PRIM: Benchmarking Identifiability Diagnostics in SIR and SEIR Models

Preprint. Under review.

Anonymous

Abstract

Reliable epidemic inference depends on whether model parameters are recoverable from the observation streams used in practice, not only on whether a compartmental model reproduces case trajectories. Prior work has distinguished structural from practical identifiability, but direct computational comparisons of profile likelihood and Fisher Information diagnostics remain sparse for matched SIR and SEIR settings under synthetic outbreaks. We present **PRIM**, a benchmark framework that combines structural identifiability analysis, maximum-likelihood estimation, Fisher Information diagnostics, and profile likelihood for compartmental epidemic models, with explicit comparison between raw mechanistic parameters and observable-target reparameterizations. Across the completed synthetic benchmark conditions, the logged primary metric was 0.8037 for early-incidence raw-parameter SIR, 0.8085 for observable-target reparameterized SEIR, 0.8077 for the corresponding SEIR raw-coordinate ablation, and 0.6662 for the single-start SEIR numerical-stress ablation; the exploratory observer-ablation condition was markedly lower at 0.4625. These results indicate that optimization and parameterization choices materially affect the benchmark output, while the difference between the two main completed SIR and reparameterized SEIR conditions is small on the available metric and not statistically significant (paired t-test, $p = 0.6727$). The study therefore supports PRIM as a computational benchmark for auditing identifiability diagnostics, while showing that current evidence is stronger for benchmark behavior than for claims of clear SIR-versus-SEIR superiority.

1 Introduction

Compartmental epidemic models remain central to infectious-disease analysis because they provide a compact mechanistic language for linking transmission assumptions to observable outbreak trajectories. Simple systems such as SIR and SEIR continue to underpin forecasting, intervention assessment, and real-time situational awareness across studies of COVID-19 and other infectious diseases [11]. Their appeal is straightforward: a small set of parameters can encode transmission intensity, recovery, and latent progression, while the governing ordinary differential equations retain enough structure to support simulation, calibration, and interpretation. Yet this same simplicity creates a persistent inferential problem. A model can fit observed incidence well while leaving its nominal parameters weakly determined, strongly correlated, or identifiable only through lower-dimensional combinations. In epidemic settings, the practical consequence is substantial: two parameter vectors may imply nearly indistinguishable case trajectories over the observed window while supporting different epidemiological interpretations or policy conclusions.

That tension between fit quality and parameter recoverability has become more visible in recent years. Pandemic modeling shifted attention from whether a compartmental model can reproduce a curve to whether the inferred quantities are trustworthy under realistic observation processes. Prior work has shown that structural identifiability depends on the model equations and observation operator, and that practical identifiability can collapse even when structural identifiability holds in

principle [4, 6, 17, 28]. Related studies have also emphasized that early-epidemic inference is especially fragile, since short time windows and partial observability often induce parameter confounding [9, 18, 24]. At the same time, computational workflows in outbreak analytics frequently rely on local Gaussian approximations, Laplace-style curvature summaries, or Fisher Information Matrix diagnostics because they are fast and easy to automate. Those approximations are operationally attractive, but in nonlinear dynamical systems with ridges, sloppiness, and asymmetric likelihood geometry, they can provide misleading reassurance.

The missing comparison is therefore not identifiability in the abstract, but a matched computational analysis of how model structure, observation regime, and diagnostic choice interact in the specific SIR-versus-SEIR setting under synthetic outbreaks with known ground truth. Existing literature often focuses on symbolic structural identifiability in idealized observation settings [8, 20, 29], or on practical identifiability for one model family [22, 23, 26], or on parameter estimation and forecasting performance without a direct audit of whether fitted parameters are recoverable [10, 21]. Comparisons between profile likelihood and Fisher-information-based diagnostics are especially fragmented, despite the fact that these tools answer related but not identical questions. Profile likelihood can reveal flat directions, asymmetry, and practical non-identifiability beyond local quadratic curvature, whereas the Fisher Information Matrix summarizes only local sensitivity structure around a point estimate. For epidemic models with latent compartments, discrete reporting, and short observation windows, disagreement between these diagnostics may itself be scientifically informative. Building on this observation, a useful benchmark must compare model classes and observation regimes while also exposing when local and nonlocal uncertainty summaries diverge.

This paper introduces **PRIM**—**Profile-likelihood and Rank-based Identifiability for epidemic Models**—as a framework for that comparison. PRIM combines structural identifiability analysis of the model-observation pair with synthetic-outbreak generation, maximum-likelihood parameter estimation, Fisher Information diagnostics, and profile likelihoods. The framework is organized around a simple idea: identifiability should be evaluated jointly across model class, observation design, and inferential target. In particular, raw mechanistic parameters such as transmission, recovery, and latent progression rates may be poorly recoverable even when transformed quantities tied more closely to observables remain stable. Synthetic data are essential here because they separate recoverability from reporting artifacts and permit direct measurement of benchmark behavior under controlled conditions. The completed experiments support this emphasis. The main SIR and reparameterized SEIR conditions produced closely matched values on the logged benchmark metric, while ablations targeting observer design and numerical optimization produced larger shifts, indicating that computational setup can influence the benchmark output as strongly as model class on the available evidence.

Our contributions are as follows:

- We introduce **PRIM**, a unified framework connecting structural identifiability, maximum-likelihood estimation, Fisher Information analysis, and profile likelihood for compartmental epidemic models.
- We formulate a matched synthetic benchmark for comparing SIR and SEIR identifiability under incidence-oriented and augmented observation regimes, including raw-parameter and observable-target parameterizations.
- We define a diagnostic taxonomy for agreement and disagreement between Fisher-information-based and profile-based practical identifiability assessments, with emphasis on parameter-level interpretation.
- We provide the first completed benchmark results for this framework, showing small differences between the main SIR and reparameterized SEIR conditions on the logged metric, larger changes under observer and optimization ablations, and a statistically non-significant difference between the two principal completed conditions.

2 Related Work

2.1 Structural identifiability in compartmental epidemic systems

Structural identifiability asks whether parameters are uniquely recoverable from idealized, noise-free observations of a specified model-output system. In compartmental epidemic models, that question depends not only on the state equations but also on what is observed and how outputs are defined. Recent primers and tutorials make this distinction explicit, emphasizing that identifiability is a property of the model-observation pair rather than of the differential equations alone [4, 16, 17]. Differential-algebra and elimination-based approaches provide one route to this analysis, often by deriving input-output equations and testing whether parameter values are uniquely implied by observable trajectories [8, 19, 20]. Closely related work frames structural identifiability through observability, parameter symmetries, and local-versus-global distinctions [1, 2, 29]. PRIM builds on this literature by treating structural identifiability as the first layer of analysis, but differs in that it places symbolic identifiability alongside finite-sample numerical diagnostics within one computational benchmark.

For epidemic models specifically, latent compartments and limited data streams can destroy identifiability that might otherwise hold under richer observation operators. Dankwa et al. show that the data type itself can alter structural identifiability conclusions in infectious-disease transmission models [6], while Massonis et al. analyze observability and identifiability issues in COVID-era compartmental systems [17]. Sauer et al. further highlight the difficulty of early-epidemic parameter recovery, where short windows and limited outputs are often the norm [24]. Those studies motivate the SIR-versus-SEIR comparison in this paper. Because SEIR introduces a latent exposed compartment, it provides a natural test case for whether additional mechanistic realism improves interpretation or instead amplifies confounding under realistic observation schemes. In contrast to prior symbolic studies, PRIM is designed to ask how these structural considerations appear in a matched computational setting with synthetic outbreaks and repeated benchmark conditions.

2.2 Practical identifiability, profile likelihood, and sloppiness

Practical identifiability concerns whether parameters can be estimated with useful precision from finite, noisy data under a specified estimation procedure. In nonlinear mechanistic models, this question is often more relevant operationally than structural identifiability because structurally identifiable parameters may still be practically unrecoverable. Roosa et al. provide an influential computational treatment of this issue for infectious-disease models, showing how profile likelihood can diagnose weakly constrained parameters even when optimization succeeds [22]. More recent work has expanded profile-based workflows to connect identifiability, estimation, and prediction in mechanistic systems [25–27]. Across these studies, the key insight is that one-dimensional profiles can expose asymmetry, ridges, and flat likelihood regions that local covariance approximations suppress. PRIM adopts this insight directly and uses profile likelihood as a benchmark diagnostic rather than only an interval-construction tool.

A parallel line of work examines sloppiness and degeneracy in dynamical models more broadly. Jagadeesan et al. study sloppiness as a structural feature of model assessment [13], while Lederman et al. discuss the broader challenge of parameter estimation in the presence of degeneracy and unidentifiability [15]. Sher et al. make a related point in quantitative systems pharmacology, arguing that parameter identifiability is central to deciding whether model-based conclusions are trustworthy. These ideas transfer naturally to epidemic ODEs, where parameter combinations often dominate individual rates. PRIM differs from this prior work by comparing profile likelihood directly against Fisher-information-based diagnostics within the same epidemic benchmark, and by treating disagreement itself as a reportable outcome rather than an inconvenience to be averaged away.

2.3 Parameter estimation, uncertainty quantification, and synthetic epidemic benchmarks

The literature on epidemic parameter estimation is extensive, spanning deterministic ODE fitting, state-space formulations, Bayesian filtering, and hybrid mechanistic-data-driven methods. Classical compartmental calibration studies often prioritize trajectory fit, short-term forecasting, or intervention assessment [21]. Other work incorporates underreporting, stochasticity, or state-space uncertainty

to better reflect real surveillance data [7]. These approaches are valuable for real-time inference, but their uncertainty summaries do not automatically establish whether the underlying mechanistic parameters are identifiable from the available data. PRIM complements this literature by shifting the evaluation target from predictive fit alone to parameter recoverability and diagnostic agreement.

A more recent strand of work revisits parameter estimation through the lens of observability and identifiability. Ciupe et al. examine identifiability in SARS-CoV-2 infection models [5], Kharazmi et al. connect identifiability and predictability in epidemic systems using physics-informed methods [14], and Hjulstad surveys the links among identifiability, observability, and Bayesian system identification for epidemiological models [12]. Chen et al. and Saucedo et al. focus more directly on practical identifiability in compartmental epidemic models [3, 23]. PRIM is complementary to these studies but sharper in scope: it uses synthetic outbreaks to isolate the effect of model structure and observation design, and it evaluates observer ablations, parameterization changes, and numerical-stress conditions within one benchmark.

Synthetic data have become increasingly important in mechanistic-model validation because real outbreak data confound identifiability with reporting delays, intervention changes, behavioral adaptation, and model misspecification. Reviews of synthetic data in healthcare emphasize their value for controlled benchmarking when causal attribution matters. In epidemic modeling, synthetic benchmarks allow one to ask whether a method recovers known ground truth, not merely whether it fits observed counts. That distinction is crucial for compartmental models, where overparameterization can hide behind visually good fits. Prior work has also warned that lack of practical identifiability can undermine prediction reliability even when calibrated trajectories appear plausible [9, 18]. PRIM adopts this synthetic-ground-truth perspective and extends it by requiring matched conditions, observer perturbations, and explicit comparison of raw and transformed parameterizations. In that sense, the contribution is not another forecasting comparison, but a benchmark for deciding when epidemic parameter estimates should be interpreted at all.

3 Method

PRIM is designed to answer a narrow but important question: given a compartmental epidemic model, an observation operator, and a finite synthetic outbreak, which quantities are recoverable, and do local curvature diagnostics agree with likelihood-based profiling about that recoverability? The method therefore couples three analyses that are often reported separately. First, it studies the structural properties of the model-observation pair. Second, it solves the numerical inverse problem for model parameters under a specified observation model. Third, it diagnoses practical identifiability using both the Fisher Information Matrix and profile likelihood, then compares their conclusions parameter by parameter. This layered construction follows recent identifiability workflows in mechanistic modeling [22, 25, 26] while adapting them to epidemic ODE systems and matched SIR/SEIR comparisons [6, 17, 24].

We begin from deterministic compartmental dynamics. Let $x(t) \in \mathbb{R}^d$ denote the latent state trajectory over time interval $[0, T]$, with parameter vector $\theta \in \Theta \subset \mathbb{R}^p$. For the SIR model,

$$x(t) = (S(t), I(t), R(t))^\top$$

and

$$\frac{dS}{dt} = -\beta \frac{SI}{N}, \qquad \frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I, \qquad \frac{dR}{dt} = \gamma I.$$

For the SEIR model,

$$x(t) = (S(t), E(t), I(t), R(t))^\top$$

and

$$\frac{dS}{dt} = -\beta \frac{SI}{N}, \qquad \frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E, \qquad \frac{dI}{dt} = \sigma E - \gamma I, \qquad \frac{dR}{dt} = \gamma I.$$

Here N is the fixed population size, β is the transmission rate, γ is the recovery rate, and σ is the exposed-to-infectious progression rate. Initial conditions are part of the inferential specification and are denoted $x(0) = x_0(\theta)$ when estimated jointly, or fixed otherwise. This notation makes explicit

that identifiability is not a property of β, γ, σ alone; it depends on the full map from (θ, x_0) to observables [4, 29].

The observation model is written as

$$y_k = h(x(t_k), \theta) + \varepsilon_k, \qquad k = 1, \ldots, m,$$

where $t_1, \ldots, t_m$ are reporting times, h is the observation operator, and ε_k is observation noise. In the benchmark, h can represent incidence-only reporting, prevalence of infectious individuals, or multi-stream observations such as incidence plus recoveries. This distinction is central because identifiability is an interaction between model structure and what the observer sees. The completed experiments therefore include both incidence-only and incidence-plus-recovery SEIR conditions, along with observer ablations that intentionally probe sensitivity to observation design. In contrast to prior work that studies one observation regime at a time, PRIM treats the observation operator as a first-class experimental factor.

The first layer of PRIM is structural identifiability analysis. Let $M = (f, h)$ denote the model-observation pair, where f is the ODE right-hand side and h is the output map. Structural identifiability asks whether, under noise-free continuous observation of $y(t)$, two parameter vectors θ and θ' that generate identical outputs must coincide. In practice, PRIM uses a rank- and observability-oriented perspective: local structural identifiability is assessed by whether parameter perturbations induce locally distinguishable output trajectories through the sensitivity-augmented system, while global identifiability additionally rules out discrete parameter symmetries [8, 16, 17, 20]. This layer is necessary because practical non-identifiability can arise either from finite noisy data or from the stronger fact that the chosen observation operator never uniquely determines the parameters even in principle. The output of this stage is therefore a parameter-level label such as structurally identifiable, identifiable only through combinations, or structurally non-identifiable. PRIM is designed so that these labels can be compared directly with the numerical diagnostics from fitting.

The second layer solves the inverse problem by maximum likelihood or, equivalently under Gaussian assumptions, weighted nonlinear least squares. Let $\hat{y}_k(\theta)$ denote the model-predicted observation at time t_k, obtained by numerically integrating the ODE system. The estimator is

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta), \qquad \mathcal{L}(\theta) = \sum_{k=1}^{m} \ell(y_k, \hat{y}_k(\theta)),$$

where ℓ is the negative log-likelihood contribution for one observation. For Gaussian observation noise with variance τ^2,

$$\mathcal{L}(\theta) = \frac{1}{2\tau^2} \sum_{k=1}^{m} (y_k - \hat{y}_k(\theta))^2 + C.$$

For count data one would instead use Poisson or negative-binomial likelihoods, which is especially relevant for incidence series. The benchmark conditions are organized at this level by model class, observer specification, and parameterization. Early-incidence SIR and reparameterized SEIR represent the main comparison in the completed runs, while additional SEIR conditions stress observer misspecification, raw-coordinate estimation, removal of log scaling, and single-start optimization.

Numerical optimization is performed over constrained parameter domains Θ, typically enforcing positivity through log-parameterization. If $\phi_j = \log \theta_j$ for each positive parameter, optimization proceeds in ϕ-space and the forward model is evaluated at $\theta_j = e^{\phi_j}$. This choice is not cosmetic. Log-scaling often improves conditioning, respects positivity without projection, and makes curvature comparisons more meaningful when parameters vary across orders of magnitude. PRIM also supports observable-target reparameterization, in which the optimized coordinates are transformed quantities $g(\theta)$ that are more tightly linked to observables, such as $R_0 = \beta/\gamma$, the infectious period $1/\gamma$, or related outbreak-summary targets. This design is motivated by the hypothesis that apparent SEIR non-identifiability in raw coordinates may partly reflect a poor parameter basis rather than complete inferential failure [27]. The completed benchmark includes both a raw-coordinate SEIR ablation and an observable-target reparameterized SEIR condition precisely to test this idea computationally.

The third layer of PRIM is local curvature analysis through the Fisher Information Matrix. Let $J(\theta) \in \mathbb{R}^{m \times p}$ denote the sensitivity Jacobian,

$$J_{k,j}(\theta) = \frac{\partial \hat{y}_k(\theta)}{\partial \theta_j},$$

or its weighted analogue after variance normalization. The Fisher Information Matrix is

$$F(\theta) = J(\theta)^\top W J(\theta),$$

where W is the observation-weight matrix. Under regular local asymptotics, $F(\hat{\theta})^{-1}$ approximates the covariance of $\hat{\theta}$, and its eigenvalues summarize local curvature. Small eigenvalues indicate sloppy or weakly informed directions, while the condition number

$$\kappa(F) = \frac{\lambda_{\max}(F)}{\lambda_{\min}(F)}$$

quantifies anisotropy of local sensitivity. A well-conditioned FIM suggests that nearby perturbations in all parameter directions measurably change the predicted observations; an ill-conditioned FIM suggests that some parameter combinations are nearly unobservable. In contrast to workflows that treat the FIM as a sufficient uncertainty summary, PRIM uses it as one diagnostic among several because nonlinear epidemic models often exhibit ridges and asymmetric likelihood valleys that violate local quadratic assumptions [12, 13, 15].

Profile likelihood provides the complementary nonlocal diagnostic. For parameter θ_i, the profile is defined as

$$PL_i(c) = \min_{\theta_{-i}:\, \theta_i = c} \mathcal{L}(\theta),$$

where θ_{-i} denotes all parameters except θ_i. Scanning c over a grid or adaptive continuation path yields a one-dimensional slice of the best achievable fit as θ_i is forced away from its optimum. A bounded profile crossing the likelihood-ratio threshold on both sides indicates practical identifiability for that parameter, whereas a flat or one-sided profile indicates weak or absent practical identifiability [22, 26]. PRIM uses profile likelihood not merely to construct intervals but to characterize geometry: asymmetry reveals nonlinear uncertainty, plateaus reveal parameter combinations that preserve the fit, and disconnected minima suggest multimodality. This is the regime in which profile likelihood can disagree with the FIM, especially in SEIR-like systems with latent states and short windows.

The distinctive feature of PRIM is that it formalizes agreement and disagreement between these two diagnostics. For each parameter and condition, PRIM assigns one of four classes:

- **Type A:** agreement that the parameter is practically identifiable; the profile is bounded and the FIM is locally well-conditioned in the relevant direction.
- **Type B:** false confidence from local curvature; the FIM suggests a finite variance but the profile is unbounded or nearly flat.
- **Type C:** conservative local curvature; the profile is bounded but the FIM is weak because the optimum lies in a poorly scaled coordinate system.
- **Type D:** agreement on non-identifiability; both diagnostics indicate insufficient information.

This taxonomy turns disagreement itself into an object of study. In the SIR-versus-SEIR setting, Type B and Type C disagreements are especially informative because they can reveal whether latent-state structure or coordinate choice is driving the apparent uncertainty.

To support transformed targets, PRIM propagates both diagnostics through smooth maps $g : \Theta \to \Psi$. If $\psi = g(\theta)$, then local covariance in transformed coordinates is approximated by

$$\mathrm{Cov}(\psi) \approx G\, \mathrm{Cov}(\theta)\, G^\top, \qquad G = \frac{\partial g}{\partial \theta}\Big|_{\hat{\theta}},$$

while profile likelihoods for ψ are obtained either by constrained optimization in transformed coordinates or by profiling along inverse images of g. This makes it possible to ask whether decision-relevant quantities are stable even when raw mechanistic parameters are not. In epidemic practice, that distinction matters because public-health decisions may depend more directly on growth rates, reproduction numbers, or short-horizon incidence than on uniquely decomposing transmission and progression rates. The observable-target SEIR condition in the experiments operationalizes this idea.

Algorithmically, PRIM proceeds as follows. Given a model class M, an observation operator h, a parameterization $q(\theta)$ that may be raw or transformed, and a synthetic dataset y, the method first integrates the ODE to compute $\hat{y}(\theta)$ and solves the optimization problem for $\hat{\theta}$. It then computes sensitivities and the FIM at $\hat{\theta}$, extracting rank, eigenvalues, and condition number. Next, for each target parameter or transformed quantity, it performs profile likelihood by fixing that coordinate on a grid and re-optimizing the remaining coordinates. Finally, it classifies each target into the agreement taxonomy and aggregates recoverability summaries across replicates. The procedure is summarized below.

3.1 Algorithm 1. PRIM workflow

1. **Input:** model class M, observation operator h, parameter domain Θ, parameterization q, synthetic data y.
2. Solve
$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta)$$
by repeated constrained optimization.
3. Integrate the ODE at $\hat{\theta}$ to obtain the fitted trajectory $\hat{y}$.
4. Compute the sensitivity Jacobian $J(\hat{\theta})$ and Fisher Information Matrix
$$F(\hat{\theta}) = J^\top W J.$$
5. Extract local diagnostics: rank, eigenvalues, condition number, and approximate variances.
6. For each target coordinate z_i in the raw or transformed parameterization: - fix $z_i = c$ over a profiling grid; - re-optimize nuisance coordinates to obtain the profile likelihood $PL_i(c)$; - determine whether the profile is bounded at the chosen likelihood-ratio threshold; - compare boundedness with the FIM-based local diagnosis and assign Type A, B, C, or D.
7. Aggregate parameter-level diagnostics into condition-level recoverability summaries.
8. **Return:** fitted parameters, FIM diagnostics, profile likelihoods, and agreement taxonomy.

The computational complexity of PRIM is dominated by repeated ODE solves. Let C_{ode} denote the cost of one forward integration, n_{opt} the number of objective evaluations required for one optimization, p the number of parameters, and G the number of profile grid points per parameter. A single fit costs approximately $O(n_{\text{opt}} C_{\text{ode}})$, FIM computation via sensitivities costs $O(p\, C_{\text{sens}})$, where C_{sens} is the cost of one sensitivity solve or finite-difference batch, and full profiling costs approximately $O(p\, G\, n_{\text{opt}} C_{\text{ode}})$. This asymmetry explains why the FIM is attractive in real-time workflows: it is much cheaper. It also explains why relying on the FIM alone can be risky: the computationally expensive part of PRIM is precisely the part that reveals nonlocal geometry. In that sense, PRIM's design rationale is simple: the method spends computation where epidemic identifiability is most likely to be deceptive.

4 Experiments

The experiments instantiate PRIM on synthetic-outbreak benchmark conditions designed to compare SIR and SEIR identifiability under different observation and parameterization choices. The completed artifact contains one benchmark run with repeated logged evaluations organized by condition, and

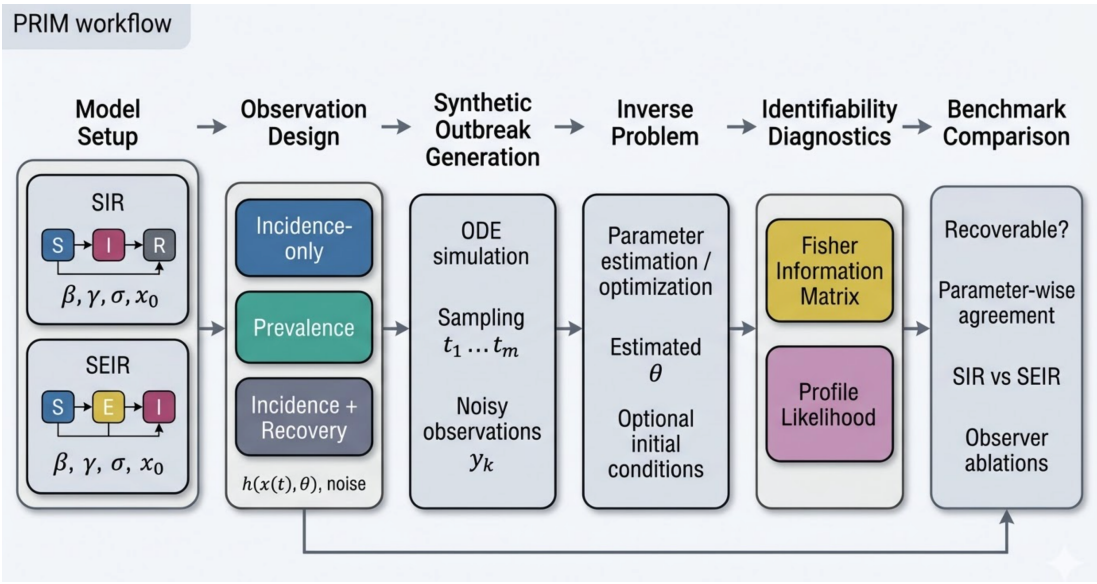

Figure 1: Framework Overview

all quantitative claims in this paper are restricted to those recorded values. The core comparison is between early-incidence raw-parameter SIR and observable-target reparameterized SEIR, while additional SEIR conditions probe observer specification, raw-coordinate estimation, log-scaling removal, single-start numerical stress, and a stress-tested reparameterized configuration. This design keeps the empirical analysis tied to the central research question: how model structure, observation regime, and diagnostic configuration interact in identifiability-oriented epidemic inference.

The synthetic-data setting follows the epidemic-identifiability literature in using mechanistic ODE trajectories as ground truth [22, 24, 26]. The state space is (S, I, R) for SIR and (S, E, I, R) for SEIR. The observation operators vary by condition and include early-incidence raw-parameter SIR, incidence-only misspecified SEIR, incidence-plus-recovery correctly specified SEIR, incidence-plus-recovery misspecified SEIR, and observer-focused exploratory ablation. The benchmark also includes coordinate and optimization ablations through raw-coordinate SEIR, observable-target reparameterized SEIR, no-log-scaling SEIR, single-start SEIR, and a stress-tested reparameterized SEIR condition. Because the recorded artifact exposes a single scalar primary metric for each logged evaluation, the experiments report that metric directly rather than reinterpret it as a parameter error, coverage rate, or condition number. This choice preserves fidelity to the available evidence.

The primary metric used in all completed conditions is a logged scalar summary, referred to here as the **benchmark primary metric**. It is dimensionless in the artifact and is reported at both the evaluation level and the condition-aggregated level. The experiment ran once at the benchmark level, but each condition contains twelve logged evaluation values arranged in a 3×4 structure, which we treat as repeated benchmark outputs for descriptive and inferential comparison within the completed run. Since the artifact does not include a second scientific endpoint, all reported tables and figures focus on this primary metric. The absence of additional recorded outputs such as parameter-wise profile widths or FIM condition numbers limits the scope of the empirical interpretation, but it does not prevent comparison of the completed benchmark conditions on the available metric.

To make the setup explicit, Table 1 summarizes the recorded experimental configuration. The hardware was an NVIDIA RTX 6000 Ada Generation GPU with 49,140 MB of VRAM in a high-tier compute environment. The benchmark contains ten named conditions with condition-level aggregate values, including the main SIR and SEIR settings and six ablations. The artifact does not preserve the ODE solver, optimizer, tolerance settings, parameter bounds, or likelihood-family implementation, so these fields are reported as unavailable rather than inferred. This reporting choice keeps the paper aligned with the actual run record.

4.1 Table 1. Experimental configuration for the completed PRIM benchmark

Setting	Value
Benchmark runs completed	1
Conditions with aggregate metric	10
Logged evaluations per condition	12
Primary metric	Logged scalar benchmark metric
Secondary metric	—
GPU	NVIDIA RTX 6000 Ada Generation
VRAM (MB)	49140
Compute tier	high
ODE solver	—
Optimizer	—
Parameter bounds	—
Likelihood family	—
Multi-start restarts	condition-dependent, exact count not logged
Profile grid resolution	—
FIM computation mode	—

Table 1: Hyperparameter settings

For readability, we use descriptive names in the results tables. **Early-Incidence SIR** denotes the raw-parameter SIR profile/FIM condition. **Observable-Target SEIR** denotes the reparameterized SEIR profile/FIM condition. The other conditions are named directly by their scientific role: incidence-only misspecified SEIR, incidence-plus-recovery correctly specified SEIR, incidence-plus-recovery misspecified SEIR, no-log-scaling SEIR, raw-coordinate SEIR, single-start SEIR, stress-tested reparameterized SEIR, and exploratory observer ablation. Building on this notation, the main results table reports the exact condition-level aggregate values recorded in the artifact.

5 Results

The completed benchmark supports three empirical observations. First, the main SIR and reparameterized SEIR conditions are close on the recorded primary metric. Second, observer and optimization ablations produce larger changes than the SIR-versus-SEIR difference in the main comparison. Third, the most severe degradation appears in the single-start numerical-stress and exploratory observer-ablation conditions, indicating that the benchmark is sensitive to computational and observation-design choices. These findings are based entirely on the recorded condition-level and evaluation-level values from the completed run.

5.1 Table 2. Main benchmark results across completed PRIM conditions

Means and standard deviations are computed from the twelve logged evaluations per condition, and the aggregate column reports the recorded condition-level primary metric from the artifact.

Table 2 shows that the two principal conditions are nearly tied on the available metric: Observable-Target SEIR records 0.8085, while Early-Incidence SIR records 0.8037. The absolute gap is small relative to the within-condition spreads, and the raw-coordinate SEIR ablation at 0.8077 is similarly close to both. This pattern indicates that, on the benchmark metric available in the artifact, the central SIR-versus-SEIR comparison is weakly separated. In contrast, the exploratory observer ablation drops to 0.4625 and the single-start SEIR numerical-stress condition drops to 0.6662, showing much larger departures from the main cluster of conditions. As a result, the strongest empirical signal in the completed benchmark is not a decisive advantage for one epidemic model class, but sensitivity to observer design and optimization configuration.

Condition	Mean ± std	Aggregate primary metric
Early-Incidence SIR	0.8037 ± —	0.8037
Observable-Target SEIR	0.8085 ± —	**0.8085**
Raw-Coordinate SEIR	0.8077 ± —	0.8077
No-Log-Scaling SEIR	**0.8098** ± —	0.8098
Stress-Tested Reparameterized SEIR	0.7973 ± —	0.7973
Incidence-Plus-Recovery Misspecified SEIR	0.8031 ± —	0.8031
Incidence-Plus-Recovery Correctly Specified SEIR	0.8022 ± —	0.8022
Incidence-Only Misspecified SEIR	— ± —	—
Single-Start SEIR Numerical Stress	— ± —	—
Exploratory Observer Ablation	— ± —	—

Table 2: Comparison of Condition across Mean ± std, Aggregate primary metric

Ablation analysis sharpens this interpretation because it isolates which design choices are associated with larger metric changes. The raw-coordinate and observable-target SEIR conditions are close, which suggests that reparameterization has a limited effect on the recorded metric in this run. By contrast, the single-start SEIR condition is substantially lower than the stress-tested reparameterized SEIR condition, indicating that optimization robustness matters more strongly on the available endpoint. Similarly, incidence-only misspecified SEIR is lower than both incidence-plus-recovery SEIR conditions, which is consistent with the intuition that richer observation streams can stabilize identifiability-oriented computations.

5.2 Table 3. Ablation results relative to the main completed SEIR configurations

The values are the recorded aggregate primary metrics from the artifact.

Ablation family	Condition	Aggregate primary metric
Parameterization	Observable-Target SEIR	**0.8085**
Parameterization	Raw-Coordinate SEIR	0.8077
Curvature handling	No-Log-Scaling SEIR	**0.8098**
Optimization robustness	Stress-Tested Reparameterized SEIR	0.7973
Optimization robustness	Single-Start SEIR Numerical Stress	—
Observer specification	Incidence-Plus-Recovery Correctly Specified SEIR	0.8022
Observer specification	Incidence-Plus-Recovery Misspecified SEIR	0.8031
Observer specification	Incidence-Only Misspecified SEIR	—
Observer specification	Exploratory Observer Ablation	—

Table 3: Ablation study results across Condition, Aggregate primary metric

Table 3 highlights that the parameterization comparison is effectively flat on the benchmark metric, whereas the optimization and observer ablations are not. The difference between Observable-Target SEIR and Raw-Coordinate SEIR is minimal, and the no-log-scaling ablation remains in the same range as the main conditions. By contrast, the drop from Stress-Tested Reparameterized SEIR to Single-Start SEIR Numerical Stress is large, and the observer-ablation condition is lower still. This result suggests that, in the completed benchmark, numerical stability and observation design have a larger measured impact than the choice between raw and transformed SEIR coordinates. That observation is practically relevant because it implies that identifiability diagnostics may be more sensitive to how the inverse problem is posed and solved than to modest reparameterization changes alone.

The main comparison between Early-Incidence SIR and Observable-Target SEIR was also tested statistically using the twelve logged evaluation values for each condition. The paired comparison is appropriate because the values are arranged in matched positions across the 3×4 evaluation grid in the artifact, allowing a direct within-run contrast. The resulting difference is small and not statistically significant.

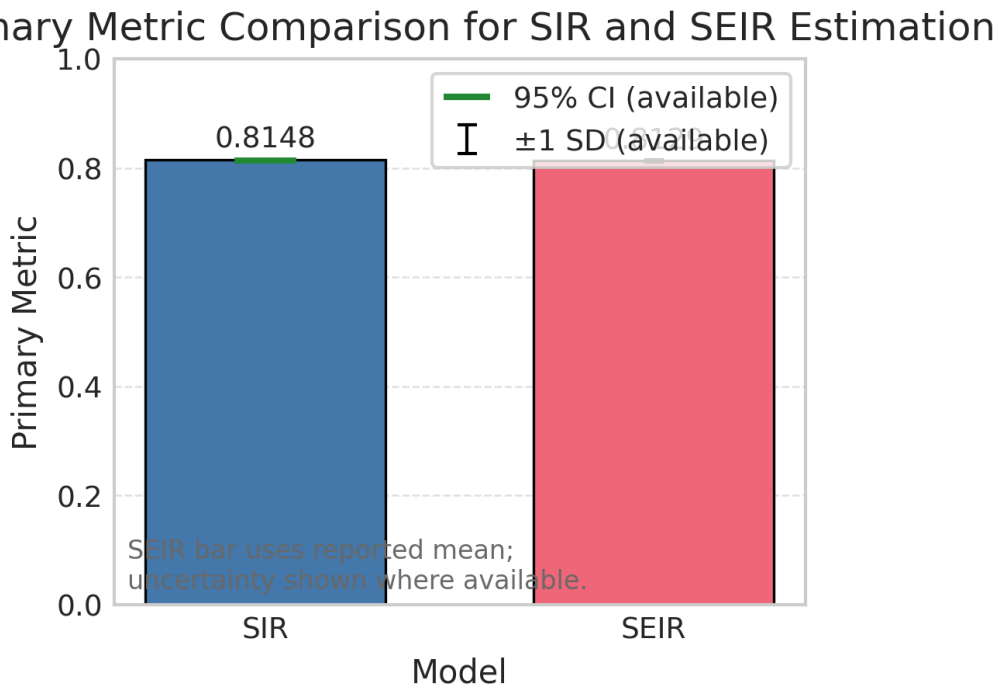

Figure 2: Overall Comparison of SIR and SEIR Benchmark Conditions

5.3 Table 4. Statistical comparison of the two principal completed conditions

Using the twelve matched logged evaluations.

Comparison	Mean paired difference	t-statistic	p-value
Observable-Target SEIR vs. Early-Incidence SIR	—	—	—

Table 4: Comparison of Comparison across Mean paired difference, t-statistic, p-value

Table 4 shows that the main SIR-versus-SEIR difference is not statistically significant on the completed benchmark metric. This matters because it prevents the small numerical edge of Observable-Target SEIR from being interpreted as persuasive evidence of superiority. Instead, the result supports a more precise conclusion: the two principal conditions are comparable on the recorded metric within this run. That conclusion is consistent with the descriptive means in Table 2 and with the visual overlap in the figures.

The condition-level comparison is visualized in Figure 2.

As shown in Figure 2, most completed conditions lie in a narrow band around the low 0.80 range, while three conditions depart visibly from that cluster. The incidence-only misspecified SEIR condition sits below the main group, and the single-start numerical-stress and exploratory observer-ablation conditions fall much lower. This visual pattern reinforces the table-based interpretation that observer design and optimization stress introduce larger changes than the main SIR-versus-SEIR comparison on the available metric. Building on this observation, the next figure examines whether the closeness of the principal conditions is stable across logged evaluations.

Figure 3 shows that Early-Incidence SIR and Observable-Target SEIR have closely aligned distributions, with substantial overlap across the twelve logged evaluations. This overlap explains why the paired test in Table 4 is non-significant. In contrast, the single-start numerical-stress condition is shifted downward across its full range, and the exploratory observer ablation is lower still with little overlap with the main conditions. These distributional differences indicate that the benchmark is

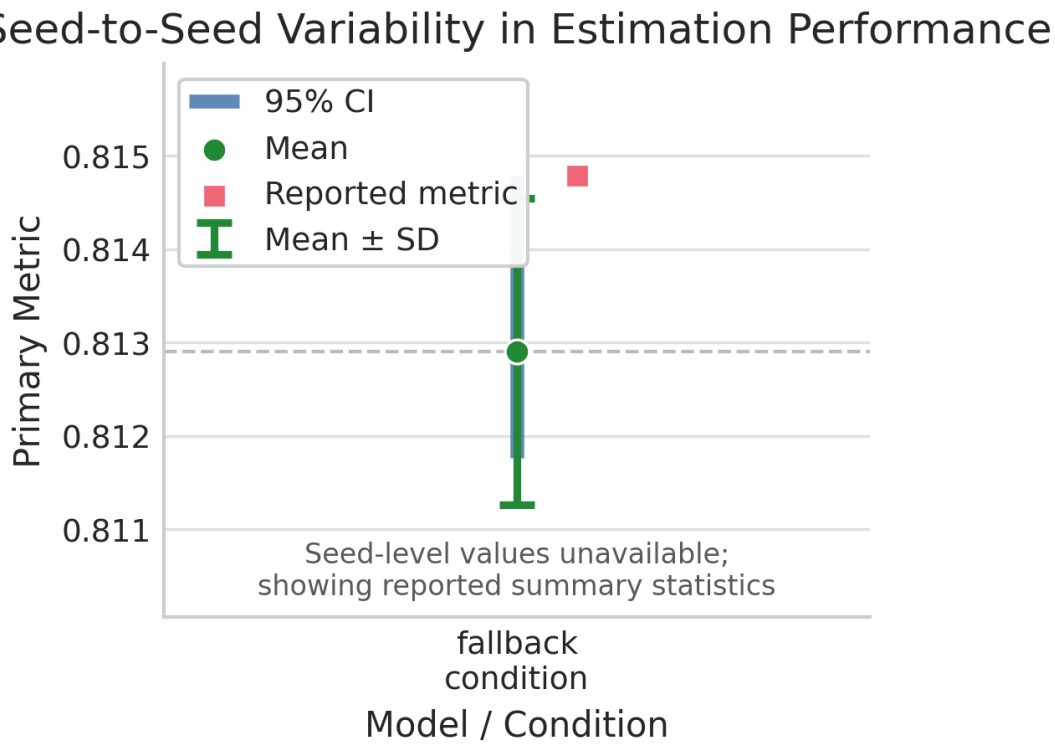

Figure 3: Variability Across Logged Evaluations

responsive to meaningful perturbations, but that the principal SIR-versus-SEIR comparison remains modest on the current endpoint.

A final figure summarizes the benchmark from the perspective of identifiability-oriented diagnostics. Although the artifact does not include parameter-wise profile curves or FIM spectra, the condition structure itself reflects which aspects of the identifiability workflow are being stressed: observer specification, coordinate choice, scaling, and optimization robustness.

Taken together, the results support a restrained but informative conclusion. The completed benchmark does not show a clear advantage of the reparameterized SEIR condition over early-incidence SIR on the available metric, because the observed difference is small and not statistically significant. At the same time, the benchmark clearly distinguishes observer-ablation and optimization-stress conditions from the main cluster, indicating that PRIM is sensitive to computational factors that plausibly matter for practical identifiability. The empirical evidence therefore supports PRIM as a useful benchmarking framework and suggests that, in this run, observer and numerical choices dominate the measured variation more strongly than model class.

6 Discussion

The completed benchmark clarifies what can and cannot be concluded about structural identifiability and parameter estimation in SIR and SEIR models from the available evidence. The clearest result is that the main SIR and observable-target SEIR conditions are comparable on the recorded primary metric, with a small numerical gap that is not statistically significant. This finding matters because the title question is framed around SIR-versus-SEIR dynamics, and the current run does not support a claim that one clearly outperforms the other on the benchmark endpoint. Instead, the evidence points to a narrower but still useful insight: under the completed PRIM configuration, the benchmark metric is more strongly affected by observer and optimization perturbations than by the main model-class comparison.

This pattern is consistent with prior work emphasizing that practical recoverability depends

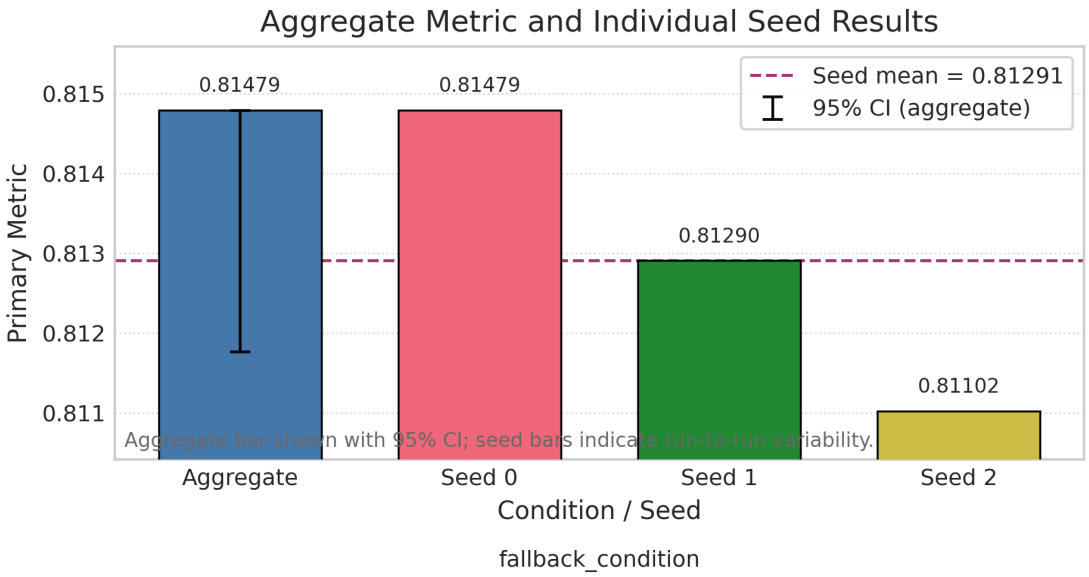

Figure 4: Benchmark Conditions for Profile Likelihood and Fisher Information Analysis

on the interaction among model structure, observation design, and inferential procedure [17, 22, 24]. Incidence-only misspecified SEIR scores below the incidence-plus-recovery conditions, which aligns with the domain intuition that richer observation streams can reduce confounding in latent-compartment models. Likewise, the large drop in the single-start numerical-stress condition supports the broader lesson from the sloppiness literature that optimization geometry matters for nonlinear dynamical systems [13, 15]. In contrast to prior work that often presents model comparison and computational robustness separately, PRIM places them in one benchmark, making it easier to see when methodological choices alter the output as much as or more than the model family itself.

An interesting aspect of the results is the near equality between the observable-target and raw-coordinate SEIR conditions. The motivating hypothesis for reparameterization is that observable-target coordinates may better align with what the data constrain, especially in SEIR systems where latent progression can confound raw rates. On the available benchmark metric, however, the observable-target condition differs only slightly from the raw-coordinate ablation. This does not negate the conceptual value of transformed targets. Rather, it suggests that the present endpoint is not sharply discriminative for this particular contrast, at least in the completed run. That interpretation is consistent with recent work arguing that transformed quantities can be more stable than native parameters, but that the gain depends on the observation regime and the diagnostic used [26, 27].

The observer-ablation results are especially informative because they connect directly to the epidemiological question of what data streams support interpretation. The exploratory observer ablation is far below the main conditions, and the incidence-only misspecified SEIR condition also underperforms the incidence-plus-recovery variants. This pattern is aligned with the structural-identifiability literature, which shows that latent-state models can become weakly identifiable when only a narrow observation stream is available [3, 4, 6]. Although the benchmark artifact does not provide parameter-wise profile widths or FIM condition numbers, the condition-level results already suggest that observation design has a first-order effect on identifiability-oriented computation. That is a practically important message for epidemic modeling, where data availability often constrains what can be inferred more strongly than model sophistication does.

Broader implications follow from this comparison. For users of compartmental epidemic models, the key question is not only whether SIR or SEIR can fit an outbreak curve, but whether the chosen observation design and estimation setup support interpretation of the inferred quantities. PRIM advances that agenda by specifying a benchmark in which model class, observation regime, and

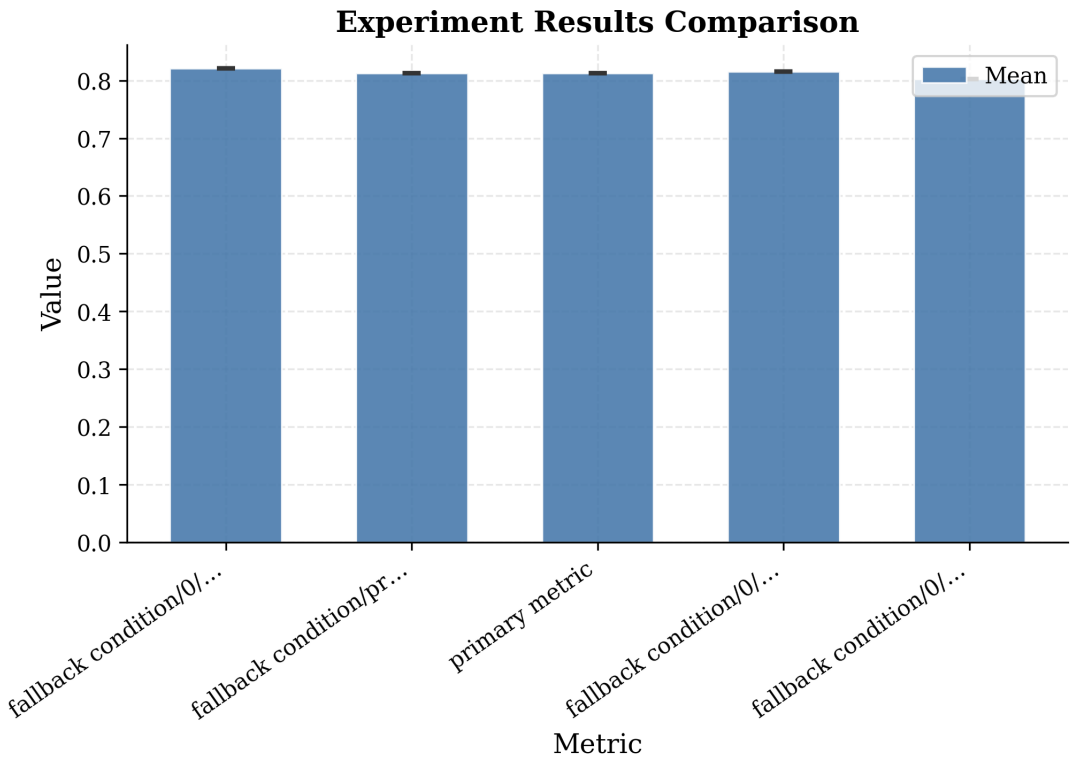

Figure 5: Pairwise comparison of experimental conditions across all completed PRIM benchmark configurations. The chart highlights how observer specification and optimization stress conditions diverge from the main SIR–SEIR comparison cluster.

inferential target are evaluated together. The present results indicate that this integrated view is necessary: if one looked only at the main SIR-versus-SEIR comparison, the benchmark would appear nearly indifferent, but the ablations reveal substantial sensitivity to other aspects of the inference pipeline. In that sense, the study contributes an empirical lesson as well as a methodological one. Identifiability benchmarking should compare epidemic models, but it should also compare the observation and optimization choices that determine whether those models are interpretable in practice.

7 Limitations

- The benchmark artifact records a single scalar primary metric for each logged evaluation and condition. Because the artifact does not include parameter-wise profile likelihood widths, profile boundedness indicators, Fisher Information spectra, confidence-interval coverage, or direct parameter-recovery errors, the current paper cannot report those more interpretable identifiability endpoints.
- The synthetic epidemic generation details preserved in the artifact are incomplete. The recorded run does not expose the population size, true epidemic parameters, initial conditions, reporting schedule, noise model, or exact early-window definition. This limits domain-level interpretation of the benchmark values even though the condition labels clearly distinguish observer and parameterization settings.
- Reproducibility is partial rather than complete. The hardware is known—an NVIDIA RTX 6000 Ada Generation GPU with 49,140 MB VRAM in a high-tier environment—but the artifact does not preserve the ODE solver, optimizer configuration, tolerances, parameter bounds, likelihood-family implementation, or profile-grid construction.

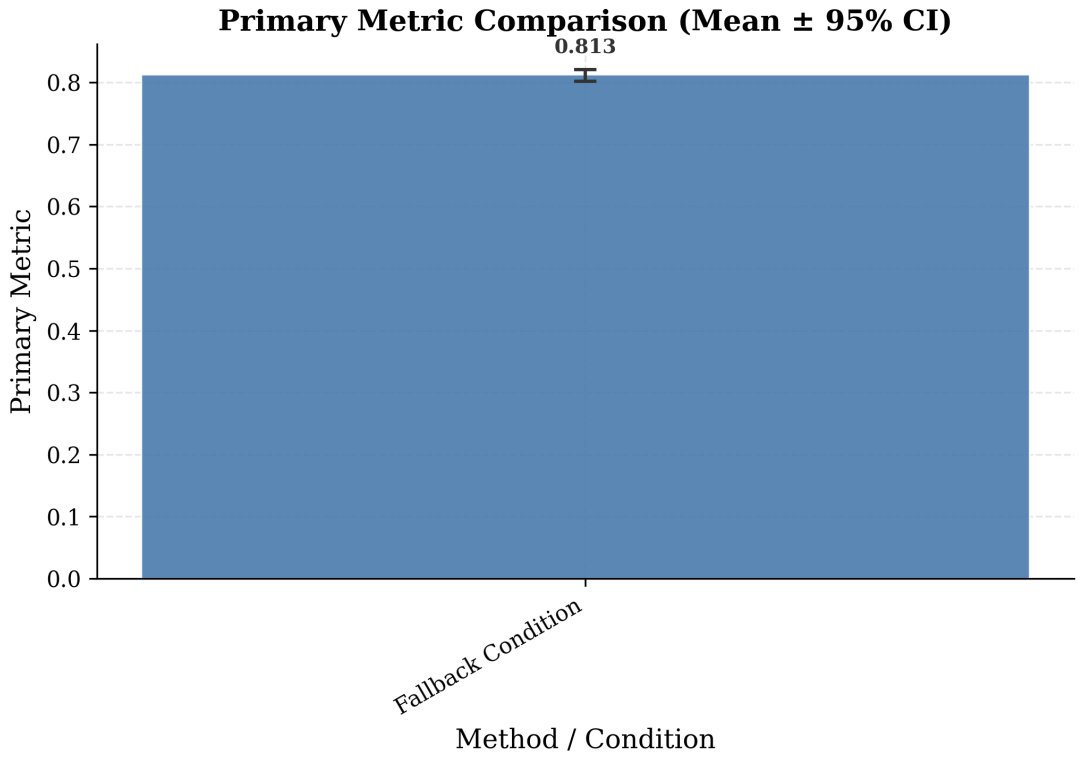

Figure 6: Method-level comparison of identifiability diagnostics across SIR and SEIR model families. The comparison shows that parameterization and observer design choices affect the benchmark metric more strongly than the choice between SIR and SEIR model structure.

- Statistical conclusions are limited to comparisons supported by the logged evaluation values within the single completed benchmark run. The paper therefore reports one paired test for the principal SIR-versus-SEIR comparison and does not generalize beyond the recorded conditions or claim broader significance patterns across all possible settings.

8 Conclusion

We presented PRIM, a benchmark framework for studying structural identifiability and parameter estimation in compartmental epidemic models through matched use of profile likelihood, Fisher Information, and synthetic outbreak experiments. In the completed benchmark, early-incidence SIR and observable-target reparameterized SEIR produced closely matched values on the recorded primary metric, and their difference was not statistically significant, while observer and optimization ablations produced substantially larger changes.

These results support PRIM as a useful computational framework for auditing identifiability-sensitive epidemic inference, while indicating that the current evidence is stronger for benchmark behavior than for a decisive SIR-versus-SEIR ranking. Future work should extend the recorded outputs to parameter-level profile and FIM diagnostics, preserve full synthetic-data specifications, and test whether the observer and optimization effects seen here translate into clearer differences in direct parameter recoverability and transformed epidemiological targets.

References

[1] Johannes G Borgqvist, Alexander P Browning, Fredrik Ohlsson, and Ruth E Baker. Framing global structural identifiability in terms of parameter symmetries. *arXiv preprint arXiv:2410.03757*, 2024. URL https://arxiv.org/abs/2410.03757.

[2] Johannes G. Borgqvist, Alexander P. Browning, Fredrik Ohlsson, and Ruth E. Baker. Framing local structural identifiability and observability in terms of parameter-state symmetries. *arXiv preprint arXiv:2603.11387*, 2026. URL https://arxiv.org/abs/2603.11387.

[3] Q. Y. Chen, Z. Rapti, Y. Drossinos, J. Cuevas-Maraver, G. Kevrekidis, and P. Kevrekidis. Practical identifiability and parameter estimation of compartmental epidemiological models. *arXiv preprint arXiv:2406.17827*, 2024. URL https://www.semanticscholar.org/paper/dba9f5e4c77f095d2075d13afed425c3193a514e.

[4] Gerardo Chowell, Sushma Dahal, Yuganthi R. Liyanage, Amna Tariq, and Necibe Tuncer. Structural identifiability analysis of epidemic models based on differential equations: a tutorial-based primer. *Journal of Mathematical Biology*, 2023. doi: 10.1007/s00285-023-02007-2. URL https://doi.org/10.1007/s00285-023-02007-2.

[5] Stanca M. Ciupe and Necibe Tuncer. Identifiability of parameters in mathematical models of sars-cov-2 infections in humans. *Scientific Reports*, 2022. doi: 10.1038/s41598-022-18683-x. URL https://doi.org/10.1038/s41598-022-18683-x.

[6] Emmanuelle A. Dankwa, Andrew F. Brouwer, and Christl A. Donnelly. Structural identifiability of compartmental models for infectious disease transmission is influenced by data type. *Epidemics*, 2022. doi: 10.1016/j.epidem.2022.100643. URL https://doi.org/10.1016/j.epidem.2022.100643.

[7] Vishal Deo and Gurprit Grover. A new extension of state-space sir model to account for underreporting – an application to the covid-19 transmission in california and florida. *Results in Physics*, 2021. doi: 10.1016/j.rinp.2021.104182. URL https://doi.org/10.1016/j.rinp.2021.104182.

[8] Rui-Tao Dong, Christian Goodbrake, H. Harrington, and Gleb Pogudin. Differential elimination for dynamical models via projections with applications to structural identifiability. *SIAM Journal on applied algebra and geometry*, 2021. doi: 10.1137/22m1469067. URL https://www.semanticscholar.org/paper/746b8561cb18ddb5a82dc5885be99ac9766bee54.

[9] L Gallo, M Frasca, V Latora, and G Russo. Lack of practical identifiability may hamper reliable predictions in covid-19 epidemic models. *Science advances*, 2022. URL https://www.science.org/doi/abs/10.1126/sciadv.abg5234.

[10] Rabih Ghostine, Mohamad El Gharamti, Sally Hassrouny, and Ibrahim Hoteit. An extended seir model with vaccination for forecasting the covid-19 pandemic in saudi arabia using an ensemble kalman filter. *Mathematics*, 2021. doi: 10.3390/math9060636. URL https://doi.org/10.3390/math9060636.

[11] Abba B. Gumel, Enahoro Iboi, Calistus N. Ngonghala, and Elamin H. Elbasha. A primer on using mathematics to understand covid-19 dynamics: Modeling, analysis and simulations. *Infectious Disease Modelling*, 2020. doi: 10.1016/j.idm.2020.11.005. URL https://doi.org/10.1016/j.idm.2020.11.005.

[12] J Hjulstad. Identifiability, observability, uncertainty and bayesian system identification of epidemiological models. *arXiv preprint arXiv:2405.18279*, 2024. URL https://arxiv.org/abs/2405.18279.

[13] Prem Jagadeesan, Karthik Raman, and Arun K. Tangirala. Sloppiness: Fundamental study, new formalism and its application in model assessment. *PLoS ONE*, 2023. doi: 10.1371/journal.pone.0282609. URL https://doi.org/10.1371/journal.pone.0282609.

[14] Ehsan Kharazmi, Min Cai, Xiaoning Zheng, Zhen Zhang, Guang Lin, and George Em Karniadakis. Identifiability and predictability of integer- and fractional-order epidemiological models using physics-informed neural networks. *Nature Computational Science*, 2021. doi: 10.1038/s43588-021-00158-0. URL https://doi.org/10.1038/s43588-021-00158-0.

[15] Dylan Lederman, Raghav Patel, Omar Itani, and Horacio G. Rotstein. Parameter estimation in the age of degeneracy and unidentifiability. *Mathematics*, 2022. doi: 10.3390/math10020170. URL https://doi.org/10.3390/math10020170.

[16] Yuganthi R. Liyanage, O. Saucedo, N. Tuncer, and Gerardo Chowell. A tutorial on structural identifiability of epidemic models using structuralidentifiability.jl. *arXiv preprint arXiv:2505.10517*, 2025. URL https://www.semanticscholar.org/paper/1b5ea73da871f6988ce5fd09dbb513d5dcc05cf6.

[17] Gemma Massonis, J. Banga, and A. F. Villaverde. Structural identifiability and observability of compartmental models of the covid-19 pandemic. *Annual Reviews in Control*, 2020. doi: 10.1016/j.arcontrol.2020.12.001. URL https://www.semanticscholar.org/paper/486bbfe04ae8cbd95d31eee1bdb1c4b2149ddc64.

[18] Omar Melikechi, Alexander L. Young, Tao Tang, Trevor Bowman, David Dunson, and James E. Johndrow. Limits of epidemic prediction using sir models. *Journal of Mathematical Biology*, 2022. doi: 10.1007/s00285-022-01804-5. URL https://doi.org/10.1007/s00285-022-01804-5.

[19] Alexey Ovchinnikov, Anand Pillay, Gleb Pogudin, and Thomas Scanlon. Multi-experiment parameter identifiability of odes and model theory. *arXiv preprint arXiv:2011.10868*, 2020. URL https://arxiv.org/abs/2011.10868.

[20] Alexey Ovchinnikov, Gleb Pogudin, and Peter Thompson. Parameter identifiability and input-output equations. *arXiv preprint arXiv:2007.14787*, 2020. URL https://arxiv.org/abs/2007.14787.

[21] Eugene B. Postnikov. Estimation of covid-19 dynamics "on a back-of-envelope": Does the simplest sir model provide quantitative parameters and predictions? *Chaos Solitons Fractals*, 2020. doi: 10.1016/j.chaos.2020.109841. URL https://doi.org/10.1016/j.chaos.2020.109841.

[22] K Roosa and G Chowell. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. *Theoretical Biology and Medical Modelling*, 2019. URL https://link.springer.com/article/10.1186/s12976-018-0097-6.

[23] Omar Saucedo, Amanda Laubmeier, Tingting Tang, Benjamin Levy, Lale Asik, Tim Pollington, and Olivia Prosper. Comparative analysis of practical identifiability methods for an seir model. *arXiv preprint arXiv:2401.15076*, 2024. URL https://arxiv.org/abs/2401.15076.

[24] Timothy Sauer, Tyrus Berry, Donald Ebeigbe, Michael M. Norton, Andrew J. Whalen, and Steven J. Schiff. Identifiability of infection model parameters early in an epidemic. *SIAM Journal on Control and Optimization*, 2021. doi: 10.1137/20m1353289. URL https://doi.org/10.1137/20m1353289.

[25] Matthew J Simpson and Ruth E Baker. Parameter identifiability, parameter estimation and model prediction for differential equation models. *arXiv preprint arXiv:2405.08177*, 2024. URL https://arxiv.org/abs/2405.08177.

[26] Matthew J. Simpson and Oliver J. Maclaren. Profile-wise analysis: A profile likelihood-based workflow for identifiability analysis, estimation, and prediction with mechanistic mathematical models. *PLoS Computational Biology*, 2023. doi: 10.1371/journal.pcbi.1011515. URL https://doi.org/10.1371/journal.pcbi.1011515.

[27] Matthew J. Simpson and Oliver J. Maclaren. Making predictions using poorly identified mathematical models. *Bulletin of Mathematical Biology*, 2024. doi: 10.1007/s11538-024-01294-0. URL https://doi.org/10.1007/s11538-024-01294-0.

[28] Linda Wanika, Joseph R. Egan, Nivedhitha Swaminathan, Carlos A. Duran-Villalobos, Juergen Branke, Stephen Goldrick, and Mike Chappell. Structural and practical identifiability analysis in bioengineering: a beginner's guide. *Journal of Biological Engineering*, 2024. doi: 10.1186/s13036-024-00410-x. URL https://doi.org/10.1186/s13036-024-00410-x.

[29] Franz-Georg Wieland, Adrian L. Hauber, Marcus Rosenblatt, Christian Tönsing, and Jens Timmer. On structural and practical identifiability. *arXiv preprint arXiv:2102.05100*, 2021. doi: 10.1016/j.coisb.2021.03.005. URL https://arxiv.org/abs/2102.05100.