An assessment of racial disparities in pretrial decision-making using misclassification models

Kimberly A. H. Webb, Sarah A. Riley, and Martin T. Wells

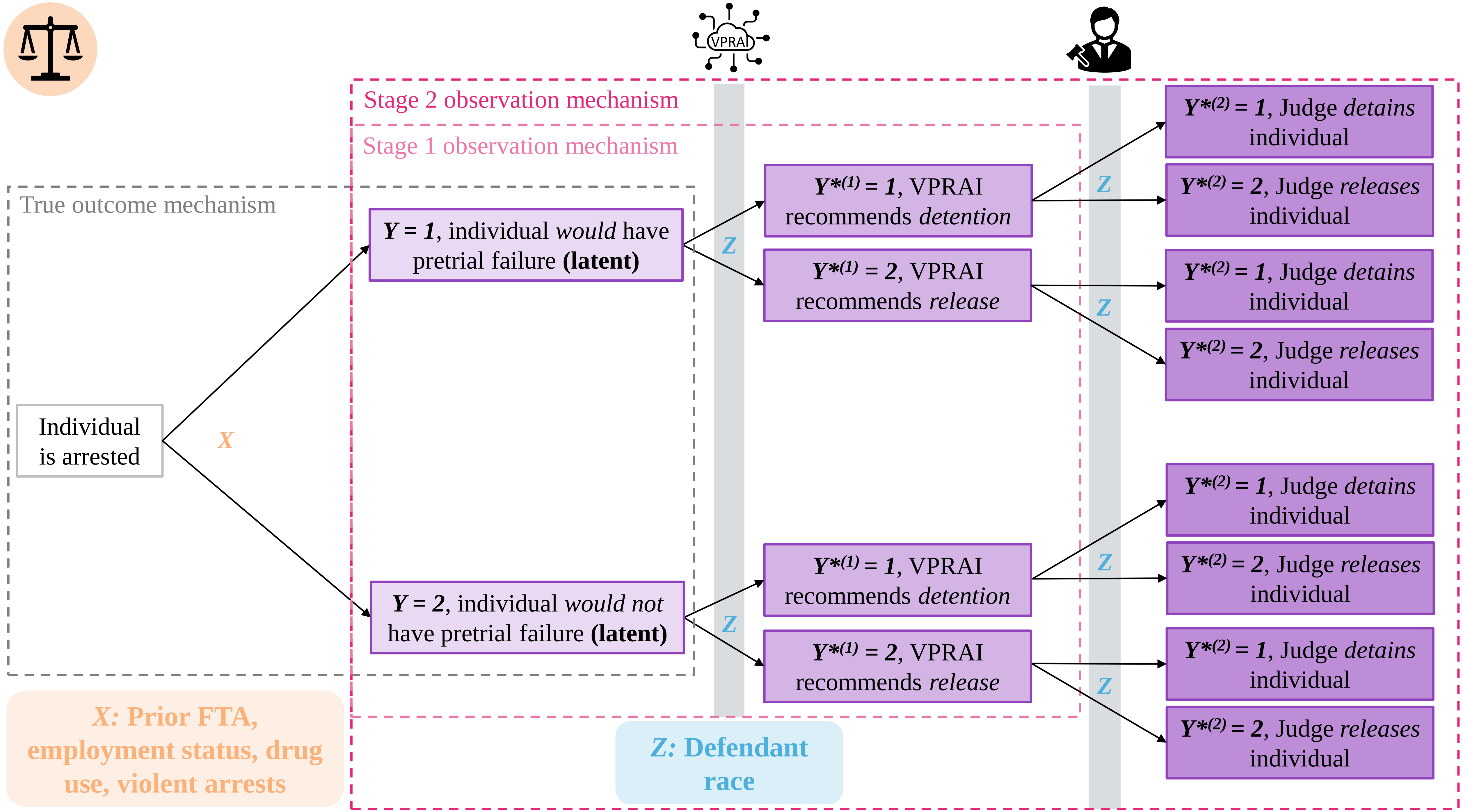
ENAR – March 25, 2025

Problem setting

- **Goal:** Study **algorithmic bias** in **pretrial risk assessment**.
  - **Pretrial risk assessment algorithms** provide an evaluation of the likelihood of **"pretrial failure"**.
  - Used by judges at arraignment to determine whether to release or detain defendants pending trial.
  - What are **risk factors** for pretrial failure? Are judges and risk assessments **accurate**? Are they **biased**?

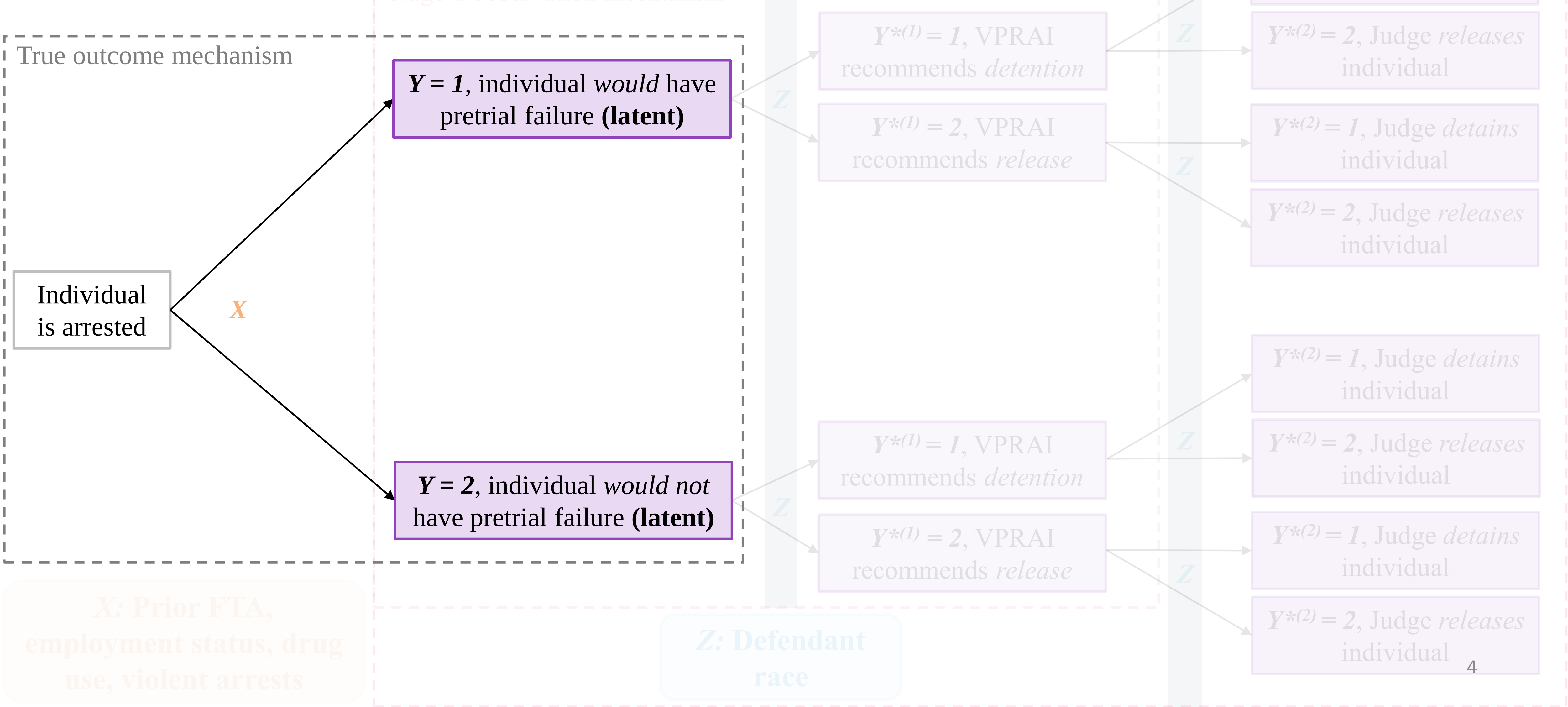
Pretrial failure: Reoffending before trial or failing to appear for trial

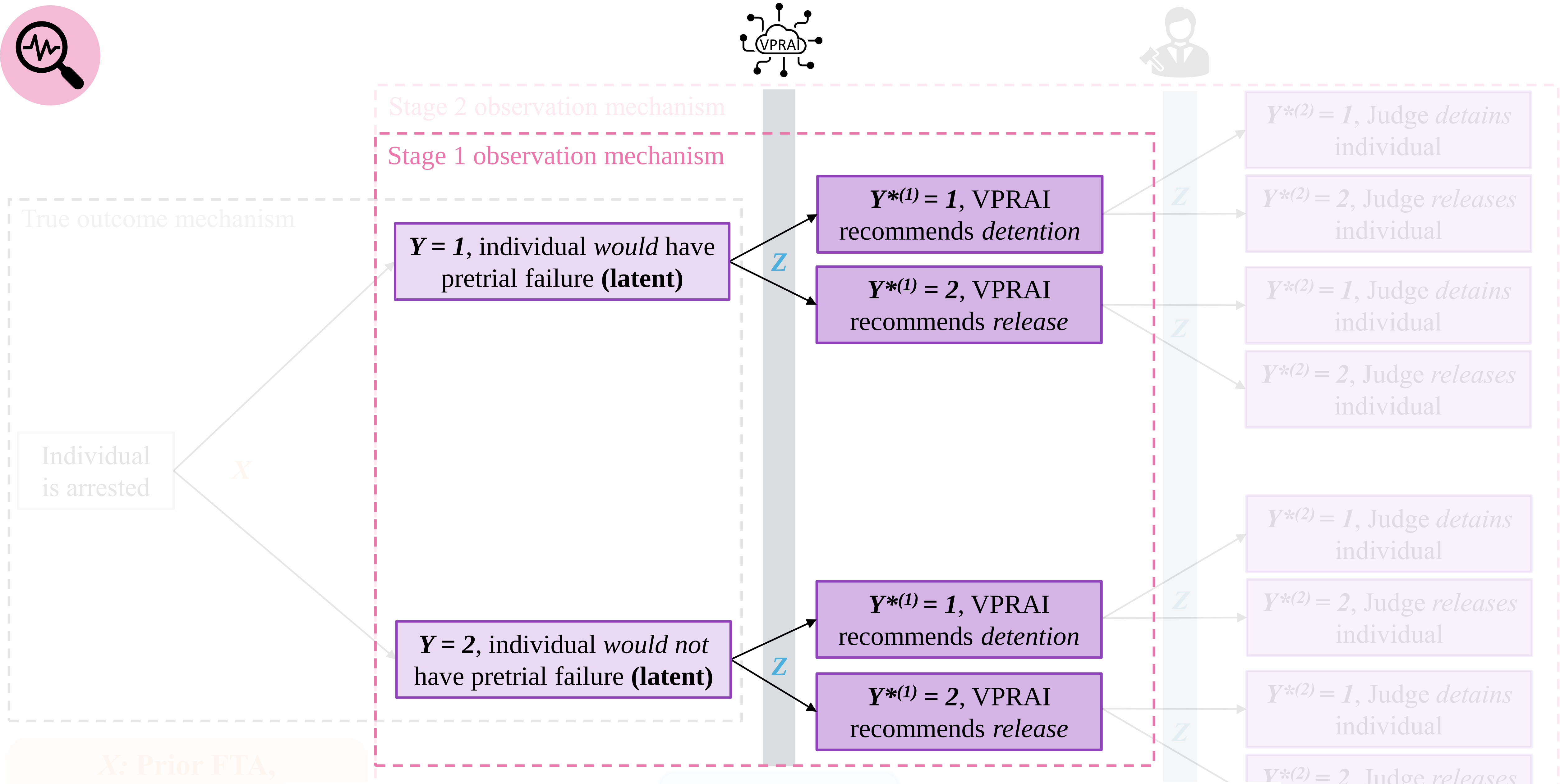
- **Method:** Develop misclassification modeling approach, incorporating the "two stage" nature of this system.



VPRAI
Stage 2 observation mechanism
Stage 1 observation mechanism
True outcome mechanism
Individual is arrested
X
Y = 1, individual would have pretrial failure (latent)
Y = 2, individual would not have pretrial failure (latent)
Z
Y*(1) = 1, VPRAI recommends detention
Y*(1) = 2, VPRAI recommends release
Y*(2) = 1, Judge detains individual
Y*(2) = 2, Judge releases individual
X: Prior FTA, employment status, drug use, violent arrests
Z: Defendant race

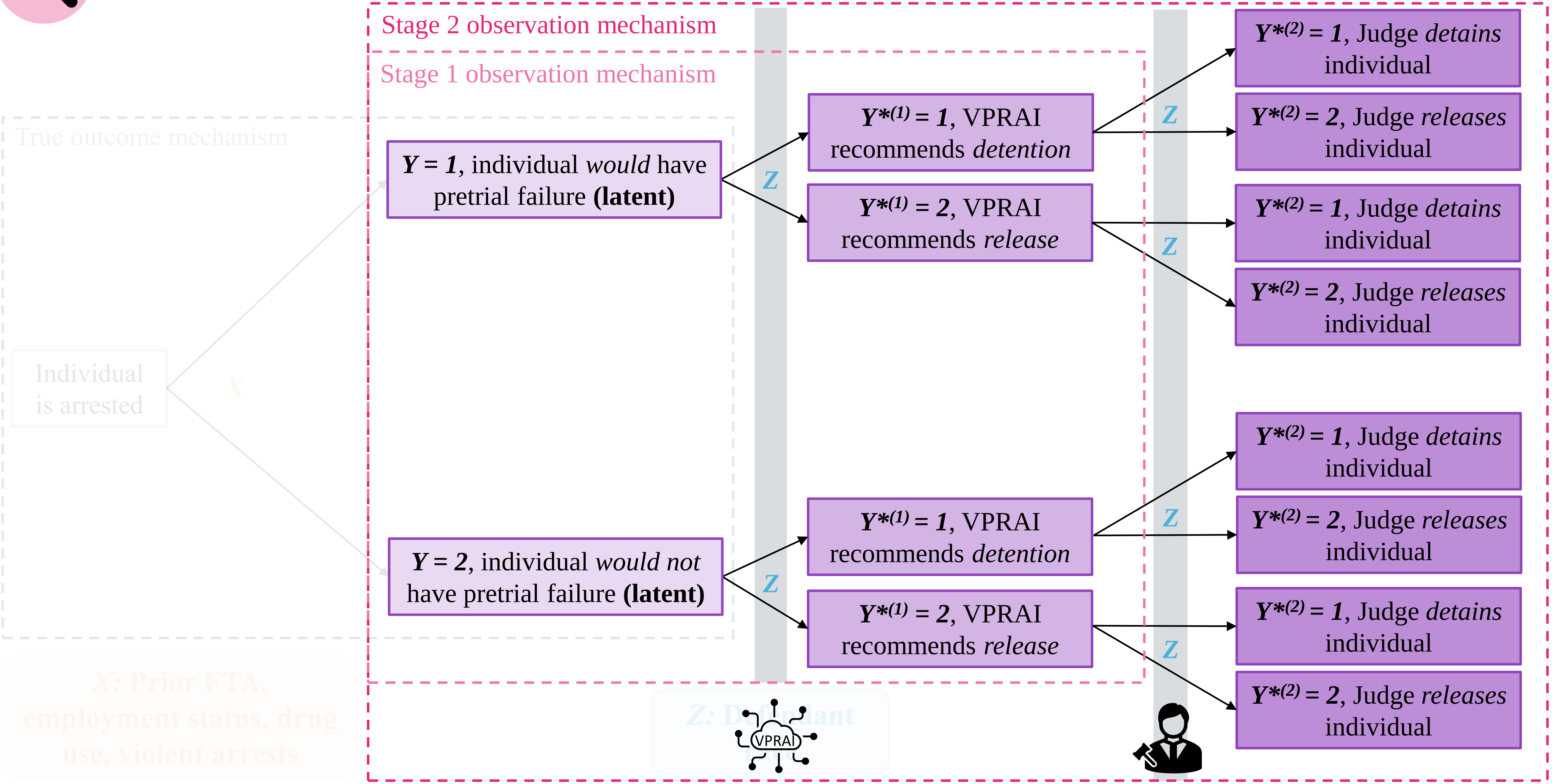
$$\operatorname{logit}\{P(Y=1|\boldsymbol{X};\boldsymbol{\beta})\} = \beta_0 + \beta_{FTA}\text{FTA} + \beta_{unemployed}\text{E} + \beta_{drug}\text{D} + \beta_{violent}\text{V}$$





$$\text{logit}\{P(Y^{*(1)} = 1 | Y = j, Z^{(1)}; \gamma)\} = \gamma_{1j0}^{(1)} + \gamma_{1j,race}^{(1)}\text{RACE}$$

$$\text{logit}\{P(Y^{*(2)} = 1 | Y^{*(1)} = k, Y = j, \mathbf{Z}^{(2)}; \boldsymbol{\gamma})\} = \gamma^{(2)}_{1kj0} + \gamma^{(2)}_{1kj,race}\text{RACE}$$



Primary interest: Estimating β

Secondary interest: Estimating γ

True outcome mechanism:

$$\text{logit}\{P(Y = 1|\boldsymbol{X};\boldsymbol{\beta})\} = \beta_0 + \beta_{FTA}\text{FTA} + \beta_{unemploymed}\text{E} + \beta_{drug}\text{D} + \beta_{violent}\text{V}$$

Stage 1 (VPRAI) observation mechanism:

$$\text{logit}\{P(Y^{*(1)} = 1|Y = j, \boldsymbol{Z}^{(\mathbf{1})};\boldsymbol{\gamma})\} = \gamma_{1j0}^{(1)} + \gamma_{1j,race}^{(1)}\text{RACE}$$

Stage 2 (Judge) observation mechanism:

$$\text{logit}\{P(Y^{*(2)} = 1|Y^{*(1)} = k, Y = j, \boldsymbol{Z}^{(\mathbf{2})};\boldsymbol{\gamma})\} = \gamma_{1kj0}^{(2)} + \gamma_{1kj,race}^{(2)}\text{RACE}$$

Primary interest: Estimating β

Secondary interest: Estimating γ

Estimation methods

- Proposed **EM algorithm**

- Bayesian methods (**MCMC**)

Complete data log-likelihood

- Y (true pretrial failure status) is a latent variable, but let's pretend we know it:

$$\ell_{complete}(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{X}, Z^{(1)}, Z^{(2)}) = \sum_{i=1}^{N} \Bigg[\boxed{\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | \boldsymbol{X}_i)\}} \quad \text{True outcome mechanism}$$

$$\text{Stage 1 (VPRAI) observation mechanism} \quad + \boxed{\sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*(1)} \log\{P(Y_i^{*(1)} = k | Y_i = j, Z^{(1)})\}}$$

$$\text{Stage 2 (Judge) observation mechanism} \quad + \boxed{\sum_{j=1}^{2} \sum_{k=1}^{2} \sum_{\ell=1}^{2} y_{ij} y_{ik}^{*(1)} y_{i\ell}^{*(2)} \log\{P(Y_i^{*(2)} = \ell | P(Y_i^{*(1)} = k, Y_i = j, Z^{(1)})\} \Bigg]}$$

Estimation: EM algorithm

Expectation Step ⟷ **Maximization Step**

$$w_{ij} = P(Y_i = j \mid Y_i^{*(2)}, Y_i^{*(1)}, \boldsymbol{X}, Z^{(1)}, Z^{(2)})$$

$$\begin{aligned} Q = \sum_{i=1}^{N} \Big[& \sum_{j=1}^{2} w_{ij} \log\{P(Y_i = j \mid \boldsymbol{X}_i)\} \\ & + \sum_{j=1}^{2}\sum_{k=1}^{2} w_{ij} y_{ik}^{*(1)} \log\{P(Y_i^{*(1)} = k \mid Y_i = j, Z^{(1)})\} \\ & + \sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=2}^{2} w_{ij} y_{ik}^{*(1)} y_{i\ell}^{*(2)} \log\{P(Y_i^{*(2)} = \ell \mid Y_i^{*(1)} = k, Y_i = j, Z^{(1)}) \Big] \end{aligned}$$

"Fill in" the latent outcome:
Given the parameters and other data, compute the probability of **pretrial failure** for each subject.

Update estimates:
Replace the ***y* terms** in the likelihood with the E-step weights and then **maximize**.

Estimation: EM algorithm

Expectation Step ⟷ **Maximization Step**

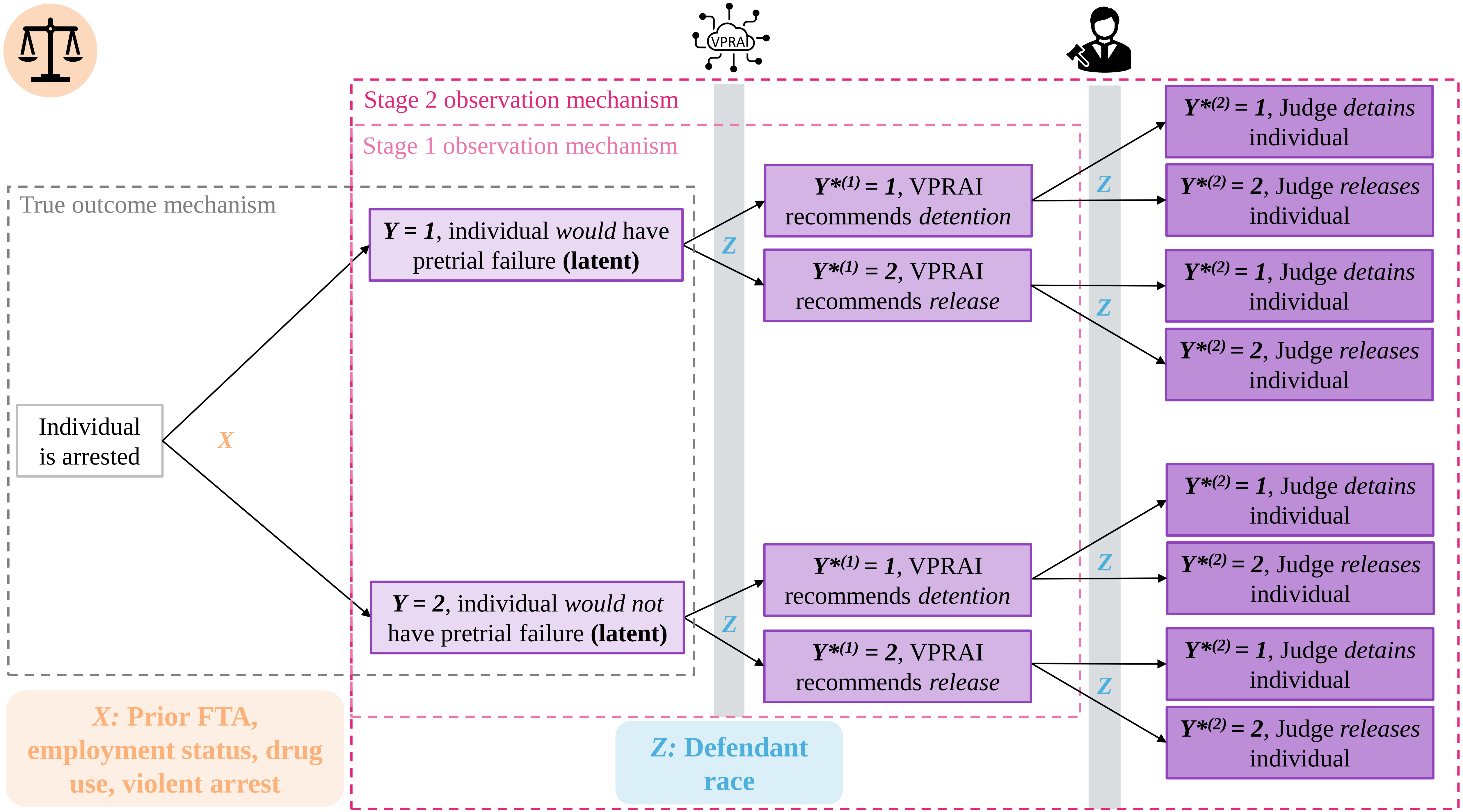
$$w_{ij} = P(Y_i = j | Y_i^{*(2)}, Y_i^{*(1)}, \boldsymbol{X}, Z^{(1)}, Z^{(2)})$$

$$\begin{aligned} Q = \sum_{i=1}^{N} \Big[& \sum_{j=1}^{2} w_{ij} \log\{P(Y_i = j | \boldsymbol{X}_i)\} \\ & + \sum_{j=1}^{2}\sum_{k=1}^{2} w_{ij} y_{ik}^{*(1)} \log\{P(Y_i^{*(1)} = k | Y_i = j, Z^{(1)}) \\ & + \sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=2}^{2} w_{ij} y_{ik}^{*(1)} y_{i\ell}^{*(2)} \log\{P(Y_i^{*(2)} = \ell | Y_i^{*(1)} = k, Y_i = j, Z^{(1)}) \Big] \end{aligned}$$

Apply the label switching correction

Estimates of β

Estimates of γ



VPRAI
Stage 2 observation mechanism
Stage 1 observation mechanism
True outcome mechanism
Individual is arrested
X
Y = 1, individual would have pretrial failure (latent)
Y = 2, individual would not have pretrial failure (latent)
Z
Y*(1) = 1, VPRAI recommends detention
Y*(1) = 2, VPRAI recommends release
Y*(2) = 1, Judge detains individual
Y*(2) = 2, Judge releases individual
X: Prior FTA, employment status, drug use, violent arrest
Z: Defendant race

Goal: Investigate (1) **risk factors for pretrial failure** and (2) the **accuracy** of both the VPRAI recommendations and judge decisions.

- Data from all admitted persons in Prince William County, VA between Jan. 2016 and Dec. 2019

- 1,990 observations in the dataset.

- 13.0% received a “detain” VPRAI recommendation, but 52.2% defendants were detained by the court ahead of their trial.

True outcome mechanism:

$$\text{logit}\{P(Y=1|\boldsymbol{X};\boldsymbol{\beta})\} = \beta_0 + \beta_{FTA}\text{FTA} + \beta_{unemployed}\text{E} + \beta_{drug}\text{D} + \beta_{violent}\text{V}$$

	EM Algorithm		**Naïve Analysis**	
	Est.	SE	Est.	SE
β_{FTA}	1.22	0.22	1.02	0.13
$\beta_{\text{unemployed}}$	0.73	0.06	0.67	0.15
β_{drug}	1.97	0.13	1.74	0.17
β_{violent}	0.28	0.02	0.26	0.03

Association between risk factors and pretrial failure is generally attenuated when misclassification in the VPRAI and judge decisions is *not* accounted for.

Stage 1 (VPRAI) observation mechanism:

$$\text{logit}\{P(Y^{*(1)} = 1 | Y = j, Z^{(\mathbf{1})}; \boldsymbol{\gamma})\} = \gamma_{1j0}^{(1)} + \gamma_{1j,race}^{(1)} \text{RACE}$$

	Estimated VPRAI Specificity P(Release | *Would not* have pretrial failure)	**Estimated VPRAI Sensitivity** P(Detain | *Would* have pretrial failure)
White defendant	100%	49.3%
Black defendant	99.3%	86.0%

Stage 2 (Judge) observation mechanism:

$$\text{logit}\{P(Y^{*(2)} = 1 | Y^{*(1)} = k, Y = j, \boldsymbol{Z}^{(2)}; \boldsymbol{\gamma})\} = \gamma^{(2)}_{1kj0} + \gamma^{(2)}_{1kj,race}\text{RACE}$$

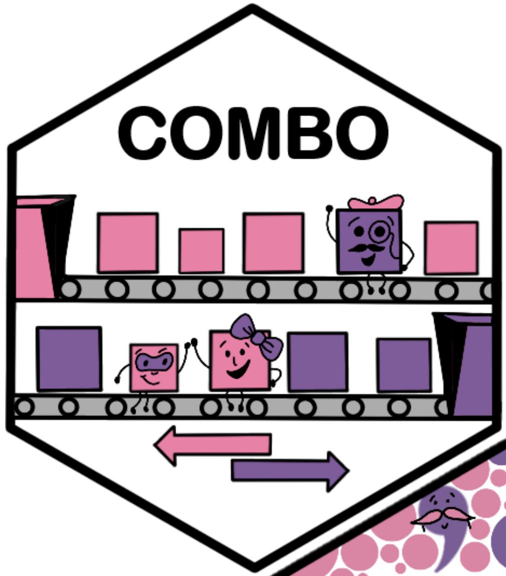
	Estimated Judge Specificity P(Release | *Would not* have pretrial failure)	**Estimated Judge Sensitivity** P(Detain | *Would* have pretrial failure)
White defendant	60.3%	76.8%
Black defendant	48.6%	88.8%

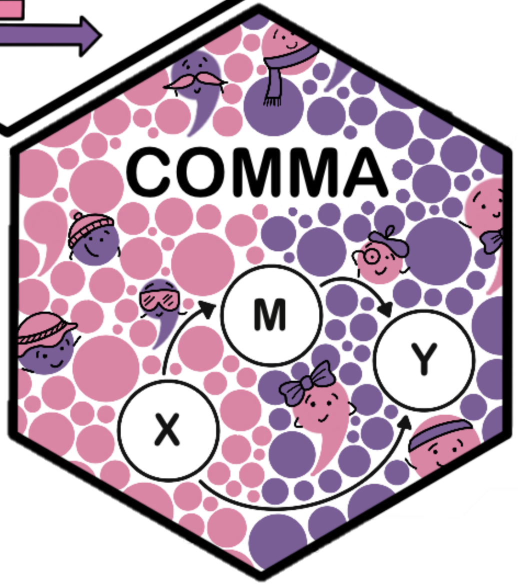
Key takeaways

- Developed new methods for handling misclassified sequential and dependent binary outcome variables.
- Used these methods to estimate misclassification rates when algorithms and judges predict pretrial failure risk.

Software

- Estimation methods for **misclassified outcomes** are available in the *COMBO* R Package on CRAN.
  - **Co**rrecting **M**isclassified **B**inary **O**utcomes
- Estimation methods for **misclassified mediators** are in the *COMMA* R Package on CRAN.
  - **Co**rrecting **M**isclassified **M**ediation **A**nalysis





Thank you!

Kimberly A. H. Webb

kimberlywebb@pitt.edu

kimhwebb.com → My "webb-site" ☺

Cornell Bowers CIS
Statistics and Data Science

University of Pittsburgh

arXiv paper:
arxiv.org/abs/2309.08599