Domain Constraints Improve Risk Prediction When Outcome Data is Missing



Sidhika Balachandar
Cornell Tech

Nikhil Garg
Cornell Tech

Emma Pierson
Cornell Tech

ML models are often trained to predict the outcome of a human decision

Medicine:

- Doctor decides whether to test a patient for disease
- Model predicts whether the patient will test positive

ML models are often trained to predict the outcome of a human decision

Medicine:

- Doctor decides whether to test a patient for disease
- Model predicts whether the patient will test positive

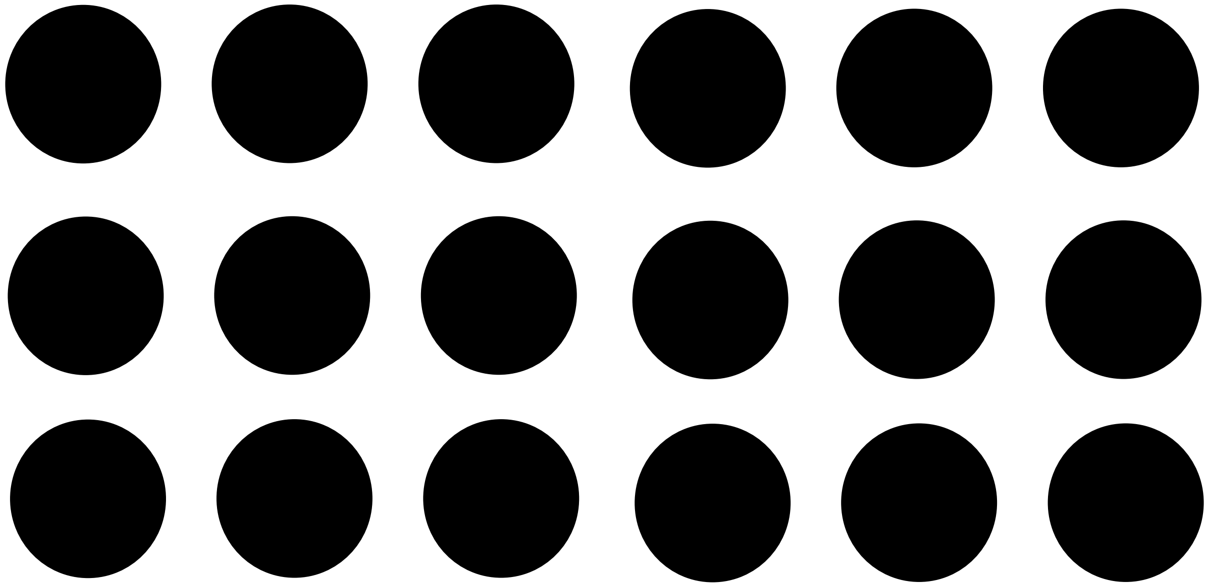
Lending:

- Creditor decides whether to grant an applicant a loan
- Model predicts whether the applicant will repay

Challenge: Selective labels problem

- Human decision censors the data

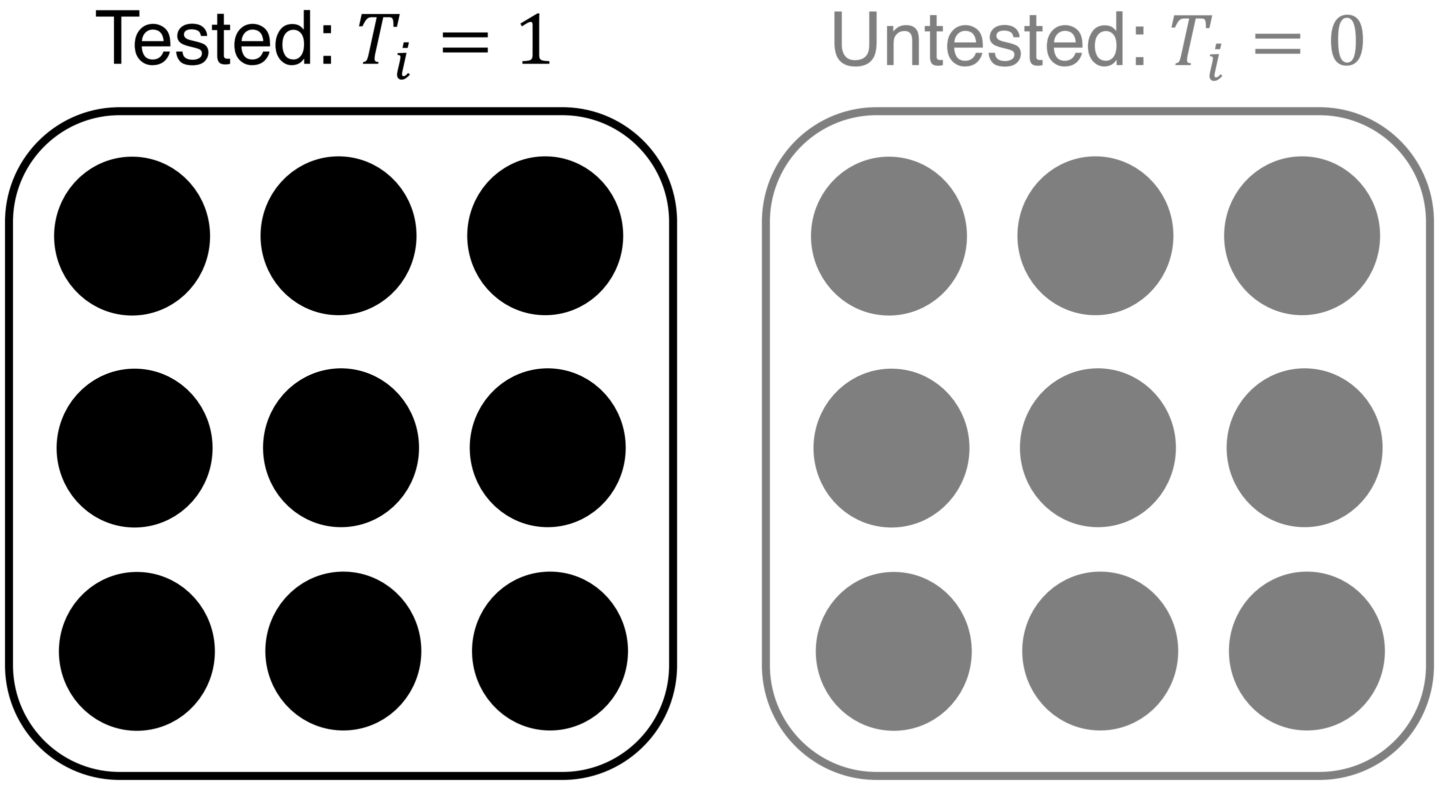
Let us index each
patient with i



Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

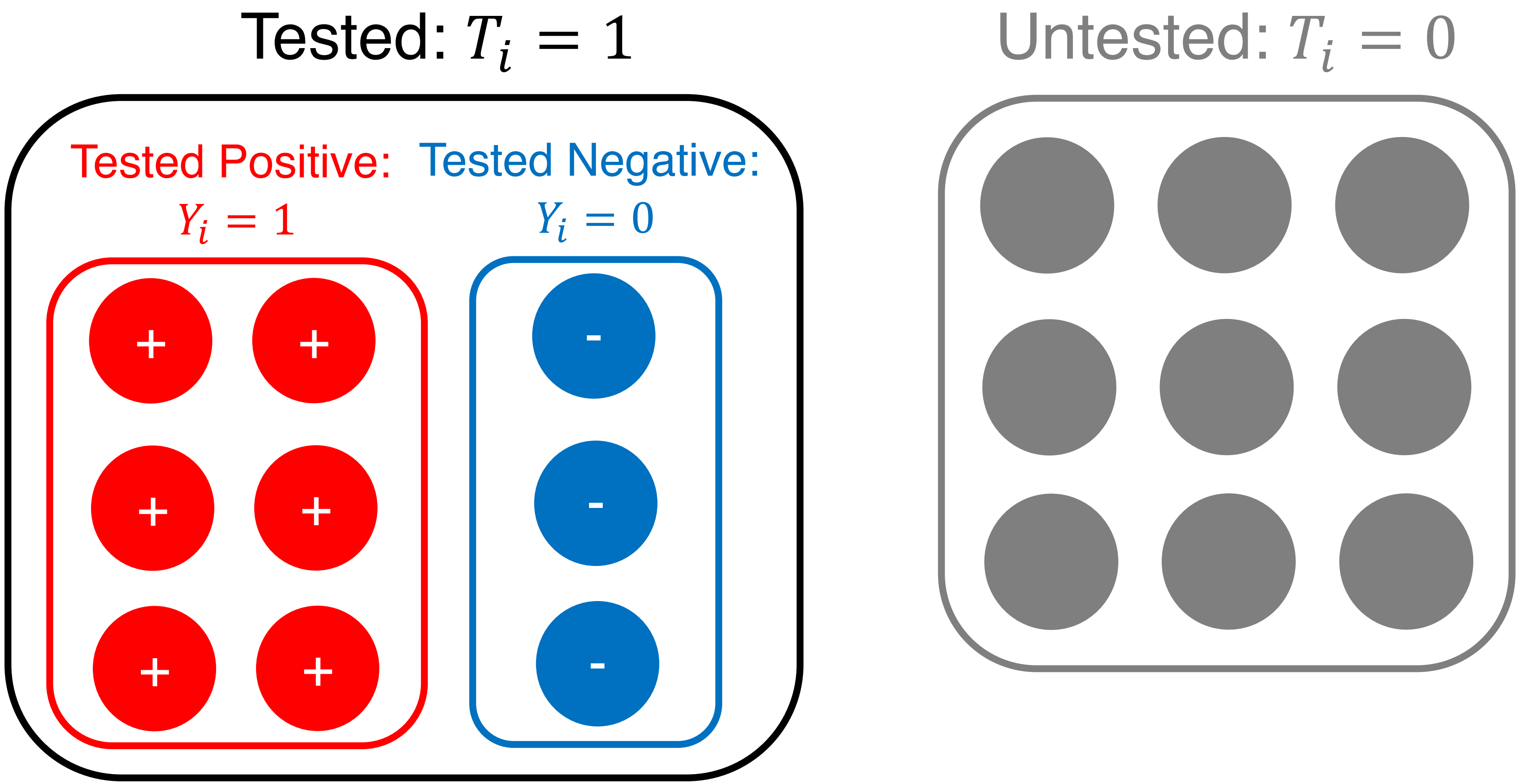
Challenge: Selective labels problem

- Human decision censors the data



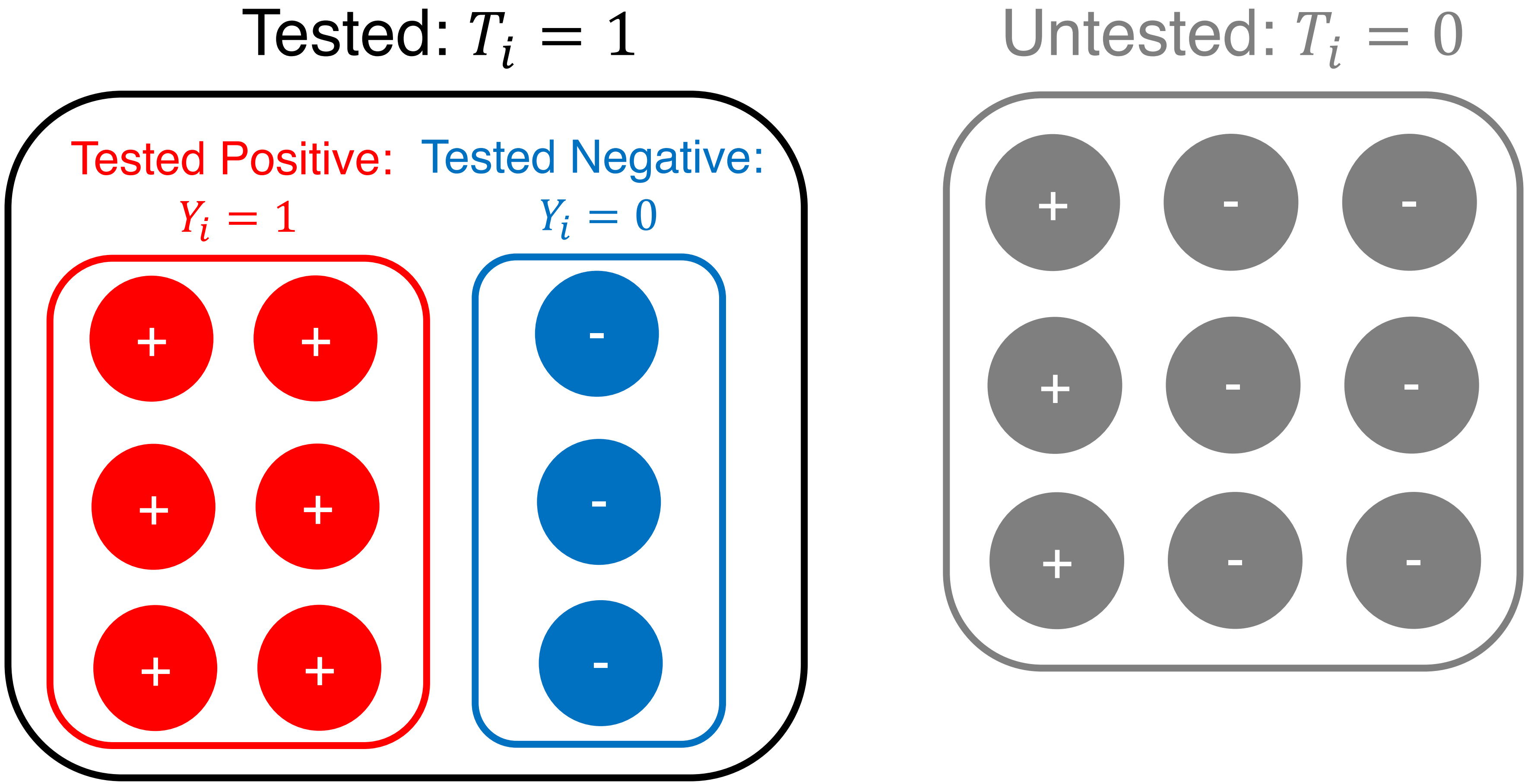
Challenge: Selective labels problem

- Human decision censors the data



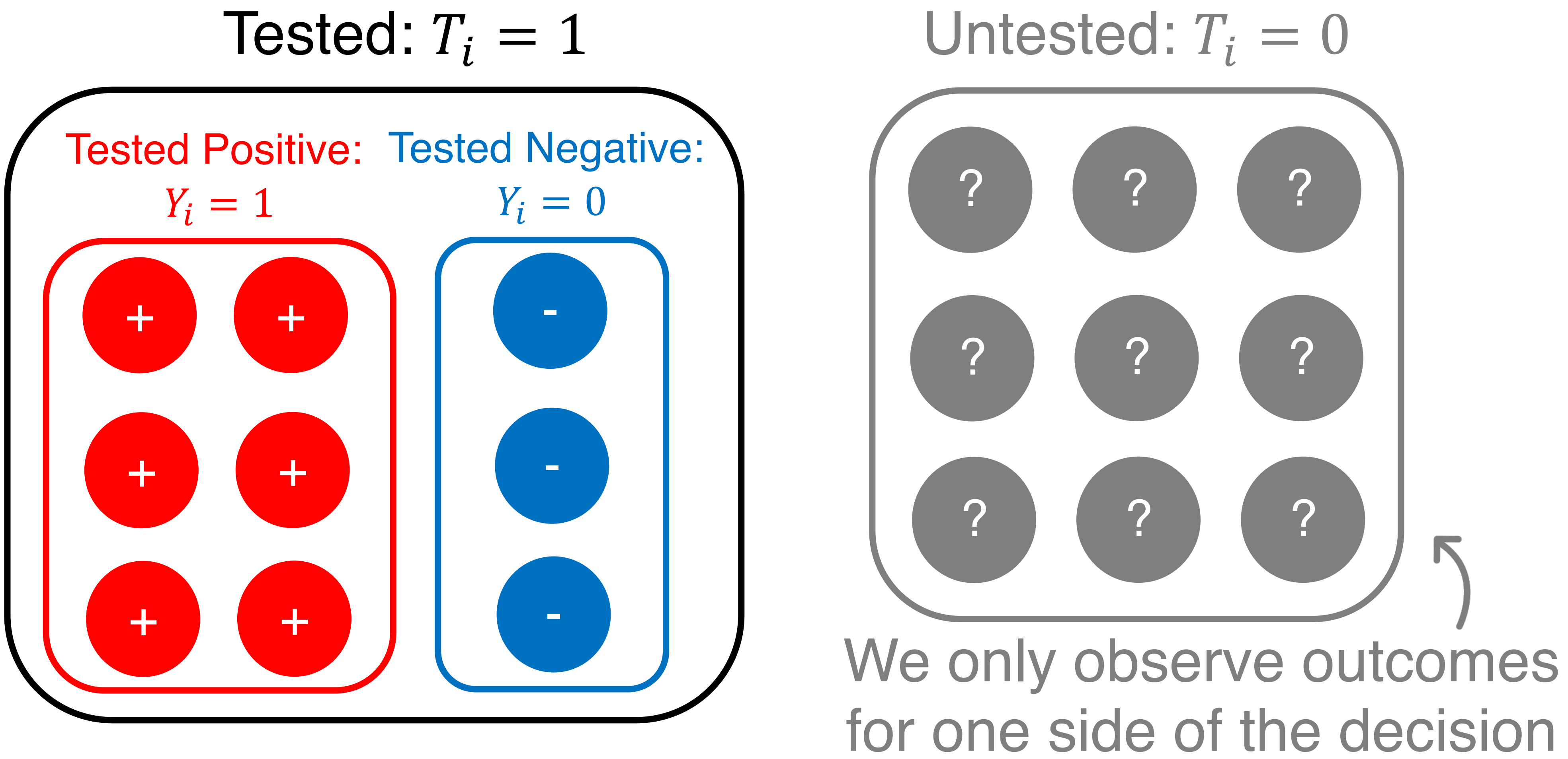
Challenge: Selective labels problem

- Human decision censors the data



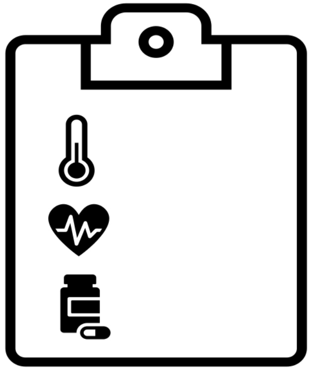
Challenge: Selective labels problem

- Human decision censors the data

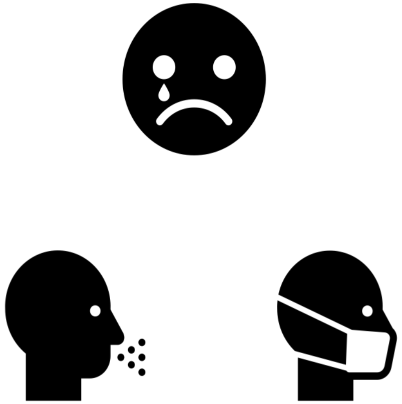


Challenge: Selective labels problem

- The tested and untested populations may differ along both observable and unobservable features



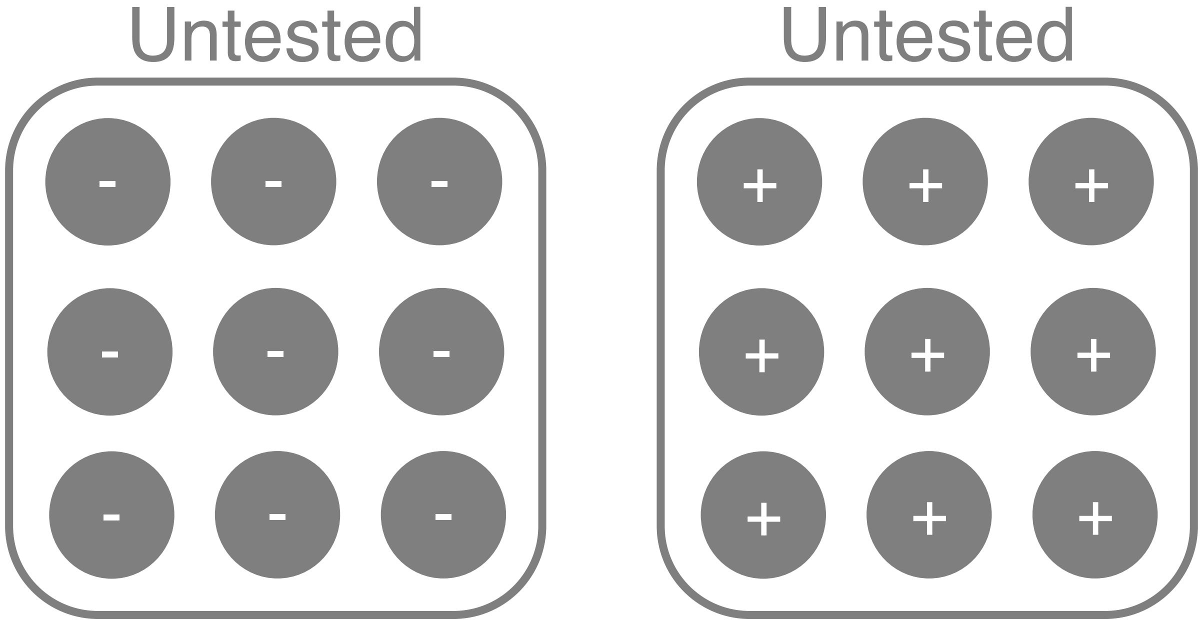
Observable features are recorded in the dataset



Unobservable features are not recorded in the dataset but impact T_i and Y_i

Challenge: Selective labels problem

- **Problem:** Without any further information, anything in between these two extremes is equally possible



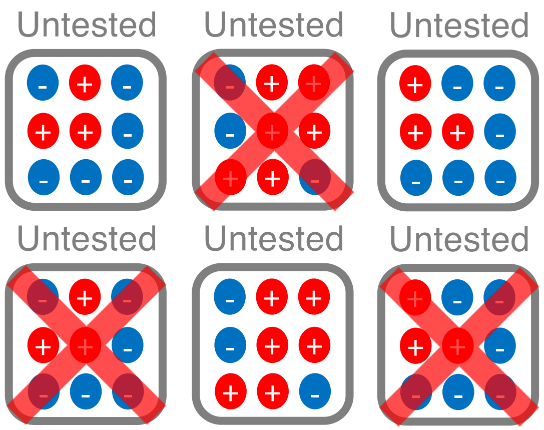
- **Solution:** Use domain constraints to restrict the possibilities

Solution: Domain constraints

Prevalence constraint: Overall fraction of patients with $Y_i = 1$ is known (perhaps approximately)

Prevalence = 33%

Prevalence = 66%



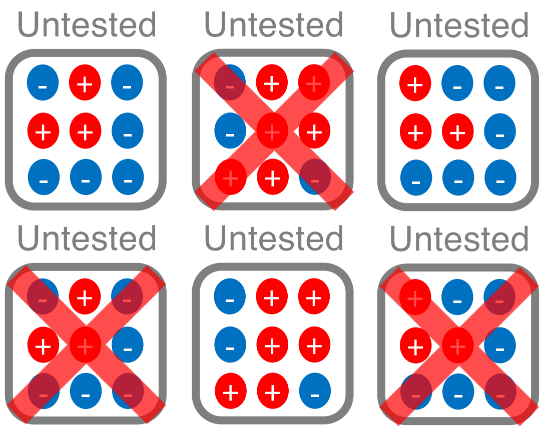
Solution: Domain constraints

Prevalence constraint: Overall fraction of patients with $Y_i = 1$ is known (perhaps approximately)

Expertise constraint: Testing allocation is not purely risk-based only along a constrained feature set

Prevalence = 33%

Prevalence = 66%



Assuming expertise constrains the functions to model $p(T_i = 1)$

Modeling goals

- Model risk $p(Y = 1)$: Accurately model risk of having disease for both tested and untested patients
- Model testing policy $p(T = 1)$: Quantify deviations from purely risk-based test allocation

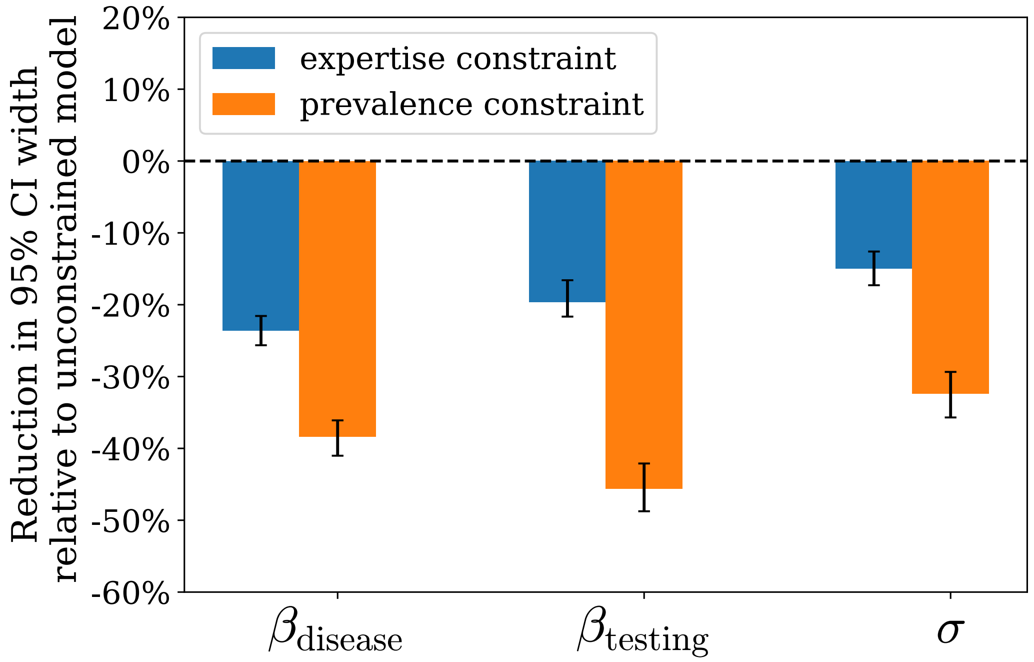
Theoretical results and synthetic experiments

- We show theoretically that the constraints never worsen the precision of parameter inference and provide conditions under which they strictly improve it

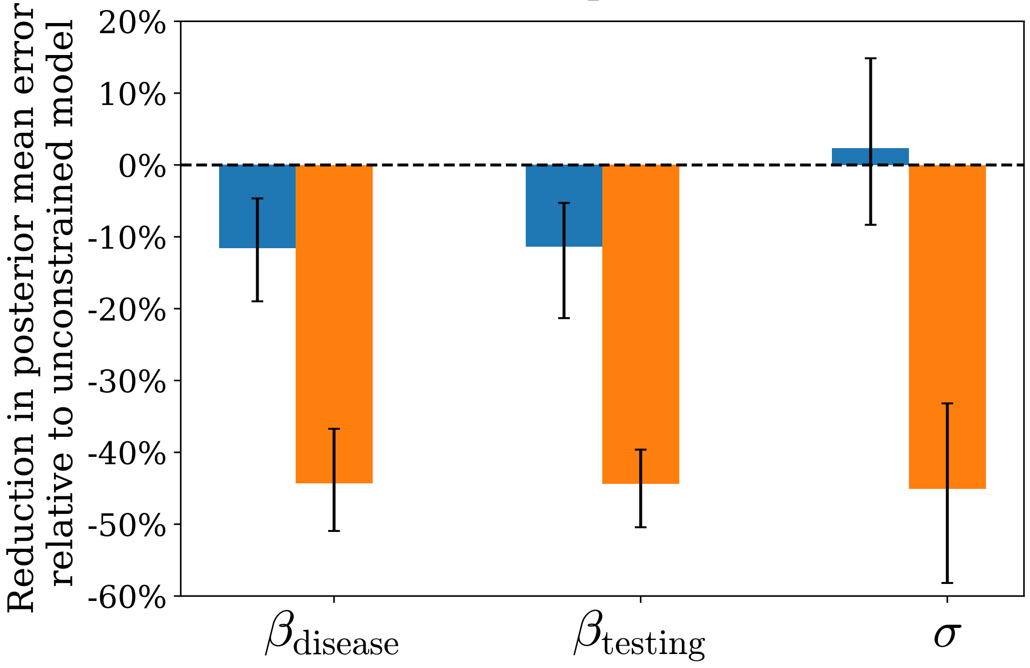
Theoretical results and synthetic experiments

- We show theoretically that the constraints never worsen the precision of parameter inference and provide conditions under which they strictly improve it

Constraints improve precision



Constraints improve accuracy

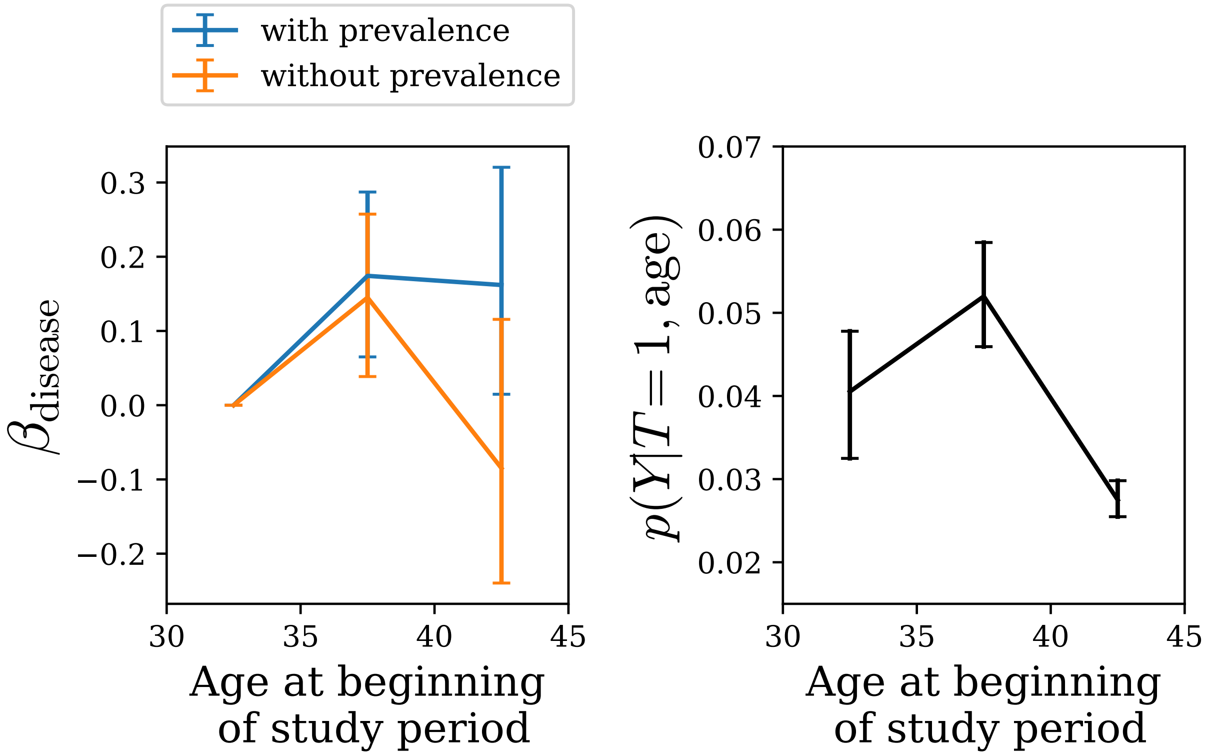


Case study: Breast cancer testing

- X: 7 health, demographic, and genetic features predictive of breast cancer
- T: tested for breast cancer?
- Y: tested positive for breast cancer?

Results

Without constraints the model learns an implausible age trend



Model validations

- Inferred risk predicts cancer diagnoses (Y_i) in both tested and untested populations

Model validations

- Inferred risk predicts cancer diagnoses (Y_i) in both tested and untested populations
- Model can identify suboptimalities in historical testing: genetic information is underused

Model validations

- Inferred risk predicts cancer diagnoses (Y_i) in both tested and untested populations
- Model can identify suboptimalities in historical testing: genetic information is underused
- … and others

Conclusions

- We describe a ***Bayesian model for selective labels settings***

Conclusions

- We describe a ***Bayesian model for selective labels settings***
- We propose the ***prevalence and expertise constraints***
  - We show theoretically and on synthetic data that the constraints improve inference.

Conclusions

- We describe a ***Bayesian model for selective labels settings***
- We propose the ***prevalence and expertise constraints***
    - We show theoretically and on synthetic data that the constraints improve inference.
- We apply our model to estimate ***breast cancer risk***
    - We show that the prevalence constraint increases the plausibility of inferences.

Conclusions

- We describe a ***Bayesian model for selective labels settings***
- We propose the ***prevalence and expertise constraints***
  - We show theoretically and on synthetic data that the constraints improve inference.
- We apply our model to estimate ***breast cancer risk***
  - We show that the prevalence constraint increases the plausibility of inferences.
- Open question: What are natural domain constraints in other selective labels domains?

Thank you!