MULTI-METHOD INFERENCE FOR CRITICAL THRESHOLDS FROM SPARSE OBSERVATIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Critical thresholds are often latent functionals rather than directly observed breakpoints. Under sparse observations, structurally different estimators can return similar numerical locations while relying on incompatible inverse mappings or targeting different quantities. We introduce a qualification-first framework that fixes the scientific target before model choice, profiles each method's native inferential object and failure modes, and uses known-truth validation to assign one of four roles: target-responsive, calibrated within a stated validation range, native diagnostic, or abstention. In a compact mechanistically aligned but observation-model-mismatched experiment at truths 0.90, 0.94, and 0.97, a hierarchical Bayesian model (HBM) responds monotonically but compresses near the upper end and is positively biased at the two lower truths (RMSE 0.032, 0.026, and 0.006); change-point detection (CPD) does not track the mechanistic target and remains native-only. Applied to 10,153 U.S. county-years from 2,212 counties during 2018–2023, the HBM posterior median unconstrained coverage crossing is 0.991, with 31.8% of posterior draws implying no feasible crossing at coverage ≤ 1 under $\eta = 0.97$. CPD returns a nonsignificant native breakpoint of 0.992 ($p = 0.366$), while a biologically informed neural network (BINN) abstains because profile support spans 0.88–0.99. The framework therefore assigns different inferential roles and performs no fusion: numerical proximity alone is not evidence that heterogeneous methods recover the same latent threshold.

1 INTRODUCTION

Many scientific questions concern a boundary separating qualitatively different latent regimes. Epidemic transmission becomes self-sustaining when effective reproduction crosses one; ecosystems can undergo regime shifts after environmental stress passes a tipping point; and engineered systems can fail beyond operating limits. In each case, the object of interest is not simply a prediction but a critical control value at which the latent system changes qualitative state.

Classical threshold regression, level-set estimation, and change-point detection address related boundary problems when the target is defined through an observable statistical representation (Hansen, 2000; Gotovos et al., 2013; Truong et al., 2020). We consider a different inverse problem in which the scientific transition itself is latent, while observations arise through nuisance mechanisms such as stochastic opportunity, measurement error, reporting, and contextual heterogeneity. The threshold is therefore a functional of an underlying system rather than necessarily an observable breakpoint (Evans and Stark, 2001; Stuart, 2010).

This distinction becomes consequential when several inferential paradigms are applied to the same data. A mechanistic Bayesian model may infer the transition through a generative model; an ordered change-point procedure may locate an abrupt feature of the observed response; and a constrained neural model may encode the transition relation while learning nuisance structure flexibly. Similar numerical outputs do not establish that these methods estimate the same quantity. Conversely, disagreement need not imply that one method is merely "wrong": it can reveal weak practical identification, model dependence, or a mismatch between native estimands.

We therefore propose a *qualification-first* multi-method framework. Its central rule is that evidence combination comes after three checks: the scientific target is fixed independently of the estimators; each method's native estimand, assumptions, uncertainty representation, and expected failures are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

explicit; and known-truth experiments determine which inferential claim the method is qualified to make. A pathway may emerge as target-responsive, calibrated within a prespecified validation range, a native diagnostic, or an abstaining method. Fusion is eligible only when at least two methods are calibrated for the same target in the relevant validation regime. Figure 1 summarizes this decision layer.

This framing is related to triangulation, which is most informative when important sources of bias differ (Lawlor et al., 2016), but it is stricter than informal agreement across analyses. It also differs from ordinary ensemble prediction: the goal is not to average models for predictive accuracy, but to decide whether heterogeneous inverse mappings support the same latent functional. Identification of the target functional may be possible even when all latent parameters are not identifiable (Raue et al., 2009; Ovchinnikov et al., 2021).

Our contributions are fourfold. First, we formalize sparse critical-threshold inference around a reference-population functional. Second, we define an operational qualification procedure that turns native estimands, assumptions, diagnostics, and known-truth validation into explicit graded inferential roles. Third, we instantiate three structurally distinct pathways—HBM, CPD, and BINN—without assuming in advance that they can be harmonized or fused. Fourth, we demonstrate role assignment and abstention on county-level measles surveillance, where the three methods produce numerically nearby locations but support different claims. The contribution is not another threshold estimator; it is an inferential decision layer that prevents numerical agreement among heterogeneous methods from being mistaken for evidence of estimand equivalence. Section 3 defines the target, Section 4 the qualification procedure, Sections 5–6 the estimators and controlled validation, Section 7 the measles case study, and Sections 8–9 the implications and limitations.

2 Related work

Thresholds and inverse inference. Threshold regression estimates unknown regime boundaries (Hansen, 2000); level-set estimation recovers boundaries of super- or sub-level regions (Gotovos et al., 2013); and change-point methods detect distributional changes in ordered sequences (Truong et al., 2020). Our target depends instead on a latent system parameterization observed only indirectly, placing the problem in the broader class of inverse inference (Evans and Stark, 2001; Stuart, 2010) and motivating functional rather than full-parameter identifiability (Raue et al., 2009; Ovchinnikov et al., 2021).

Multiple inferential pathways. Triangulation argues that convergence is most probative when approaches have different major sources of bias (Lawlor et al., 2016). But methods fit to the same observations can share preprocessing, covariates, and structural misspecification. Forecast-combination theory likewise emphasizes dependence among component errors (Wang et al., 2023). We operationalize this caution at the estimand level: a numerical output is not eligible for common-target comparison until its relation to that target has been justified and empirically qualified.

Validation, practical identification, and abstention. Known-truth simulation is a standard way to test whether an inference procedure can recover quantities generated under controlled conditions; simulation-based calibration provides a particularly explicit Bayesian example (Talts et al., 2018). Profile-based diagnostics distinguish practical localization from mere optimizer convergence (Raue et al., 2009). Our use is narrower than either literature: controlled experiments determine which scientific claim each heterogeneous estimator may support, including an explicit option to withhold a threshold when the available data do not localize it.

Scientifically informed learning. Science-informed learning can regularize inverse inference by embedding mechanistic structure in flexible models (Raissi et al., 2019; Karniadakis et al., 2021). Such constraints also create an evidential issue: behavior imposed by architecture cannot later be counted as independent evidence that the data discovered the same relationship. We therefore treat structural constraints as part of the method profile and require separate identification diagnostics for learned thresholds.

3 CRITICAL-THRESHOLD INFERENCE AS AN INVERSE PROBLEM

3.1 OBSERVED SYSTEM AND SCIENTIFIC TARGET

Conditioning on the observed design variables, let

$$\mathcal{D}_n = \{(u_i, x_i, g_i, y_i)\}_{i=1}^n, \qquad Y_i \mid u_i, x_i, g_i \sim P_{\theta^*,\nu^*}(\cdot \mid u_i, x_i, g_i), \tag{1}$$

where $u_i \in \mathcal{U} \subseteq \mathbb{R}$ is a scalar control, x_i contextual covariates, g_i a grouping variable used when a method models structured dependence, y_i the observed response, θ^* transition-relevant latent structure, and ν^* nuisance structure such as opportunity, reporting, or measurement processes. Let $r_\theta(u, x)$ be a latent transition index with critical value c. Fix a reference distribution G over context X and define

$$\bar{r}_G(u;\theta) = \mathbb{E}_{X\sim G}[r_\theta(u, X)], \qquad \tau_G(\theta) = \inf\{u \in \mathcal{U} : \bar{r}_G(u;\theta) \le c\}. \tag{2}$$

The averaging operator is a reference-standardization choice rather than a universal definition of a critical threshold; other scientific problems may require a different target functional. The reference G is part of this estimand, so changing the standardized population changes the target even if the fitted model is unchanged.

3.2 IDENTIFIABILITY AND SPARSE INFORMATION

The threshold functional is identifiable over an admissible model class if

$$P_{\theta_1,\nu_1} = P_{\theta_2,\nu_2} \quad \Longrightarrow \quad \tau_G(\theta_1) = \tau_G(\theta_2). \tag{3}$$

This is weaker than identifying every component of θ. Sparse data make practical identification difficult even when structural identification is plausible. For a neighborhood width $\delta > 0$, local support can be summarized by $N_\delta(\tau_G) = \sum_i \mathbf{1}(|u_i - \tau_G| \le \delta)$; more generally, the number of observations that constrain the transitioned regime may be far smaller than n. As a preview of the application, the primary measles panel contains more than ten thousand county-years but only 128 positive case observations.

4 QUALIFICATION-FIRST MULTI-METHOD FRAMEWORK

4.1 METHOD PROFILES AND NATIVE ESTIMANDS

For method m, define the profile

$$\mathcal{P}_m = (I_m, \psi_m, H_m, A_m, F_m, U_m, Z_m), \tag{4}$$

where I_m is the inferential pathway, ψ_m the native estimand, H_m a proposed map to the common target, A_m assumptions/inductive biases, F_m expected failure modes, U_m uncertainty representation, and Z_m diagnostics. Fitting produces $(\widehat{\psi}_m, \widehat{U}_m, z_m)$. Crucially, H_m is not assumed to exist. A method can remain scientifically useful on its native scale even when no defensible map to τ_G is available.

4.2 KNOWN-TRUTH VALIDATION AND ROLE ASSIGNMENT

For controlled setting s with known truth τ_s^* and nuisance condition ξ_s, generate replicates $\mathcal{D}_s^{(b)}$ and fit the same estimator implementation used on observational data. The framework uses graded roles rather than an undefined binary pass/fail gate. A method is *target-responsive* when a scientifically meaningful map H_m exists, its mapped output changes with τ_s^* in the expected direction over the tested grid, and method-specific diagnostics are satisfactory. It is *calibrated within the validation range* only when, in addition, its error meets a prespecified claim-specific tolerance throughout that range. In this study we use $|\widehat{\tau} - \tau^*| \le 0.03$ as the compact point-recovery tolerance; with three replicates per cell we treat this only as a screening rule, not a precise calibration estimate. A method whose native quantity is scientifically interpretable but lacks a validated map to τ_G is a *native diagnostic*. A method whose own practical-identification criterion fails *abstains*.

These roles constrain downstream claims. Target-responsive methods may be reported as model-conditional target-aligned estimates, but only methods calibrated over the relevant regime are eligible for common-target fusion. Failure of a condition changes the role rather than being silently removed.

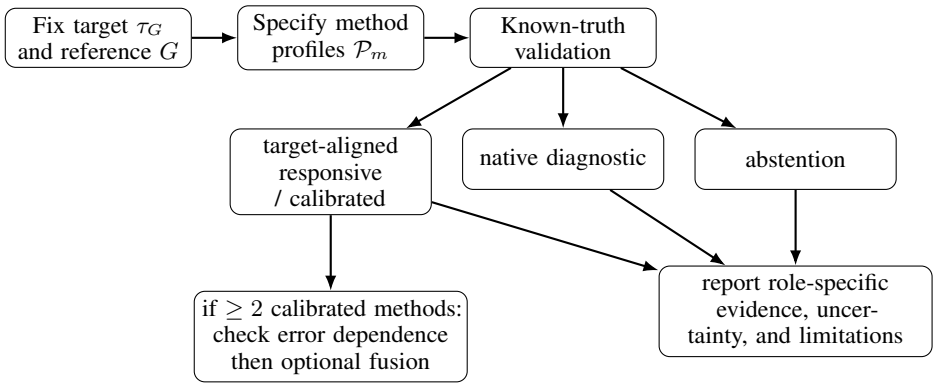

Figure 1: Qualification-first inference workflow. Known-truth validation determines the inferential role of each method. Only calibrated target estimators feed a potential fusion route; native diagnostics and abstentions remain role-specific evidence.

Table 1: Method profiles and roles after qualification. "Native" does not mean uninformative; it means that the reported quantity is not relabeled as the common mechanistic threshold.

Method	Inferential route / native estimand	Qualification finding	Role in the measles analysis
HBM	generative model $\to R_0(x) \to \tau_G$	positive recovery bias at lower truths despite good MCMC	target-responsive mechanistic pathway; not calibrated across the full tested grid and not fusion-eligible
CPD	ordered coverage-response structure $\to$ breakpoint κ_C	breakpoint does not track mechanistic truth stably	native empirical-transition diagnostic; harmonization disabled
BINN	fixed-τ profile with learned nuisance structure	flat profile / threshold nonidentification	profile-based target pathway; abstains because practical-identification support is wide

4.3 Agreement, disagreement, and fusion eligibility

Let $\mathcal{E}$ be the set of methods calibrated for τ_G under the same validation regime. Only estimates from $m \in \mathcal{E}$ are eligible for common-target agreement and potential fusion. If $|\mathcal{E}| \geq 2$, controlled replicates can be used to estimate method-specific errors and error dependence before choosing a fusion rule; defining and validating such a rule is outside the scope of the present case study. If $|\mathcal{E}| < 2$, numerical proximity among outputs is descriptive only. The framework therefore determines eligibility for fusion rather than requiring fusion as a mandatory final step.

Table 1 shows that the three methods contribute different inferential objects after validation; numerical proximity is therefore not interpreted as three votes on one parameter.

5 Estimator instantiation

5.1 Hierarchical Bayesian model

The HBM represents transmission as

$$R_0(x) = \bar{R}_0 h(x), \qquad \mathbb{E}_{G_{\text{train}}}[h(X)] = 1, \tag{5}$$

where population-level density and age-share effects and hierarchical state/county effects determine relative heterogeneity, while $\bar{R}_0$ controls the arithmetic mean scale. Vaccination enters through

$$R_{\text{e}}(v, x) = R_0(x)(1 - \eta v). \tag{6}$$

A hurdle likelihood separates whether any cases are observed from burden conditional on a positive outcome. The occurrence component uses density, age share, population, prior cases, and

year as contextual proxies for observation/importation opportunity, but deliberately excludes vaccination/transmission. Consequently, the dominant zero/nonzero pattern does not directly identify the transmission threshold through the Bernoulli component; R_{e} informs threshold inference only through positive burden. Among positive observations, $Y-1$ follows a negative-binomial model with mean increasing in a smooth positive-part transformation of $R_{\mathrm{e}}-1$. This separation is a substantive model assumption intended to distinguish exposure/opportunity from transmission conditional on observed cases. Posterior draws are standardized over fixed G and transformed to the algebraic crossing described below. Full likelihood and derivation appear in Appendix C.

5.2 Change-point detection

CPD orders county-years by observed coverage and seeks a single split maximizing reduction in an outcome-based segmentation cost. The primary signal uses random Fourier features approximating an RBF kernel on $\log(1+Y)$; tied coverage values are never split. County-block bootstrap represents breakpoint uncertainty, and significance is assessed by permuting coverage trajectories among compatible county blocks. Its native estimand is the empirical breakpoint κ_C. Controlled validation did not support a defensible mapping from κ_C to the mechanistic τ_G; we therefore interpret CPD only on its native breakpoint scale (Appendix D).

5.3 Biologically informed neural network

The BINN does not jointly optimize a free threshold with neural nuisance parameters. Instead it profiles a prespecified grid of candidate τ values. For each candidate, nuisance parameters are trained with τ fixed and evaluated in held-out county folds. The selected point minimizes mean held-out loss; a one-standard-error near-optimal support set quantifies identification. A threshold is reported only if the minimum is interior and support width is at most 0.04. This design was adopted after direct threshold optimization proved initialization-sensitive (Appendix F).

6 Controlled known-truth validation

6.1 Design

We evaluate the HBM and CPD at three known mechanistic thresholds to test target responsiveness and characterize CPD on its native scale. For each $\tau^* \in \{0.90, 0.94, 0.97\}$, three evaluation replicates contain 500 counties observed for eight years. The generator targets a 0.987 zero fraction and uses the same mechanistic relation as equation 6; latent truth columns are saved separately and never passed to estimators. Its observation model is an opportunity indicator multiplied by an ordinary negative-binomial count, not the HBM hurdle likelihood. The experiment is therefore mechanistically aligned but observation-model-mismatched. With three replicates per cell it is a compact validation screen rather than a high-precision Monte Carlo benchmark.

6.2 Recovery results

Table 2 summarizes the controlled validation. HBM estimates move monotonically with the known truth, from 0.932 at $\tau^* = 0.90$ to 0.974 at $\tau^* = 0.97$, and all nine HBM fits pass the prespecified sampler and posterior-predictive diagnostics. Responsiveness is nevertheless compressed near the upper end: the mean changes only from 0.966 at truth 0.94 to 0.974 at truth 0.97. Point recovery is also nonuniform: the HBM is positively biased by 0.032 and 0.026 at the two lower truths, whereas RMSE falls to 0.006 at 0.97. All nine reported credible intervals contain the generating truth, but three replicates per cell are insufficient to interpret this as a calibrated coverage estimate. CPD's native mean moves from 0.868 to 0.850 to 0.829 as the mechanistic truth increases, reinforcing the decision not to harmonize its breakpoint to τ_G. Native breakpoint detection and mechanistic-threshold validity are distinct: a detectable empirical split need not track the latent target. Its discrepancy statistic is therefore descriptive only and must not be read as common-target RMSE. Figure 2 shows the wide HBM intervals and weak separation between the 0.94 and 0.97 cells alongside CPD's opposite-direction movement.

Table 2: Compact known-truth validation study (three replicates per threshold). HBM is evaluated on the common-target scale. CPD remains on its native breakpoint scale, so its final column is a descriptive RMS discrepancy from the varied mechanistic truth, not a validated common-target RMSE.

Truth	HBM mean	HBM bias	HBM RMSE	diag. pass	CPD native mean	native–truth RMS
0.90	0.932	+0.032	0.032	3/3	0.868	0.050
0.94	0.966	+0.026	0.026	3/3	0.850	0.092
0.97	0.974	+0.004	0.006	3/3	0.829	0.142

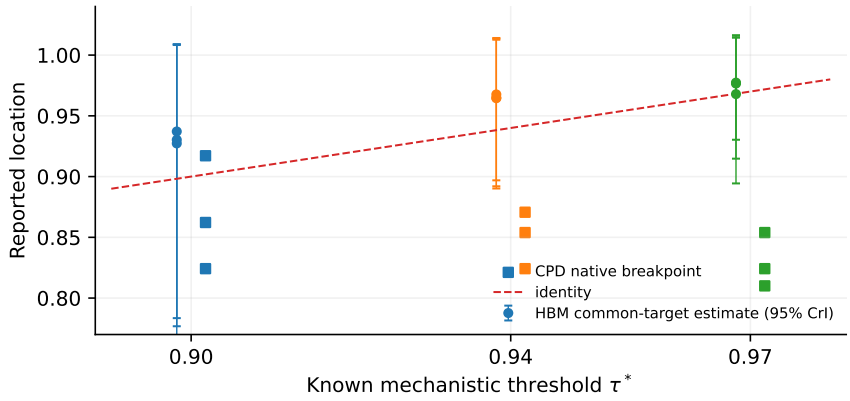

Figure 2: Per-replicate controlled results. HBM circles include posterior 95% credible intervals and are compared with the identity line on the common-target scale. CPD squares are native breakpoints shown only as a responsiveness diagnostic; their opposite/downward movement across the tested truths is evidence against harmonization.

7 Measles Case Study

The case study tests how the framework assigns inferential roles when observations are sparse and the methods return superficially similar locations. We ask what the HBM implies for its model-defined reference-standardized crossing, whether CPD provides a statistically supported empirical breakpoint, and whether the BINN profile localizes a threshold strongly enough to report one. The goal is role assignment and uncertainty characterization, not forced numerical consensus.

7.1 Target, Data, and Cohort Construction

Let $v \in [0, 1]$ denote observed MMR coverage and let η denote the multiplicative protection parameter used by the model bridge in equation 6. For fixed reference distribution G, define the unconstrained algebraic crossing

$$\tau_G^{\text{raw}} = \frac{1 - 1/\mathbb{E}_{X\sim G}[R_0(X)]}{\eta}, \tag{7}$$

and the feasibility indicator $F_G = \mathbf{1}(0 \le \tau_G^{\text{raw}} \le 1)$. If $F_G = 0$ because $\tau_G^{\text{raw}} > 1$, the physical coverage domain contains no modeled crossing; values above one are not interpreted as vaccination coverages. We set $\eta = 0.97$ as a model assumption motivated by the approximately 97% effectiveness reported for two MMR doses by the U.S. CDC (CDC, 2026), while recognizing that clinical effectiveness is not identical to a transmission-blocking parameter. Sensitivity to η is therefore an unresolved model sensitivity rather than an empirically estimated uncertainty component. Measles R_0 varies substantially across settings (Guerra et al., 2017), and the classical herd-immunity relation is itself an idealization (Fine et al., 2011).

Figure 3 distinguishes the broad source union, analytic cohort, and fixed reference distribution. The canonical source-union panel contains 25,398 county-years across 2016–2024. The primary threshold analysis is stricter: eligibility requires contemporaneous coverage and covariate support, yielding

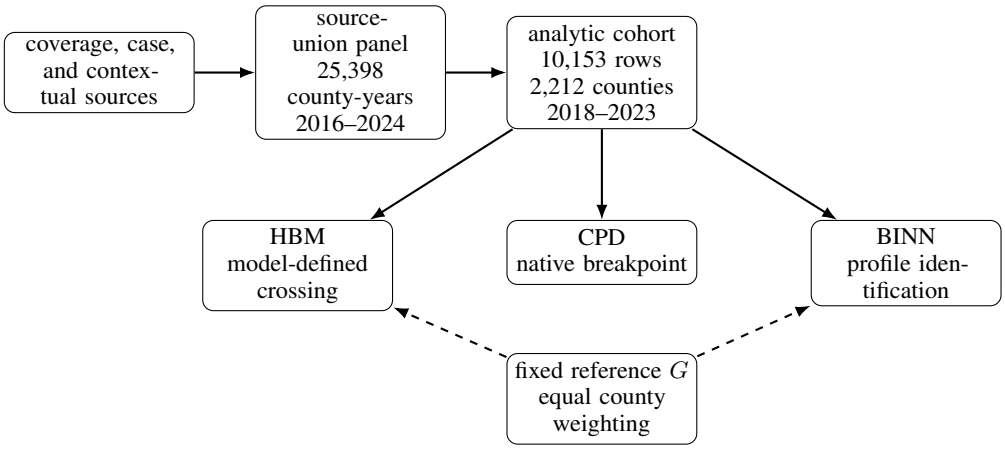

Figure 3: Case-study construction. The source-union panel is distinct from the analytic cohort. HBM and BINN use the fixed reference G (dashed links), whereas CPD remains on its native ordered-response scale.

10,153 county-years from 2,212 counties in 2018–2023. Only 128 county-years have reported positive cases, giving observed zero fraction 0.98739. Coverage-source conflicts are retained with provenance rather than averaged, and the primary cohort uses no interpolation, winsorization, or forward filling. Reported positives, inferred zeros, and reconciled complete state-years remain distinguishable in the data build (Appendix G). The reference G is the included 2,212-county cohort with equal county weighting, not a U.S. population-weighted national reference.

7.2 Historical results

HBM yields a posterior median reference-standardized mean $R_0 = 25.87$. This is above the often-cited 12–18 range for measles, although published estimates vary widely across settings (Guerra et al., 2017); here it should be interpreted as a model- and cohort-conditional scale that may reflect sparse positive counts, parameterization, prior information, and observation-model assumptions rather than as a national R_0. The posterior median unconstrained crossing is $\widehat{\tau}_G^{\mathrm{raw}} = 0.9911$, with a 95% posterior interval $[0.9275, 1.0188]$ on the unconstrained algebraic scale. Sampling diagnostics are strong: crossing $\widehat{R} = 1.0019$, maximum core $\widehat{R} = 1.0059$, bulk ESS 2,389, minimum BFMI 0.903, zero divergences, and passing posterior predictive checks. Under $\eta = 0.97$, 31.8% of posterior draws imply $\tau_G^{\mathrm{raw}} > 1$, i.e., no feasible crossing at coverage ≤ 1 under the fitted model; the corresponding feasible-crossing probability is 68.2%.

CPD returns native breakpoint $\widehat{\kappa}_C = 0.99186$ with county-block bootstrap interval $[0.8596, 0.99186]$. All 500 bootstrap fits succeed, but the permutation test is nonsignificant ($p = 0.366$). The breakpoint is therefore reported as an observable transition diagnostic, not as a mechanistic herd-immunity estimate.

BINN's cross-validated profile has its minimum at 0.98 and all three refits return the same selected candidate, but near-optimal support spans $[0.88, 0.99]$ (width 0.11), exceeding the prespecified 0.04 identification criterion. The model therefore abstains from a threshold estimate. Table 3 summarizes the three outputs and, critically, their different inferential status. Figure 4 places them on a shared horizontal display scale only to show why visual proximity must not be confused with estimand equivalence. This separates optimization stability from inferential identification: a stable best candidate can coexist with insufficient curvature to localize the threshold.

8 Discussion

The case study demonstrates why a qualification step is necessary even when several methods appear to "agree." HBM, CPD, and BINN all place their best-fitting location near the upper end of observed coverage, yet the evidential content differs sharply. HBM supplies a computationally well-mixed

Table 3: Historical method outputs. Intervals have different meanings and are intentionally not pooled.

Method	Reported location	Interval/support	Framework interpretation
HBM	raw crossing 0.9911	raw 95% CrI $[0.9275, 1.0188]$	model-defined target; responsive but not fully calibrated; 31.8% of draws imply no feasible crossing at $v \leq 1$
CPD	$\widehat{\kappa}_C = 0.9919$	bootstrap $[0.8596, 0.9919]$	native breakpoint; $p = 0.366$; no validated map to τ_G
BINN	no threshold reported	profile support $[0.88, 0.99]$	profile minimum 0.98, but practical identification fails

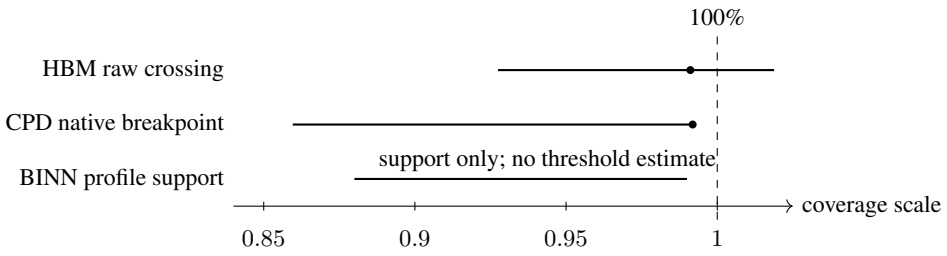

Figure 4: Historical outputs on a common horizontal display scale, but not a common inferential scale. HBM shows an interval for the unconstrained algebraic crossing, CPD a bootstrap interval for its native breakpoint, and BINN only a profile support set because it abstains. The dashed line marks the feasibility boundary at 100% coverage.

posterior for a mechanistic reference-population threshold, and the controlled validation shows that this estimator responds to changes in the generating threshold; however, it also shows positive bias at truths 0.90 and 0.94. CPD finds no statistically supported empirical transition in the historical data and, across the tested threshold sweep, its native breakpoint does not track the mechanistic truth. BINN's optimization is stable while its threshold profile remains too flat for identification. Treating these three locations as votes would erase exactly the information supplied by their different inferential structures.

The framework therefore changes the interpretation of disagreement and agreement. A failed harmonization is not repaired by averaging. A nonsignificant native breakpoint can still show where an ordered-response objective is maximized, but it does not become a latent threshold. A learned model can remain useful for prediction while abstaining from a structural parameter. Conversely, a computationally excellent posterior is not automatically scientifically calibrated. Here, all nine controlled HBM fits pass diagnostics and all nine credible intervals include the known truth, yet point recovery is compressed upward at the lower two truths. Moreover, the historical estimate 0.991 lies above the highest controlled truth 0.97. We therefore do not use the controlled biases to post-hoc correct the historical posterior and do not describe the historical point estimate as externally calibrated.

The posterior probability of no feasible crossing is also substantively important. It should not be converted into a deceptively precise "100%" estimate. The target is conditional on the chosen G, the simplified bridge $R_{\mathrm{e}} = R_0(1 - \eta v)$, ecological county-year aggregation, and the hurdle observation model. The result is therefore a model-conditional unconstrained crossing plus a feasibility probability, not a universal deterministic herd-immunity constant.

Limitations. The historical data are extremely sparse: only 128 positive county-years enter the primary analysis, and the primary HBM specification deliberately prevents the much larger zero/nonzero mass from directly identifying transmission through the occurrence component. County-level aggregation cannot represent within-county clustering of susceptibility, dose history, waning, contact-network heterogeneity, importation/local-transmission distinctions, or all differences in ascertainment. Coverage measurements are heterogeneous across sources, and the primary zero policy includes assumed as well as reconciled zeros. The protection bridge fixes $\eta = 0.97$; prior sensitivity for the transmission

scale, quantitative η sensitivity, zero-policy sensitivity, and alternative reference weighting were not available in the supplied analysis outputs and are therefore not claimed here. The controlled study uses only three replicates at each of three truths and an observation model that is not likelihood-matched to the HBM; it establishes directional responsiveness, not high-precision calibration or robustness to misspecification. Its grid stops at $\tau^* = 0.97$, so the historical median raw crossing 0.991 is partly extrapolative. Finally, the BINN profile grid is designed as a practical-identification screen rather than a high-resolution confidence procedure.

Future work. The immediate extensions are likelihood-matched HBM validation, larger nuisance and misspecification grids, prior/η/reference-weight/zero-policy sensitivities, denser support checks near the upper coverage boundary, and a fusion experiment only if at least two methods achieve calibrated target status over the same regime.

Broader methodological implication. Multi-method inference is useful only when methodological heterogeneity is made part of the inferential procedure. The framework is therefore less an ensemble recipe than a disciplined gate: define the target, characterize native objects, test recoverability, and preserve ambiguity when the evidence does not justify consensus. This logic can transfer to other latent tipping-point problems in which controlled generators and method-specific failure modes can be specified transparently.

9 Conclusion

We formulated sparse critical-threshold inference as estimation of an estimator-independent latent functional and proposed a qualification-first multi-method procedure. In the measles analysis, HBM is the only method that directly targets the model-defined reference crossing and shows monotone but compressed responsiveness under a mechanistically aligned, observation-model-mismatched generator; CPD remains a nonsignificant native breakpoint with no validated map to the latent target; and BINN abstains because its profile is not practically identified. No multi-method fusion is performed. Numerical proximity is not evidence of estimand equivalence, and validation limits should constrain rather than silently justify extrapolation.

AI use statement

AI tools were used for retrieval and discovery of related scientific papers, to assist in formalizing the framework and proving mathematical claims, and to help properly format the paper according to the given style specifications. The authors thoroughly reviewed all AI-assisted work and take full responsibility for the final content, claims, code, and artifacts.

Ethics statement

This study uses aggregate county-level vaccination, demographic, and measles surveillance data and no individual-level records or human-subject intervention. Relevant concerns are ecological interpretation, unequal surveillance completeness, uncertainty in vaccination measurements, and the risk of stigmatizing communities using model-derived quantities. We therefore avoid deterministic labels for counties, do not interpret aggregate geographic estimates as labels of individual communities as "safe", "unsafe", or "noncompliant", and distinguish a model-conditional transmission threshold from a guarantee about outbreak occurrence.

Reproducibility statement

Appendices A–H provide the derivations, simulator specification, estimator definitions, qualification history, data-construction rules, diagnostics, and reported software versions. The anonymous supplementary archive contains the machine-readable controlled-validation outputs used for Table 2 and Figure 2, together with the script that regenerates their manuscript summary and figure. Reconstruction of the observational panel additionally requires source-specific provenance and data-build inputs that are not redistributed in the submitted archive.

REFERENCES

S. N. Evans and P. B. Stark. Inverse problems as statistics. Technical Report 609, University of California, Berkeley, 2001.

P. Fine, K. Eames, and D. L. Heymann. Herd immunity: A rough guide. *Clinical Infectious Diseases*, 52(7):911–916, 2011.

A. Gotovos, N. Casati, G. Hitz, and A. Krause. Active learning for level set estimation. In *IJCAI*, 2013.

F. M. Guerra et al. The basic reproduction number (R_0) of measles: A systematic review. *The Lancet Infectious Diseases*, 17(12):e420–e428, 2017.

B. E. Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

G. E. Karniadakis et al. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

D. A. Lawlor, K. Tilling, and G. Davey Smith. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, 45(6):1866–1886, 2016.

A. Ovchinnikov, A. Pillay, G. Pogudin, and T. Scanlon. Computing all identifiable functions of parameters for ODE models. *Systems & Control Letters*, 157:105030, 2021.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear PDEs. *Journal of Computational Physics*, 378:686–707, 2019.

A. Raue et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. arXiv:1804.06788, 2018.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

X. Wang, R. J. Hyndman, F. Li, and Y. Kang. Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 39(4):1518–1547, 2023.

U.S. Centers for Disease Control and Prevention. Measles Vaccine Recommendations: MMR efficacy against measles. Updated March 3, 2026. Accessed September 26, 2026.

A MATHEMATICAL DERIVATIONS AND INFERENTIAL CLAIMS

A.1 REFERENCE-STANDARDIZED THRESHOLD UNDER MULTIPLICATIVE PROTECTION

Assume the latent transition index takes the form

$$r_\theta(v, x) = a_\theta(x)(1 - \eta v), \qquad 0 < \eta \leq 1, \tag{8}$$

with critical value $c > 0$. Let $\bar{a}_G = \mathbb{E}_G[a_\theta(X)]$. Then

$$\bar{r}_G(v; \theta) = \bar{a}_G(1 - \eta v). \tag{9}$$

If $\bar{a}_G > 0$, the unconstrained crossing is

$$\tau_G^{\text{raw}} = \frac{1 - c/\bar{a}_G}{\eta}. \tag{10}$$

For measles, $a_\theta(x) = R_0(x)$ and $c = 1$. Define $F_G = \mathbf{1}(0 \leq \tau_G^{\text{raw}} \leq 1)$. The implementation lower-truncates negative algebraic crossings at zero but does not upper-clip values above one. A draw with $\tau_G^{\text{raw}} > 1$ therefore has $F_G = 0$: no feasible crossing exists on the physical coverage domain under the assumed η and fitted latent transmission scale. We retain the raw value only as an algebraic diagnostic, not as a vaccination coverage.

Proof. Because v is constant with respect to X, $\mathbb{E}_G[a_\theta(X)(1 - \eta v)] = (1 - \eta v)\bar{a}_G$. Solving $(1 - \eta v)\bar{a}_G = c$ yields equation 10. Monotonicity in v gives uniqueness whenever $\eta \bar{a}_G > 0$. □

A.2 WHY THE REFERENCE POPULATION IS PART OF THE ESTIMAND

Let G_1 and G_2 be two reference distributions with $\bar{a}_{G_1} \neq \bar{a}_{G_2}$. From equation 10,

$$\tau_{G_1}^{\text{raw}} - \tau_{G_2}^{\text{raw}} = \frac{c}{\eta}\left(\frac{1}{\bar{a}_{G_2}} - \frac{1}{\bar{a}_{G_1}}\right). \tag{11}$$

Hence changing G generally changes the threshold even with identical model parameters. This is why primary and sensitivity fits use the same saved reference table rather than allowing each analytic cohort to silently redefine its target.

A.3 FUNCTIONAL IDENTIFIABILITY

Proposition 1 (threshold identifiability). Let $\mathcal{M}$ be an admissible set of (θ, ν). If equation 3 holds for all pairs in $\mathcal{M}$, then there exists a well-defined functional T of the observable distribution such that $T(P_{\theta,\nu}) = \tau_G(\theta)$ on $\mathcal{M}$.

Proof. Define $T(P)$ by selecting any $(\theta, \nu) \in \mathcal{M}$ such that $P_{\theta,\nu} = P$ and setting $T(P) = \tau_G(\theta)$. equation 3 guarantees that this value does not depend on which representative is selected, so T is well defined. □

The proposition clarifies why full recovery of θ is unnecessary: only equivalence classes of latent parameters that change τ_G matter for this target.

A.4 NUMERICAL AGREEMENT DOES NOT IMPLY ESTIMAND EQUIVALENCE

Proposition 2 (agreement is not harmonization). Let two methods estimate native functionals $\psi_1(P)$ and $\psi_2(P)$. Equality $\widehat{\psi}_1(D) = \widehat{\psi}_2(D)$ on one realized dataset does not imply the existence of maps H_1, H_2 such that $H_1(\psi_1(P)) = H_2(\psi_2(P)) = \tau_G(P)$ throughout a validation class.

Proof by counterexample. Take $\psi_1(P) = \tau_G(P)$ and $\psi_2(P) = \tau_G(P) + b(P)$, where $b(P_0) = 0$ for the observed distribution P_0 but $b(P_1) \neq 0$ for another admissible P_1. The native values agree at P_0, yet the identity map cannot harmonize ψ_2 at P_1. A valid harmonization requires structure beyond observed numerical equality, exactly what controlled variation in τ_G and nuisance conditions is designed to test. □

B CONTROLLED-EXPERIMENT SPECIFICATION

B.1 GENERATOR

For each replicate, $n_c = 500$ counties are observed for eight years. County context is drawn once per county, then repeated over years. Let $X_c \in \mathbb{R}^2$, county effect a_c, and true coverage V_{ct}. Raw transmission is

$$R_{0,ct}^{\text{raw}} = \exp(0.15X_{c1} + 0.08X_{c2} + a_c). \tag{12}$$

With reference weights w_{ct} that assign equal total weight to counties and equal weight to their observed years, the simulator rescales

$$R_{0,ct} = \frac{R_{0,ct}^{\text{raw}}}{\sum_{ct} w_{ct} R_{0,ct}^{\text{raw}}} \cdot \frac{1}{1 - \eta\tau^*}. \tag{13}$$

Therefore $\sum w_{ct} R_{0,ct} = 1/(1 - \eta\tau^*)$, and substituting into equation 10 with $c = 1$ gives exactly $\tau_G = \tau^*$. By construction, the standardized latent threshold equals τ^*.

Effective transmission is $R_{e,ct} = R_{0,ct}(1 - \eta V_{ct})$. Under the primary mechanistic generator, amplification is the smooth positive part

$$a_{ct} = \frac{1}{8} \log\{1 + \exp(8(R_{e,ct} - 1))\}. \tag{14}$$

An opportunity process creates zeros. This opportunity-times-count observation model is not the HBM hurdle likelihood: observed positivity also depends on the ordinary negative-binomial probability of a nonzero count. Conditional burden follows a negative-binomial distribution with mean

$$\mu_{ct} = \exp\{-0.7 + 1.2a_{ct} + \log(N_{ct}/10^5)\}. \tag{15}$$

The opportunity intercept is solved numerically so that the expected nonzero fraction matches the requested target. The realized zero fraction is saved, not forced.

B.2 EVALUATION DESIGN

The compact study varies only $\tau^* \in \{0.90, 0.94, 0.97\}$, with three evaluation replicates per truth. All methods see only the simulated observable table; latent R_0, R_e, true coverage, and opportunity variables are persisted separately. HBM uses 1,200 tuning draws, 1,000 retained draws, four chains, and target acceptance 0.97. CPD uses 128 RBF features, 200 county-block bootstraps, and 199 permutations. The same random seed identifies paired data across methods within a replicate.

B.3 METRICS AND INTERPRETATION

For HBM we report bias, RMSE, median absolute error, $|\widehat{\tau} - \tau^*| \leq 0.03$, interval coverage when available, and sampler/PPC diagnostics. Across the nine HBM fits, all sampler/PPC diagnostics pass and all reported credible intervals contain the known truth. Point recovery is less uniform: at truths 0.90, 0.94, and 0.97, mean estimates are 0.932, 0.966, and 0.974 with biases $+0.032$, $+0.026$, and $+0.004$ and RMSEs 0.032, 0.026, and 0.006. Only one of three 0.90 replicates lies within 0.03 of truth, versus all three replicates in each of the other cells. With three replicates per cell these quantities are descriptive recovery checks, not precise calibration estimates. For CPD, because harmonization is disabled, we report discrepancy between the native breakpoint and mechanistic truth only as a descriptive responsiveness diagnostic; its native means (0.868, 0.850, 0.829) do not track the increasing mechanistic truth and cannot support a common-estimand error claim.

C HIERARCHICAL BAYESIAN MODEL

C.1 MEAN-NORMALIZED TRANSMISSION PARAMETERIZATION

For county-year i, let $x_i^{(R)}$ contain standardized log density and age share, with state $s(i)$ and county $c(i)$. The relative log multiplier is

$$\ell_i = \beta_R^\top x_i^{(R)} + \sigma_s z_{s(i)} + \sigma_c z_{c(i)}, \qquad z_s, z_c \sim \mathcal{N}(0, 1). \tag{16}$$

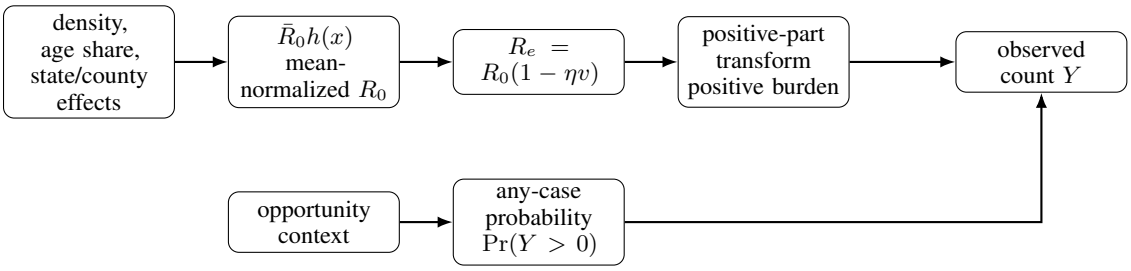

Figure 5: Primary HBM structure. Transmission enters the positive-burden branch through R_e; the occurrence branch is context-driven and excludes vaccination/transmission. The two branches jointly define the hurdle likelihood.

Writing $m_i = \exp(\ell_i)$ and $\bar{m} = \sum_i w_i m_i$, the model uses

$$R_{0,i} = \exp(\mu_R)\frac{m_i}{\bar{m}}. \tag{17}$$

Hence the arithmetic weighted mean over the training reference is exactly $\exp(\mu_R)$ for every posterior draw. This separation prevents variance in log-scale heterogeneity from implicitly moving the global mean transmission scale.

Figure 5 summarizes the two-branch likelihood.

C.2 Hurdle observation likelihood

Let $Z_i = \mathbf{1}(Y_i > 0)$. The primary occurrence model is

$$Z_i \sim \text{Bernoulli}(\pi_i), \qquad \text{logit}(\pi_i) = \alpha_A + q_i^\top \beta_A, \tag{18}$$

where q_i contains standardized context/opportunity features. Vaccination-driven transmission is intentionally absent from equation 18 in the primary specification. Define

$$T_i = \frac{1}{k}\log\{1 + \exp(k(R_{e,i} - 1))\}, \qquad k = 8. \tag{19}$$

For positive observations,

$$Y_i - 1 \sim \text{NegBin}(\mu_i, \phi), \qquad \log \mu_i = \alpha_B + \gamma T_i + \log(N_i/10^5), \quad \gamma > 0. \tag{20}$$

Priors are $\mu_R \sim \mathcal{N}(\log 12, 1)$, $\beta_R \sim \mathcal{N}(0, 0.5^2)$, half-normal priors for hierarchical scales and γ, and a half-normal prior for dispersion. Posterior unconstrained-crossing draws are computed by evaluating R_0 on fixed reference G, taking the weighted arithmetic mean, and applying equation 10; feasibility is recorded separately.

C.3 Historical diagnostics

The historical HBM uses four chains, 1,500 tuning draws and 1,500 retained draws per chain. Threshold rank-normalized $\widehat{R} = 1.0019$, maximum core $\widehat{R} = 1.0059$, bulk ESS for the threshold is 2,388.8, minimum BFMI is 0.903, and divergences are zero. Posterior predictive checks cover both the observed zero fraction and the positive-count 95th percentile. These diagnostics establish computational adequacy conditional on the HBM; they do not substitute for known-truth validation.

D Change-point detection

Let observations be sorted by coverage, yielding signal vectors $z_{(1)}, \ldots, z_{(n)}$. For candidate cut k, define within-segment squared-error cost $C(1:k) + C(k+1:n)$. Because the total sum of squared norms is constant across k, minimizing this cost is equivalent to maximizing

$$G(k) = \frac{\|S_k\|^2}{k} + \frac{\|S_n - S_k\|^2}{n-k} - \frac{\|S_n\|^2}{n}, \tag{21}$$

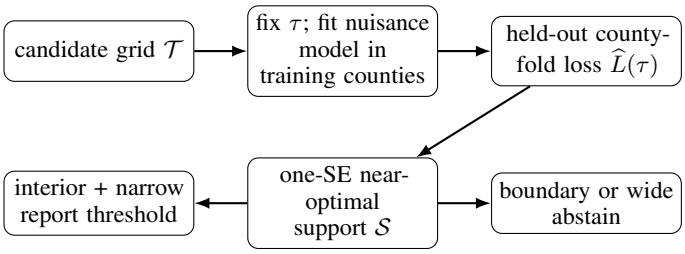

Figure 6: BINN threshold profiling. Optimization stability and threshold identification are separated: a reproducible profile minimum is insufficient if the near-optimal support remains wide.

where $S_k = \sum_{i \le k} z_{(i)}$. The implementation evaluates only cuts that do not split tied coverage values and that satisfy minimum segment size and fraction constraints.

For the RBF variant, z_i is a deterministic random Fourier feature vector of $\log(1 + Y_i)$, avoiding an $O(n^2)$ dense kernel matrix. County-block bootstrap preserves within-county temporal dependence. The permutation null exchanges whole coverage trajectories among counties within compatible state/year-pattern blocks, leaving outcome trajectories unchanged. The historical result has 2,192 movable county blocks, 500 successful bootstraps, no bootstrap failures, and $p = 0.366$.

Qualification across known mechanistic thresholds showed that the native RBF breakpoint did not track τ_G stably enough to justify an inverse calibration. We therefore retain the breakpoint only as a native empirical transition and do not harmonize it to τ_G.

E BINN PROFILE IDENTIFICATION

The BINN parameterizes the reference mean transmission scale through a fixed candidate τ and learns only relative heterogeneity/context nuisance structure at that candidate. For profile grid $\mathcal{T}$, county-fold validation risk is

$$\widehat{L}(\tau) = \frac{1}{K} \sum_{k=1}^{K} L(\widehat{\omega}_{-k}(\tau); \mathcal{D}_k), \qquad \tau \in \mathcal{T}, \tag{22}$$

where $\widehat{\omega}_{-k}(\tau)$ denotes nuisance parameters trained outside fold k with τ fixed. The best candidate is $\widehat{\tau}_{\text{prof}} = \arg\min_{\tau \in \mathcal{T}} \widehat{L}(\tau)$. Define tolerance

$$\epsilon = \max\{\text{SE}[\widehat{L}(\widehat{\tau}_{\text{prof}})], 10^{-4}\} \tag{23}$$

and support

$$\mathcal{S} = \{\tau \in \mathcal{T} : \widehat{L}(\tau) \le \widehat{L}(\widehat{\tau}_{\text{prof}}) + \epsilon\}. \tag{24}$$

A threshold is declared identified only if the minimum is not on the grid boundary and $\max \mathcal{S} - \min \mathcal{S} \le 0.04$. This support is an identification diagnostic, not a frequentist confidence interval.

Figure 6 summarizes the profiling and abstention rule. The historical grid is $\{0.88, 0.90, 0.92, 0.94, 0.96, 0.98, 0.99\}$. Its minimum is 0.98 but $\mathcal{S} = [0.88, 0.99]$, so the model reports no threshold estimate. The learned probability is also checked for monotonic nonincrease as coverage increases at fixed context, and reference normalization error is required to be negligible.

F QUALIFICATION HISTORY AND ROLE ASSIGNMENT

Role-assignment criteria were fixed using qualification experiments before interpreting the historical case study. Table 4 summarizes the evidence: joint BINN threshold optimization was strongly initialization-dependent, motivating fixed-threshold profiling; CPD variants did not track the mechanistic target stably, motivating native-only status; and HBM reparameterization improved target responsiveness. The compact validation then showed monotone HBM response but measurable

Table 4: Selected qualification outcomes and resulting design decisions.

Issue	Evidence	Final decision
BINN scale nonidentification	low/high initialization produced widely separated thresholds under the same truths	disable joint threshold optimization; use fixed-τ cross-validated profile and abstention rule
CPD harmonization	RBF/binary/incidence native breakpoints did not show sufficiently stable mechanistic-threshold mapping	retain CPD as native empirical-transition diagnostic; block calibration/fusion
HBM occurrence sensitivity	free transmission term in occurrence produced severe nonconvergence/divergences in qualification	primary occurrence model excludes transmission; alternative occurrence coupling is sensitivity-only
HBM scale interpretation	recovery changed with heterogeneous log-R_0 scale	separate arithmetic mean $\bar{R}_0$ from mean-one heterogeneity and standardize threshold over fixed G
HBM threshold validation	all 9 fits pass diagnostics; biases $+0.032, +0.026, +0.004$ across truths 0.90, 0.94, 0.97	retain HBM as target-responsive, but not calibrated/fusion-eligible across the full grid or validated above 0.97

positive bias and upper-end compression under a mechanistically aligned yet observation-model-mismatched generator. HBM is therefore treated as target-responsive and model-conditional, not as calibrated for fusion or extrapolation.

G Data provenance, quality control, and analytic support

Table 5: Case-study data construction and provenance fields represented in the supplied analysis archive.

Variable family	Resolution	Analytic role	Quality / provenance handling
MMR coverage	county-year proportion	control v	source-priority field retained; newer source preferred and legacy values fill otherwise missing support; source conflicts retained and auditable; no interpolation, forward fill, or winsorization in primary cohort; anomalous values quarantined
Measles cases	county-year count	observed response	reported positives, inferred zeros, and reconciled complete state-years distinguished; state totals used for reconciliation but not duplicated as county outcomes
Population	county-year	exposure / offset	canonical county/year join; no silent carry in primary contemporaneous cohort
Density, age share	county-year	transmission context	joined with provenance; transformations fit on training data only
State/county IDs	geographic hierarchy	random effects / grouped folds	canonical identifiers; duplicate county-year keys are fatal
Year	annual	temporal context	source union spans 2016–2024; strict primary threshold cohort spans 2018–2023

Table 5 records the construction rules represented in the supplied archive. The canonical processed panel has 25,398 rows. The primary case-study frame has 10,153 rows, 128 positives, and 2,212 counties. The fixed reference table has the same 10,153 rows and 2,212 counties. The broad panel and the analytic frame are therefore not interchangeable descriptions of sample size or time coverage.

Zero classification. A county-year with no allocated reported cases can enter under an explicit assumed-zero policy, but reconciled zero state-years are separately flagged when state totals and county allocations support completeness. This prevents "zero" from being silently equated with "not observed." The primary analysis uses the declared assumed-zero scenario. A verified-zero sensitivity was not present in the supplied results and is therefore a limitation/future analysis rather than a reported sensitivity.

Coverage conflicts. Where multiple sources provide different numeric coverage for the same county-year, the chosen source follows the stored priority field rather than averaging incompatible measurements. Conflict flags and original source values remain available for audit. The reference distribution uses the same finalized primary rows as the primary fit. The supplied manuscript archive does not contain the source-specific retrieval dates/URLs needed to reconstruct every upstream raw source; those metadata must be included in the anonymous data-provenance manifest accompanying submission.

H Reproducibility details

The historical manifest records NumPy 2.4.6, pandas 3.0.6, SciPy 1.17.1, PyMC 5.28.5, ArviZ 0.23.4, and Torch 2.14.0, together with configuration hash, model-source hashes, runner hash, input-data hash, analytic row counts, included years, and fixed-reference dimensions. The Linux environment pins the PyMC/PyTensor stack through conda-forge and OpenBLAS.

Every canonical run writes: (i) a configuration snapshot; (ii) a run manifest; (iii) the exact analysis and reference tables; (iv) method-specific result JSON; (v) posterior/profile/resampling artifacts as applicable; and (vi) a scalar summary CSV. Controlled runs additionally persist the simulated observable data and latent truth in separate files, making it possible to verify that estimators do not receive ground-truth columns.

The submitted supplementary archive preserves the controlled-validation summaries, replicate-level outputs, and the exact table/figure regeneration helper used for the reported controlled results. The appendices specify the mathematical estimands, observation models, diagnostics, and decision rules. End-to-end reconstruction of the observational panel additionally depends on the external source/provenance manifest and data-build assets, which are not redistributed here.