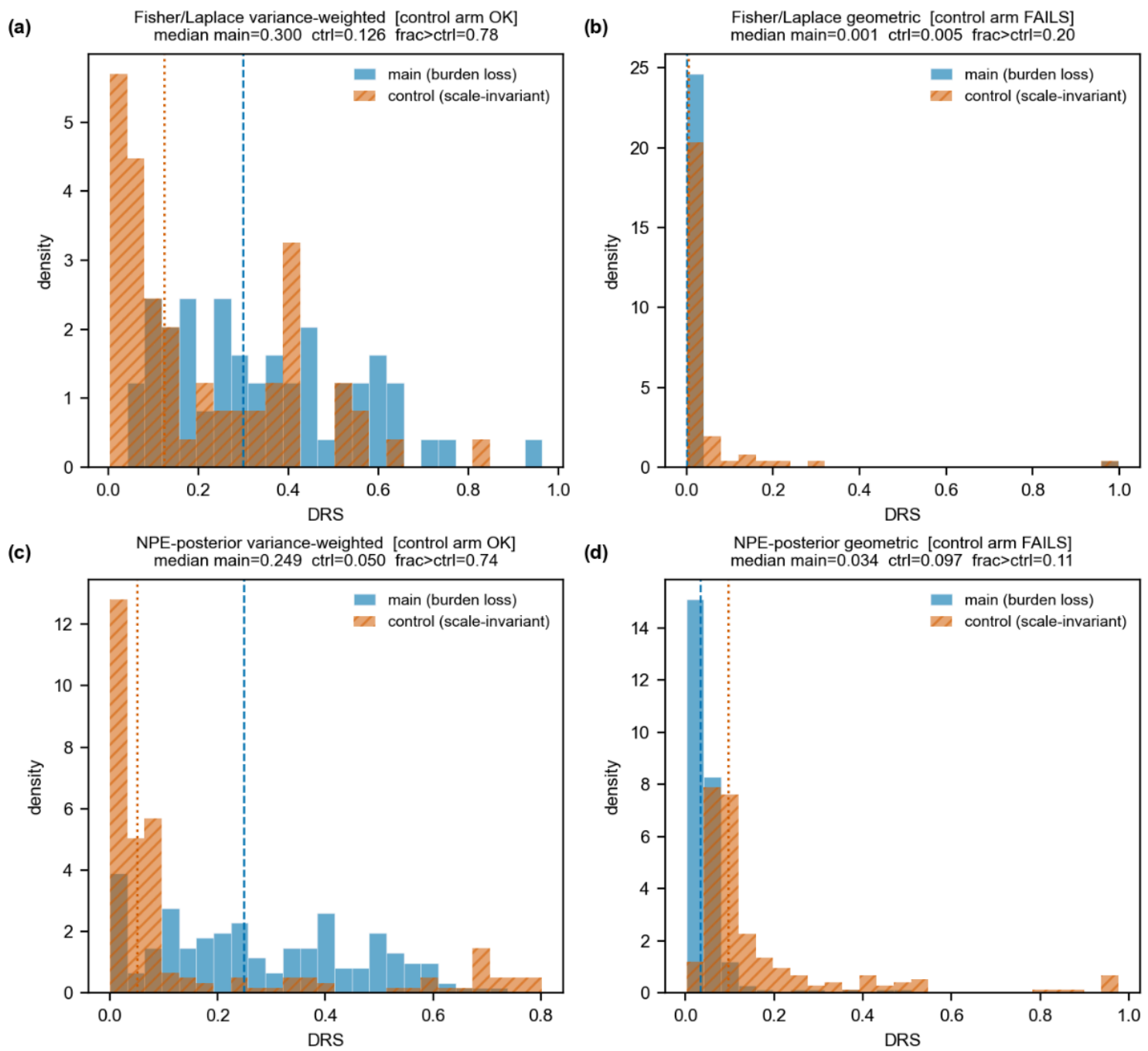


Figure 1

Variance-weighted DRS separates the arms; geometric DRS does not

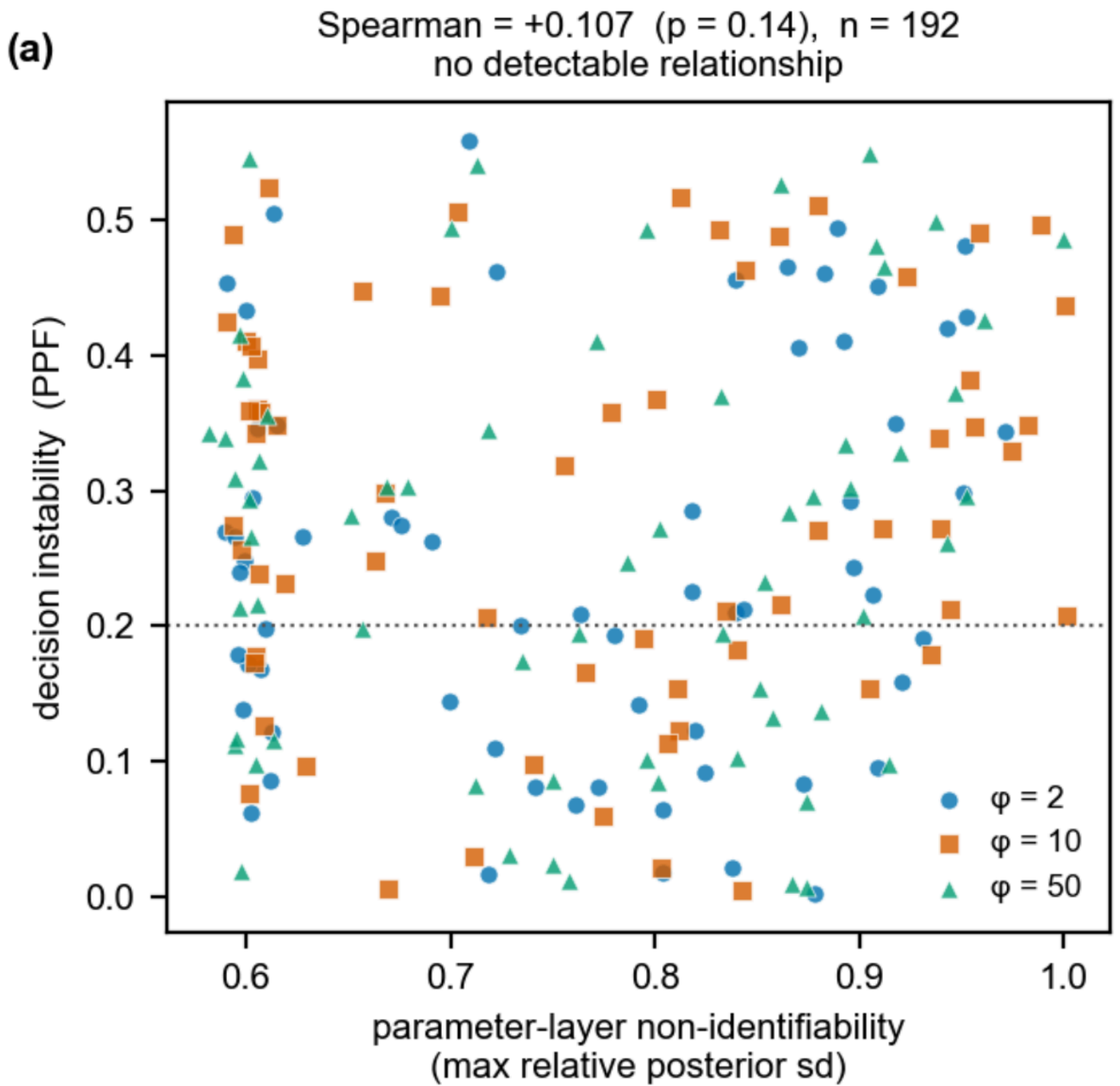


Figure 2

Parameter-level non-identifiability does not predict decision instability

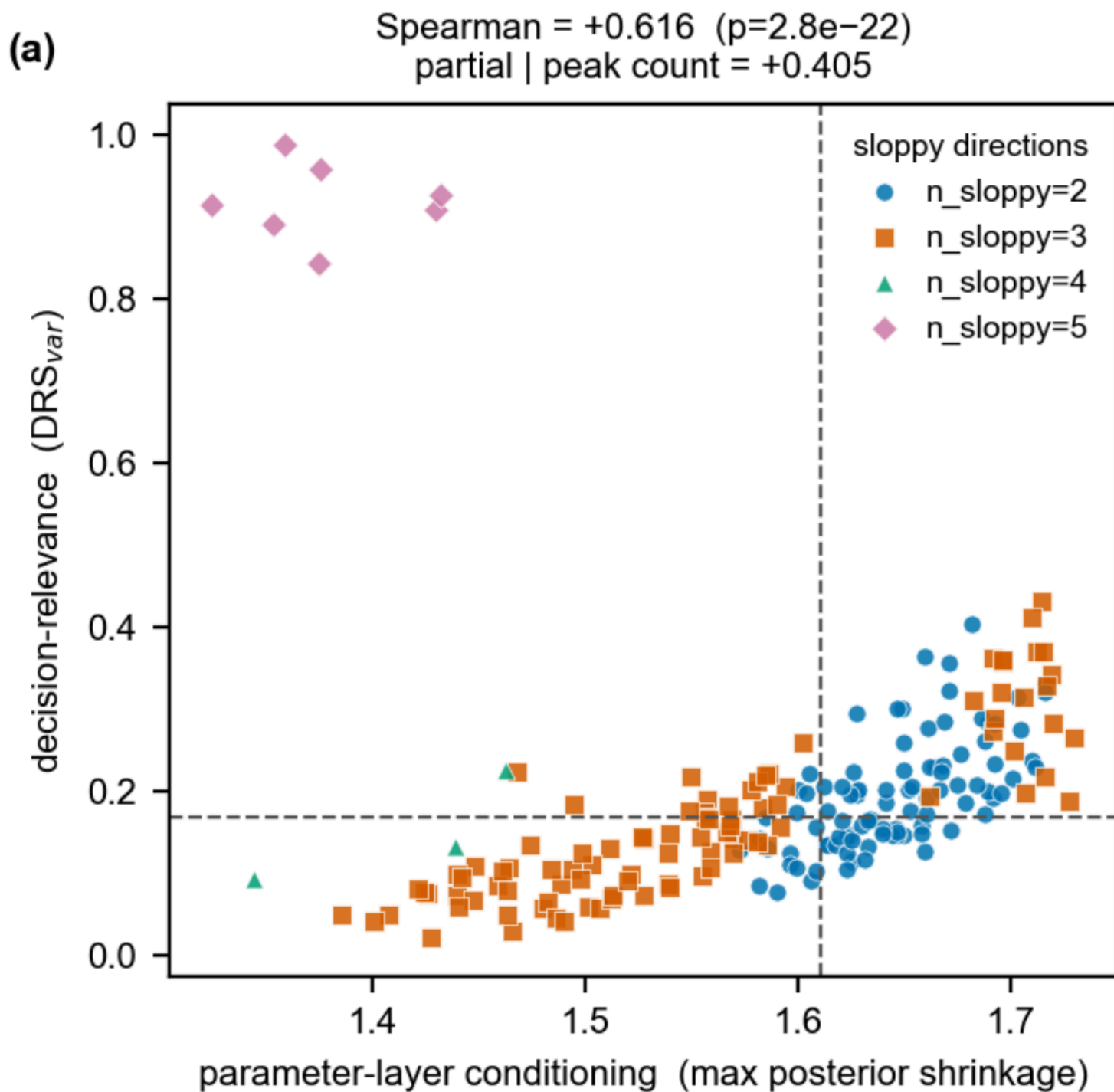


Figure 3

Conditioning informs but does not determine decision-relevance

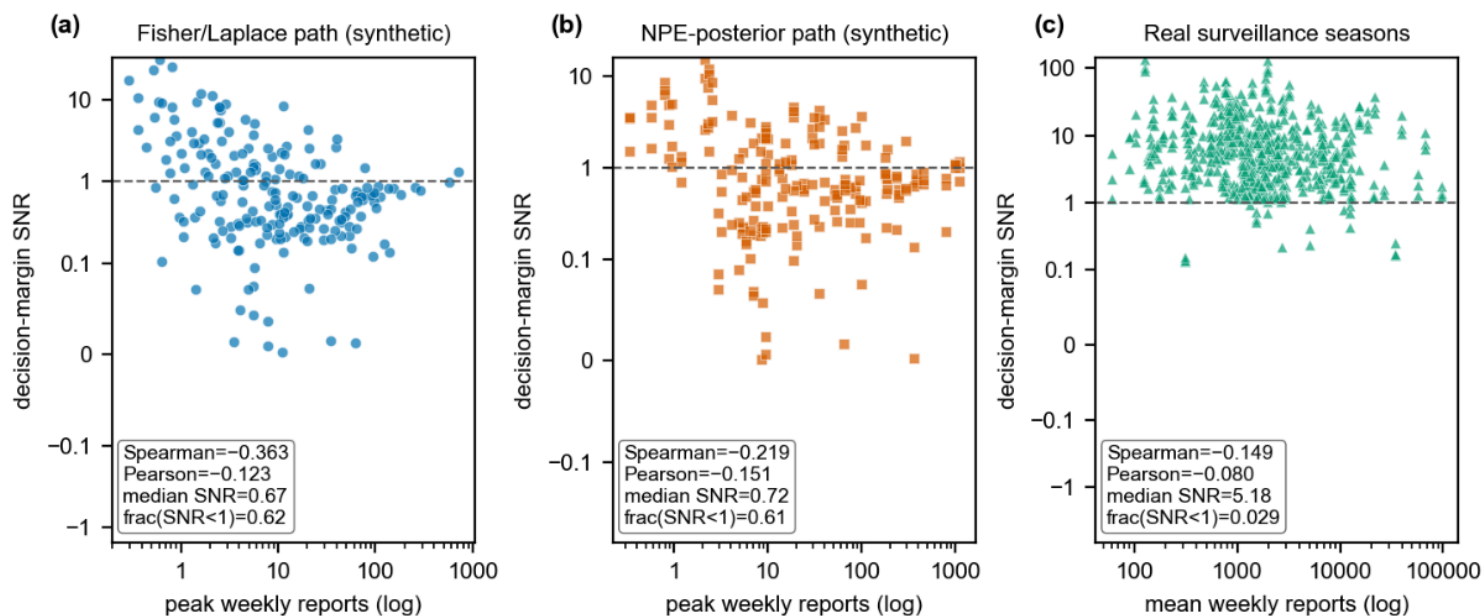


Figure 4

Margin signal-to-noise falls as case counts rise, in synthetic and real data

Supplementary Files

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

- [Additionalfile1.tif](#)