

# Simulate the Effect of Detection Uncertainty on Prevalence Estimation Using a Mechanistic and Bayesian Approach

Simulate the Effect of Detection Uncertainty on Prevalence Estimation Using a  
Mechanistic and Bayesian Approach

Kulig, Schäfer, Addo, Wilczek, Lange, Hensel, Jung

2026-10-05

## Table of contents

|          |                                                          |          |
|----------|----------------------------------------------------------|----------|
| <b>1</b> | <b>Methodological Setup of the Current Analysis</b>      | <b>3</b> |
| 1.1      | Scenario Setup                                           | 3        |
| 1.2      | Global Simulation Parameters                             | 3        |
| 1.3      | Diagnostic Target Performance                            | 3        |
| <b>2</b> | <b>Descriptive Statistics</b>                            | <b>5</b> |
| 2.1      | Pre-Flight Check: Parameter Synchronization              | 5        |
| 2.2      | Free and Tuned Model Parameters                          | 5        |
| 2.3      | Visualizing Diagnostic Assumptions (Sensitivity)         | 5        |
| 2.4      | Assessment of the Wasserstein Distance                   | 7        |
| <b>3</b> | <b>Analytical Statistics</b>                             | <b>9</b> |
| 3.1      | Unpacking the Mechanistic Transfer Function              | 9        |
| 3.2      | Unpacking the Bayesian Framework                         | 9        |
| 3.3      | Impact of Epistemic Uncertainty on Interval Estimation   | 11       |
| 3.4      | Density Overlay: Naive Baseline vs. Robust Bootstrapping | 11       |

|          |                                                                          |           |
|----------|--------------------------------------------------------------------------|-----------|
| <b>4</b> | <b>Interpretation and Consequences</b>                                   | <b>12</b> |
| 4.1      | The Precision Trap of the Naive Baseline . . . . .                       | 12        |
| 4.2      | Mechanistic Correction . . . . .                                         | 12        |
| 4.3      | Bayesian Stabilization . . . . .                                         | 13        |
| 4.4      | Model Synchronization and the Role of the Wasserstein Distance . . . . . | 13        |
| 4.5      | The Role of Cluster Effects . . . . .                                    | 14        |
| 4.6      | Consequences for Sample Size Planning . . . . .                          | 14        |
| 4.7      | Limitations . . . . .                                                    | 14        |
| <b>5</b> | <b>References</b>                                                        | <b>16</b> |

# 1 Methodological Setup of the Current Analysis

This section outlines the foundational parameters governing the simulated field visit and the statistical boundaries applied in the bootstrap workflows.

## 1.1 Scenario Setup

- **Total Population ( $N$ ):** 5000 animals
- **Sample Size ( $n$ ):** 100 animals sampled
- **Cluster Structure ( $k$ ):** 10 Clusters (Derived cluster size  $m \approx 10$  animals/cluster)
- **Observed Cases in the Field:** 14 (Apparent Prevalence = 14 %)
- **Auto-Tuning (Pre-Flight Synchronization): Active**
  - **Direction:** Bayesian prior fitted to the mechanistic model (`bayes_to_mech`)
  - **Matching criterion:** Location only (expected sensitivity)
  - **Free (reference) model:** Mechanistic –  $\mu_s = 2.70$ ,  $\sigma_s = 0.60$ ,  $s_0 = 2.50$ ,  $k = 3.000$  (effective  $Se$ : mean 59.56 %, SD 29.12 %; certainty for the Rogan-Gladen correction: 100 pseudo-animals)
  - **Tuned model:** Bayesian – adjusted parameters: `Se_alpha`, `Se_beta` (prior strength `alpha` + `beta` kept constant)

## 1.2 Global Simulation Parameters

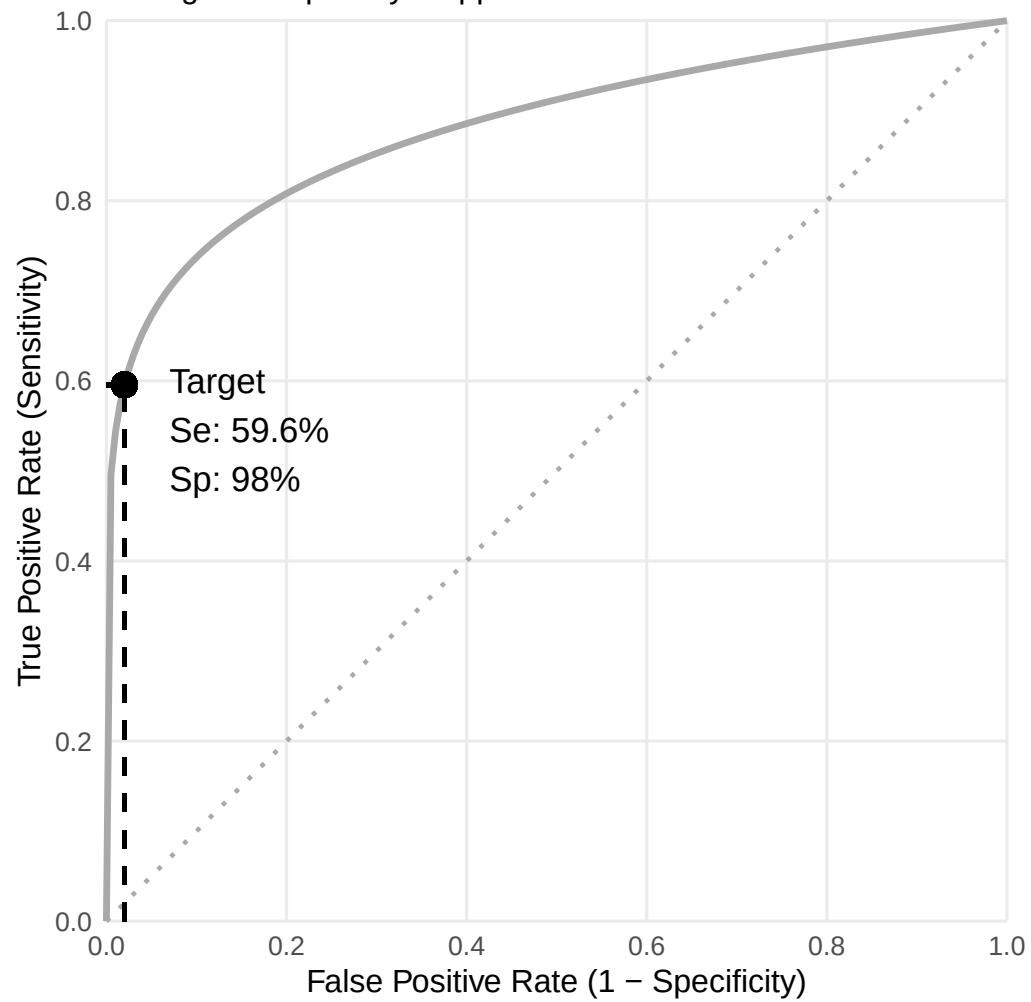
- **Targeted Confidence / Credible Level:** 95 %
- **Bootstrap Iterations ( $B$ ):** 10,000 Resampling steps

## 1.3 Diagnostic Target Performance

The target diagnostic capability configured for this simulation is visualized in the Receiver Operating Characteristic (ROC) space below. The point indicates the expected relationship between the True Positive Rate (Sensitivity) and the False Positive Rate ( $1 - \text{Specificity}$ ) before the variance of cluster effects is introduced.

## Diagnostic Target (ROC Space)

Configured capability mapped to an idealized ROC curve



## 2 Descriptive Statistics

### 2.1 Pre-Flight Check: Parameter Synchronization

To prevent “Apples-to-Oranges” comparisons between the mechanistic and Bayesian frameworks, the underlying diagnostic capabilities of both models must be synchronized.

The mechanistic framework derives its *Effective Sensitivity* dynamically from the biological expressivity of the disease. The Bayesian framework, on the other hand, relies on a pre-defined *Prior Sensitivity*.

#### Auto-Tuning Active

The models have been mathematically synchronized prior to execution.

- **Direction:** Bayesian prior fitted to the mechanistic model (`bayes_to_mech`)
- **Matching criterion:** Location only (expected sensitivity)
- **Free (reference) model:** Mechanistic; **tuned model:** Bayesian
- **Optimizer:** Closed-form solution (mean matched, prior strength kept).

### 2.2 Free and Tuned Model Parameters

The table lists the parameters of both models as configured (*before*) and as used in the bootstrap simulation (*after*). The rows of the free (reference) model remain unchanged; adjusted rows are printed in bold. Input parameters are set in the configuration file, derived quantities are the resulting mean and standard deviation of the sensitivity distribution.

### 2.3 Visualizing Diagnostic Assumptions (Sensitivity)

The following table performs a diagnostic cross-check to verify whether both models expect the same detection rate (location) and the same uncertainty about it (spread) before the bootstrap simulation starts. All values are computed exactly from the parametric distributions (logit-normal for the mechanistic model, Beta for the Bayesian prior).

# Model Parameters Before and After Auto-Tuning

Matching criterion: mean | direction: bayes\_to\_mech

| Parameter                                                  | Type           | Configured (before) | Simulation (after) | Difference    |
|------------------------------------------------------------|----------------|---------------------|--------------------|---------------|
| Mechanistic (free model)                                   |                |                     |                    |               |
| Symptom mean (s_mean)                                      | input          | 2.700               | 2.700              | 0.000         |
| Symptom SD (s_sd)                                          | input          | 0.600               | 0.600              | 0.000         |
| Threshold (s0)                                             | input          | 2.500               | 2.500              | 0.000         |
| Steepness (k)                                              | input          | 3.000               | 3.000              | 0.000         |
| Effective Se: mean (%)                                     | derived        | 59.557              | 59.557             | 0.000         |
| Effective Se: SD (%)                                       | derived        | 29.123              | 29.123             | 0.000         |
| Certainty of Se_eff for the RG correction (pseudo-animals) | input          | 100.000             | 100.000            | 0.000         |
| Bayesian (tuned model)                                     |                |                     |                    |               |
| <b>Se prior alpha</b>                                      | <b>input</b>   | <b>60.000</b>       | <b>59.557</b>      | <b>-0.443</b> |
| <b>Se prior beta</b>                                       | <b>input</b>   | <b>40.000</b>       | <b>40.443</b>      | <b>0.443</b>  |
| Prior strength (alpha + beta)                              | derived        | 100.000             | 100.000            | 0.000         |
| <b>Se prior: mean (%)</b>                                  | <b>derived</b> | <b>60.000</b>       | <b>59.557</b>      | <b>-0.443</b> |
| <b>Se prior: SD (%)</b>                                    | <b>derived</b> | <b>4.875</b>        | <b>4.883</b>       | <b>0.009</b>  |
| Shared (both models)                                       |                |                     |                    |               |
| Sp prior alpha                                             | input          | 98.000              | 98.000             | 0.000         |
| Sp prior beta                                              | input          | 2.000               | 2.000              | 0.000         |
| Sp prior: mean (%)                                         | derived        | 98.000              | 98.000             | 0.000         |
| Sp prior: SD (%)                                           | derived        | 1.393               | 1.393              | 0.000         |

Even if location and spread agree, the two distributions can still differ in their shape (e.g. skewness or bimodality).

- **Mechanistic Sensitivity (Green):** This is an *emergent distribution*. It arises because the true symptom severity varies among the diseased animals, leading to varying individual detection probabilities. With normally distributed expressivity and a logistic detection function, it is exactly a logit-normal distribution (Aitchison & Shen, 1980).
- **Bayesian Prior (Red):** This represents the prior belief about the observer’s overall capability, modeled via a Beta distribution.

We quantify the overlap of both distributions with the **Wasserstein distance** ( $W_1$ ) (Panaretos & Zemel, 2019; Villani, 2009). In one dimension,  $W_1$  equals the average absolute distance between corresponding quantiles of the two distributions,  $W_1 = \int_0^1 |F_{Mech}^{-1}(u) - F_{Prior}^{-1}(u)| du$ . It is therefore measured on the same scale as the sensitivity itself:  $W_1 = 0.01$  means that the quantiles differ on average by one percentage point. A value close to 0 indicates an “Apples-to-Apples” scenario.

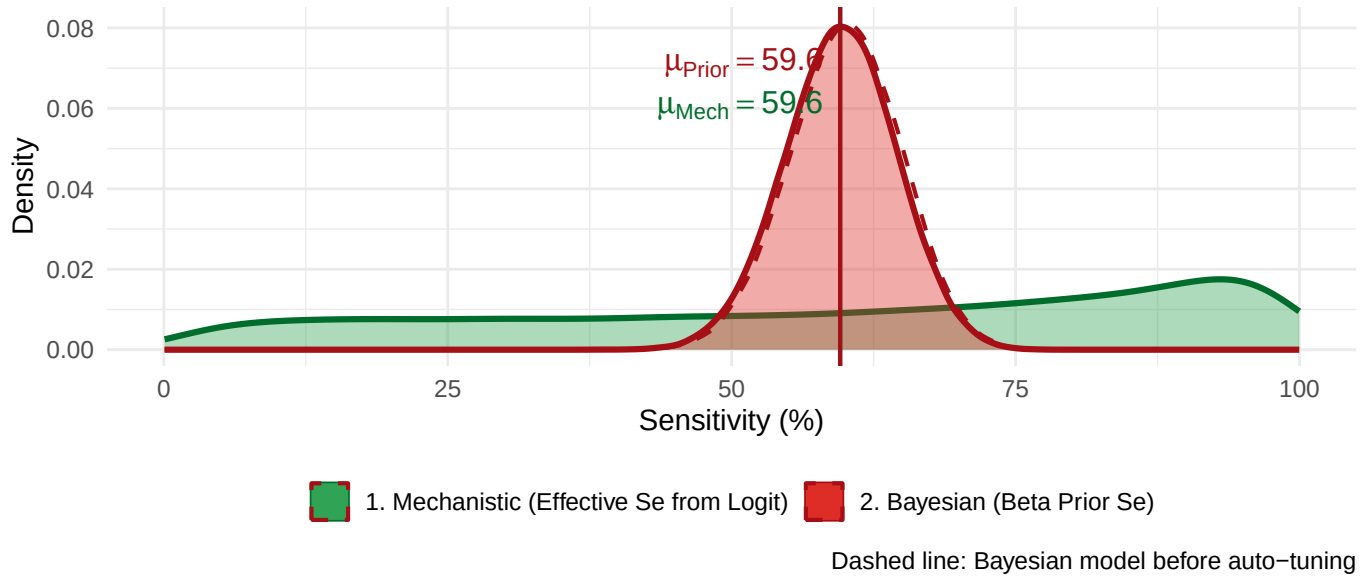
## Diagnostic Alignment Check

Comparing location and spread of the sensitivity assumptions

| Diagnostic Metric                      | Mechanistic Model (%) | Bayesian Prior (%) | Absolute Diff. (pp) |
|----------------------------------------|-----------------------|--------------------|---------------------|
| Expected Diagnostic Sensitivity (Mean) | 59.56                 | 59.56              | 0.00                |
| Spread of Sensitivity (SD)             | 29.12                 | 4.88               | 24.24               |

## Comparison of Diagnostic Sensitivity Assumptions

Wasserstein distance  $W_1 = 0.2144$  (Poor agreement); before tuning: 0.2140



### Resulting Distribution Characteristics:

- **Mechanistic Distribution:** Mean = 59.56 %, Standard Deviation = 29.12 %
- **Bayesian Prior Distribution:** Mean = 59.56 %, Standard Deviation = 4.88 %
- **Shape Alignment (Wasserstein  $W_1$ ):** 0.2144

## 2.4 Assessment of the Wasserstein Distance

The numerical value of  $W_1$  is assessed on the probability scale with the thresholds defined in the configuration (excellent  $\leq 1.0$  pp, good  $\leq 2.5$  pp, acceptable  $\leq 5.0$  pp, poor above). These thresholds are pragmatic conventions of this framework, not universal standards: a difference of one percentage point in the assumed sensitivity changes a Rogan-Gladen prevalence estimate only marginally compared to the sampling error of typical herd studies.

To see *why* two distributions differ, the squared quadratic Wasserstein distance is split exactly into a location, a scale and a shape component (Panaretos & Zemel, 2019):

$$W_2^2 = \underbrace{(\mu_{Mech} - \mu_{Prior})^2}_{\text{location}} + \underbrace{(\sigma_{Mech} - \sigma_{Prior})^2}_{\text{scale}} + \underbrace{2 \sigma_{Mech} \sigma_{Prior} (1 - \rho_q)}_{\text{shape}},$$

where  $\rho_q$  is the correlation between the two quantile functions. Mean-only tuning removes the location component, moment tuning additionally removes the scale component, and only the shape component remains.

## Assessment of Model Synchronization

Distances in percentage points; components as share of  $W_2$  squared

|                                   | W1 (pp) | W2 (pp) | Location (%) | Scale (%) | Shape (%) | Rating                |
|-----------------------------------|---------|---------|--------------|-----------|-----------|-----------------------|
| Configured (before tuning)        | 21.40   | 24.44   | 0.0          | 98.4      | 1.5       | Poor agreement        |
| Used in simulation (after tuning) | 21.44   | 24.43   | 0.0          | 98.5      | 1.5       | <b>Poor agreement</b> |

$W_1 = 0.2144$ : Poor agreement

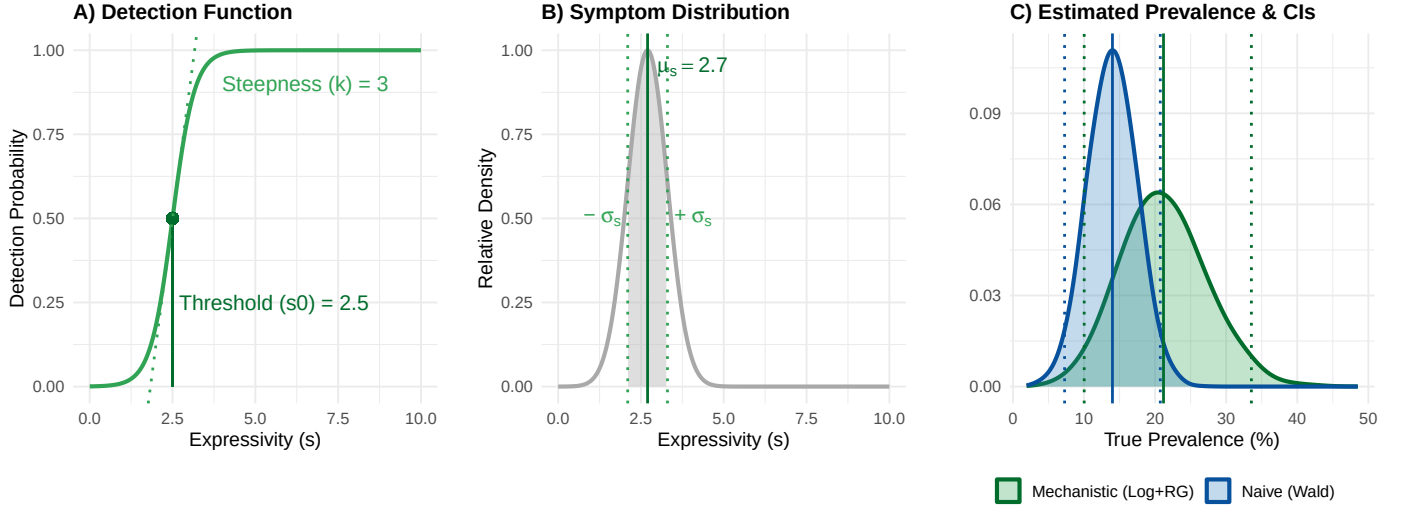
$W_1$  exceeds 5.0 pp. The two frameworks assume substantially different diagnostic capabilities. Any difference between their intervals is confounded with this mismatch, so the comparison is not apples-to-apples. Auto-tuning changed  $W_1$  from 21.40 pp (poor agreement) to 21.44 pp. The remaining difference is mainly due to a different width of the distributions (scale) (98.5 % of  $W_2^2$ ).

**Conceptual caveat.** The two distributions do not describe the same kind of uncertainty. The mechanistic distribution describes the *variability between animals* (aleatoric: some animals show strong symptoms and are almost always detected, others are almost never detected), while the Beta prior describes the *uncertainty about the herd-level sensitivity* (epistemic). Matching the means is always justified, because the effective sensitivity is the expected individual detection probability. Matching the spread or the entire distribution equates both kinds of uncertainty and can therefore lead to extreme parameter values (e.g. an almost flat detection function or a U-shaped Beta prior). The criteria **moments** and **wasserstein** should therefore be interpreted as a sensitivity analysis rather than as the default calibration (see Section 4.4). The tuning never alters the Rogan-Gladen correction of the mechanistic model directly: it uses its own epistemic distribution centred on  $Se_{eff}$  (Section 4.2), so a U-shaped Bayesian prior cannot break the mechanistic interval.

### 3 Analytical Statistics

#### 3.1 Unpacking the Mechanistic Transfer Function

The mechanistic model derives the effective field sensitivity dynamically through the intersection of a technological threshold (Plot A) and the biological symptom distribution (Plot B). Plot C demonstrates the resulting interval expansion via the Rogan-Gladen core compared to the naive Wald baseline.

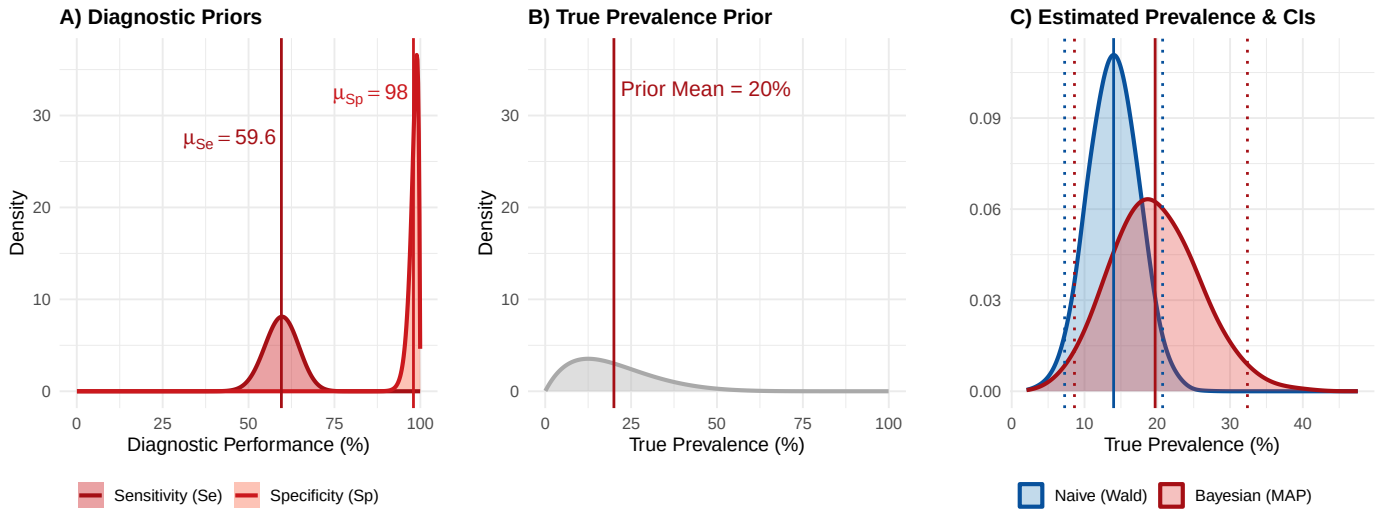


Parameters governing the mechanistic uncertainty:

- **Clinical Expressivity (Biology):** Mean  $\mu_s = 2.70$ , Standard Deviation  $\sigma_s = 0.60$
- **Detection System (Observer):** Threshold  $s_0 = 2.50$ , Steepness  $k = 3.000$
- **Resulting Effective Sensitivity:**  $Se_{eff} = 59.56\%$
- **Sensitivity used in the Rogan-Gladen correction:**  $Se \sim \text{Beta}(59.56, 40.44)$ , centred on  $Se_{eff}$  with a certainty of 100 pseudo-animals (independent of the Bayesian prior)
- **Specificity (Noise/False Alarms):**  $Sp \approx 98\%$

#### 3.2 Unpacking the Bayesian Framework

The Bayesian framework incorporates epistemic uncertainty directly through Beta priors for the diagnostic parameters (Plot A) and the true herd prevalence (Plot B). Plot C shows the distribution of the Maximum A Posteriori (MAP) estimates across the bootstrap iterations.



**Parameters governing the Bayesian uncertainty** (after auto-tuning, see Section 2.2):

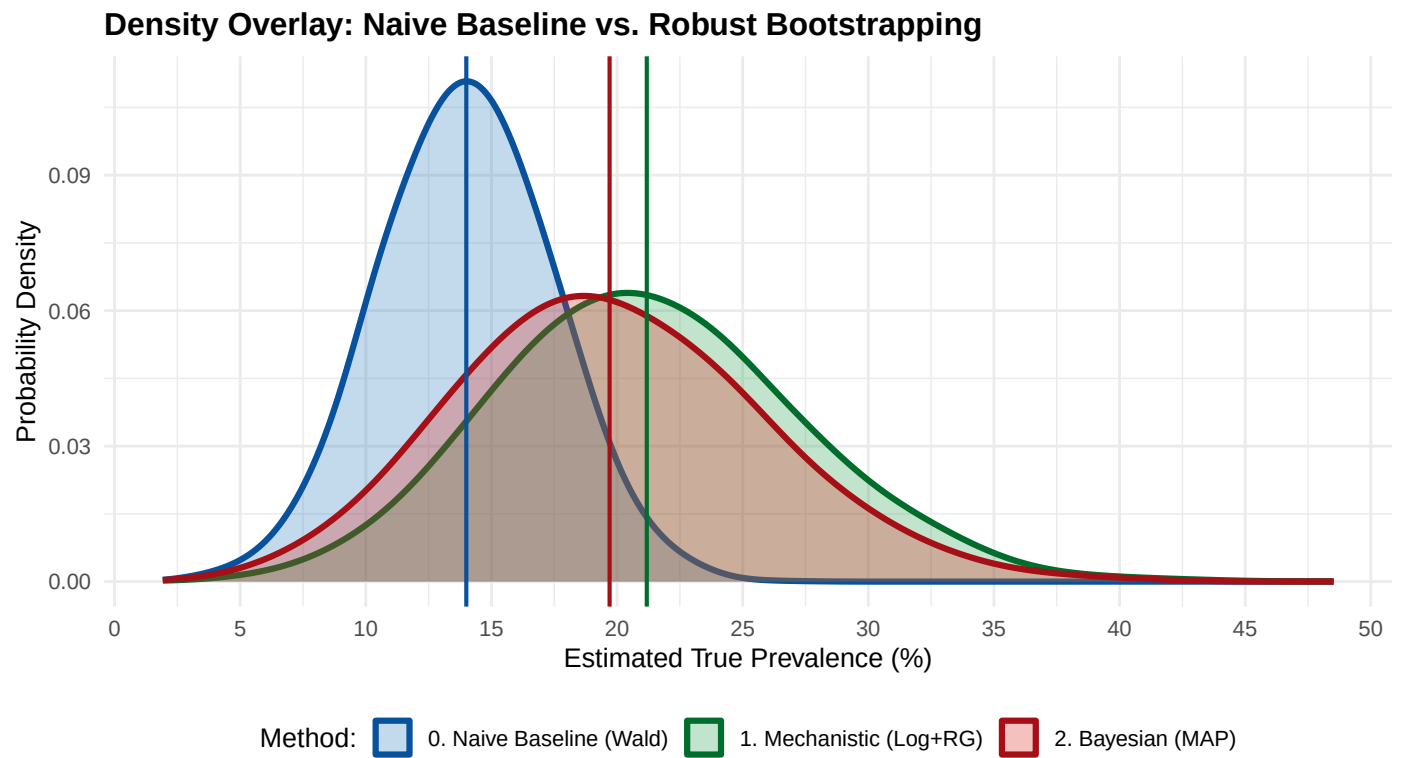
- **Diagnostic Priors:** Sensitivity  $\sim \text{Beta}(59.56, 40.44)$ , Specificity  $\sim \text{Beta}(98, 2)$
- **Herd Prevalence Prior:** True Prevalence  $\sim \text{Beta}(2, 8) \rightarrow \text{Prior Mean} = 20\%$
- **Resulting Bayesian MAP Estimate:**  $\hat{\pi}_{MAP} = 19.7\%$

### 3.3 Impact of Epistemic Uncertainty on Interval Estimation

The resulting interval estimations are quantified below, comparing the error-free assumption with the two uncertainty-aware frameworks.

### 3.4 Density Overlay: Naive Baseline vs. Robust Bootstrapping

The final density plot overlays all three framework pathways, highlighting the massive structural differences in parameter spread when cluster effects and diagnostic uncertainties are holistically integrated.



### *Impact of Epistemic Uncertainty on Interval Estimation*

Targeted Confidence / Credible Level: 95 %

| Applied Framework Method                 | Estimated Prev. (%) | Lower Bound (%) | Upper Bound (%) | Interval Width (%) |
|------------------------------------------|---------------------|-----------------|-----------------|--------------------|
| 0. Naive Baseline (Wald/FPC)             | 14.00               | 7.27            | 20.73           | 13.47              |
| 1. Mechanistic (Logistic + Rogan-Gladen) | 21.18               | 10.03           | 33.55           | 23.52              |
| 2. Bayesian (Bootstrap-MAP)              | 19.70               | 8.62            | 32.39           | 23.77              |

## 4 Interpretation and Consequences

The comparative analysis shows how strongly the uncertainty of a prevalence estimate is underestimated when classical frequentist methods are naively applied to error-prone veterinary field data. The following sections explain the mechanisms behind the differences between the three intervals in Table “Impact of Epistemic Uncertainty on Interval Estimation” and derive consequences for sample size planning. All numbers refer to the current configuration.

### 4.1 The Precision Trap of the Naive Baseline

The naive Wald interval has a width of 13.47 pp. It rests on two assumptions that do not hold in the field. First, it assumes a perfect classification ( $Se = Sp = 1$ ) and therefore estimates the *apparent* prevalence (14.00 %), not the true prevalence (Greiner & Gardner, 2000; Rogan & Gladen, 1978). Second, the Wald interval itself is known to have erratic and often too low coverage for small samples and for proportions close to 0 or 1; score or exact intervals are preferable even when classification is perfect (Brown et al., 2001). The finite population correction only accounts for sampling without replacement from a herd of  $N = 5000$  animals; it does not address misclassification or clustering.

The resulting interval is narrow because every source of uncertainty except binomial sampling error is set to zero. It is mathematically correct within its axioms but describes a precision that a field study cannot achieve.

### 4.2 Mechanistic Correction

The mechanistic pathway combines two building blocks.

**Rogan-Gladen correction.** The true prevalence is obtained from the apparent prevalence by  $\hat{\pi} = (p_{app} + Sp - 1)/(Se + Sp - 1)$  (Rogan & Gladen, 1978). The denominator is Youden’s index  $J = Se + Sp - 1$ , here  $J = 0.576$  (point estimate  $\hat{\pi} = 20.85\%$ ). Because the standard error of  $p_{app}$  is divided by  $J$ , the sampling uncertainty alone is inflated by the factor  $1/J = 1.74$  (Greiner & Gardner, 2000). A weak diagnostic test therefore amplifies the sampling error, and estimates outside  $[0, 1]$  have to be truncated, which creates point masses at the boundaries and asymmetric percentile intervals (Lang & Reiczigel, 2014).

**Uncertainty about Se and Sp.** The mechanistic pathway separates two kinds of uncertainty. The variability *between animals* (aleatoric) is generated by the logistic detection function when the test results are simulated. For the correction, only the *uncertainty about the mean* sensitivity  $Se_{eff}$  (epistemic) matters. In each bootstrap iteration,  $Se$  is therefore drawn from Beta(59.56, 40.44), a distribution centred on  $Se_{eff} = 59.56\%$  whose certainty corresponds to a validation study with 100 diseased animals, and  $Sp$  is drawn from its Beta prior. The mechanistic pathway is thus independent of the Bayesian sensitivity prior; both frameworks share only the specificity assumption. Unlike the sampling error, this component does not decrease with larger sample sizes (Lang & Reiczigel, 2014): with an infinitely large sample, the Rogan-Gladen interval would still be 10.87 pp wide (Section 4.6).

**Logistic transfer function.** The detection probability of an individual animal depends on its symptom expressivity. With normally distributed expressivity, the individual detection probabilities follow a logit-normal distribution (Aitchison & Shen, 1980), whose mean is the effective sensitivity. Animals with weak symptoms are systematically missed. The “true prevalence” therefore always refers to the case definition implied by the expressivity scale.

Overall, the mechanistic interval is 23.52 pp wide, i.e. 1.75 times as wide as the naive interval.

### 4.3 Bayesian Stabilization

The Bayesian pathway combines a binomial likelihood for the number of test-positive animals,  $x \sim \text{Bin}(n, \pi Se + (1 - \pi)(1 - Sp))$ , with Beta priors for  $Se$ ,  $Sp$  and  $\pi$  (Branscum et al., 2005; Joseph et al., 1995; Lewis & Torgerson, 2012). Instead of truncating impossible estimates, the prior restricts  $\pi$  to  $[0, 1]$  and pulls estimates from the boundary in a continuous manner, which avoids the truncation artefacts of the Rogan-Gladen estimator (Gelman et al., 2013).

Two properties have to be kept in mind when interpreting the result:

- **Prior influence.** The prevalence prior  $\text{Beta}(2, 8)$  has a mean of 20.0 % and carries the information of about 10 pseudo-observations, compared with  $n = 100$  animals in the sample. The data dominate, but the estimate is pulled noticeably towards the prior mean when the observed prevalence is far from it.
- **Non-identifiability.** With a single test applied to a single population,  $Se$ ,  $Sp$  and  $\pi$  cannot all be estimated from the data. The data only inform the apparent prevalence; the split into true prevalence and misclassification is determined by the priors on  $Se$  and  $Sp$  (Joseph et al., 1995; Lewis & Torgerson, 2012). The result is therefore conditional on these priors and should be accompanied by a prior sensitivity analysis.

**Implementation note.** The present tool approximates the posterior uncertainty by bootstrap resampling of clusters combined with draws of  $Se$  and  $Sp$  and a MAP optimization in each iteration. This is not a full posterior from Markov chain Monte Carlo (MCMC) sampling. The spread of the MAP estimates does not contain the curvature of the posterior and may therefore deviate from a proper credible interval. For confirmatory analyses, a full MCMC implementation (e.g. JAGS or Stan) is recommended (Gelman et al., 2013; Lewis & Torgerson, 2012).

The Bayesian interval is 23.77 pp wide, i.e. 1.76 times as wide as the naive interval.

### 4.4 Model Synchronization and the Role of the Wasserstein Distance

A fair comparison of the mechanistic and the Bayesian pathway requires that both assume the same diagnostic capability; otherwise differences between the intervals are confounded with different input assumptions. The Wasserstein distance is the minimal “transport cost” between two probability distributions (Villani, 2009) and, in one dimension, has a direct interpretation as the average distance between quantiles (Panaretos & Zemel, 2019). It is therefore a natural, scale-preserving measure of model agreement.

In the current configuration, auto-tuning is active. Before tuning,  $W_1$  was 21.40 pp (poor agreement); in the simulation it is 21.44 pp (poor agreement). Section 2.4 decomposes the remaining difference into location, scale and shape.

The choice of the matching criterion is a methodological decision:

- **mean** aligns only the expected sensitivity. This is always defensible, because the effective sensitivity of the mechanistic model is by definition the mean individual detection probability. The distributions may still differ strongly in spread.
- **moments** and **wasserstein** additionally align the spread or the entire distribution. As explained in Section 2.4, the mechanistic distribution describes the heterogeneity between animals, whereas the Beta prior describes the uncertainty about a herd-level parameter. Equating both can require extreme parameter values. These criteria are best used as a sensitivity analysis showing how much of the difference between the intervals is due to the diagnostic assumptions.

## 4.5 The Role of Cluster Effects

Animals within the same pen or compartment share housing, management and infection pressure and are therefore more similar than animals from different clusters. This reduces the information content of the sample by the design effect  $\text{deff} = 1 + (m - 1)\rho$ , where  $m$  is the cluster size and  $\rho$  the intraclass correlation (Dohoo et al., 2003; Kish, 1965). The cluster bootstrap used in this tool resamples entire clusters and thereby preserves the within-cluster correlation (Davison & Hinkley, 1997; Field & Welsh, 2007). With only  $k = 10$  clusters, however, the bootstrap distribution is coarse, and bootstrap intervals based on few independent units tend to be less reliable (Davison & Hinkley, 1997).

The synthetic field data used here distribute the 14 positive animals randomly across the clusters, so any estimated intraclass correlation is a chance fluctuation of the random allocation ( $\hat{\rho} = 0.035$ ,  $\widehat{\text{deff}} = 1.31$ ), not a real cluster effect. In real herds, intraclass correlations of several percent are common. With  $m = 10$  animals per cluster, an intraclass correlation of  $\rho = 0.05$  already gives  $\text{deff} = 1.45$ , and  $\rho = 0.10$  gives  $\text{deff} = 1.90$ . The required sample size increases by the same factor. The intervals shown in this report are therefore a lower bound of the uncertainty that has to be expected under real cluster effects.

## 4.6 Consequences for Sample Size Planning

**1. Inflation of the nominal sample size.** For the sampling component, the interval width scales approximately with  $1/\sqrt{n}$ . A naive comparison of the widths would suggest inflating the sample size by the factor  $(W_{\text{Mech}}/W_{\text{Wald}})^2 = 3.05$ . This is too simple, because part of the width does not depend on  $n$ .

**2. Irreducible uncertainty.** If the apparent prevalence were known exactly ( $n \rightarrow \infty$ ), the uncertainty about  $Se$  and  $Sp$  would still produce an interval of 10.87 pp for the true prevalence (15.25 % to 26.13 %). Assuming that the sampling and the diagnostic variance components add up, the sample size needed to reach a target width  $w$  is approximately

$$n_{\text{req}} \approx n \cdot \frac{W_{\text{Mech}}^2 - W_{\infty}^2}{w^2 - W_{\infty}^2}, \quad w > W_{\infty}.$$

To reach the precision suggested by the naive Wald interval (13.47 pp) with the mechanistic model, about  $n_{\text{req}} \approx 690$  animals instead of 100 would be required (factor 6.90), before accounting for cluster effects.

**3. Improving the diagnostic information.** The irreducible width can only be reduced by more precise knowledge of  $Se$  and  $Sp$  (in this tool: a larger `se_eff_strength` and a stronger specificity prior), e.g. through a validation study with a reference standard or a latent class analysis with several tests (Branscum et al., 2005; Greiner & Gardner, 2000). A higher Youden index  $J$  additionally reduces the inflation factor  $1/J$  of the sampling error.

**4. Accounting for clustering.** The sample size must additionally be multiplied by the design effect. Increasing the number of clusters is more efficient than increasing the number of animals per cluster (Dohoo et al., 2003).

**5. Transparent reporting.** The assumptions on  $Se$ ,  $Sp$ , the prior on  $\pi$ , the auto-tuning settings and the cluster structure should be reported together with the estimate, because they determine a substantial part of the interval width.

## 4.7 Limitations

- The field data are synthetic and generated backwards from the observed number of cases. They contain no real cluster structure (Section 4.5).
- Specificity is assumed to be the same for all healthy animals and is not part of the auto-tuning. Conditional independence of the classification errors is assumed (Greiner & Gardner, 2000).

- The logistic detection function and the normal distribution of expressivity on a 0 to 10 scale are model assumptions that should be validated with field data.
- The Bayesian pathway is a bootstrap-MAP approximation, not a full MCMC posterior (Section 4.3).
- All bootstrap results are subject to Monte Carlo error, which decreases with the number of iterations ( $B = 10,000$ ). A fixed random seed (`set.seed()`) is recommended for reproducible reports.

## 5 References

The methodological foundations of this framework are based on the following literature (generated from `References.bib`, formatted with `apa.csl`):

- Aitchison, J., & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2), 261–272. <https://doi.org/10.2307/2335470>
- Branscum, A. J., Gardner, I. A., & Johnson, W. O. (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*, 68(2–4), 145–163. <https://doi.org/10.1016/j.prevetmed.2004.12.005>
- Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101–133. <https://doi.org/10.1214/ss/1009213286>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802843>
- Dohoo, I. R., Martin, W., & Stryhn, H. (2003). *Veterinary epidemiologic research*. AVC Inc.
- Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3), 369–390. <https://doi.org/10.1111/j.1467-9868.2007.00593.x>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC. <https://doi.org/10.1201/b16018>
- Greiner, M., & Gardner, I. A. (2000). Epidemiologic issues in the validation of veterinary diagnostic tests. *Preventive Veterinary Medicine*, 45(1–2), 3–22. [https://doi.org/10.1016/S0167-5877\(00\)00114-8](https://doi.org/10.1016/S0167-5877(00)00114-8)
- Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3), 263–272. <https://doi.org/10.1093/oxfordjournals.aje.a117428>
- Kish, L. (1965). *Survey sampling*. Wiley.
- Lang, Z., & Reiczigel, J. (2014). Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. *Preventive Veterinary Medicine*, 113(1), 13–22. <https://doi.org/10.1016/j.prevetmed.2013.09.015>
- Lewis, F. I., & Torgerson, P. R. (2012). A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerging Themes in Epidemiology*, 9, 9. <https://doi.org/10.1186/1742-7622-9-9>
- Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6, 405–431. <https://doi.org/10.1146/annurev-statistics-030718-104938>
- Rogan, W. J., & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1), 71–76. <https://doi.org/10.1093/oxfordjournals.aje.a112510>
- Villani, C. (2009). *Optimal transport: Old and new*. Springer. <https://doi.org/10.1007/978-3-540-71050-9>