What To Do When Your LLM is Not State of the Art

When (not) to Worry About Misclassification and How to Correct for It in Social Science Applications

James Bisbee [1] Arthur Spirling[2]

[1]Vanderbilt University
[2]Princeton University

March 14, 2025

Motivation: LLMs are everywhere!

- Growing use of Large Language Models for **annotation & coding**.
  - *Published: Wang et al. (2024); Goel et al. (2023); Li (2024); Nasution & Onan (2024); Druart et al. (2024); Rathje et al. (2024); Tornberg (2024); Wu et al. (2024)...; Working papers: Sapkota et al.; Ye et al.; Rouzegar & Makrehchi; Cheng et al...etc.*

Motivation: LLMs are everywhere!

- Growing use of Large Language Models for **annotation & coding**.
  - *Published: Wang et al. (2024); Goel et al. (2023); Li (2024); Nasution & Onan (2024); Druart et al. (2024); Rathje et al. (2024); Tornberg (2024); Wu et al. (2024)...; Working papers: Sapkota et al.; Ye et al.; Rouzegar & Makrehchi; Cheng et al...etc.*
- (Rapidly) evolving tech $\Rightarrow$ concerns about robustness and **reproducibility**.
  - 4 new versions of ChatGPT between July 2024 and today

Motivation: LLMs are everywhere!

- Growing use of Large Language Models for **annotation & coding**.
  - *Published: Wang et al. (2024); Goel et al. (2023); Li (2024); Nasution & Onan (2024); Druart et al. (2024); Rathje et al. (2024); Tornberg (2024); Wu et al. (2024)...; Working papers: Sapkota et al.; Ye et al.; Rouzegar & Makrehchi; Cheng et al...etc.*
- (Rapidly) evolving tech $\Rightarrow$ concerns about robustness and **reproducibility**.
  - 4 new versions of ChatGPT between July 2024 and today
- **Trade-offs** between open-source transparency and proprietary model performance.
  - Replication norms in tension with state-of-the-art models (Bisbee et al. 2024; Spirling et al. 2024)

Motivation: LLMs are everywhere!

- Growing use of Large Language Models for **annotation & coding**.
  - *Published: Wang et al. (2024); Goel et al. (2023); Li (2024); Nasution & Onan (2024); Druart et al. (2024); Rathje et al. (2024); Tornberg (2024); Wu et al. (2024)...; Working papers: Sapkota et al.; Ye et al.; Rouzegar & Makrehchi; Cheng et al...etc.*
- (Rapidly) evolving tech $\Rightarrow$ concerns about robustness and **reproducibility**.
  - 4 new versions of ChatGPT between July 2024 and today
- **Trade-offs** between open-source transparency and proprietary model performance.
  - Replication norms in tension with state-of-the-art models (Bisbee et al. 2024; Spirling et al. 2024)
- Do we always need State-of-the-Art (SOTA)?

Research Question

Main Questions

- How problematic is not using State-of-the-Art (SOTA) LLMs for social science applications?
- Can misclassification bias from inferior models be effectively corrected?

Our Context

- Annotation of unstructured (typically text) data to extract binary variable of theoretical interest
  - I.e., populist party manifestos, offensive social media posts, etc.

Our Context

- Annotation of unstructured (typically text) data to extract binary variable of theoretical interest
  - I.e., populist party manifestos, offensive social media posts, etc.
- Measurement error here is a **harder problem**
  - Many other ways to use LLMs (continuous latent variables, e.g. Wu et al. 2024; synthetic data e.g. Argyle et al. 2023, etc.)
  - Outcome variable use-cases already well-developed in the methods literature

Our Context

- Annotation of unstructured (typically text) data to extract binary variable of theoretical interest
  - I.e., populist party manifestos, offensive social media posts, etc.
- Measurement error here is a **harder problem**
  - Many other ways to use LLMs (continuous latent variables, e.g. Wu et al. 2024; synthetic data e.g. Argyle et al. 2023, etc.)
  - Outcome variable use-cases already well-developed in the methods literature
- **Contribution**: Use general methods for misclassification corrections to put structure on LLM performance

Misclassification Problem

- Binary classification task

Misclassification Problem

- Binary classification task
- Definitions:
  - True treatment D^* vs Observed treatment D: $D_i = D_i^* + U_i$

$$U_i = \begin{cases} -1 \text{ if } D_i^* = 1 \\ 1 \text{ if } D_i^* = 0 \\ 0 \text{ otw} \end{cases}$$

Misclassification Problem

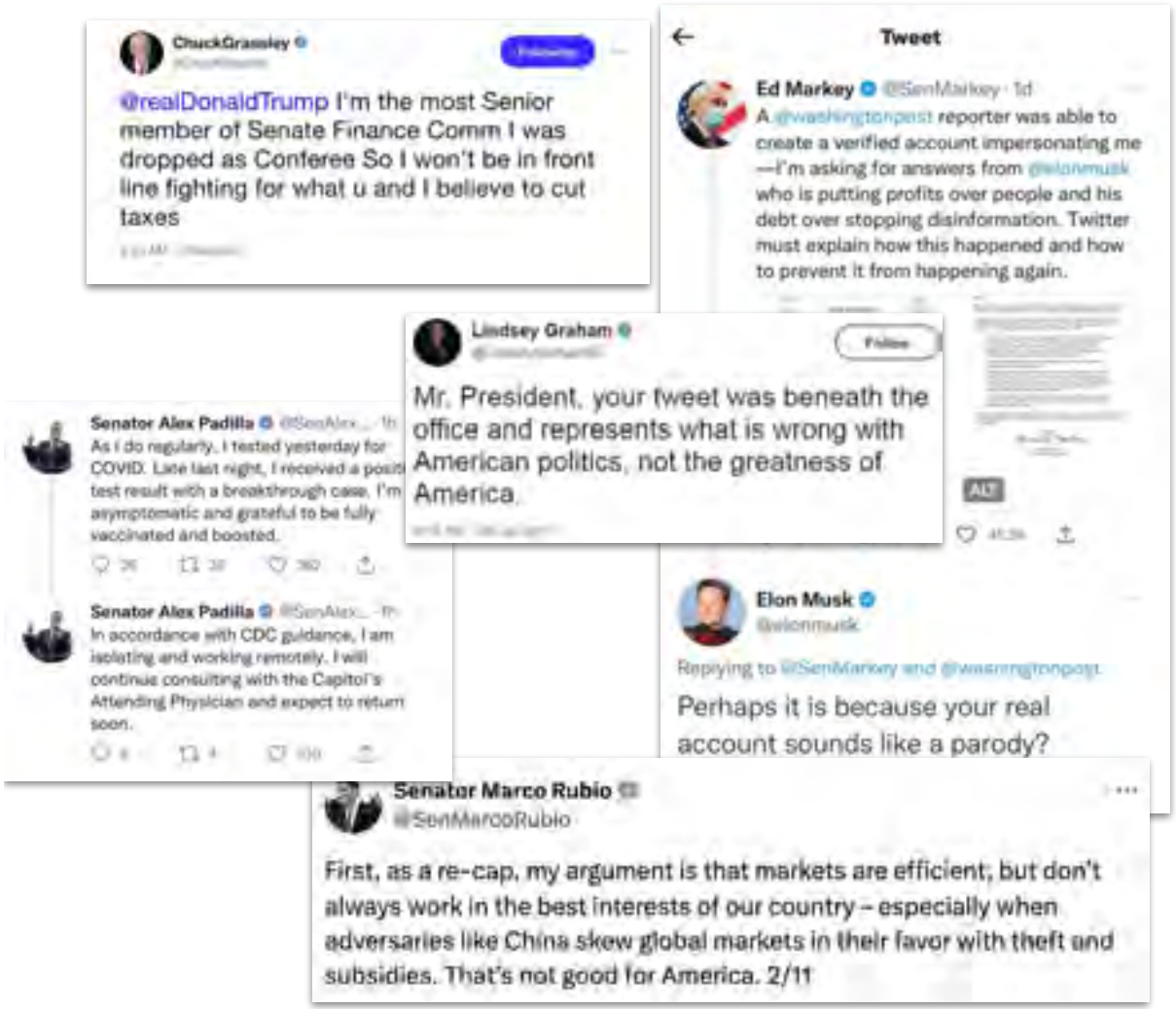
- Binary classification task
- Definitions:
  - True treatment D^* vs Observed treatment D: $D_i = D_i^* + U_i$

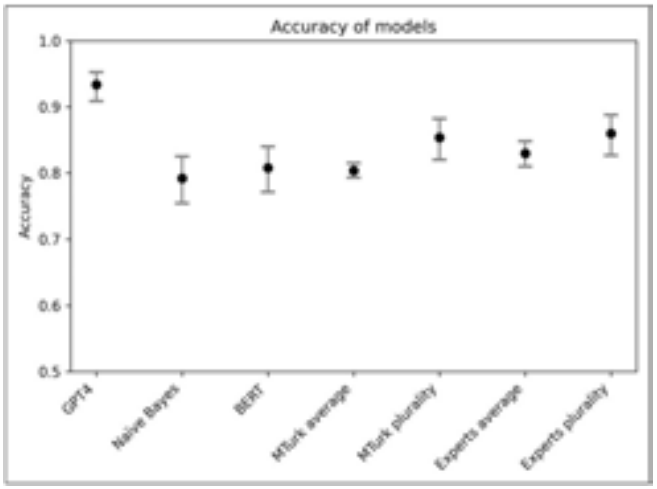
$$U_i = \begin{cases} -1 \text{ if } D_i^* = 1 \\ 1 \text{ if } D_i^* = 0 \\ 0 \text{ otw} \end{cases}$$

  - Misclassification rate: $R_{method} = \frac{\sum m_i}{N}$

$$m_i = \begin{cases} 0 \text{ if } D_i^* = D_i \\ 1 \text{ if } D_i^* \neq D_i \end{cases}$$

Misclassification Problem: Törnberg (2024) example





Bias due to Misclassification

Misclassification causes attenuation bias in OLS estimates:

$$Y_i = c + \beta D_i^* + X_i'\gamma + \epsilon_i$$
$$Y_i = \hat{c} + \hat{\beta} D_i + X_i'\hat{\gamma} + \hat{e}_i$$

Bias due to Misclassification

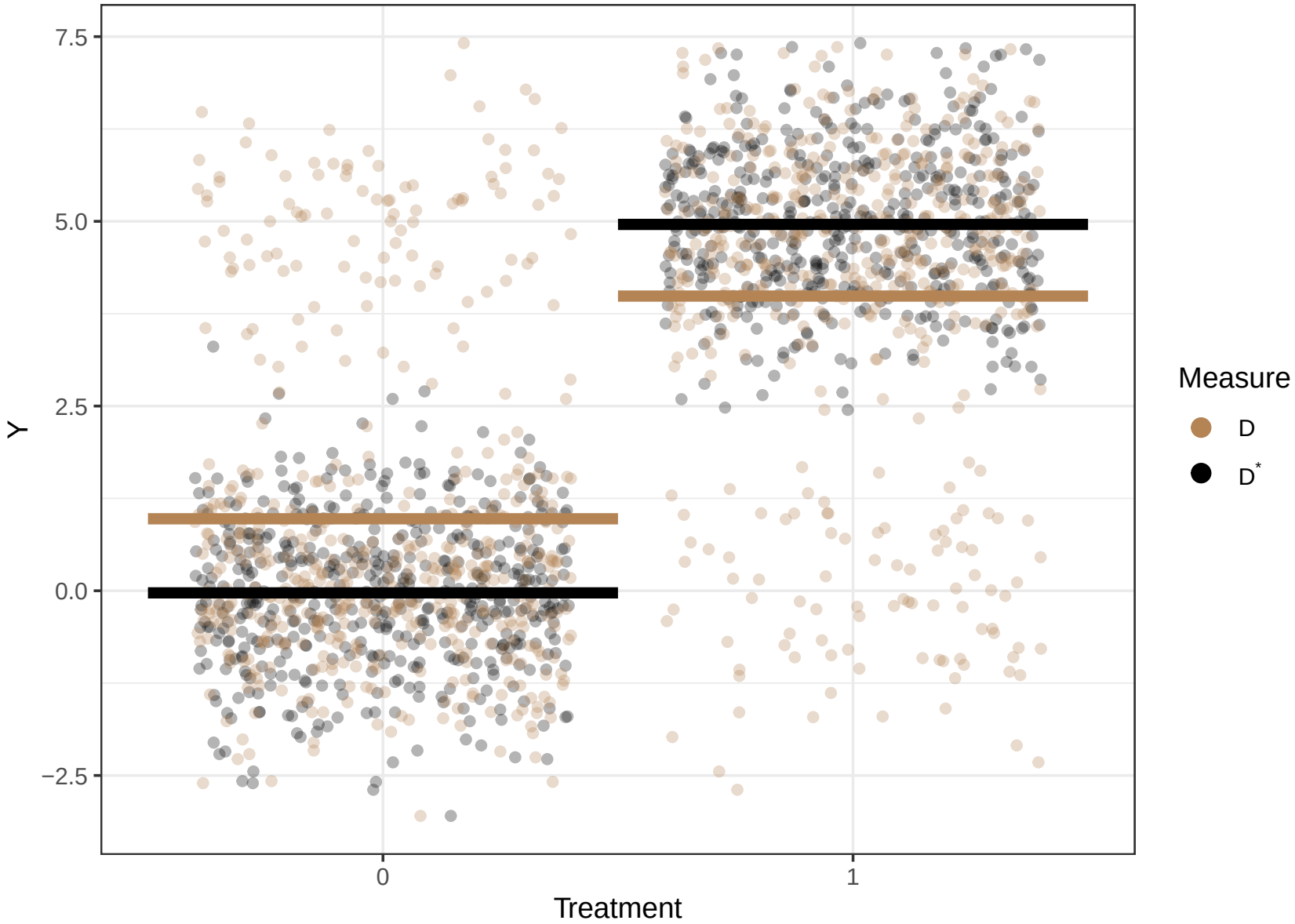
Misclassification causes attenuation bias in OLS estimates:

$$Y_i = c + \beta D_i^* + X_i'\gamma + \epsilon_i$$
$$Y_i = \hat{c} + \hat{\beta} D_i + X_i'\hat{\gamma} + \hat{e}_i$$

Attenuation bias towards zero (Cochran, 1968)

$$\begin{aligned}
\beta &= E[Y|D^* = 1] - E[Y|D^* = 0] \\
\hat{\beta} &= E[Y|D = 1] - E[Y|D = 0] \\
&= E[Y|D_m^* = 0, D_a^* = 1] - E[Y|D_m^* = 1, D_a^* = 0]
\end{aligned}$$

Attenuation Bias



Assumptions

- Symmetric D^*: $\pi_{D^*} = 0.5$
- No correlation between X and D^*: $\rho_{X,D^*} = 0$
- No differential misclassification $Pr(m_i = 1) \perp\!\!\!\perp X, D^*, Y$: $\rho_{m,D^*} = 0$

Correcting the Bias: Nguimkeu et al. (2021)

- Bias Adjusted Least Squares (BALS) for $\rho_{X,D^*} \neq 0$ and $\rho_{ME,D^*} = 0$
- Requires estimation of misclassification probabilities:

$$\alpha_0 = \frac{FP}{TN + FP}, \quad \alpha_1 = \frac{FN}{FN + TP}$$

- I.e., Törnberg (2024)
  - D^* & D: True and predicted partisanship
  - R_{GPT4}: proportion incorrectly labeled via ChatGPT 4.0
  - α_0 (α_1): proportion of Dems (Reps) labeled Rep (Dem)
  - $\rho_{m,D^*} = 0$: misclassification uncorrelated with partisanship

If you fall asleep...

- Spoiler: **misclassification is rarely going to overturn conclusions**

If you fall asleep...

- Spoiler: **misclassification is rarely going to overturn conclusions**
- **But** need a "gold standard" sample to evaluate sensitivity
  - Can do this manually, or rely on bleeding edge LLM
  - Look at (1) skew in D^*, (2) misclassification rate of at-scale solution R_m, (3) correlation between m_i and D^*, Y, and X.
  - Use our forthcoming `R` package!

If you fall asleep...

- Spoiler: **misclassification is rarely going to overturn conclusions**
- **But** need a "gold standard" sample to evaluate sensitivity
  - Can do this manually, or rely on bleeding edge LLM
  - Look at (1) skew in D^*, (2) misclassification rate of at-scale solution R_m, (3) correlation between m_i and D^*, Y, and X.
  - Use our forthcoming `R` package!
- Roadmap:
  1. Simulation evidence: how bad can it be? (black and red)
  2. Simulation evidence: can we fix it? (teal and tomato)
  3. Calibration evidence: what does it look like "in the wild"?

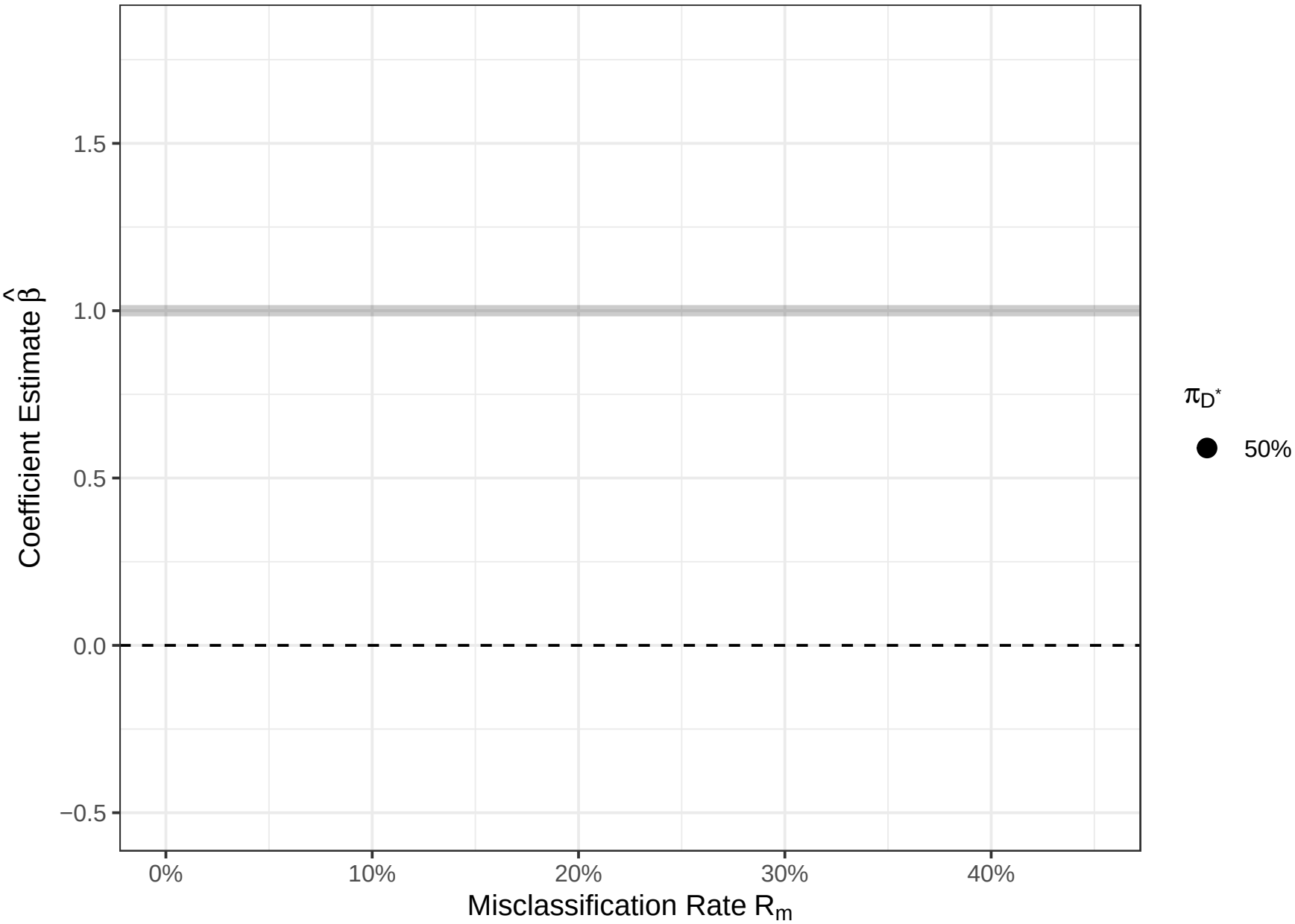
Simulation Evidence

- RQ1: How problematic is not using SOTA?
- RQ2: Can bias be corrected?

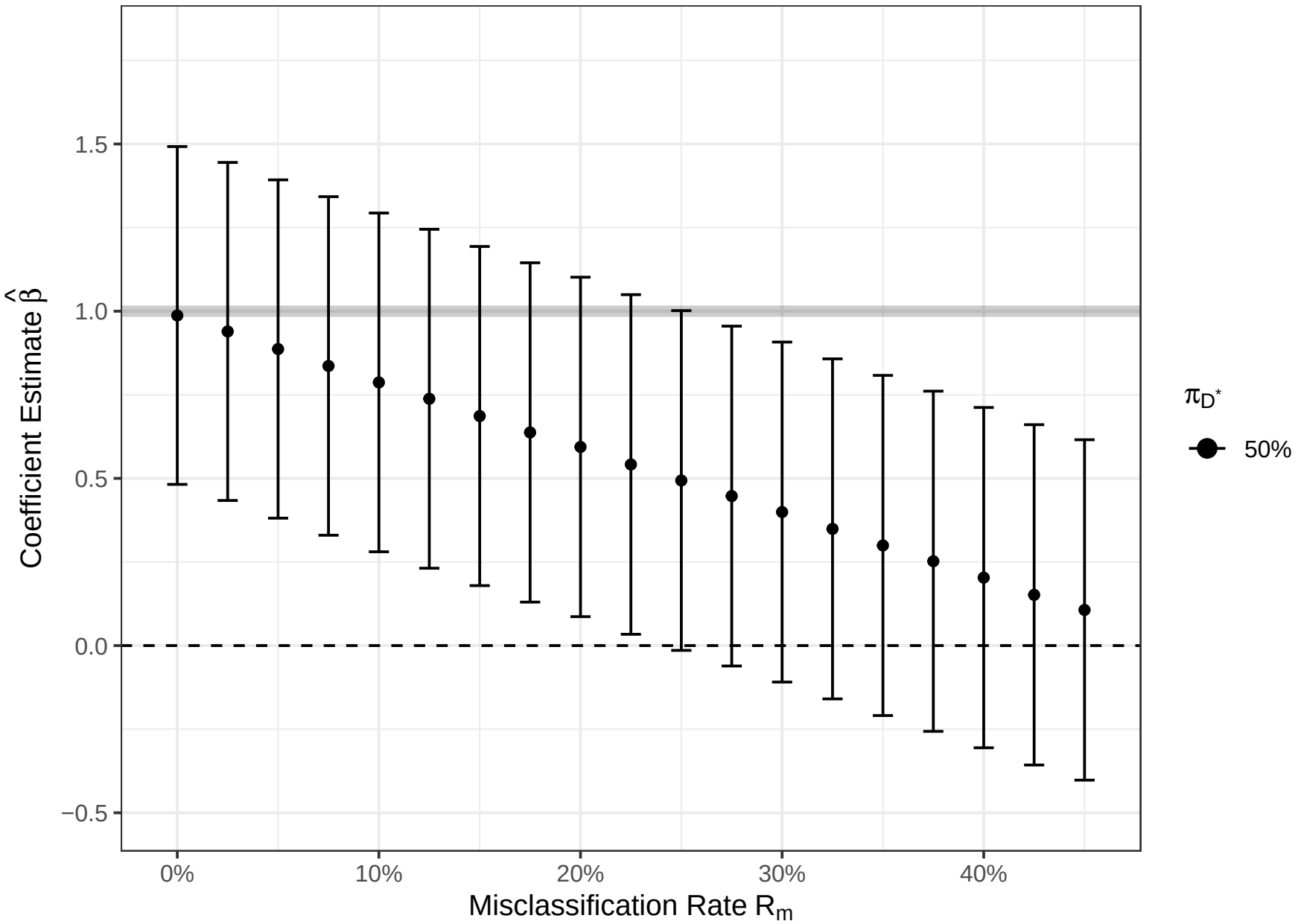
Simulations + calibration

- $Y = c + \beta D^* + X'\gamma + \varepsilon$
- Hyperparameters: R_{method}, π_{D^*}, ρ_{X,D^*}, ρ_{m,D^*}

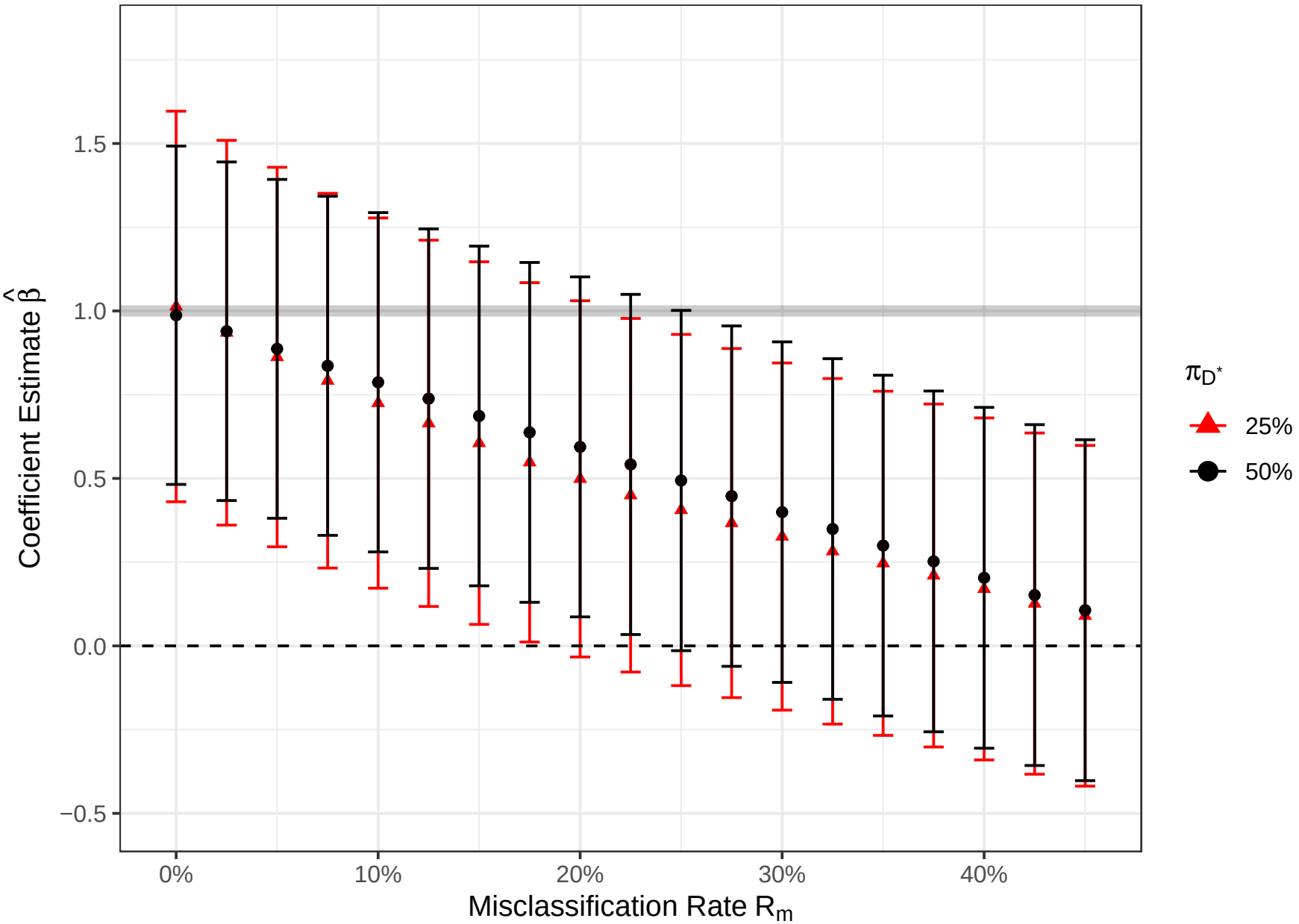
How bad can it be?



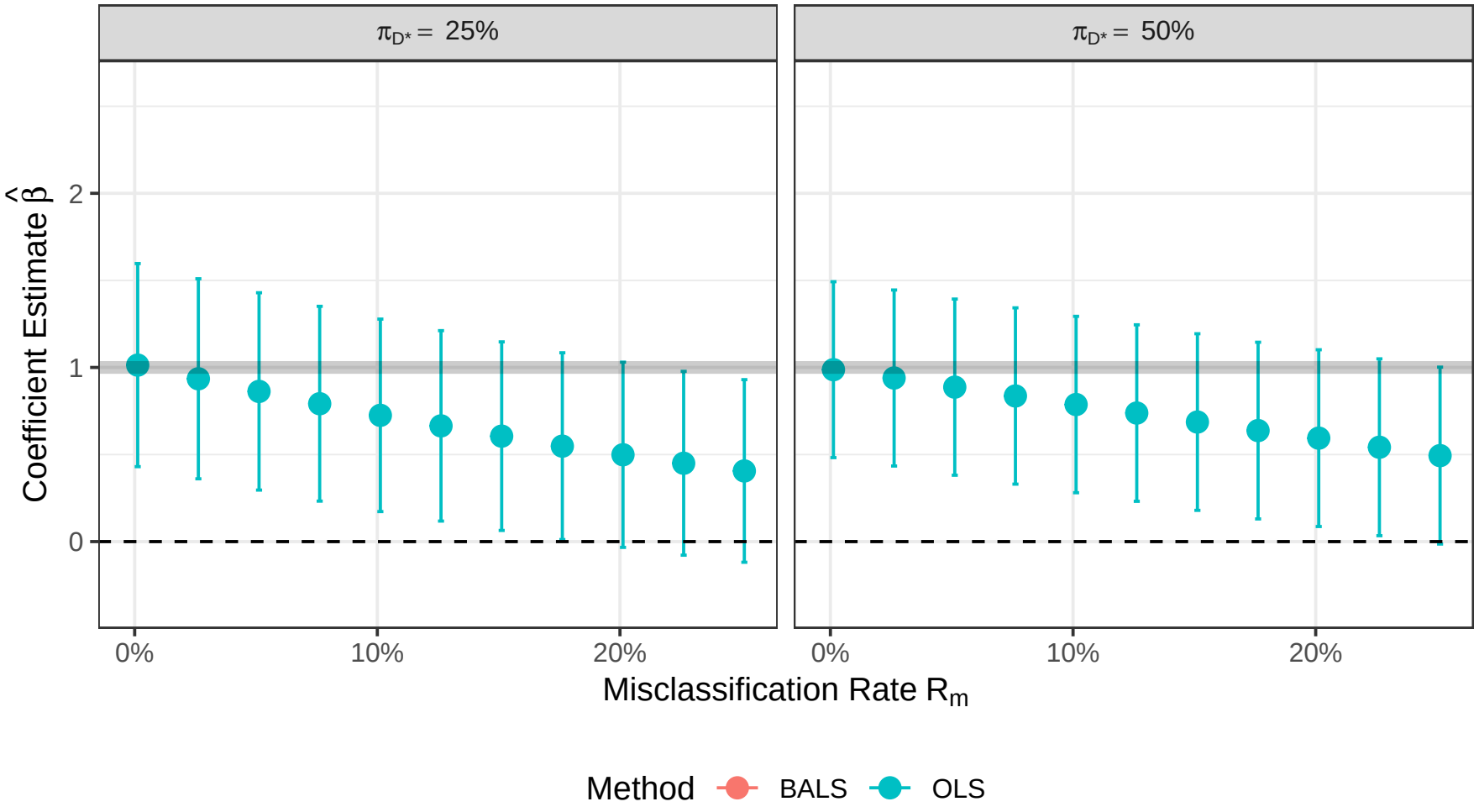
How bad can it be?



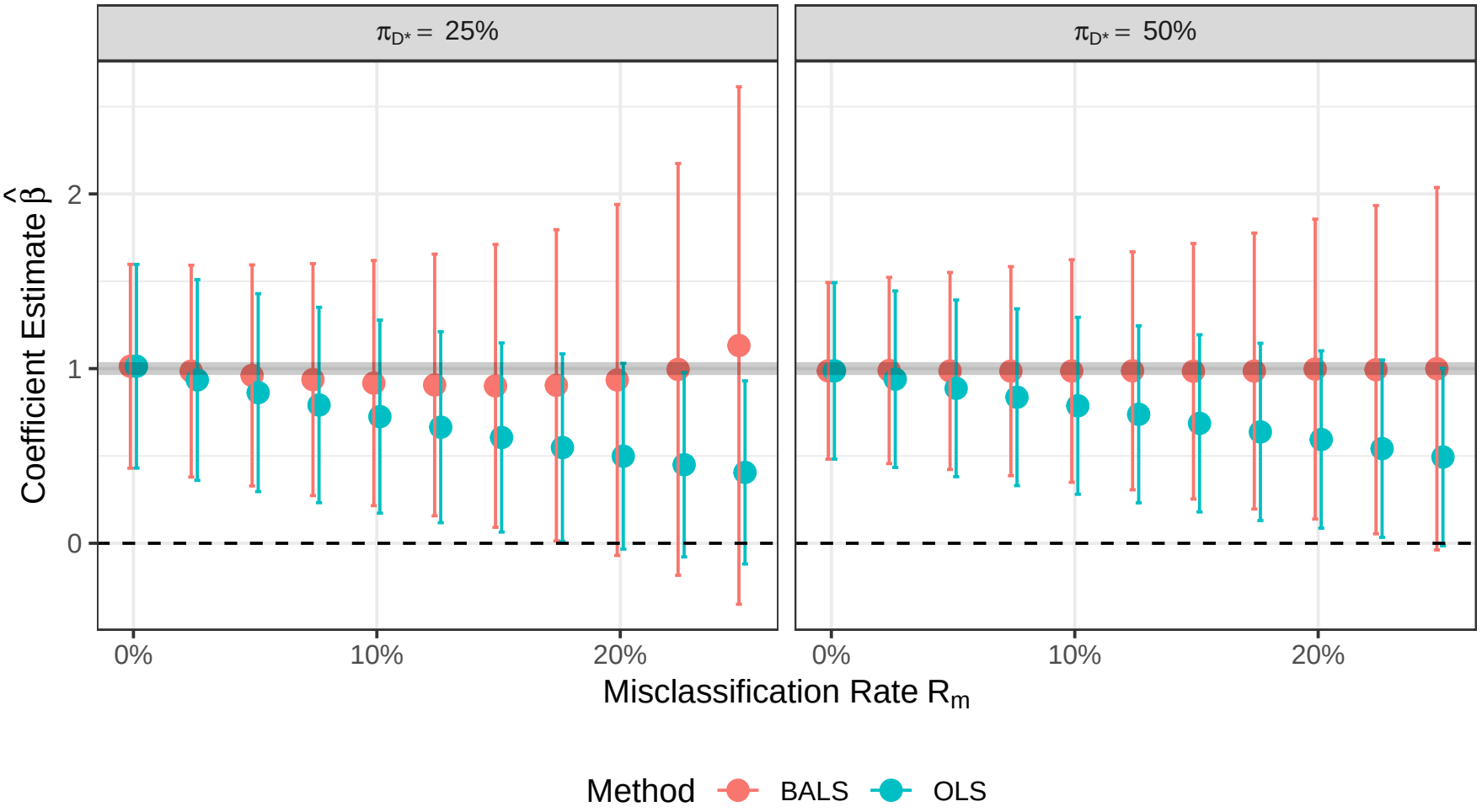
How bad can it be?



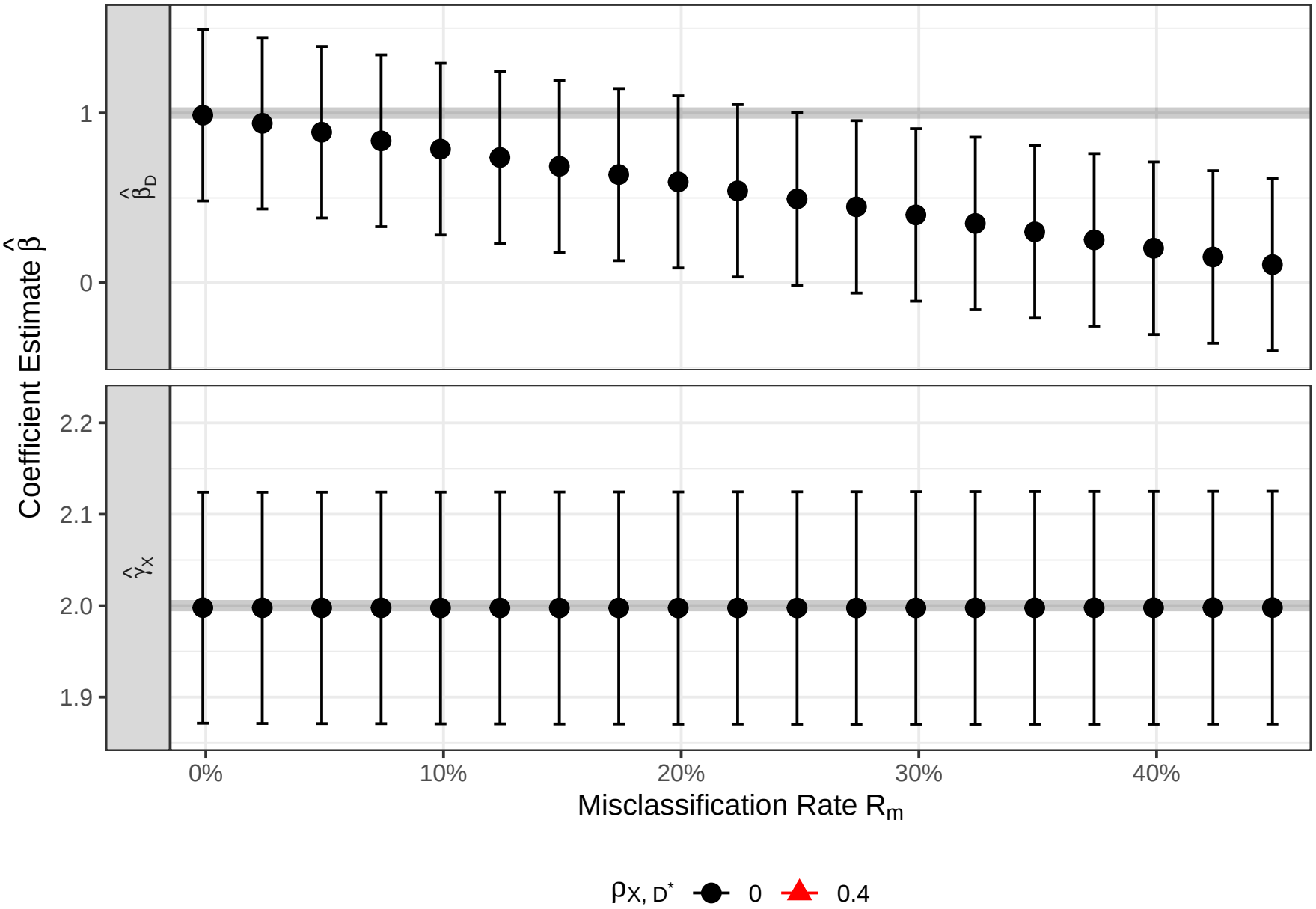
Is it correctable?



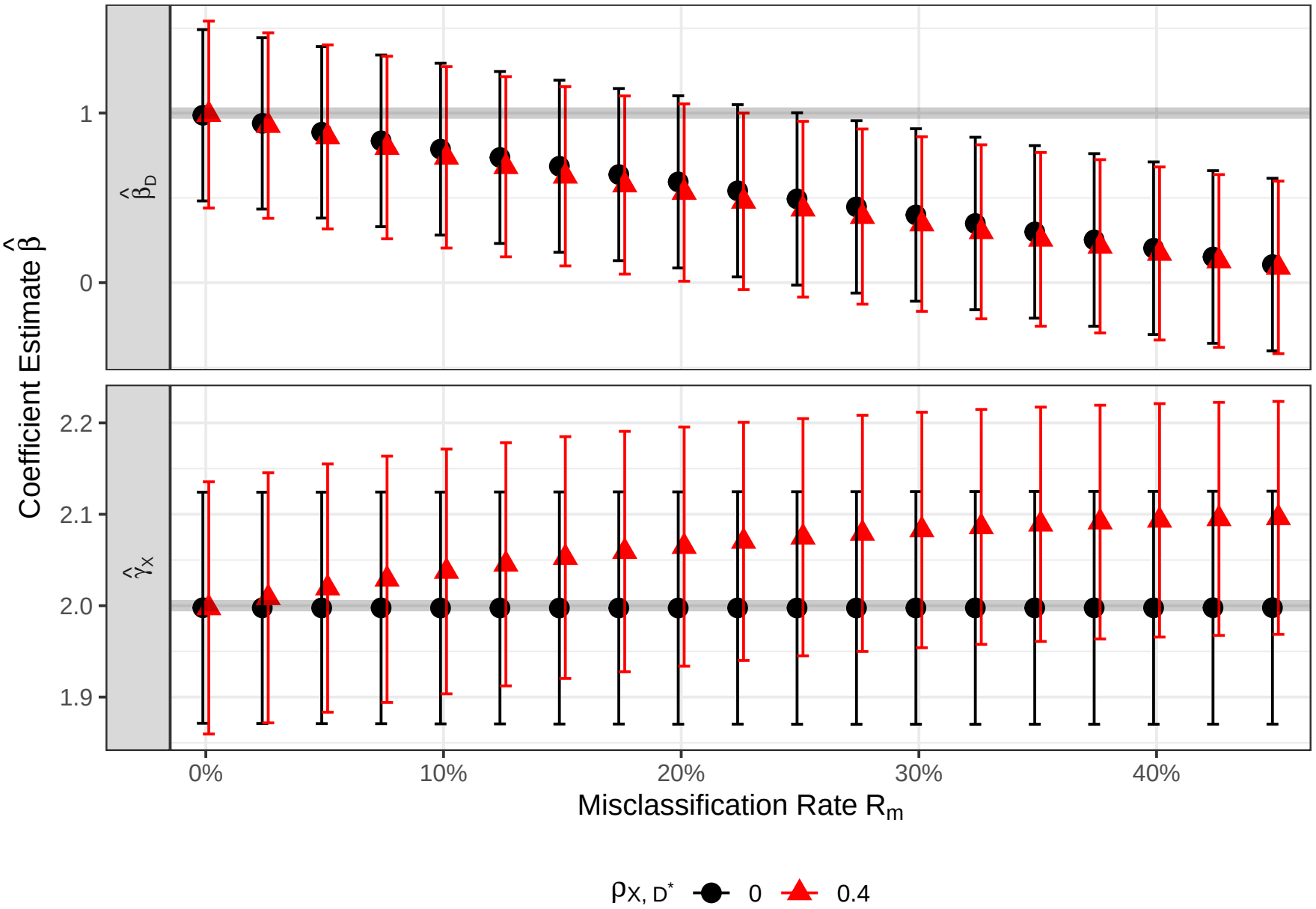
Is it correctable?



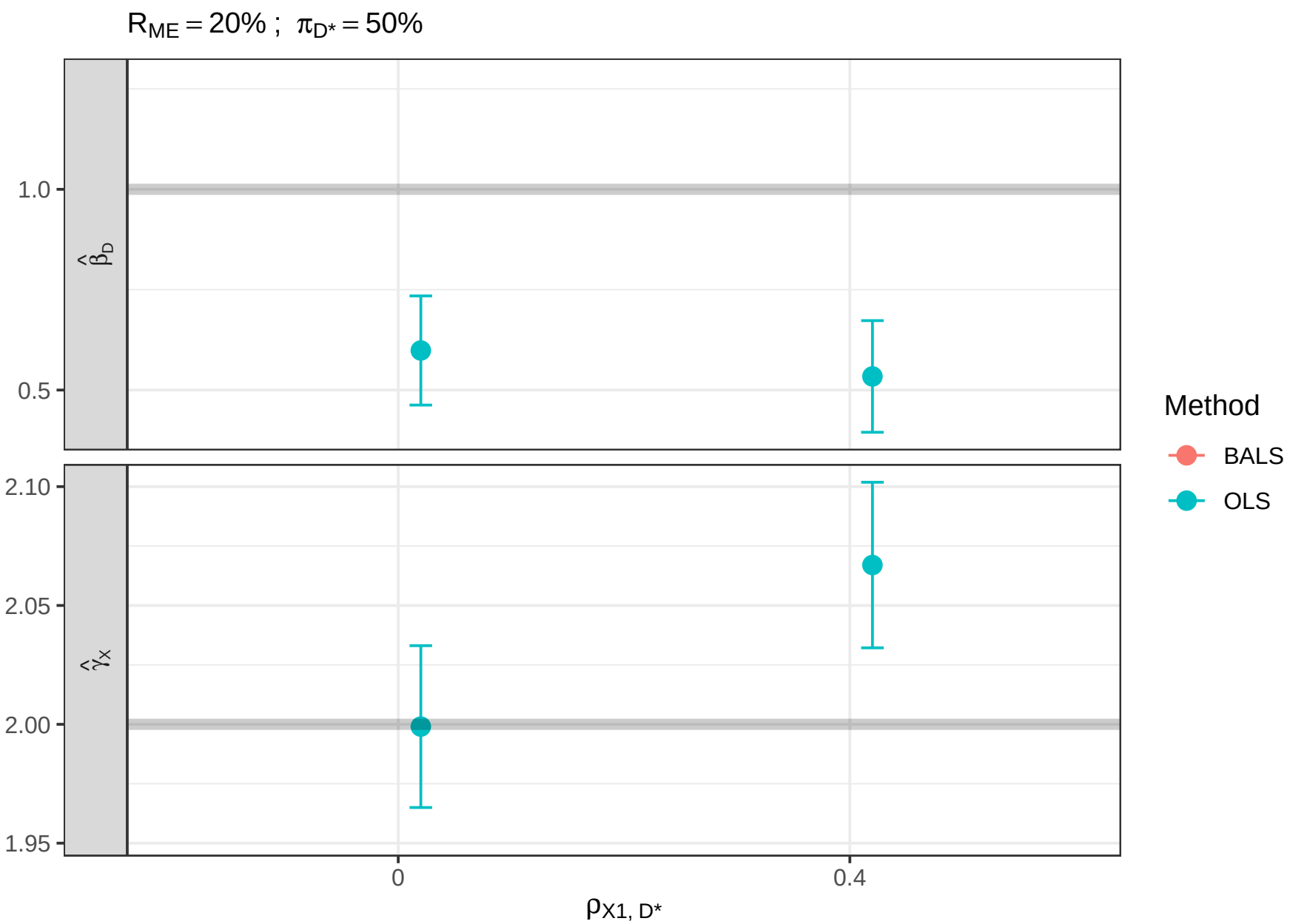
Allowing for correlated controls



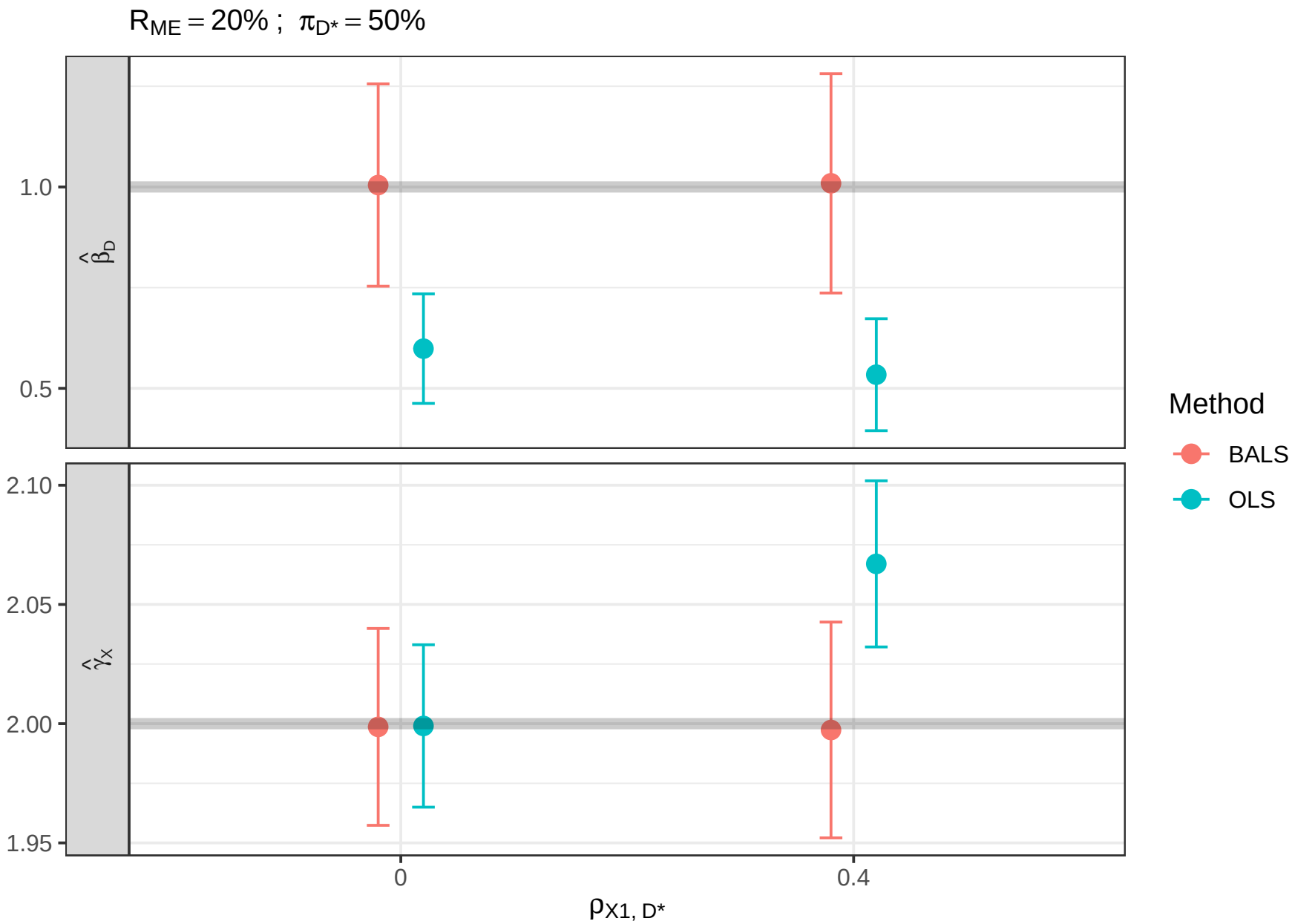
Allowing for correlated controls



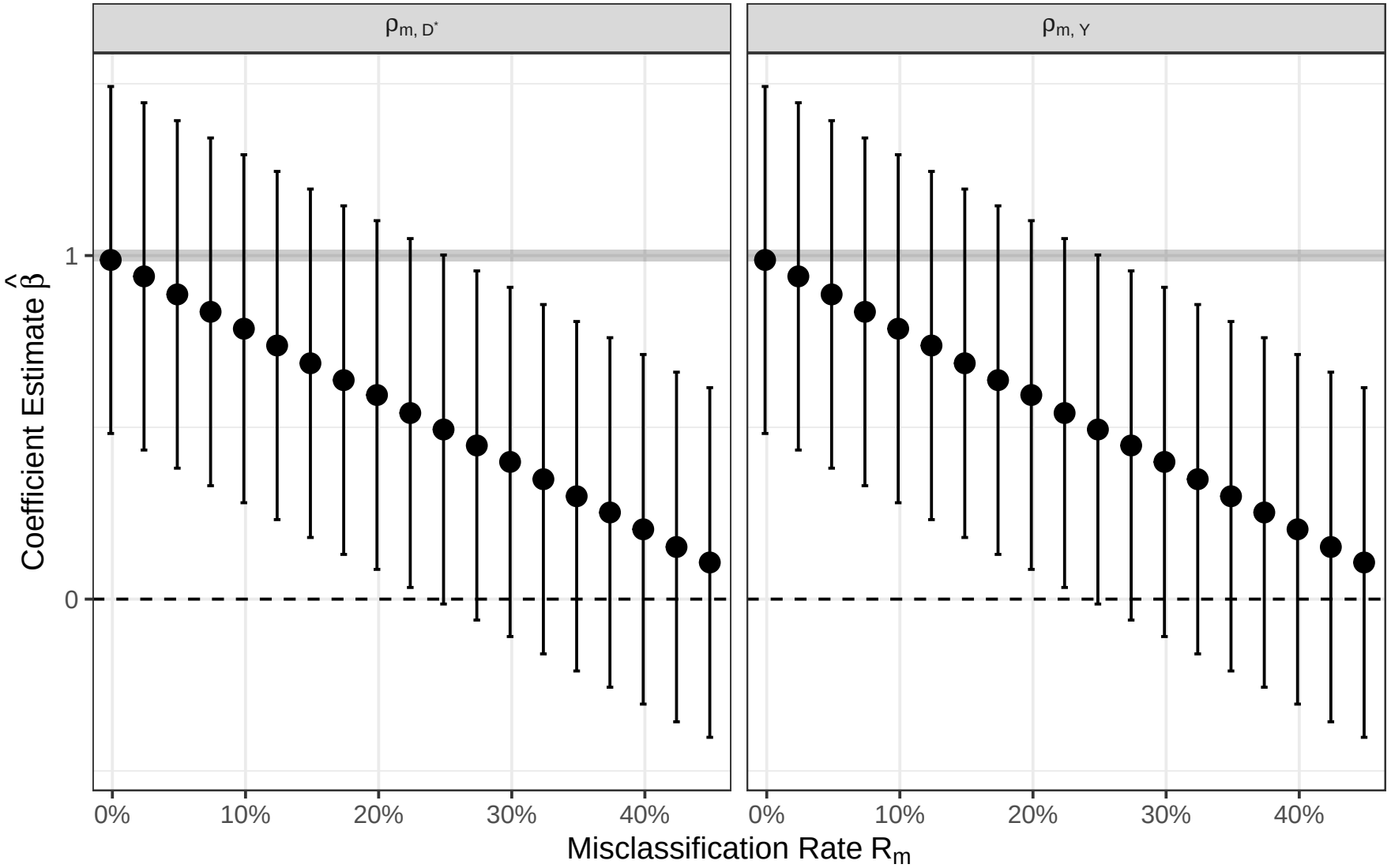
Fixing Correlated Controls



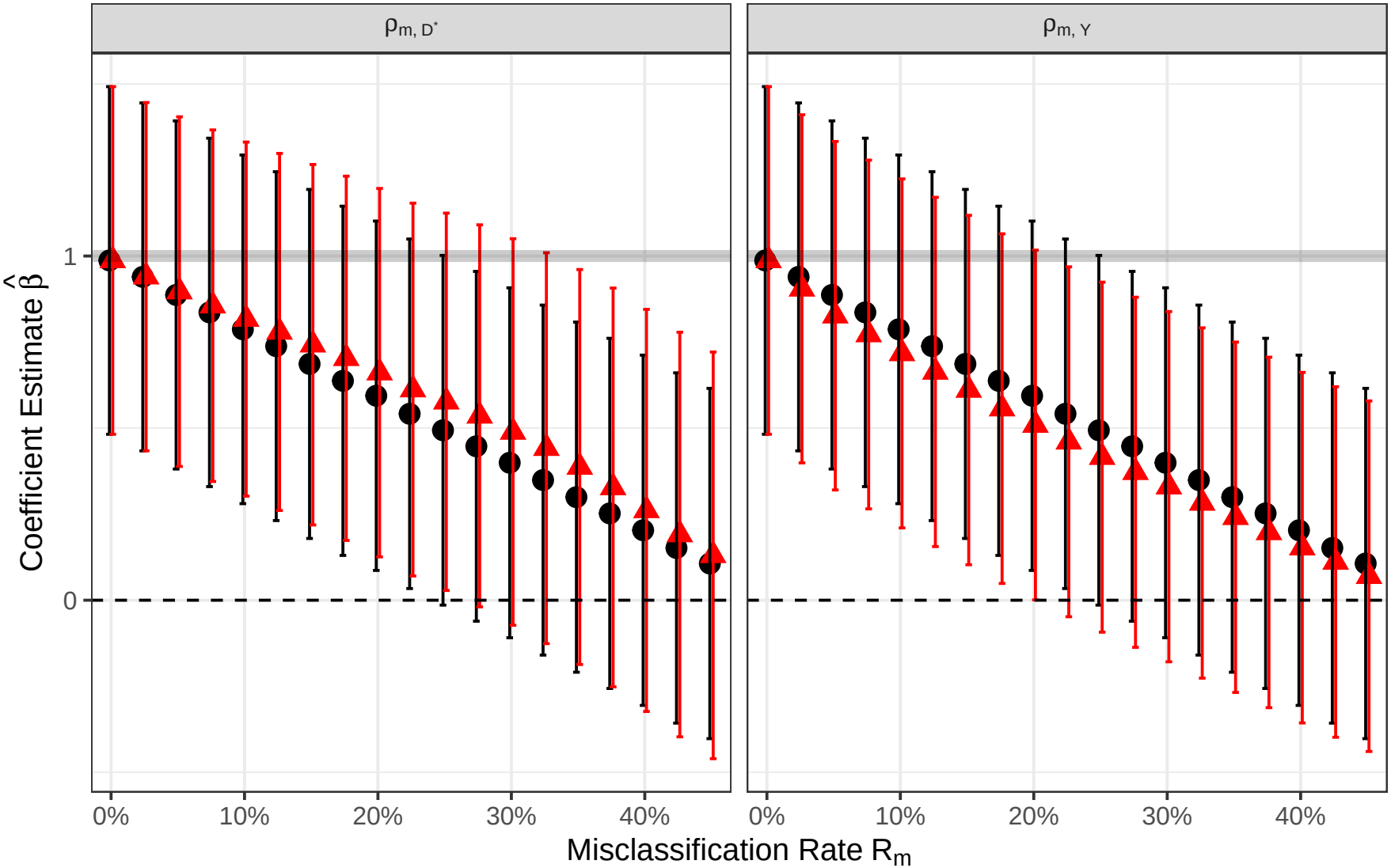
Fixing Correlated Controls



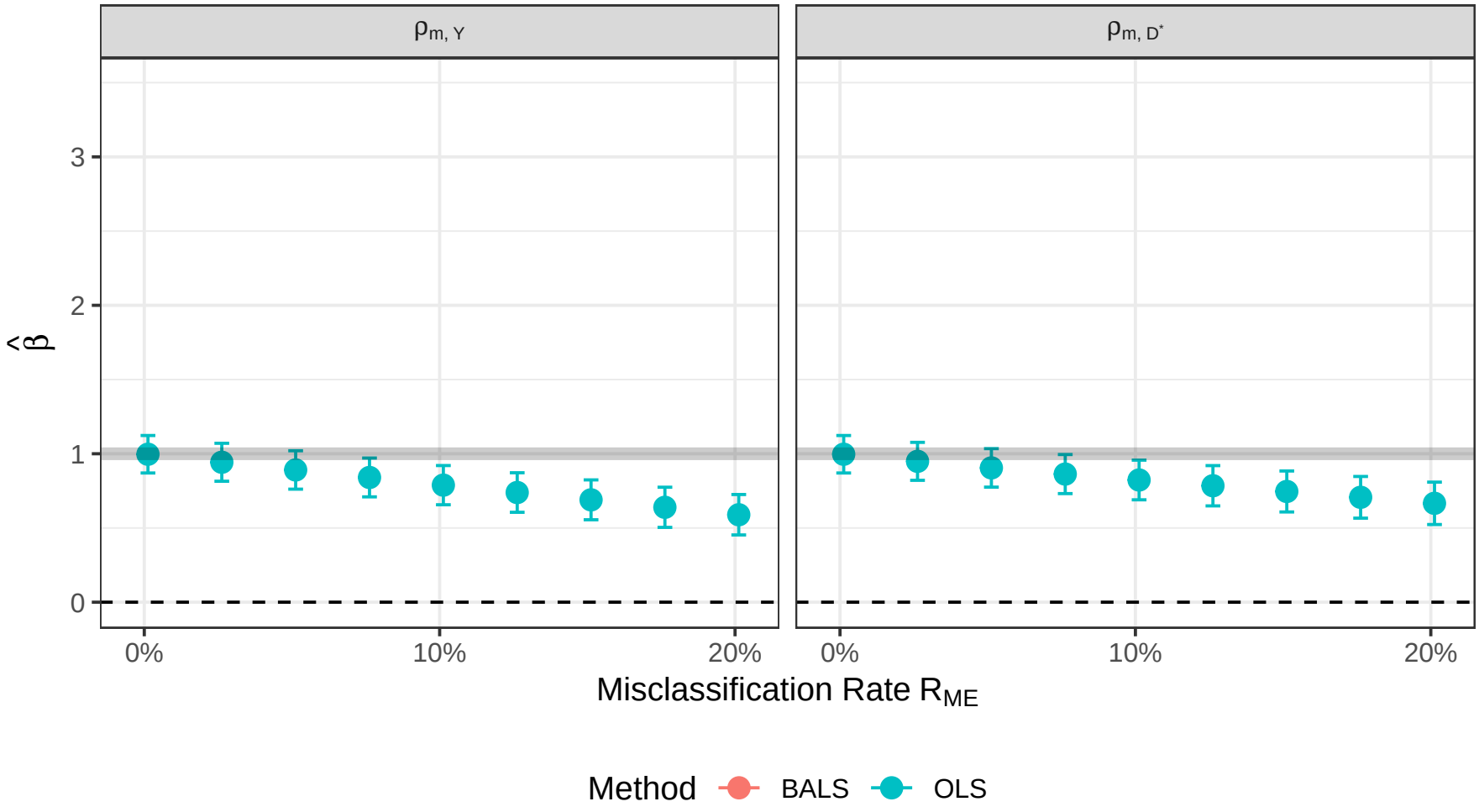
Allowing for differential misclassification



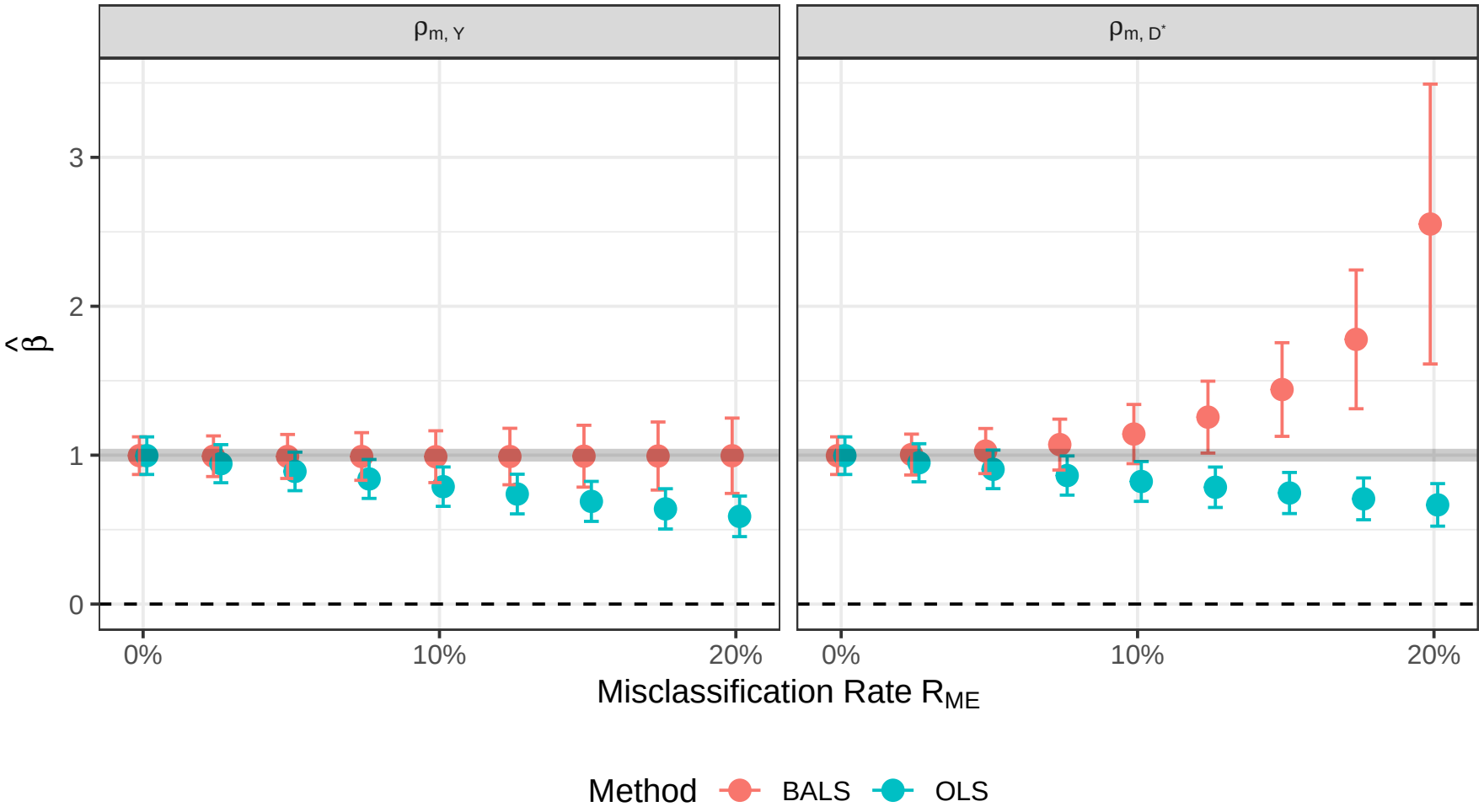
Allowing for differential misclassification



Fixing differential misclassification



Fixing differential misclassification



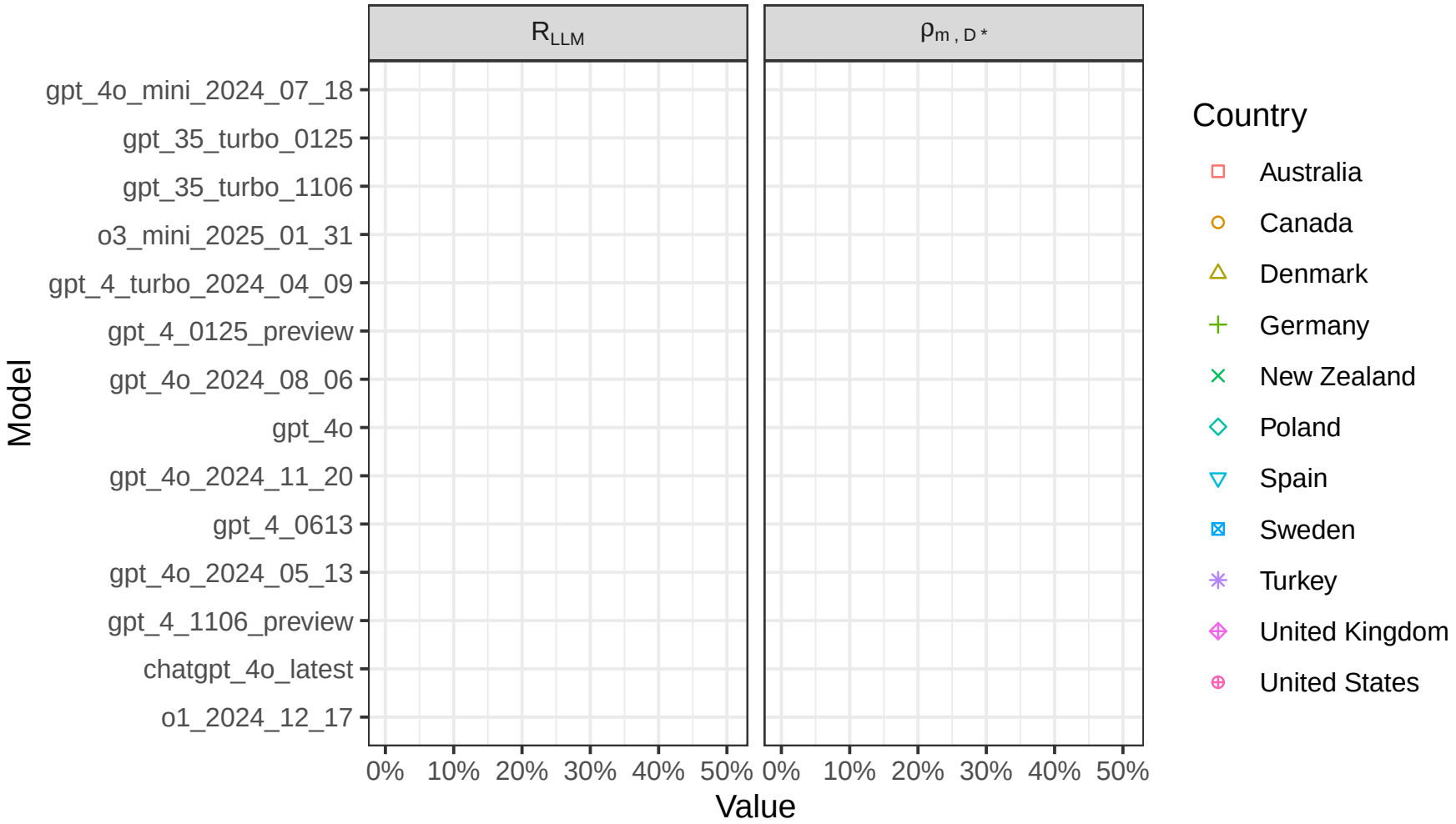
Simulation Evidence

- Simulations reveal:
  - Misclassification causes attenuation bias in $\hat{\beta}$
  - BALS correction largely removes bias
  - Moderate sensitivity to **reasonable(?)** violations of assumptions

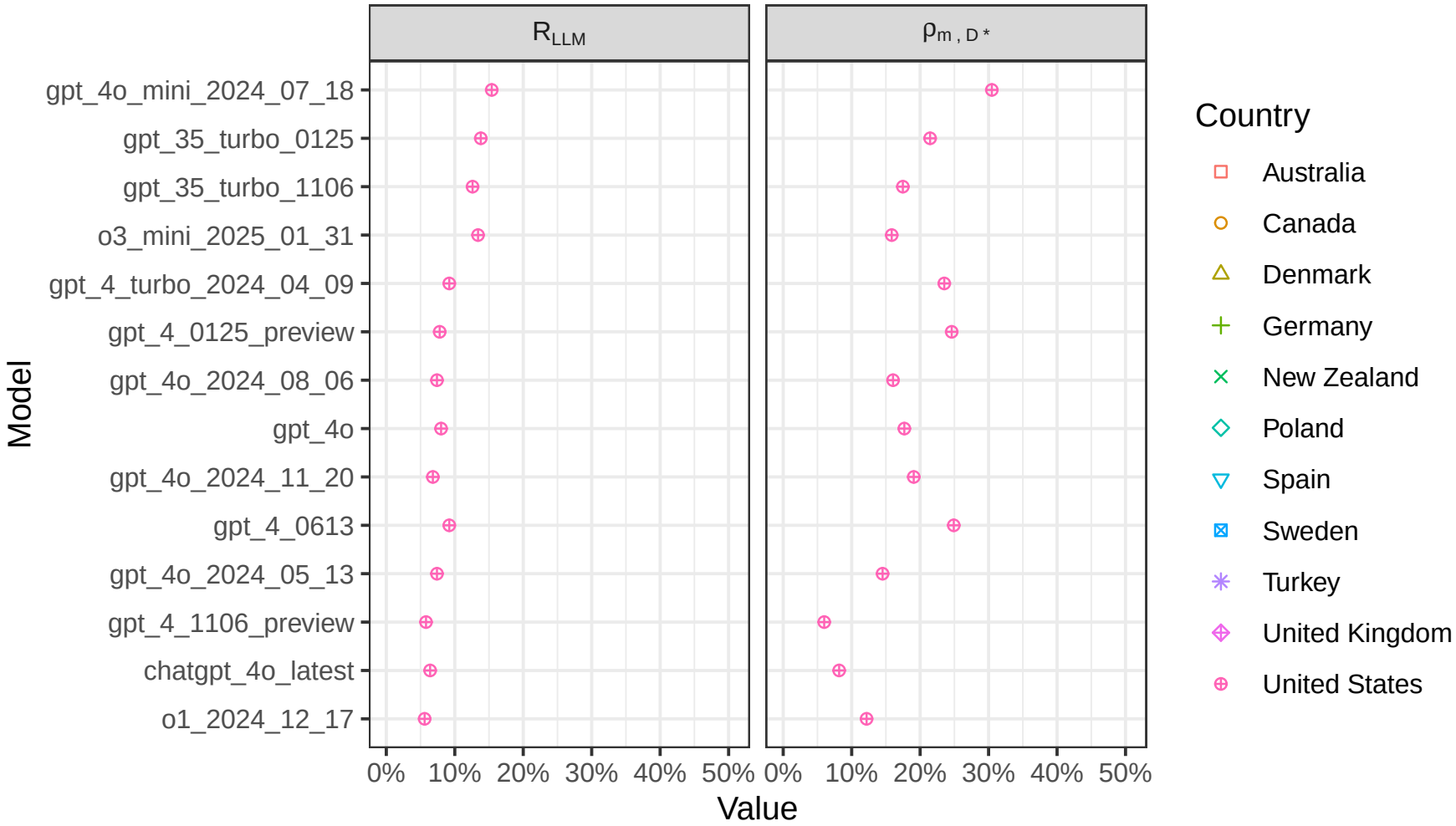
Replication-based calibration

- Törnberg (2024): annotating account partisanship using tweets
  - 11 countries annotated for two dominant parties
  - $\pi_{D^*} = 0.5$ and $\rho_{m,D^*} \approx 0$, but for other outcomes $\rho_{m,Y} \neq 0$!
- Rathje et al. (2024): sensitivity of conclusions to misclassification
  - ChatGPT 4.0 used to annotate social media posts for offensiveness
  - Ground truth data coded by humans
  - $\pi_{D^*} = 0.28$ and $\rho_{ME,D^*} \neq 0$!

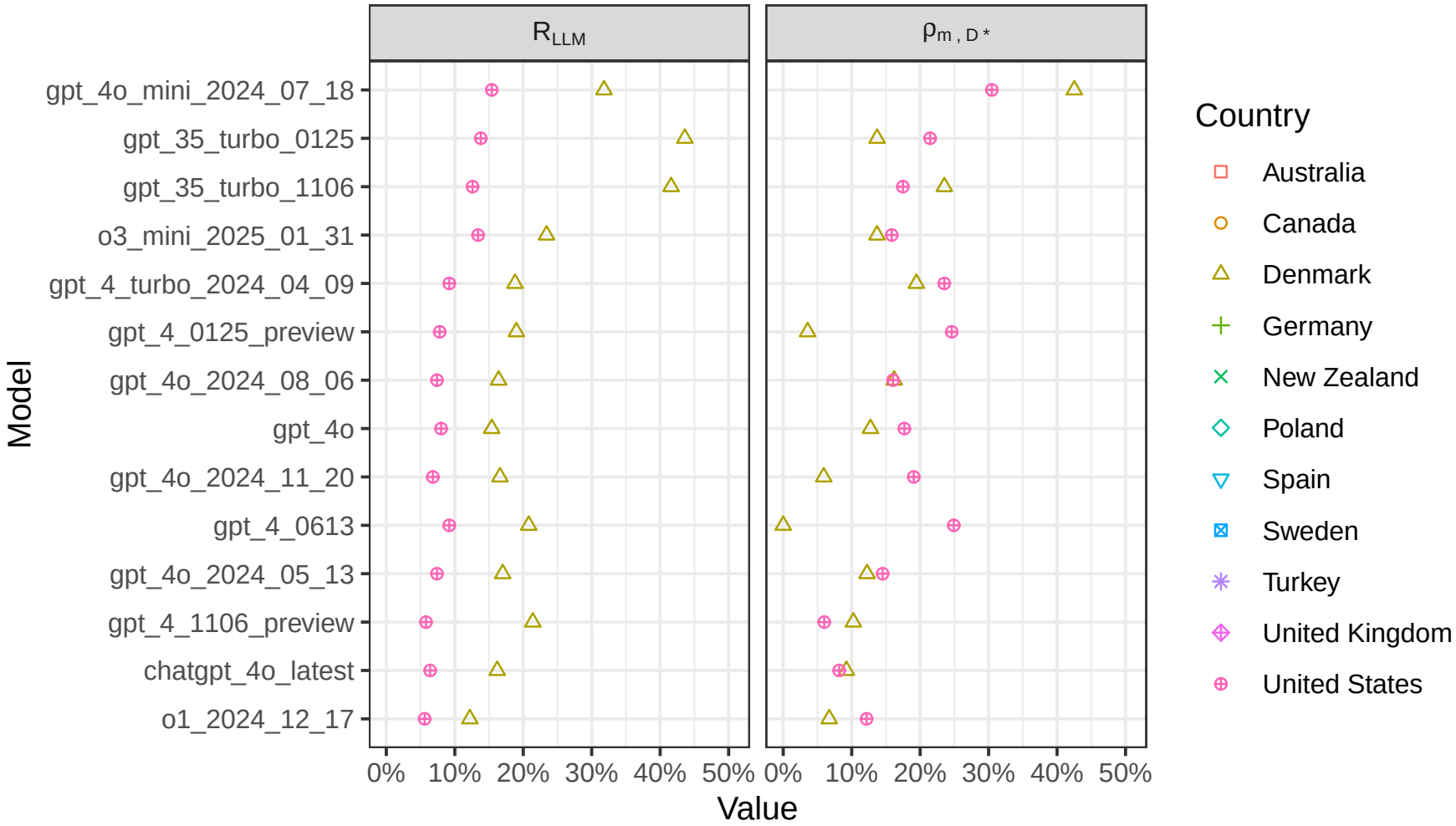
Case Study: Partisanship from Tweets (Törnberg, 2024)



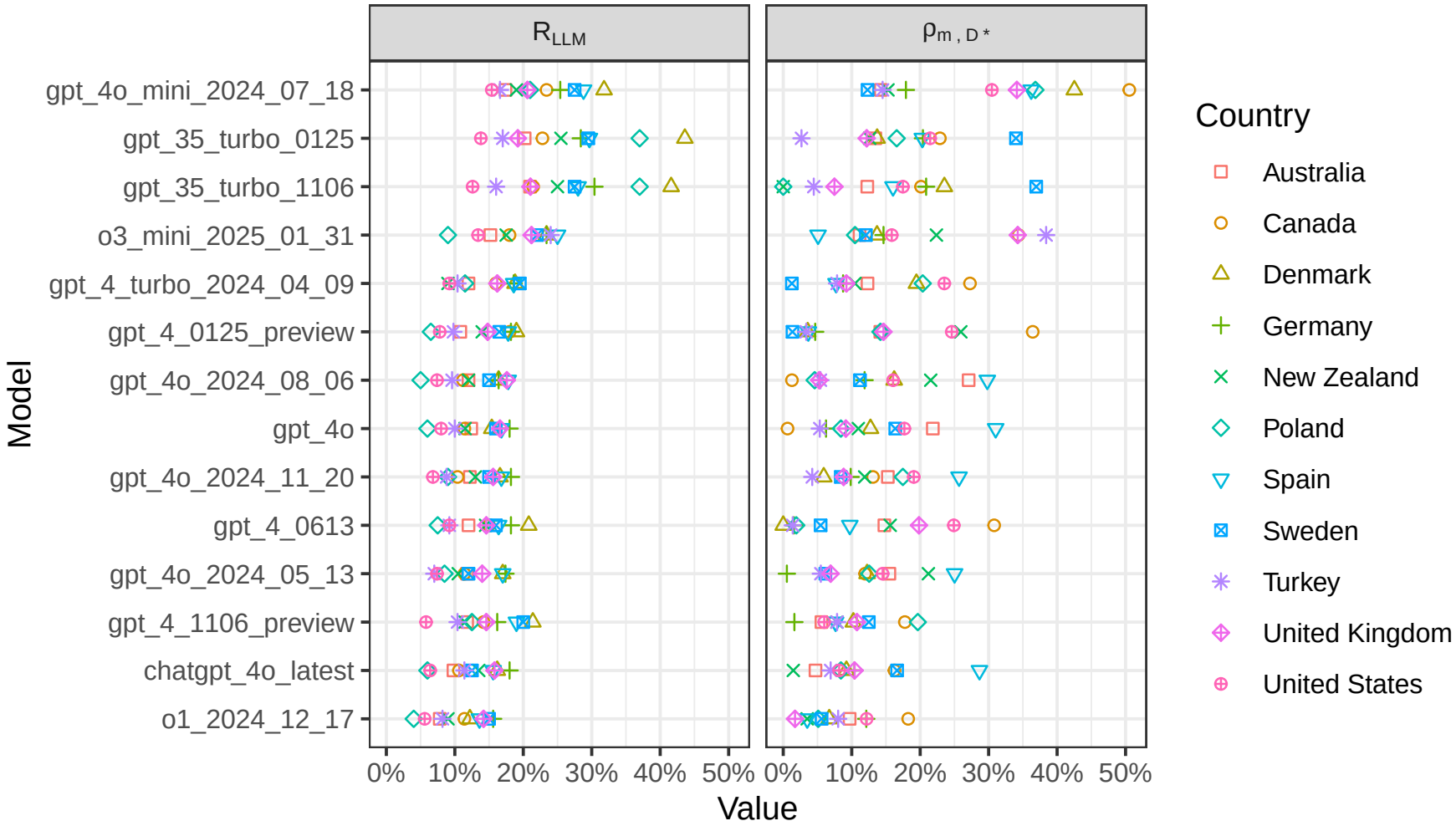
Case Study: Partisanship from Tweets (Törnberg, 2024)



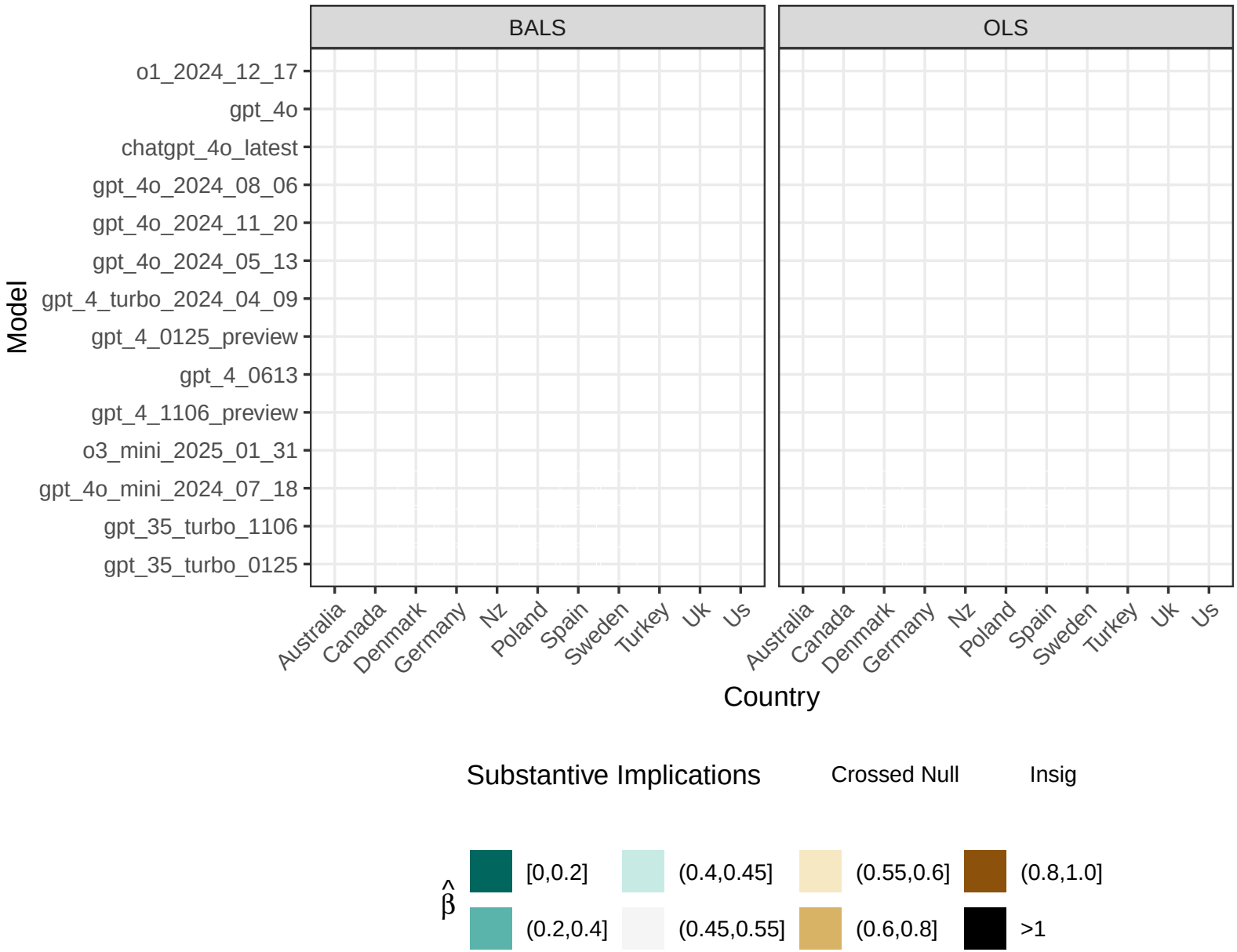
Case Study: Partisanship from Tweets (Törnberg, 2024)



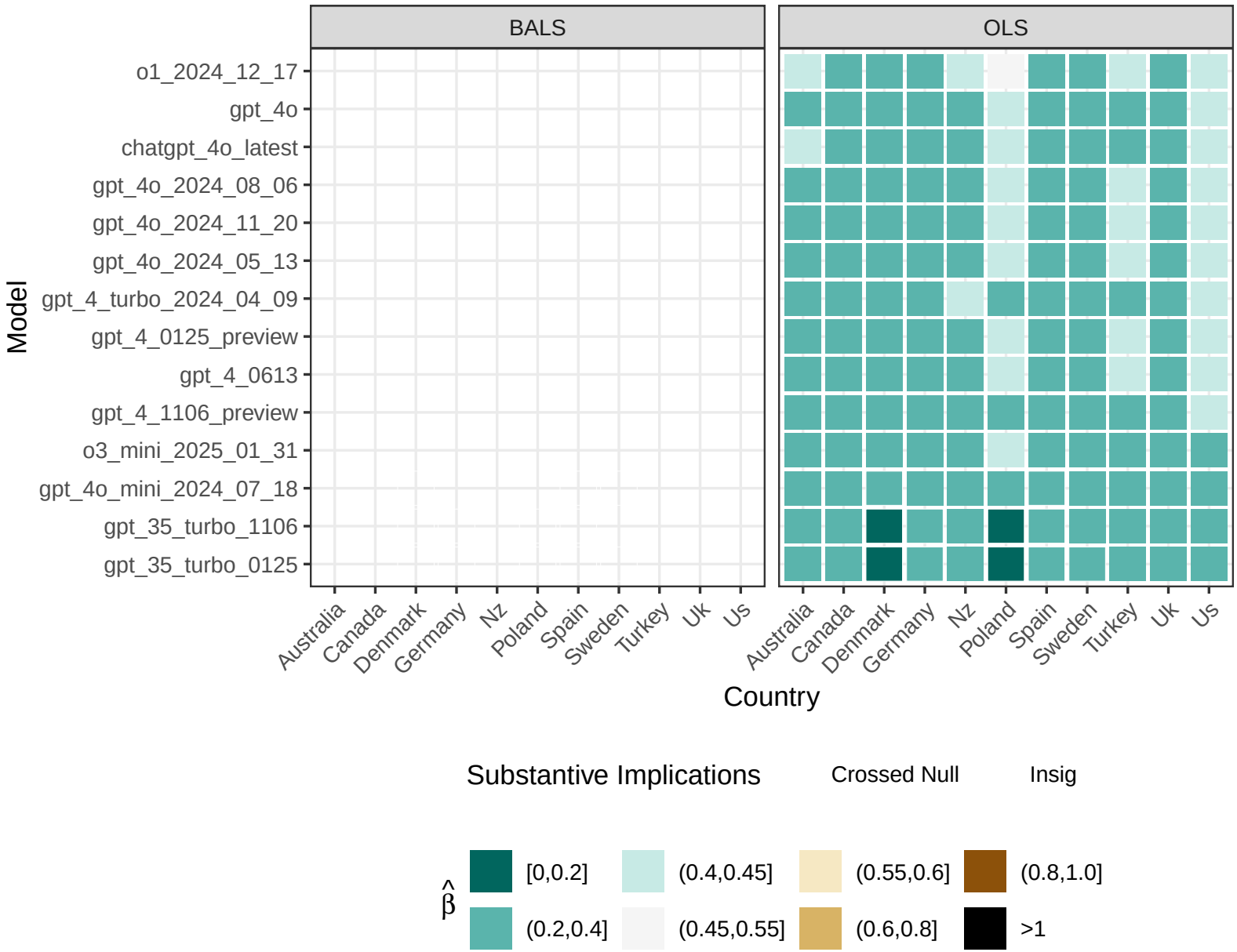
Case Study: Partisanship from Tweets (Törnberg, 2024)



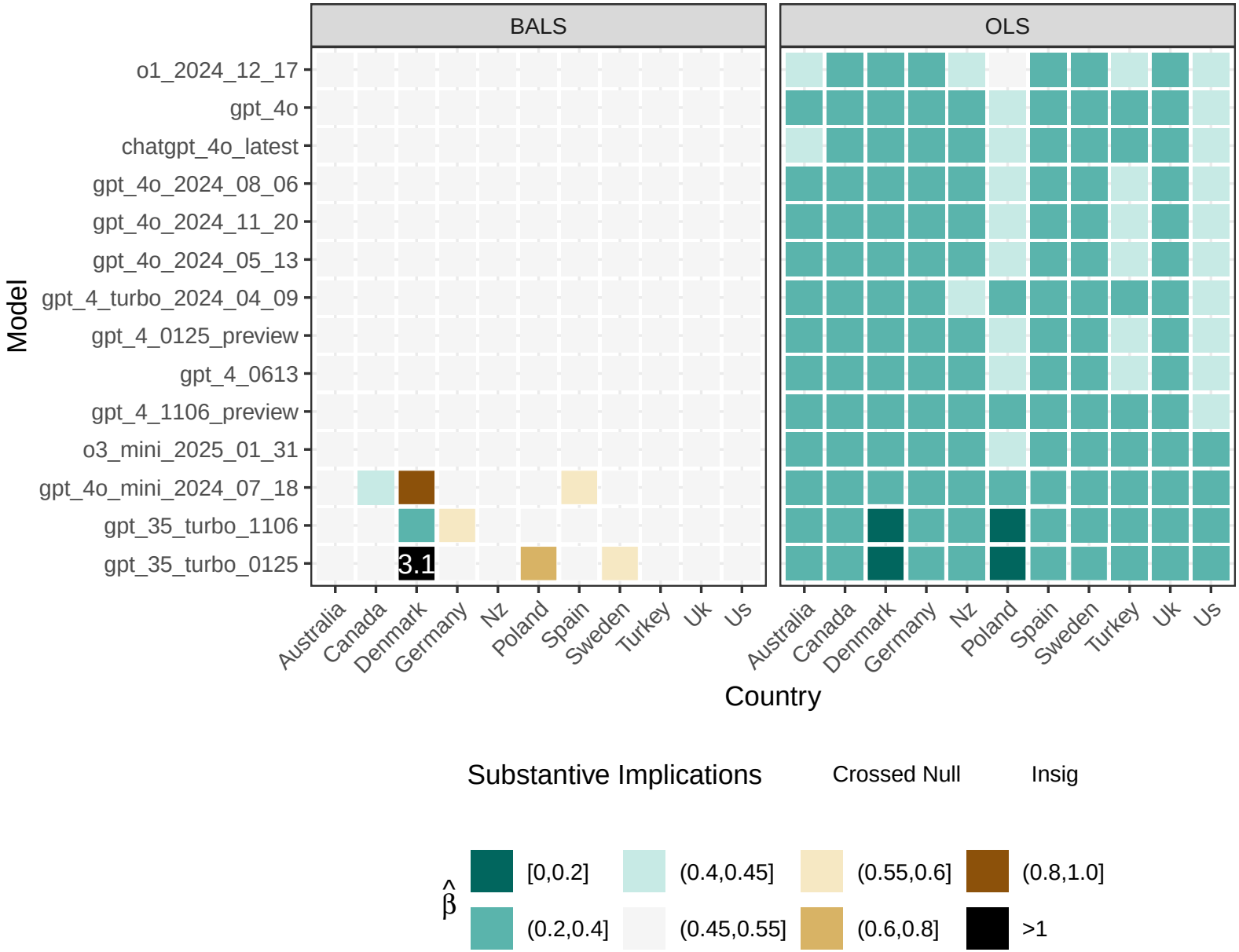
Case Study: Partisanship from Tweets (Törnberg, 2024)



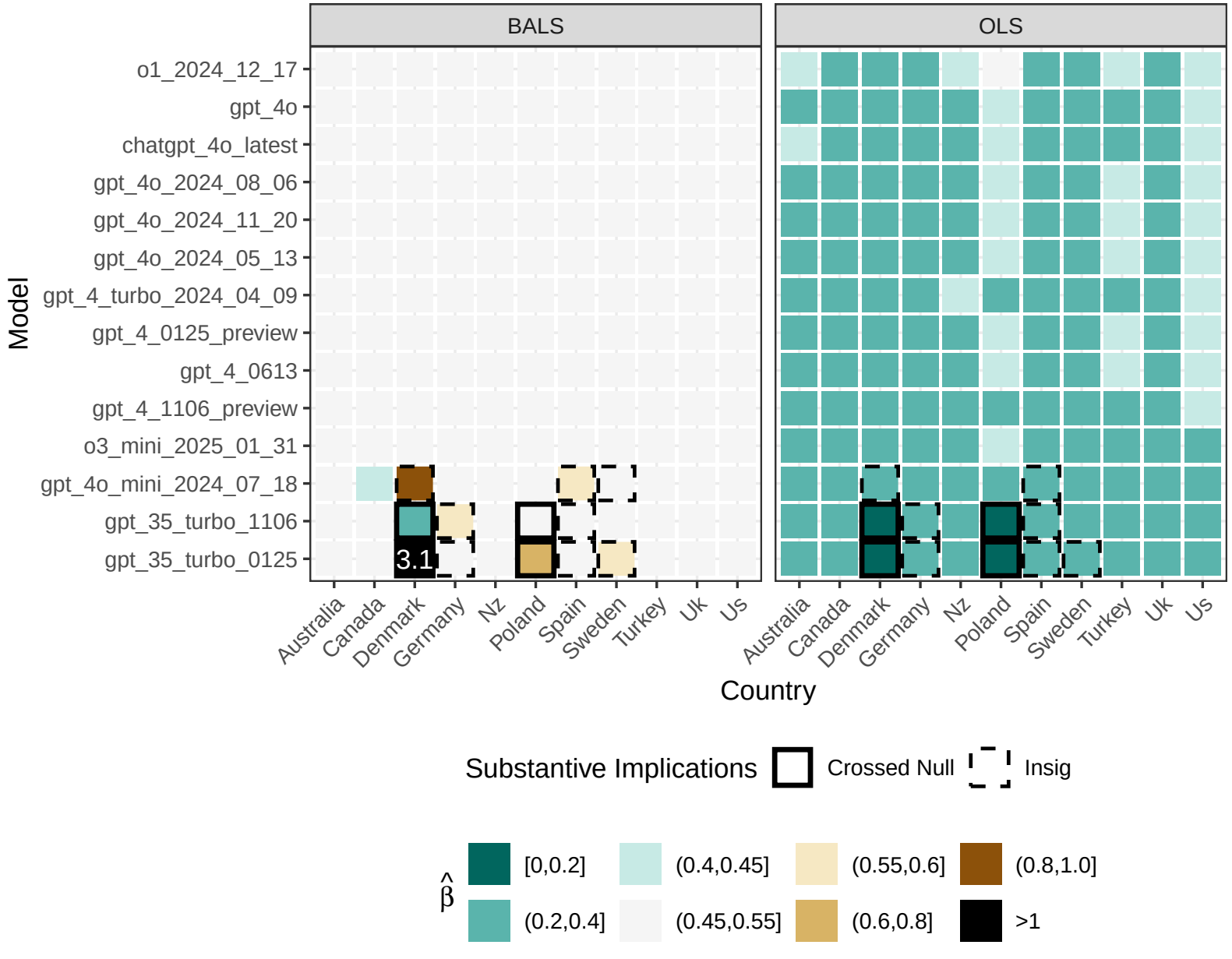
Case Study: Partisanship from Tweets (Törnberg, 2024)



Case Study: Partisanship from Tweets (Törnberg, 2024)



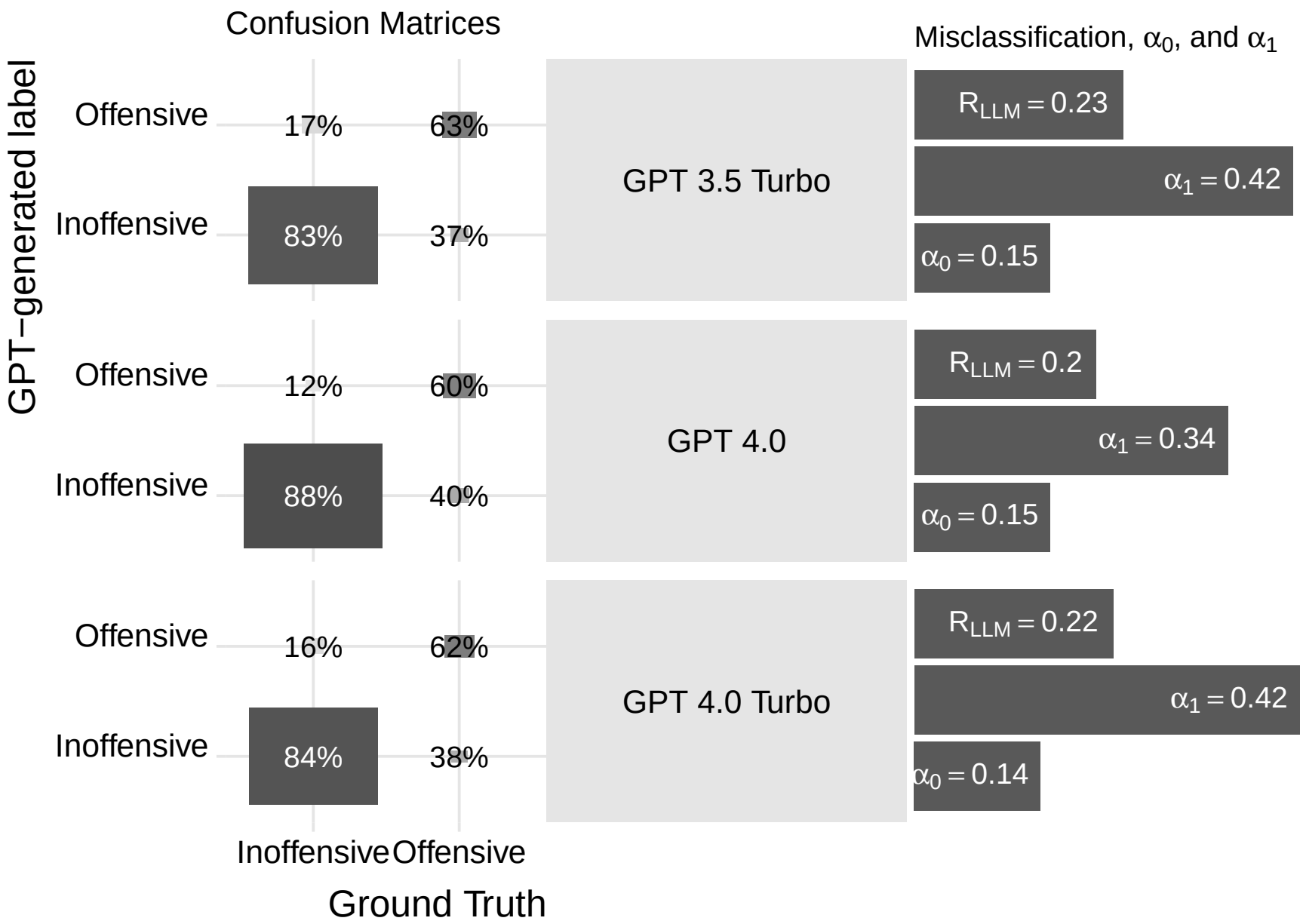
Case Study: Partisanship from Tweets (Törnberg, 2024)



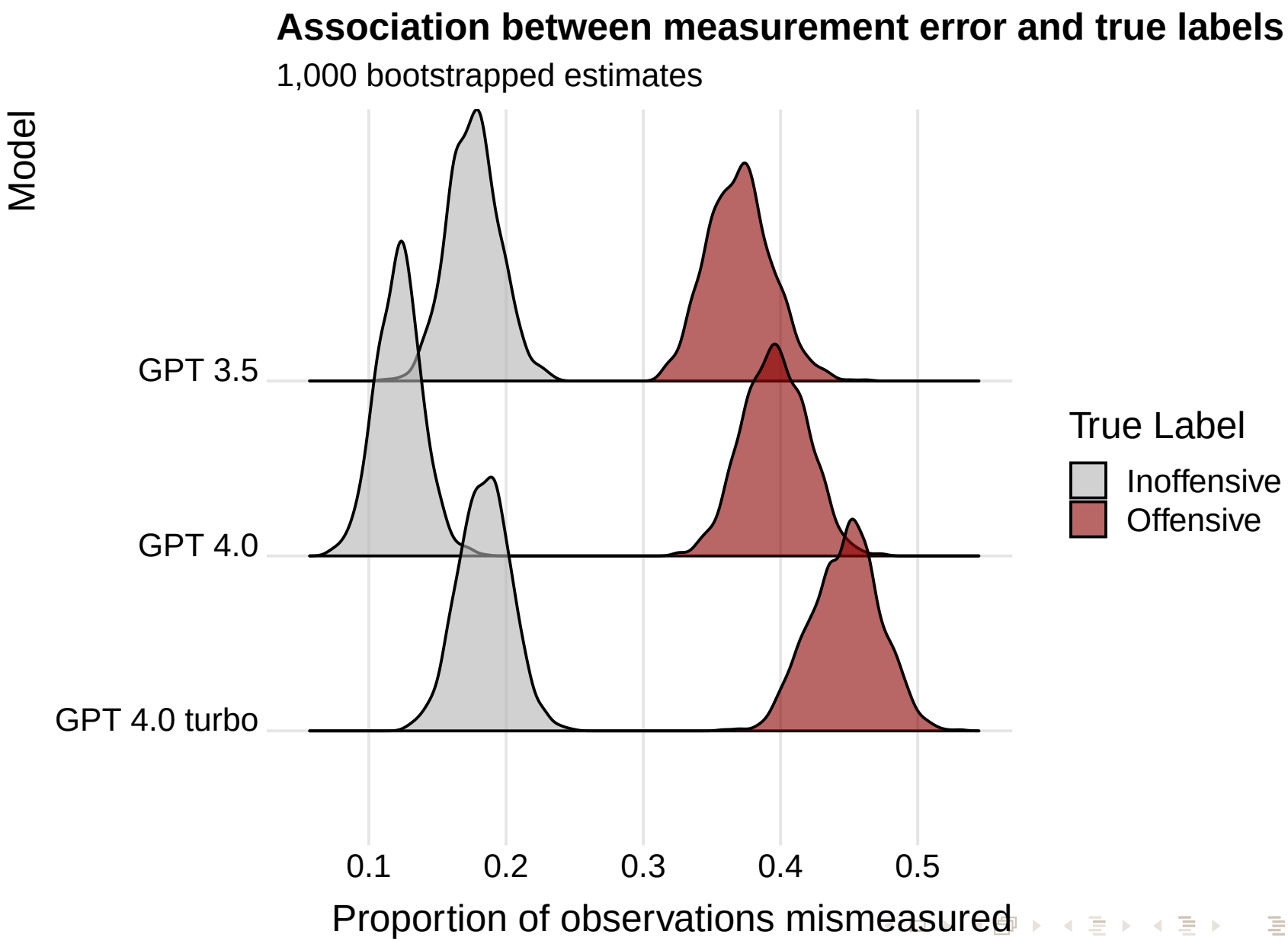
Case Study: Partisanship from Tweets (Törnberg, 2024)

- ChatGPT predicting politicians' partisanship
- Replication with multiple LLMs shows small variation in misclassification
- BALS effective in correcting bias across LLMs **for outcome uncorrelated with misclassification by design**
    - Need to be attentive to specific application

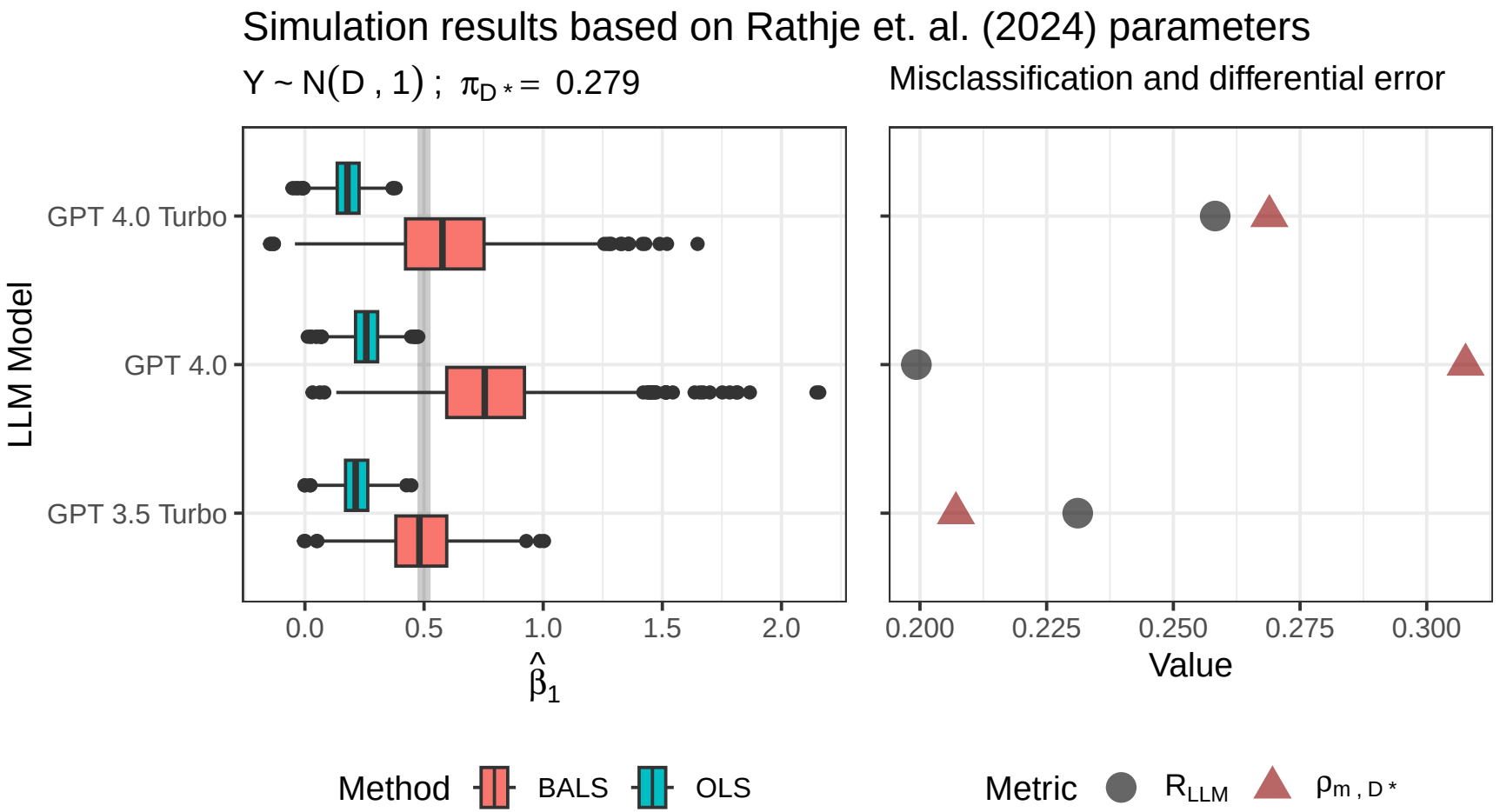
Case Study: Offensiveness (Rathje et al., 2024)



Case Study: Offensiveness (Rathje et al., 2024)



Case Study: Offensiveness (Rathje et al., 2024)



Case Study: Offensiveness (Rathje et al., 2024)

- Annotation of offensiveness in social media posts
- LLM performance similar, but substantial differential misclassification
- Results highlight conditions where BALS correction might fail or inflate estimates

Summary of Results

- Not using SOTA is often acceptable
- Misclassification generally leads to moderate attenuation
- BALS can effectively correct bias
- Differential misclassification can complicate corrections

Summary of Results

- Not using SOTA is often acceptable
- Misclassification generally leads to moderate attenuation
- BALS can effectively correct bias
- Differential misclassification can complicate corrections

Practical Recommendations: **Gold standard sample required!**

1. Check distribution of true treatment (π_{D^*})
2. Examine correlation with controls (ρ_{X,D^*})
3. Characterize misclassification rates and differential error (R_m, ρ_{m,D^*})
4. Apply BALS to correct estimates
5. Use our forthcoming `R` package!

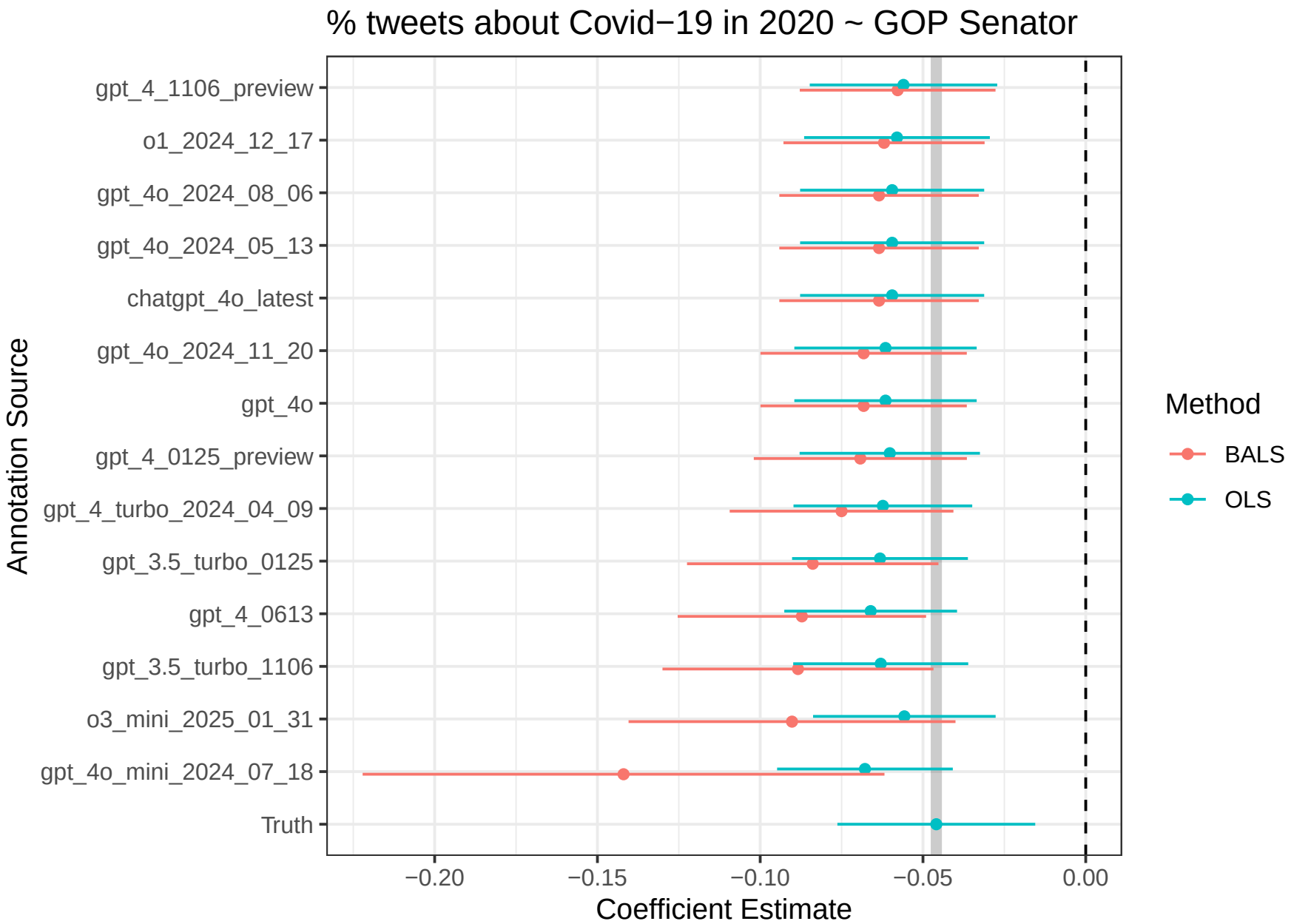
Broader Implications

- Arguments against open-source models on accuracy grounds overstated
- Robustness checks and bias corrections more important than model choice alone

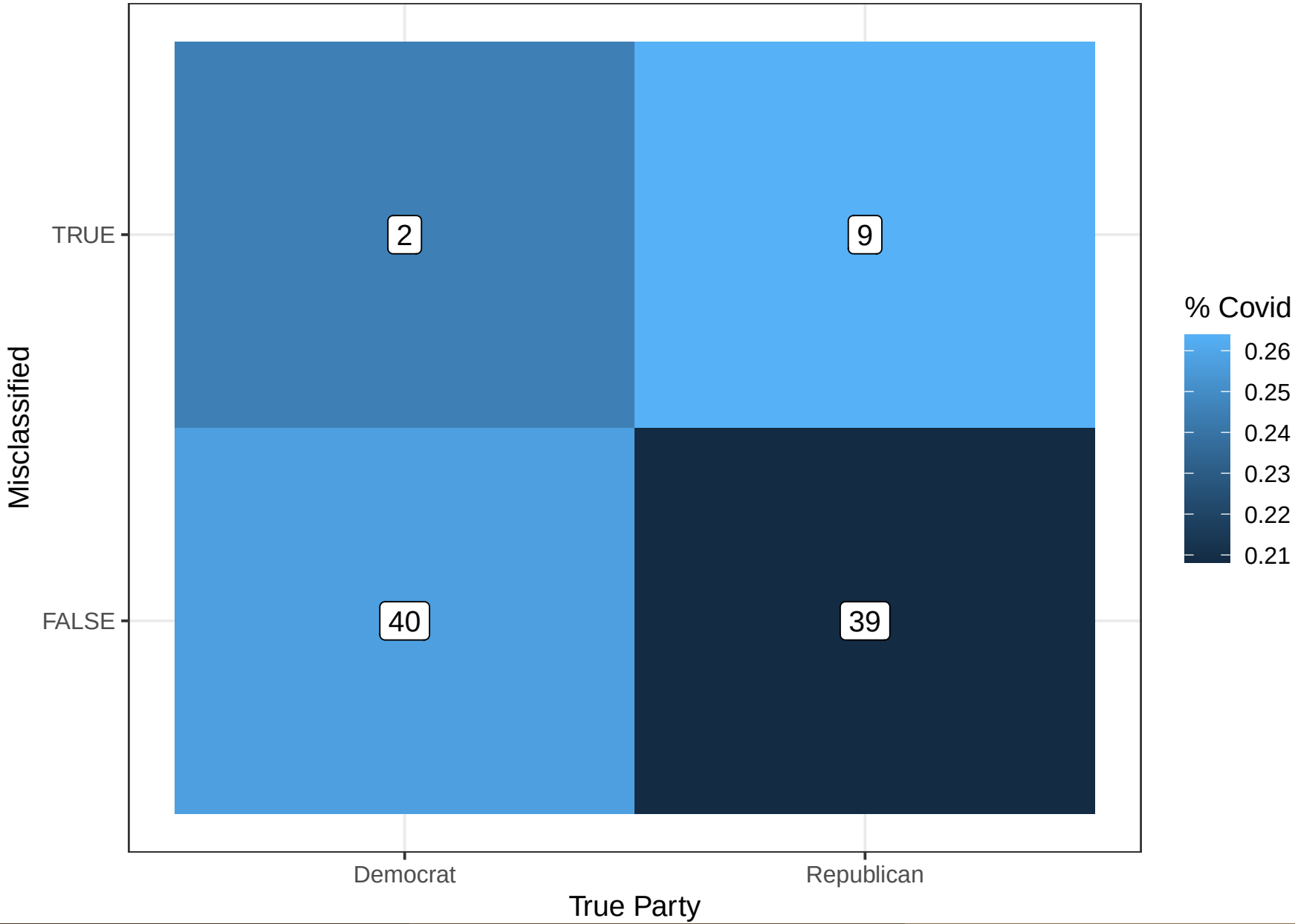
Thank You!

james.h.bisbee@vanderbilt.edu
www.jamesbisbee.com

Differential wrt Outcome



Differential wrt Outcome



Differential wrt Outcome