# where are we…

**updates using pompjax, and the empirical prevalences from the literature.**

# ABM inferences

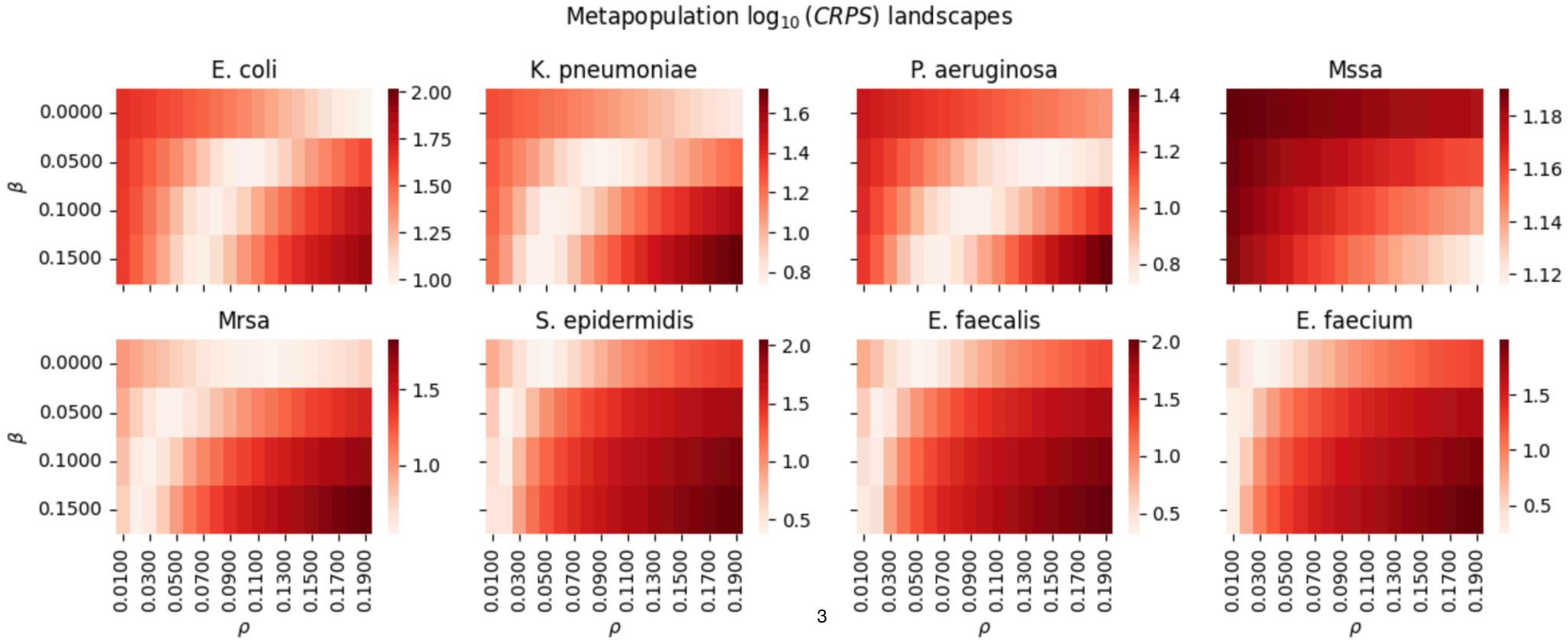
## recap

- Process model at daily time scales.
- Same study period as before 01/Feb/2020 to 28/Feb/2021 (weird drop in culture after feb 2021)
- we decided to drop *C. albicans* and include *E. faecalis,* also in the top 15 of microorganism detected.

# ABM inferences

## Inference on synthetic data.

- as the definition of $\rho$ is a little blurry and hard to find references I did some grid search and computed the fit to the hospital level data to ballpark the range for the different importation rates (AMRO prevalences).



# ABM inferences

## Inference on synthetic data.

- Using the ranges in the grid search, I selected values that already minimize the CRPS to the hospital-level observations and ran inferences on that.
  - I checked the convergence plots and the inferences seem to be working/converging.
- The empirical prevalences range from 0.4% (MRSA), to 60.3% (E. coli), so in general, we are covering a huge range of importations from all the AMRO to test in synthetic inferences.
  - Only MRSA and *P. aeruginosa* have importations lower than 15%
- **Results:** in general it seems that the inference is working very well (truths are falling in the posterior - even shrinking the variance). However, for low importation rates (lower than 15%) the system struggles in both $\beta$ and $\rho$.
  - Next slides shows the posterior for the 10 scenarios for all the AMROs. Prior ranges for the inference are shown as the limit of the x-axis.
- **Next:** If we want to report CI maybe I could remove the shrinking between iterations and see how those posterior look like.

# ABM inferences

## Shrinking vs Not shrinking

- It seems the not shrinking version is definitely working better, both captures more often the truth value and also provide a correct uncertainty.
  - Results differ from the ones Rami obtained with the ABM, high importations led to a more certain estimation of $\beta$ and low importations to a more uncertain estimation.
  - Note, that the limits in the axes are not the prior ranges so in general the posterior is also converging towards the truth even when biased.
- Convergences plots also look to be reaching asymptotically the truth value.
- Also, as the covariance estimated by the EAKF is conserved between iterations that relation is shown in the posterior estimate (not in the original IF implementation).

# Synthetic scenarios - identifiability

## Two setting of synthetic inferences

1. Inferences on truths tuple $(\rho, \beta)$ that reproduced the hospital-level observations of each microorganism as quantified with the Continuous Ranked Probability Score (CRPS) - grid searches in slide 3. (**Slide 7-15**)

2. Inferences with two levels of importations $\gamma = [25\,\%, 50\%]$, and different combinations of $\rho, \beta$. For a total of 24 ‘not constrained’ (to the hospital level) synthetic inferences. (**Slide 16-15**)

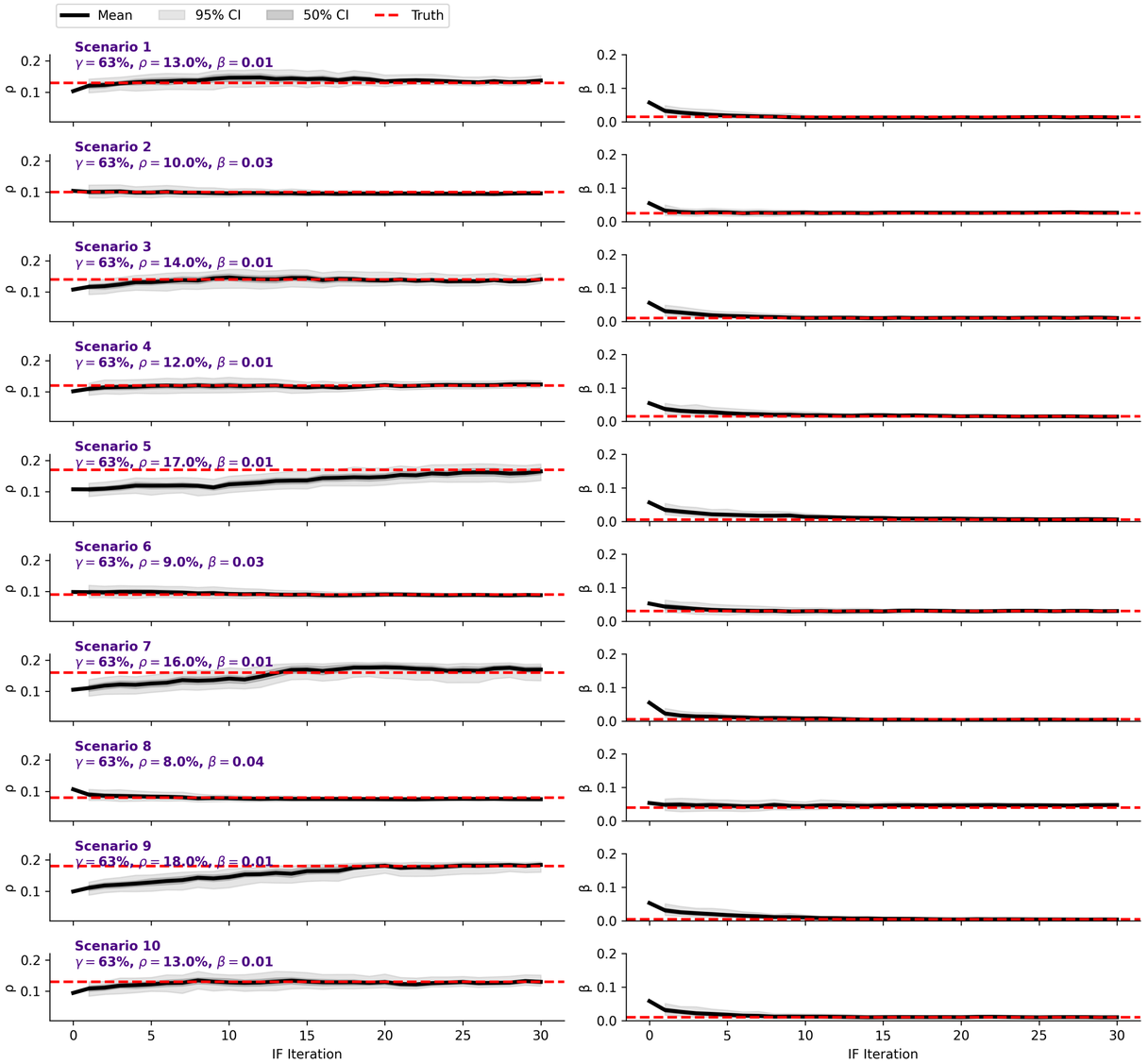
   - $\rho \in [1\,\%, 5\,\%, 10\,\%, 18\%]$
   - $\beta \in [0.01, 0.05, 0.1]$

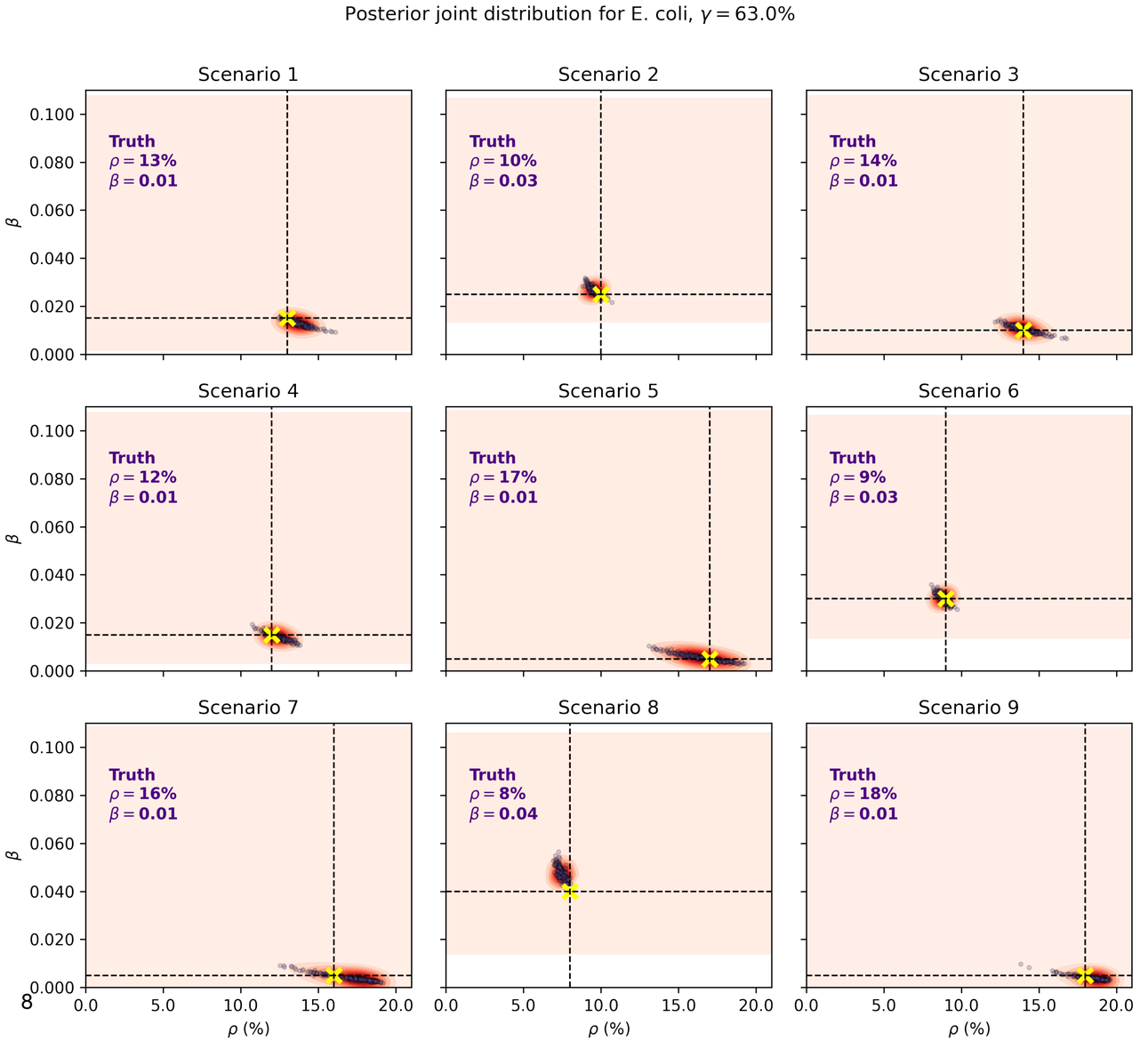
# inferences with truths that reproduced the synthetic observations

Very challenging IMO 🙄

# Inferences on simulated observations

## E. coli



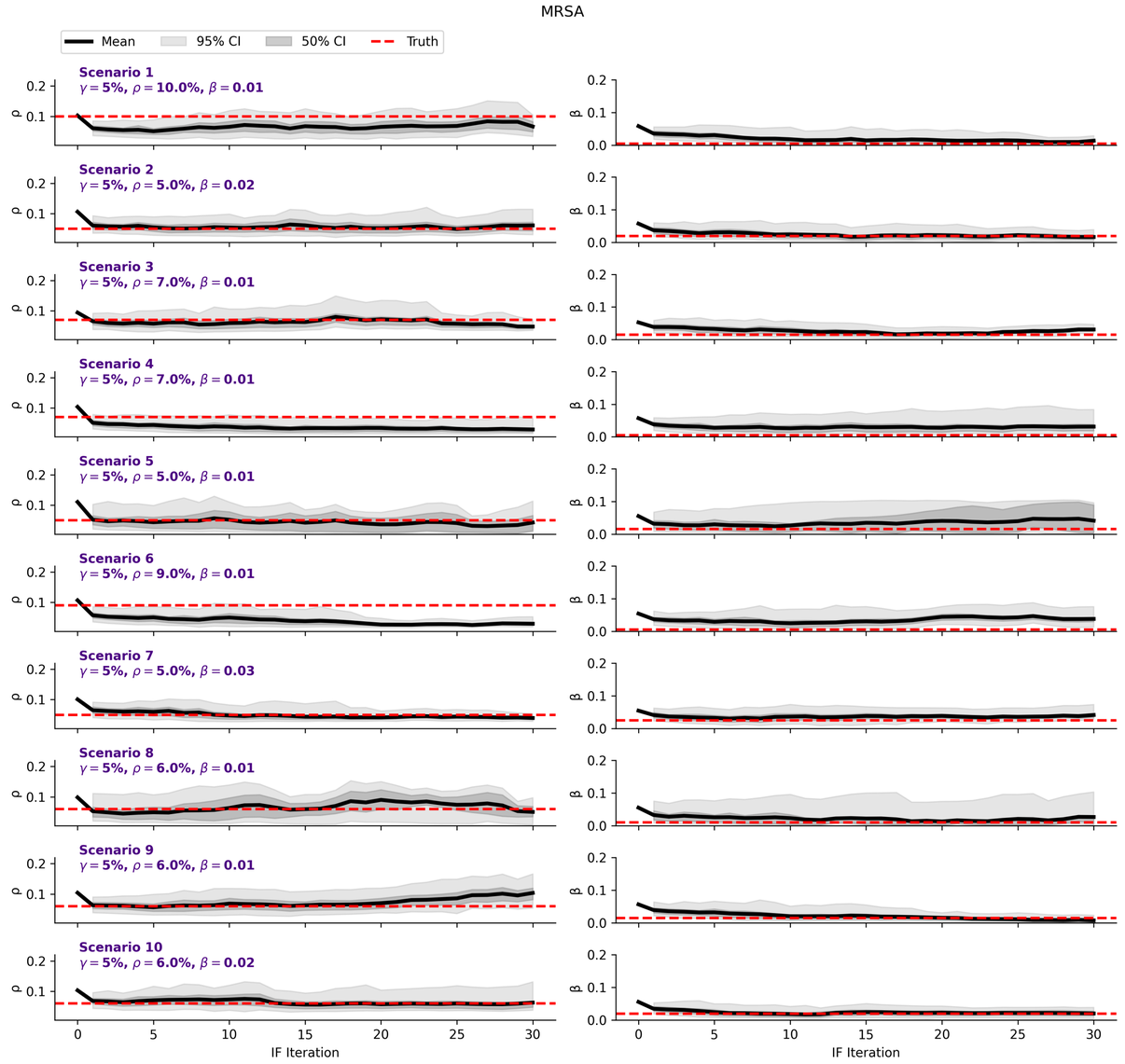



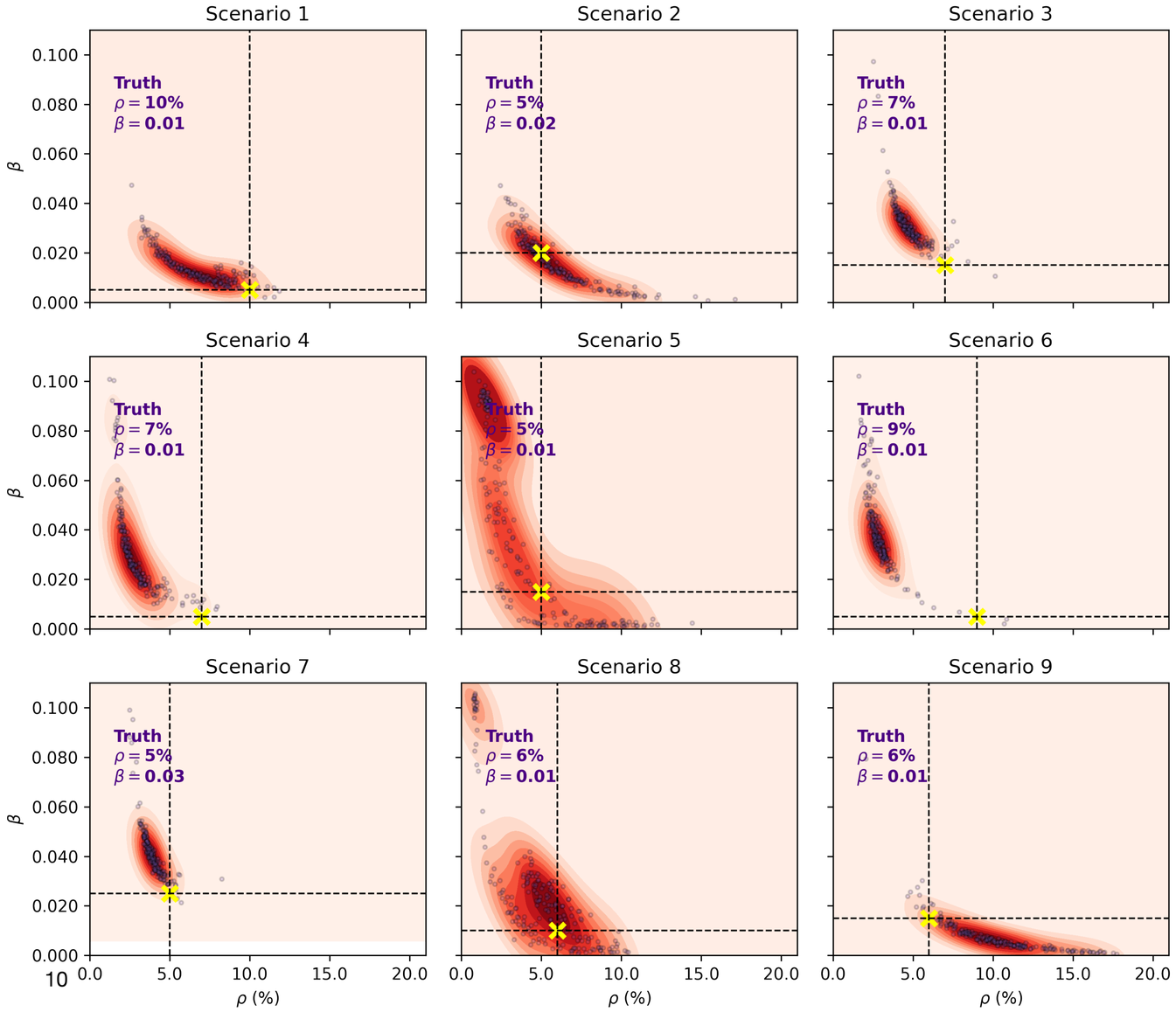
# Inferences on simulated observations
## MSSA

**Re-running with the new prevalences**

# Inferences on simulated observations

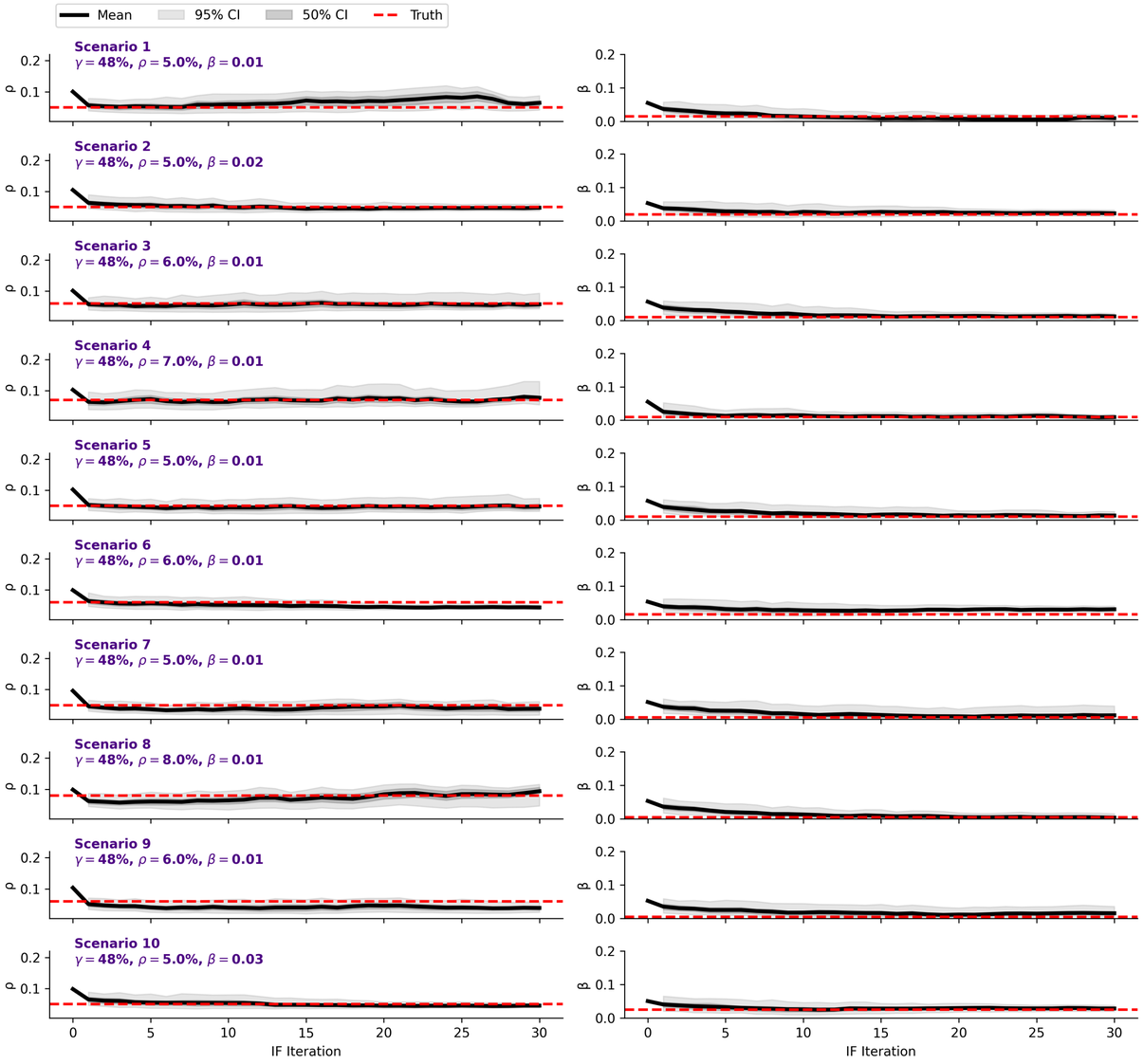
## MRSA



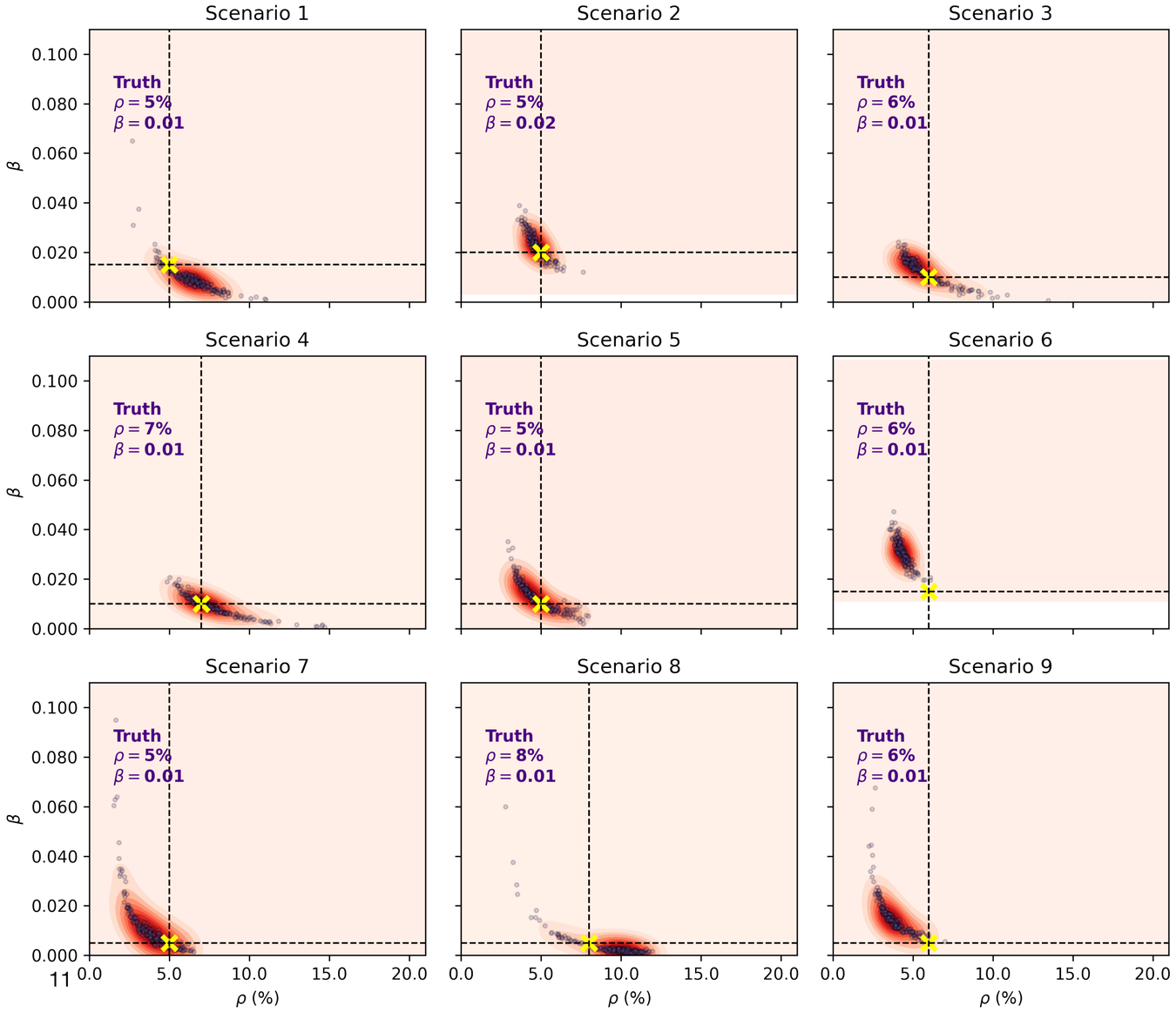



# Inferences on simulated observations

## E. faecalis

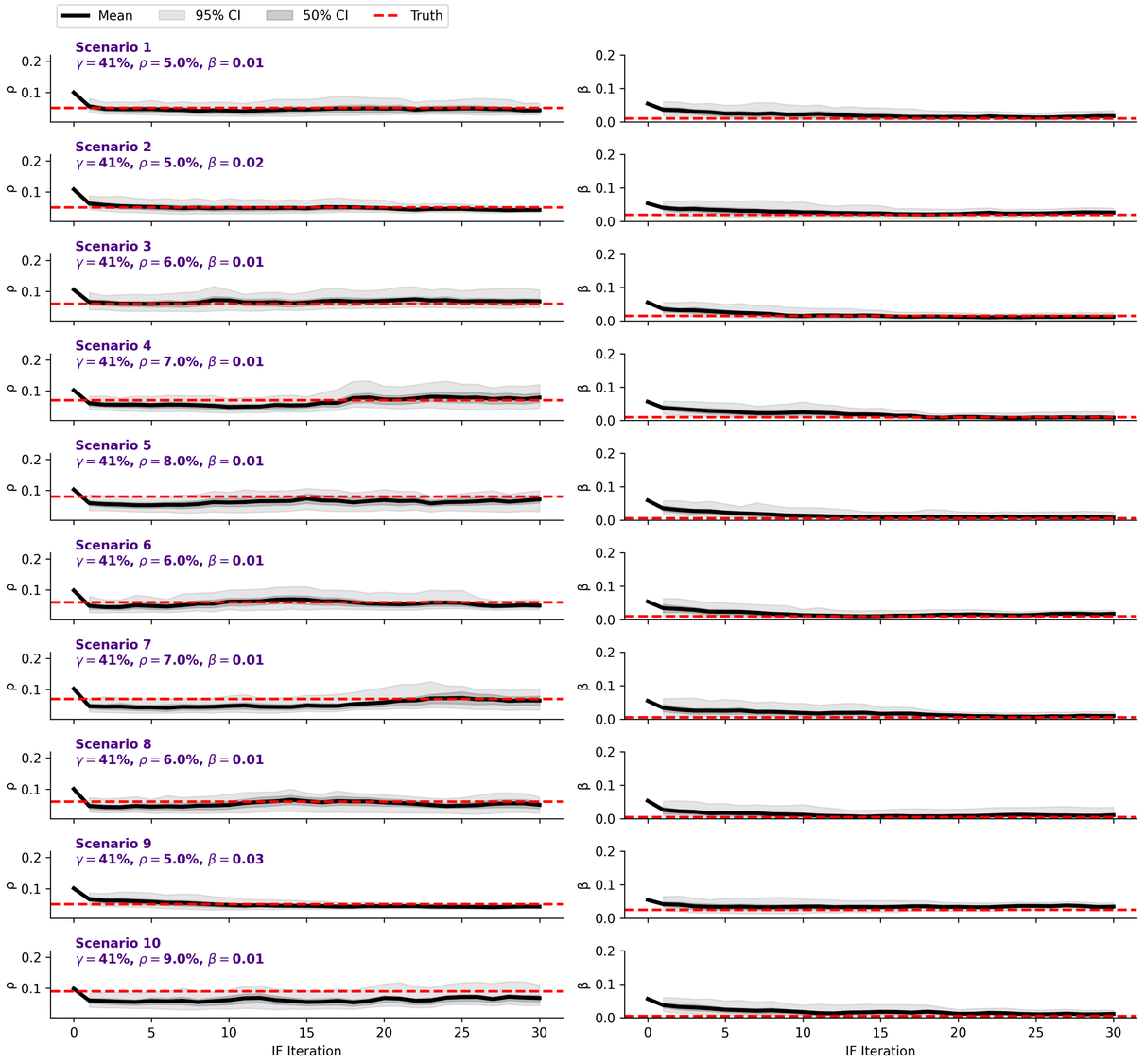



Posterior joint distribution for E. faecalis, $\gamma = 47.6\%$

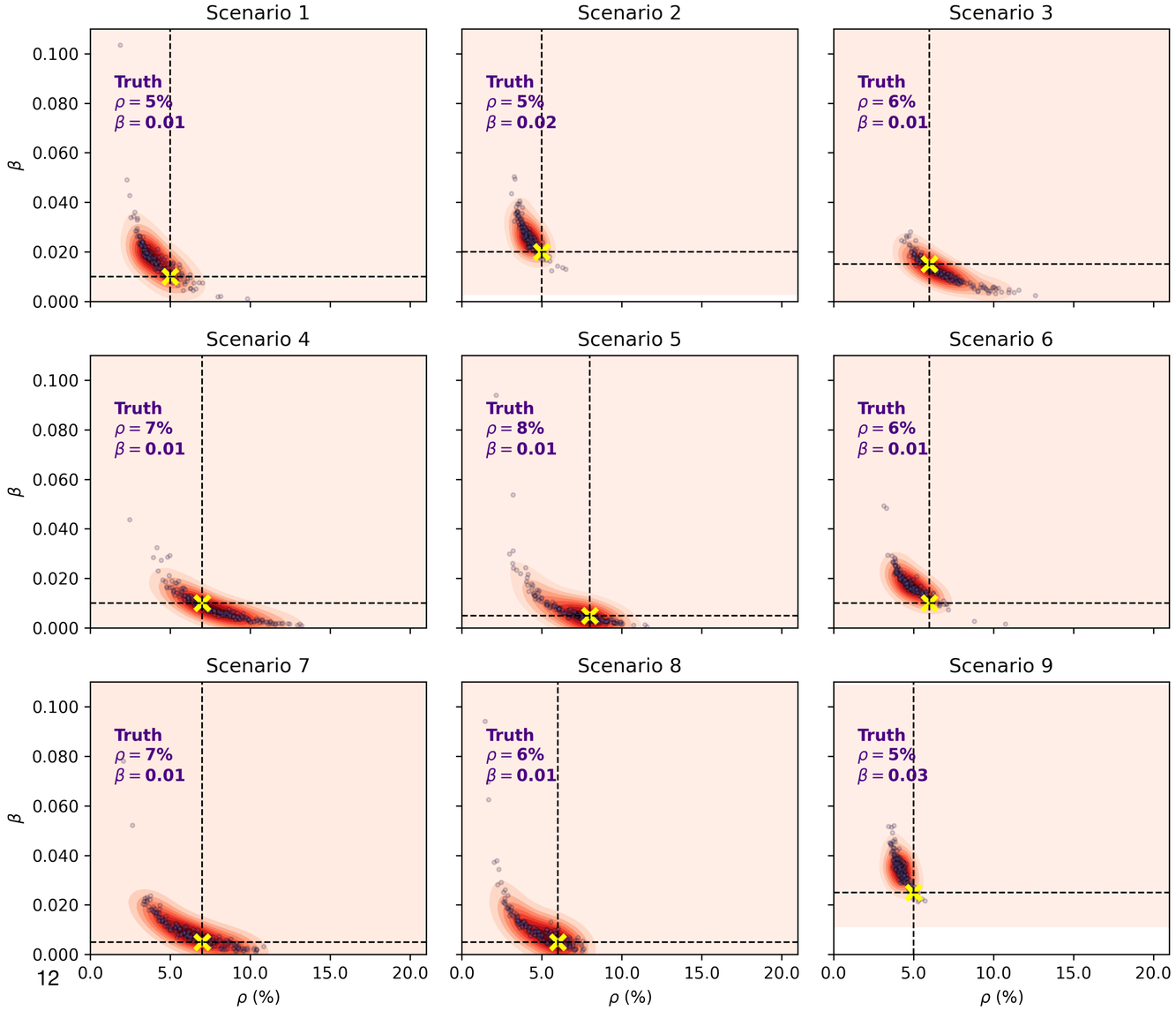


# Inferences on simulated observations

## E. faecium

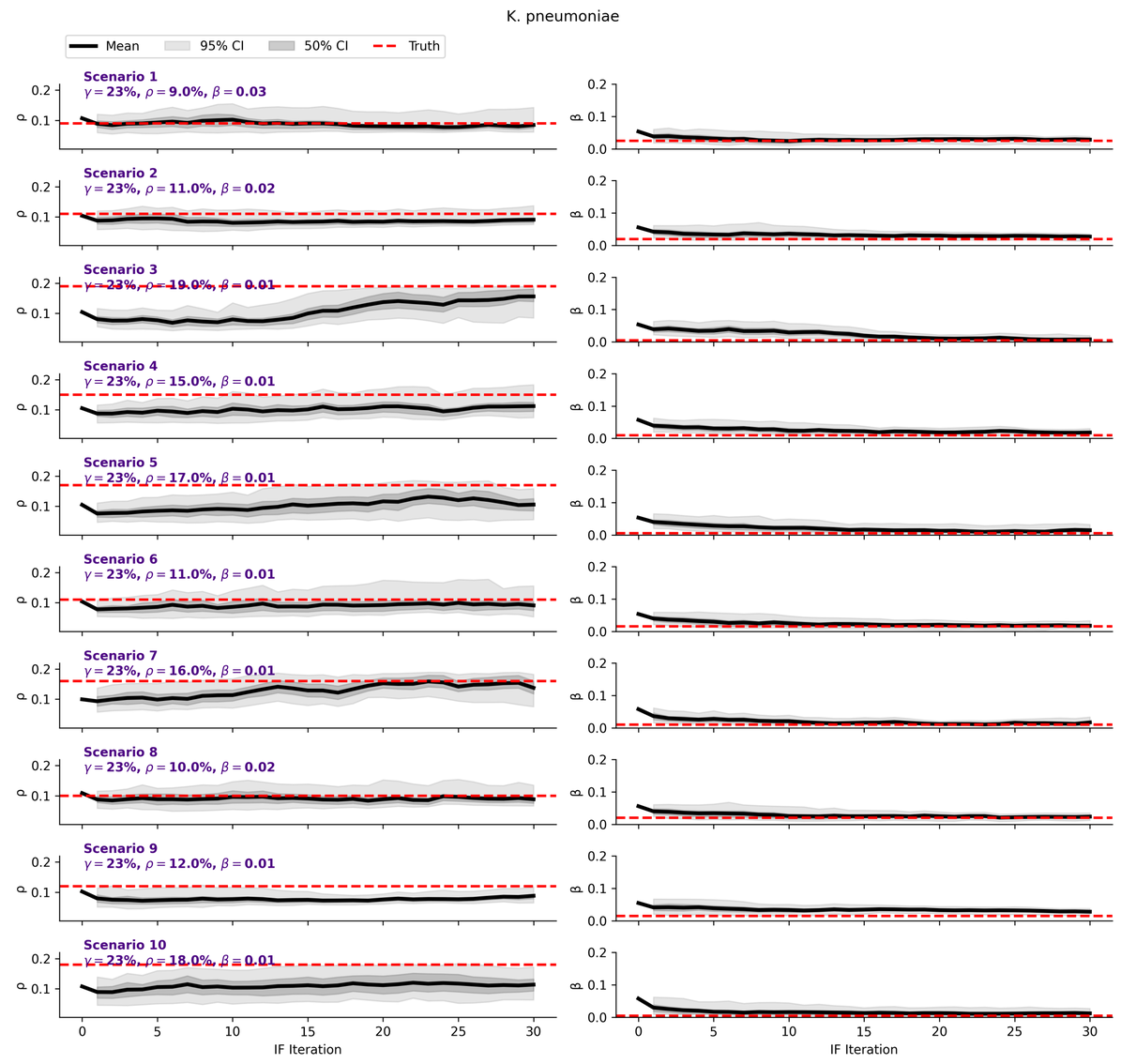


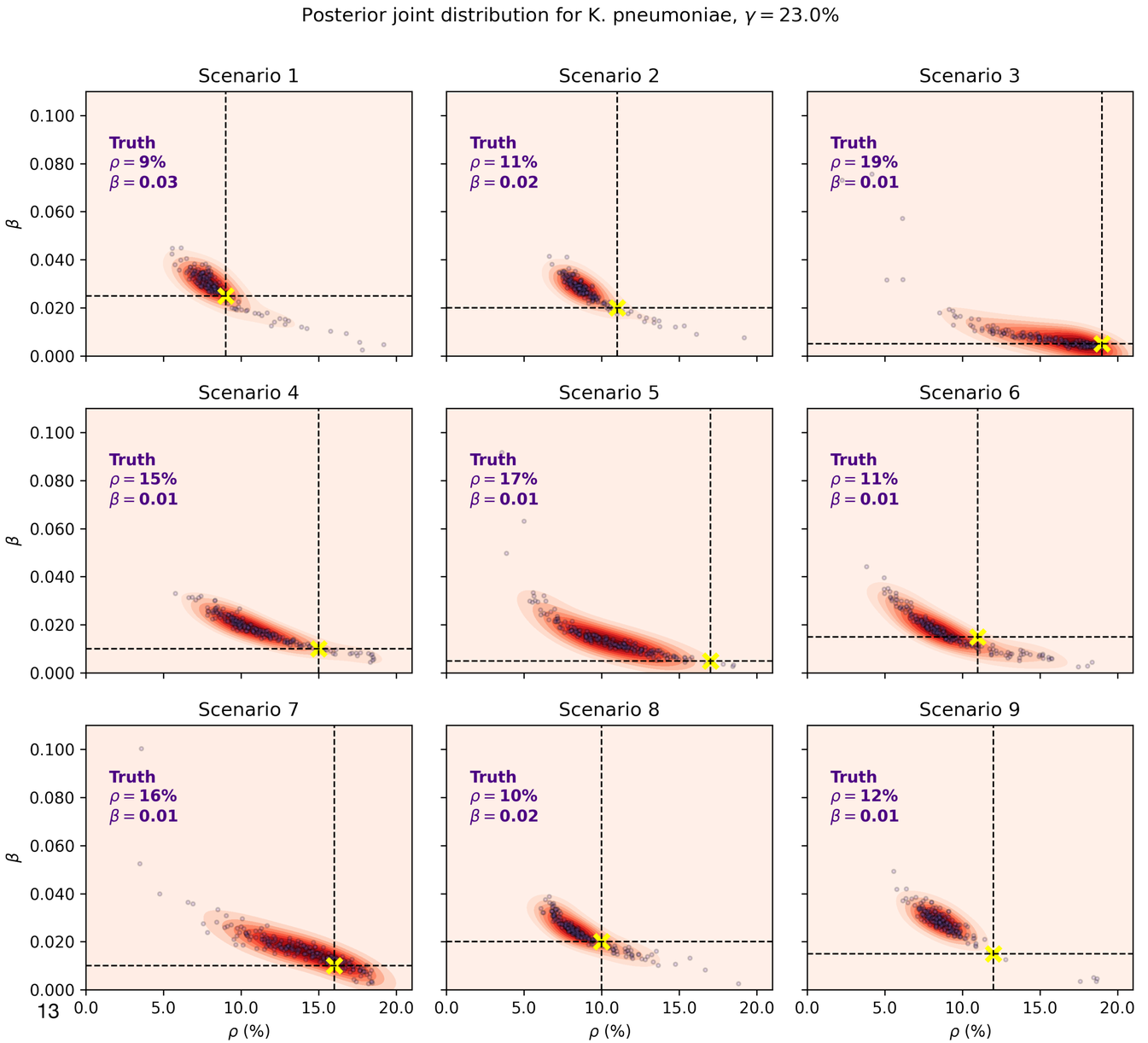

Posterior joint distribution for E. faecium, $\gamma = 40.6\%$



# Inferences on simulated observations

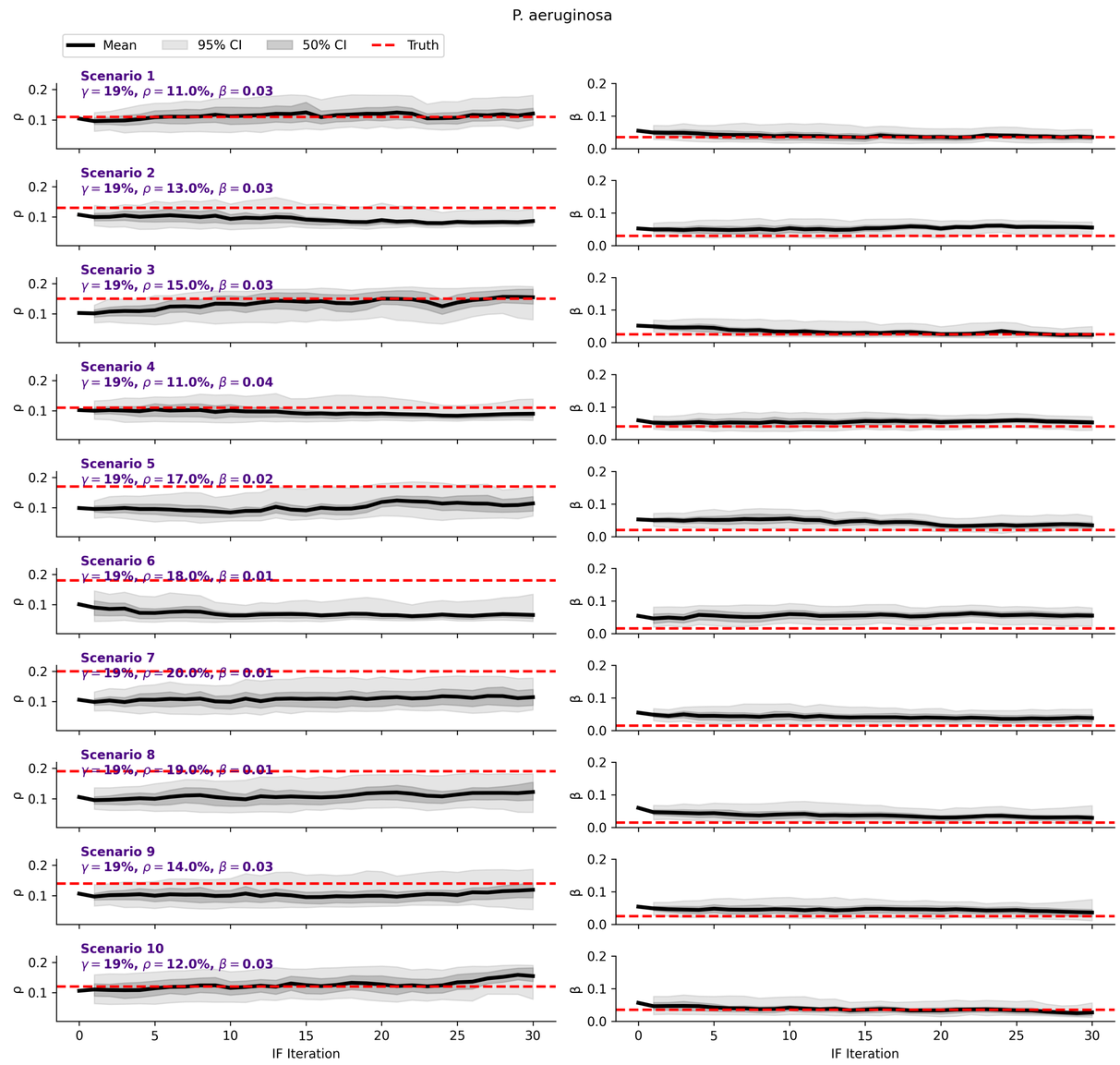
## K. pneumoniae



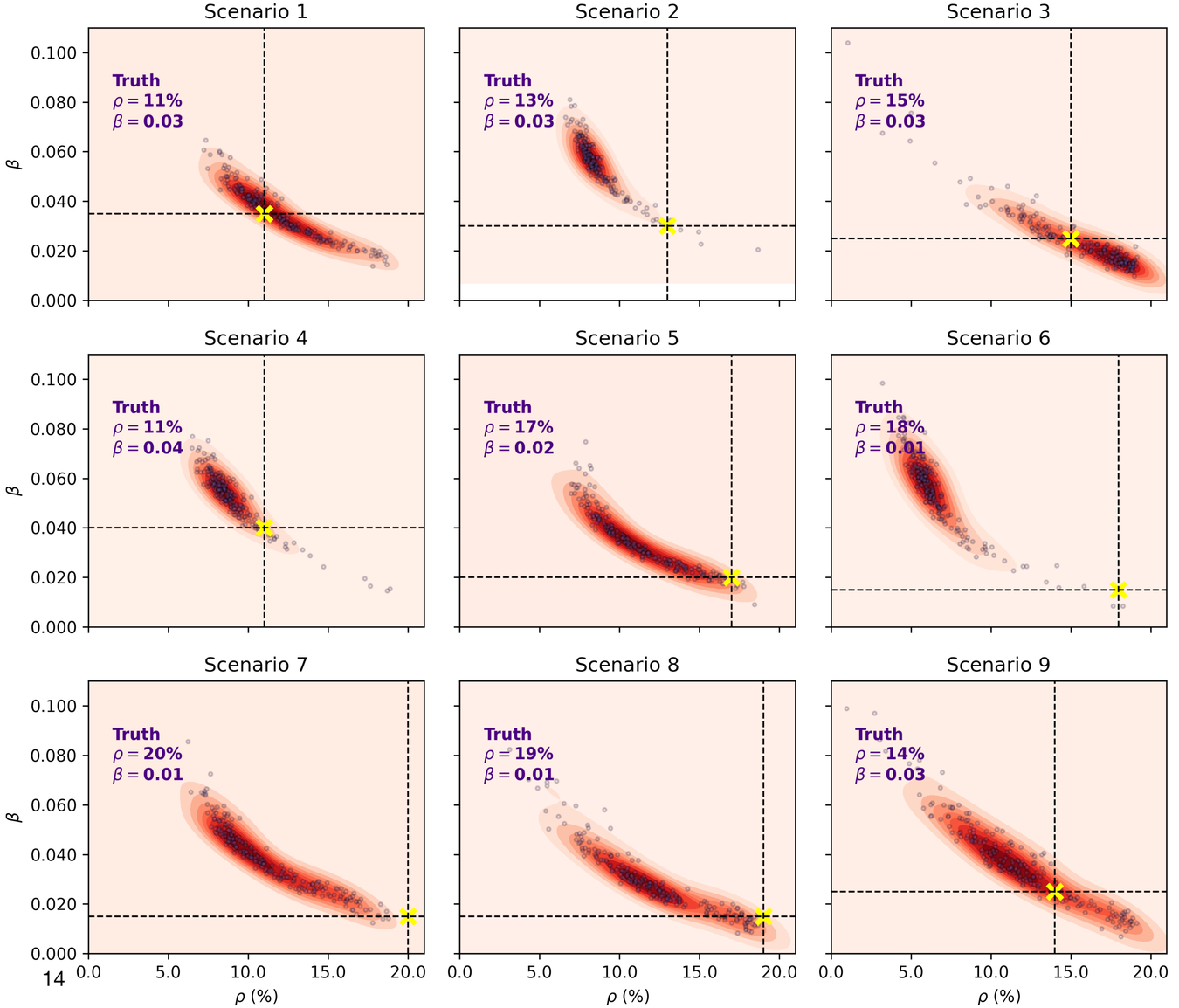



# Inferences on simulated observations

## P. aeruginosa

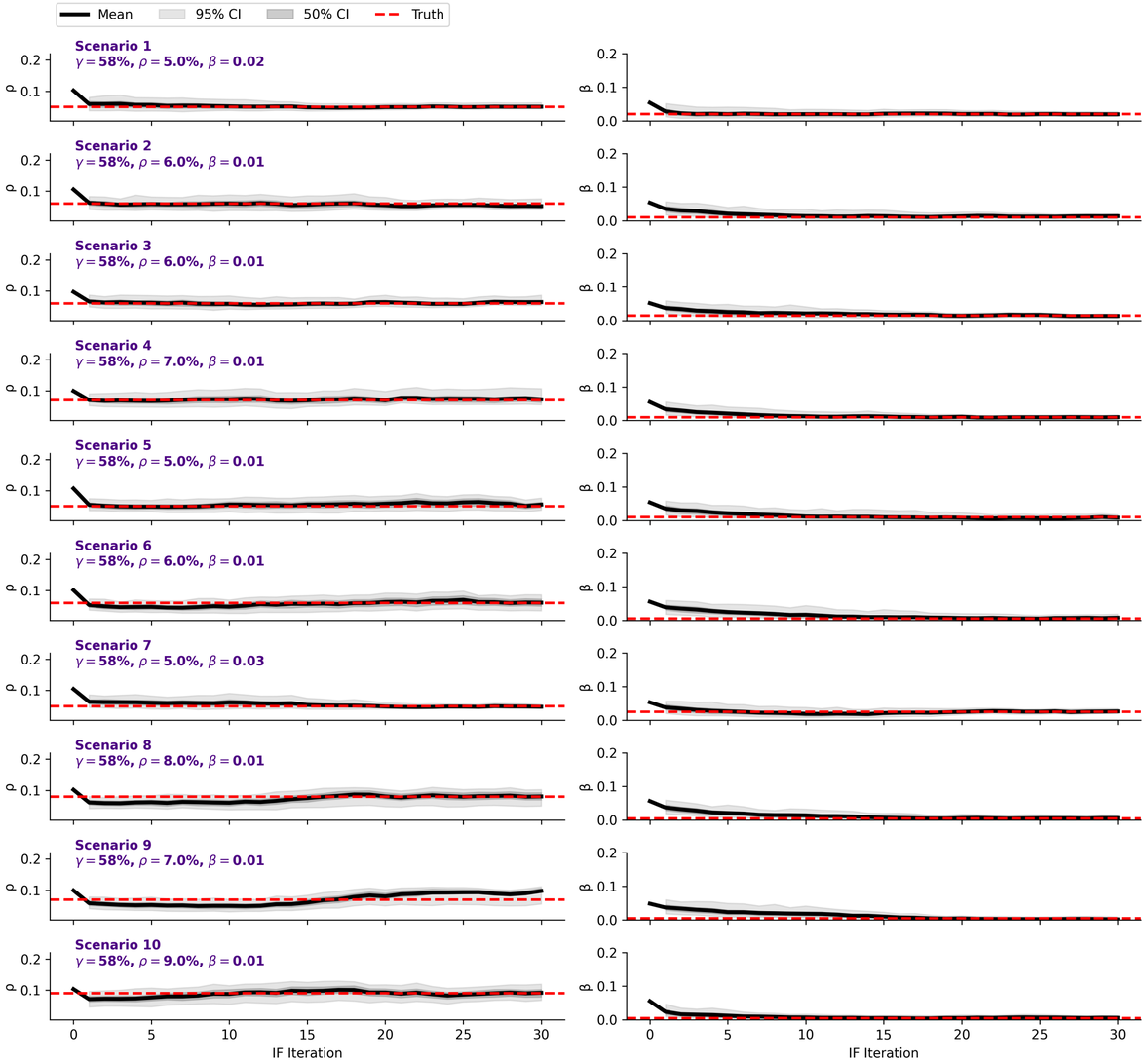



Posterior joint distribution for P. aeruginosa, $\gamma = 18.8\%$

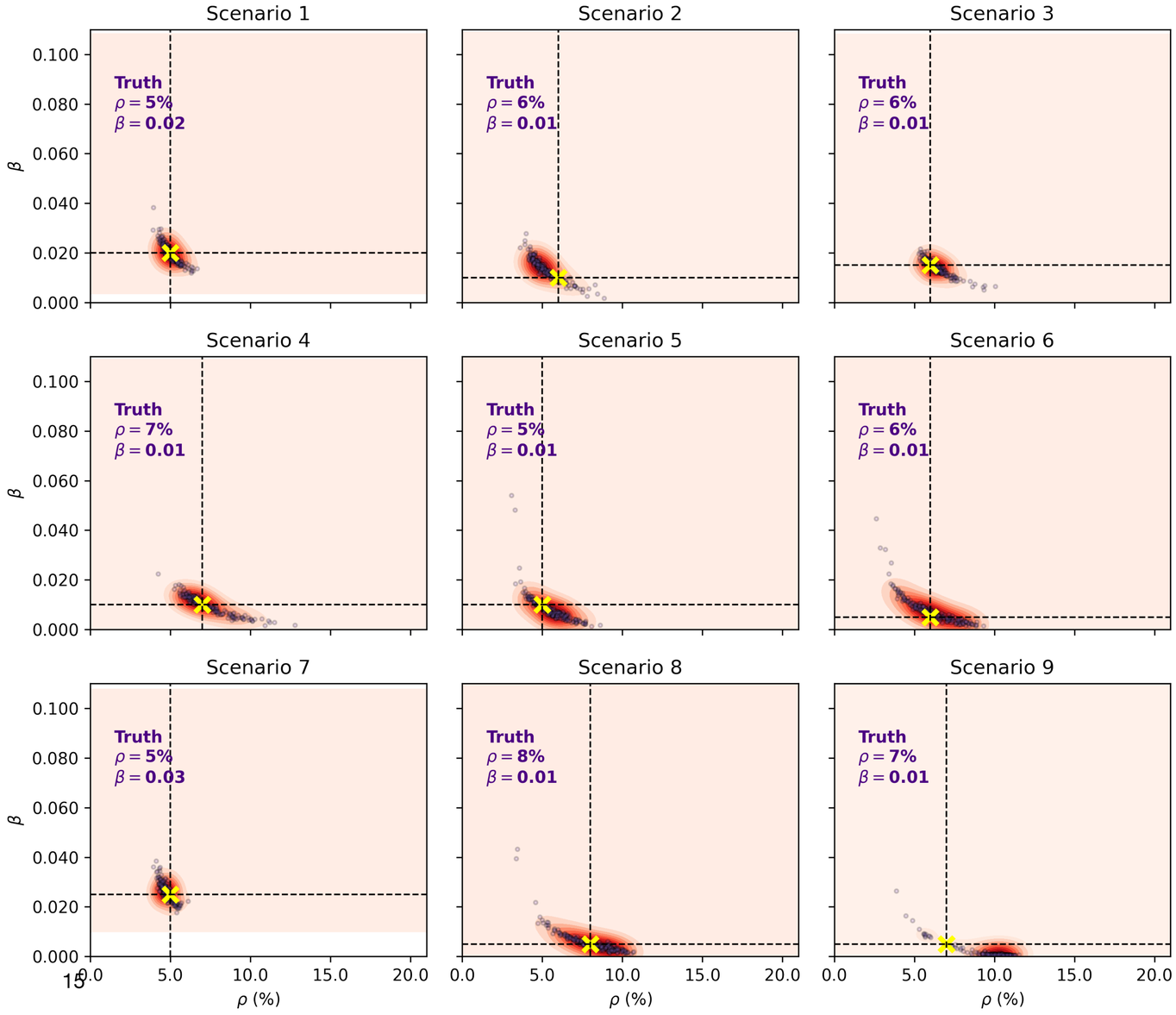


# Inferences on simulated observations

## S. epidermis




Posterior joint distribution for S. epidermidis, $\gamma = 58.0\%$

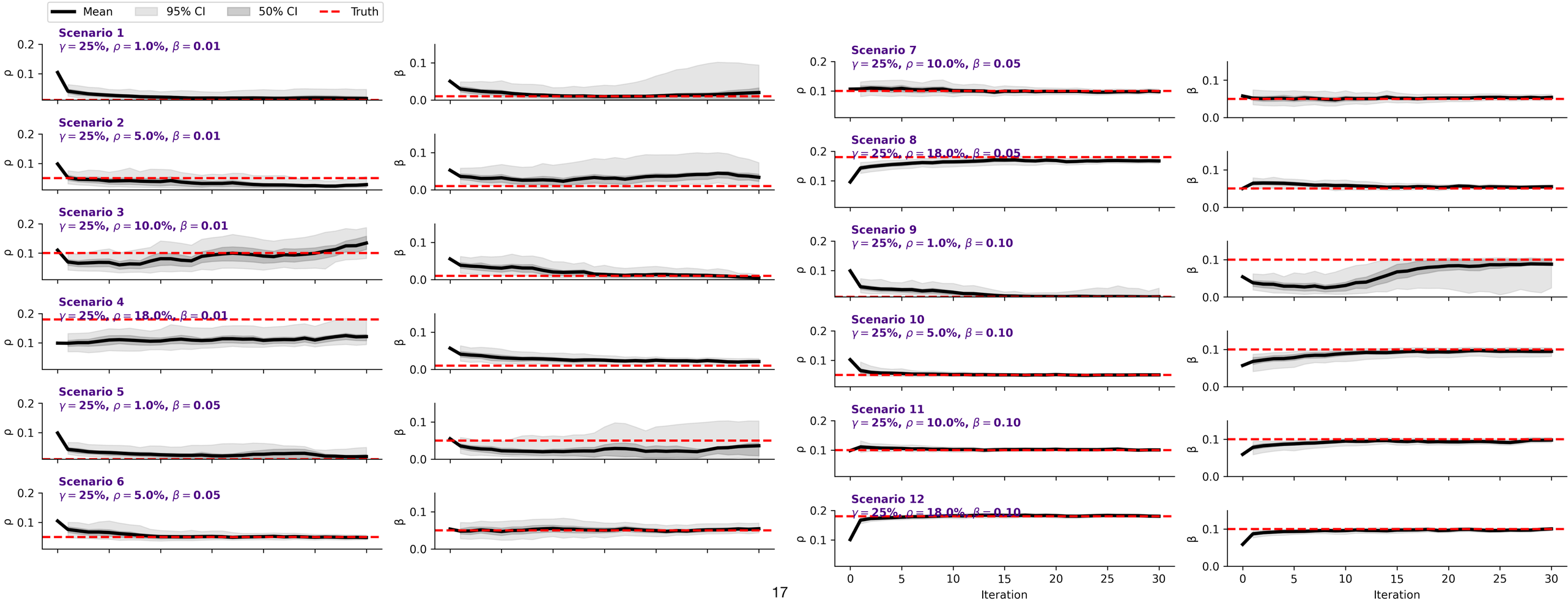


# Inferences on a broader range

## Jeff’s request

# Broad synthetic inferences

## Convergence plots - $\gamma = 25\,\%$

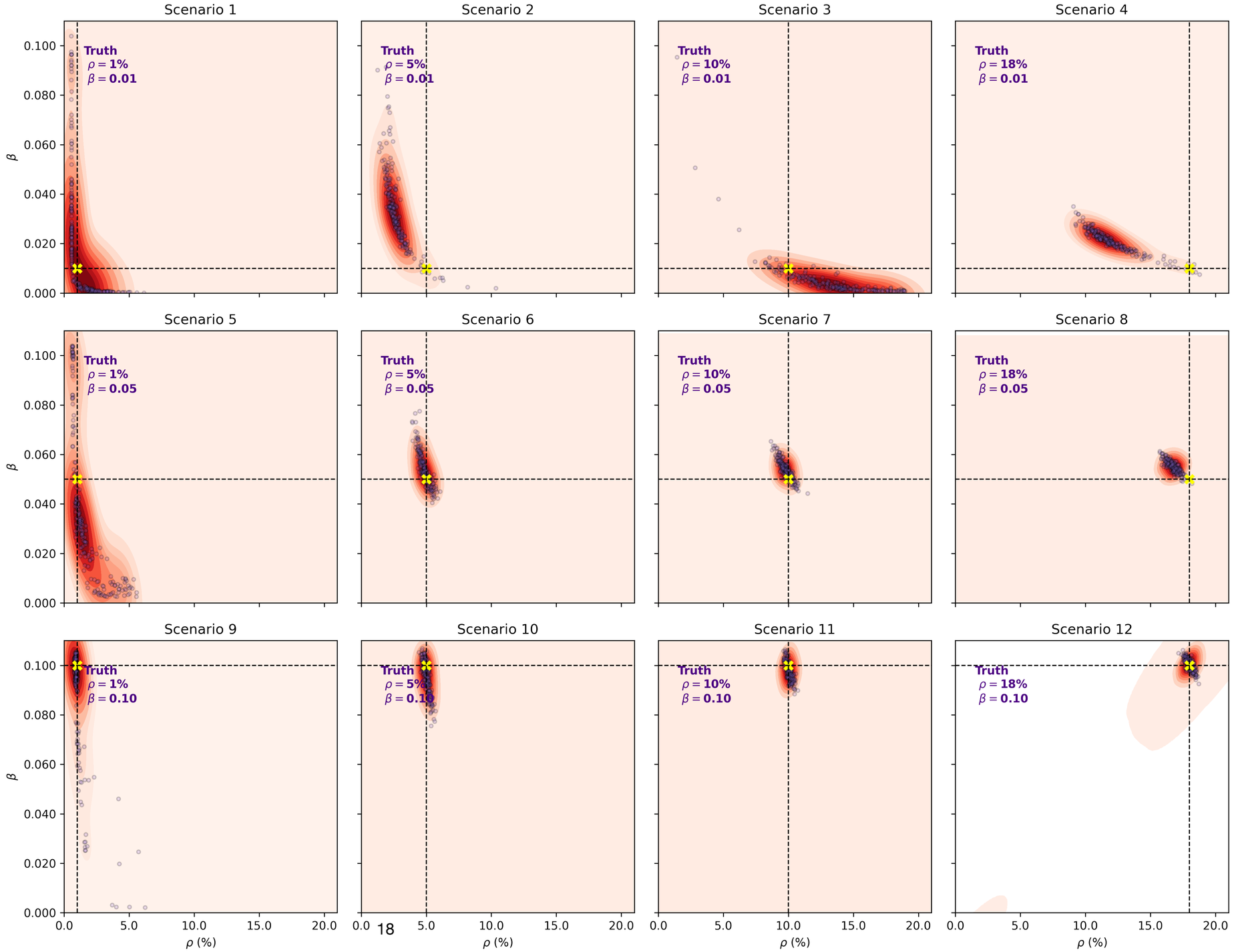


# Broad synthetic inferences

## Joint posterior

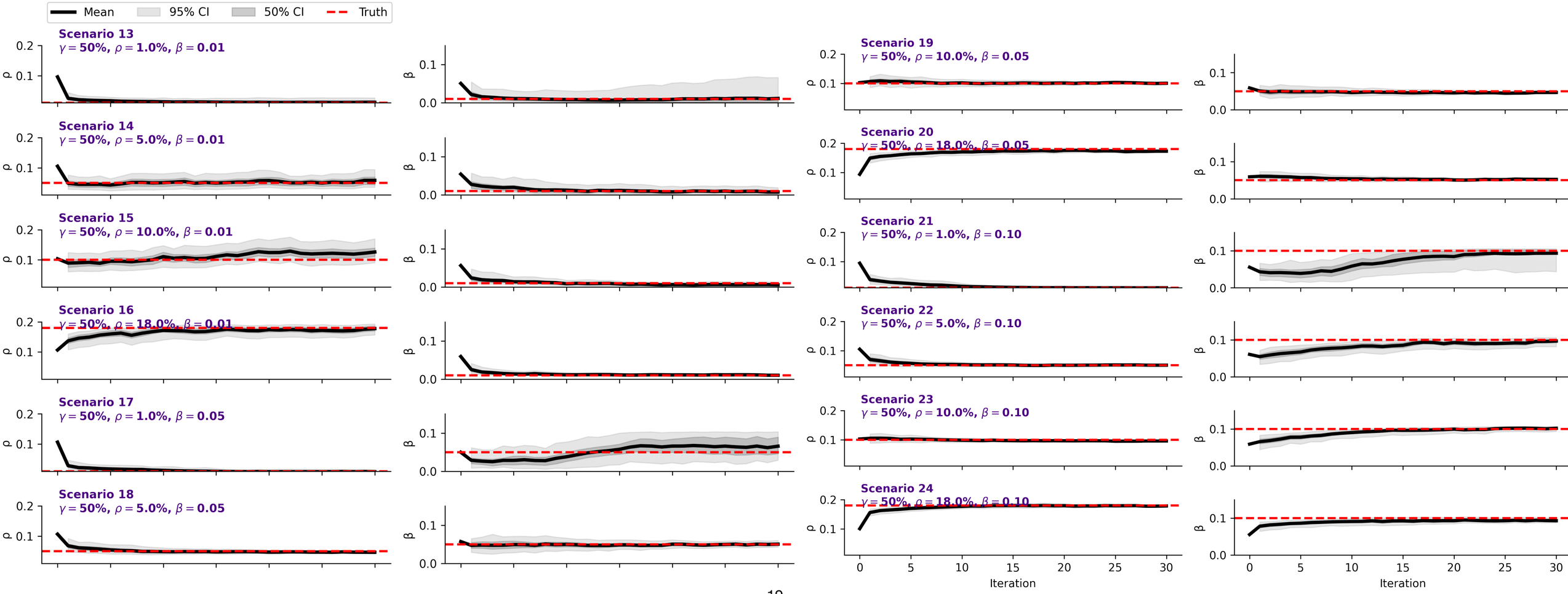
$\gamma = 25\,\%$

- Prior ranges shown in the axis.
- $\beta$ increases per row
- $\rho$ varies per column
- Truth as X cross and intersection of lines.



# Broad synthetic inferences

## Convergence plots - $\gamma = 50\,\%$

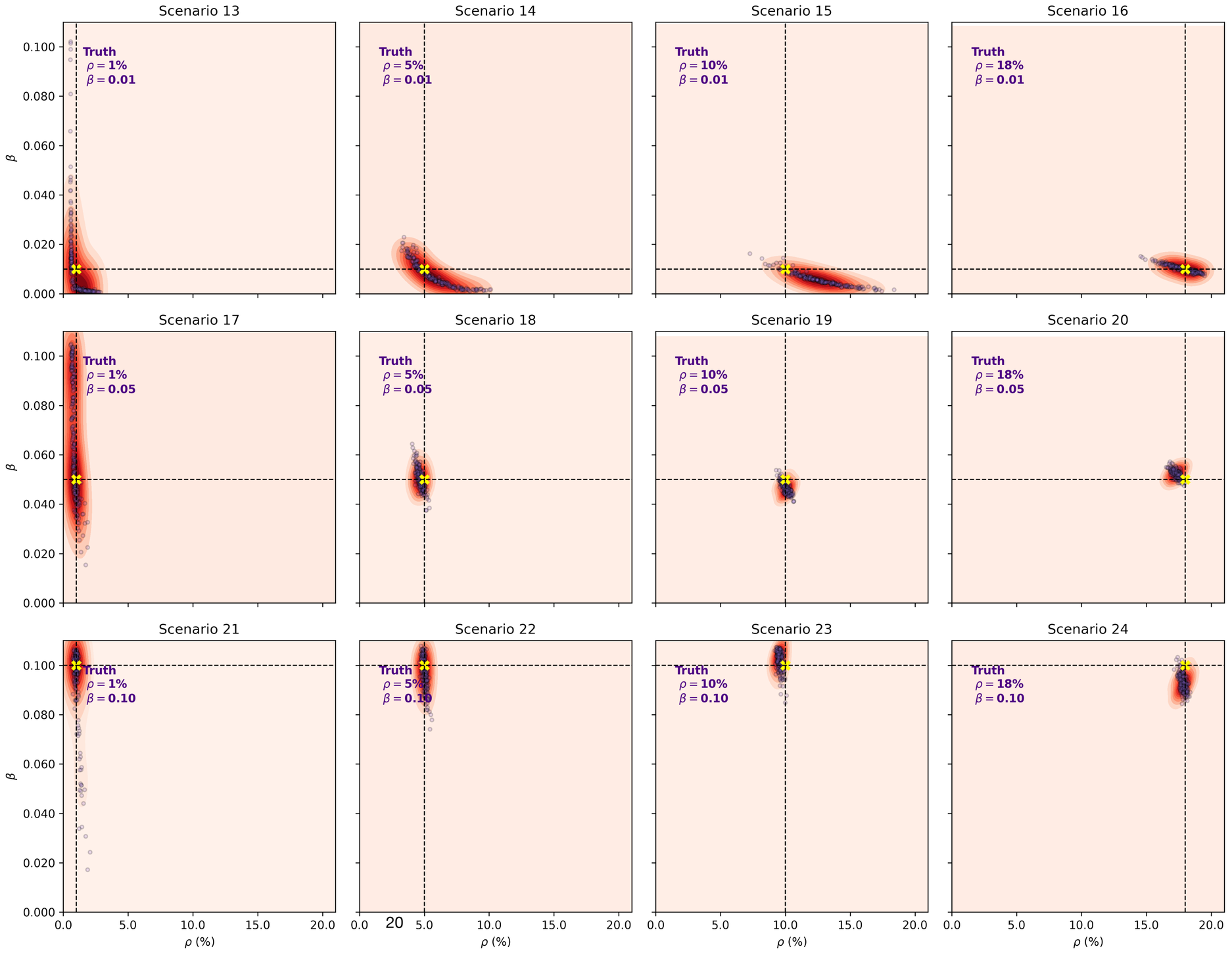


# Broad synthetic inferences

## Joint posterior

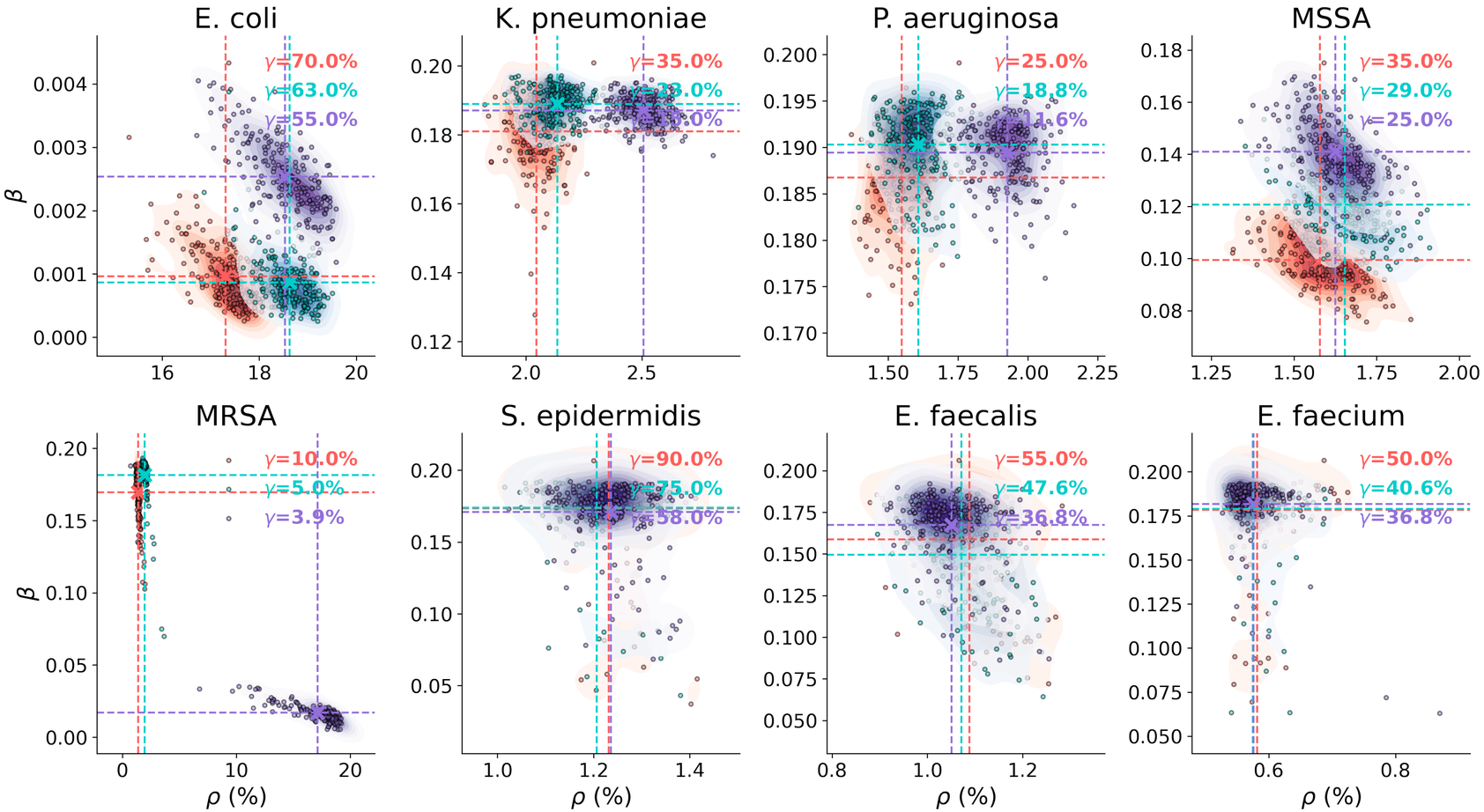
$\gamma = 50\,\%$

- Prior ranges shown in the axis.
- $\beta$ increases per row
- $\rho$ varies per column
- Truth as X cross and intersection of lines.



# Inferences with different prevalences

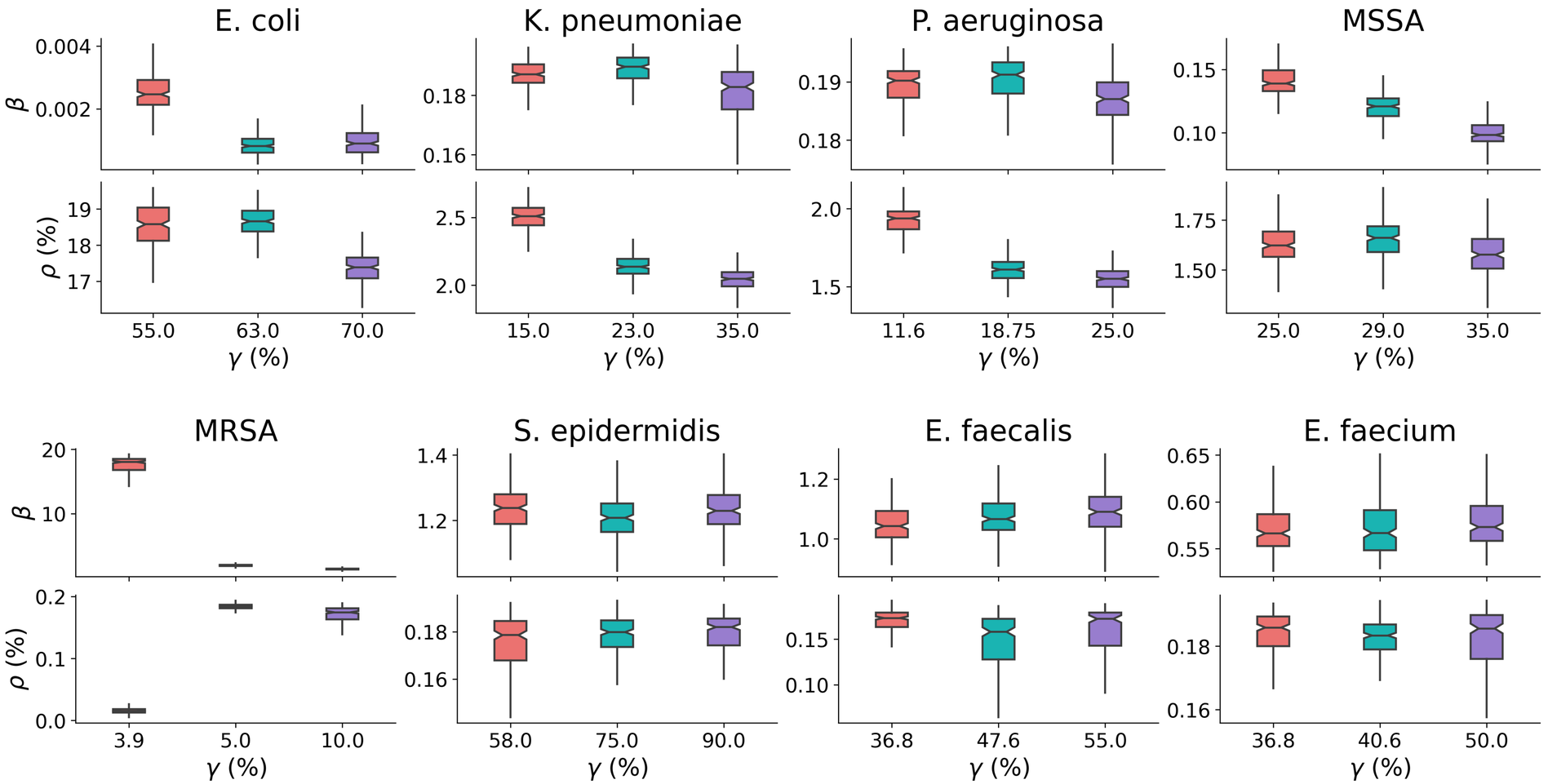
## Posterior joint distribution




- Each posterior inference with a different $\gamma$ is color coded.
- Mean estimate is highlighted with the intersections of the dashed lines.
- Prior range not showed in the axis.

# Inferences with different prevalences

## Posterior marginals for different prevalence levels




- Each posterior inference with a different $\gamma$ is color coded.
- Posterior marginal, whiskers shows the 5 and 97.5 quantiles, boxes the 50%.
- X-axis is not uniform

# How measure distance between the different posterior pdfs?

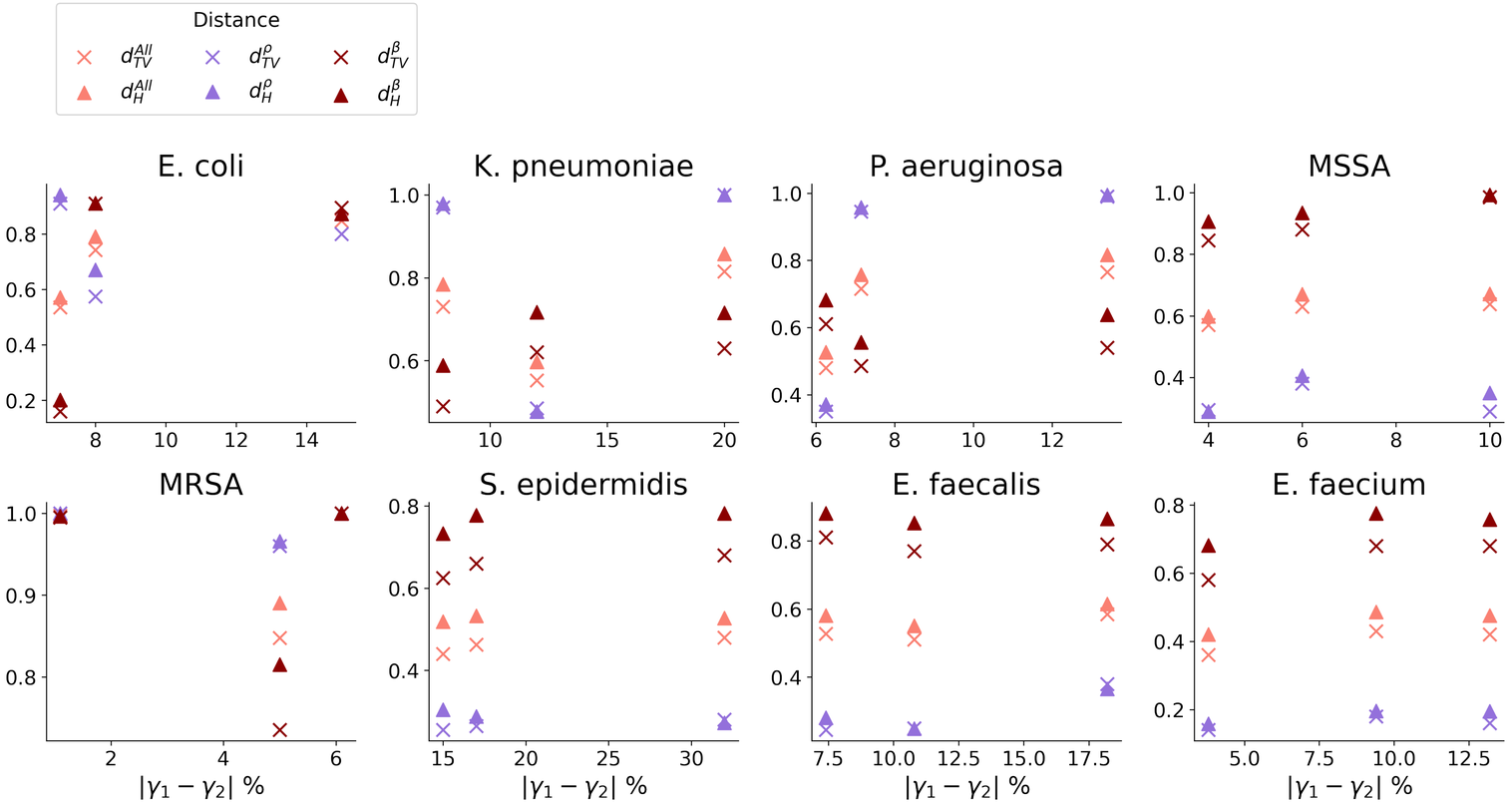
## Univariate distance would explain trade-offs in bias between $\rho$ and $\beta$

- Total variation between pdfs $\pi(\theta)$ and $\pi^{\star}(\theta)$
  - $d_{tv}\left(\pi, \pi^{\star}\right) = \frac{1}{2}\int |\,\pi(\theta) - \pi^{\star}(\theta)\,| d\theta$
  - $d_{tv}\left(\pi, \pi^{\star}\right) = \frac{1}{2}||\,\pi - \pi^{\star}\,||_{L_1}$
- Hellinger distance
  - $d_H\left(\pi, \pi^{\star}\right) = \sqrt{\frac{1}{2}\int |\sqrt{\pi(\theta)} - \sqrt{\pi^{\star}(\theta)}\,| d\theta}$
  - $d_H\left(\pi, \pi^{\star}\right) = \frac{1}{\sqrt{2}}||\sqrt{\pi} - \sqrt{\pi^{\star}}\,||_{L_2}$

- Both distances are bounded between 0, 1.
- The lower the value the similar the distances.
- Ref: Inverse problems and data assimilation (Stuart).

# How measure distance between the different posterior pdfs?

## Univariate distance would explain trade-offs in bias between $\rho$ and $\beta$




- X-axis shows distance between the a pair of prevalences.
- y-axis shows the distances (previous slide), between a pair of inferences
  - Red is the distance of the posterior marginals for $\beta$
  - Purple is the distance of the posterior marginals for $\rho$
  - Salmon is the mean distance.
- Trade-offs between $\beta$ and $\rho$ are evident, I
  - Low $d^{\rho} \leftrightarrows$ high $d^{\beta}$
  - high $d^{\rho} \leftrightarrows$ low $d^{\beta}$
- **Problem:** Distances do not consider convergence with respect to the prior range, just a distance between posterior.

# Why are some inferences biased?

## My guess: Stochasticity of the system a.k.a. Monte Carlo error

- Each synthetic inference is conducted on a single stochastic trajectory $y^i = [y_1^i, y_2^i, \ldots, y_T^i]$.
- We know the distribution of possible $y_i$, so if we know how far was $y_i$ from the distribution of possible trajectories $Y$, we could explain biases in the inference?
  - Not need to compute a lot of inferences (expensive with the ABM).
  - We can understand the impact of the parameters on the bias too, more than knowing how they impact the uncertainty in the estimates.

# Distances

- Distance between a posterior estimate $\Theta$ and truth $\theta_T$ (point)
  - The posterior can be parametrized with a vector of means $\mu_\Theta$, and a covariance matrix $C_\Theta$
    - Distance between posterior and truth: Just a norm scaled with the covariance (Mahalanobis distance (?) according to Wikipedia)
      - $D_1 = |\mu_\Theta - \theta_T|_{C_\Theta} = \sqrt{(\mu_\Theta - \theta_T)^T C_\Theta^{-1} (\mu_\Theta - \theta_T)}$
    - Distance between inferred observation and set of possible observations (continuou ranked probability score).
      - $D_2 = CRPS(Y, y_i)$

# What should we expect?

- The more common the observation (lower CRPS) the lower the distance between the posterior and the truth (less bias)
    - Probably depends on importation rates $\gamma$, other guess: the higher the prevalence less error in the posterior (?).

# Result

$\hat{\theta}$ : Mean estimate, $\theta_T$ : Truth, $C$ : Covariance estimate

$$D=\sqrt{(\hat{\theta}-\theta_T)^T C^{-1}(\hat{\theta}-\theta_T)}$$