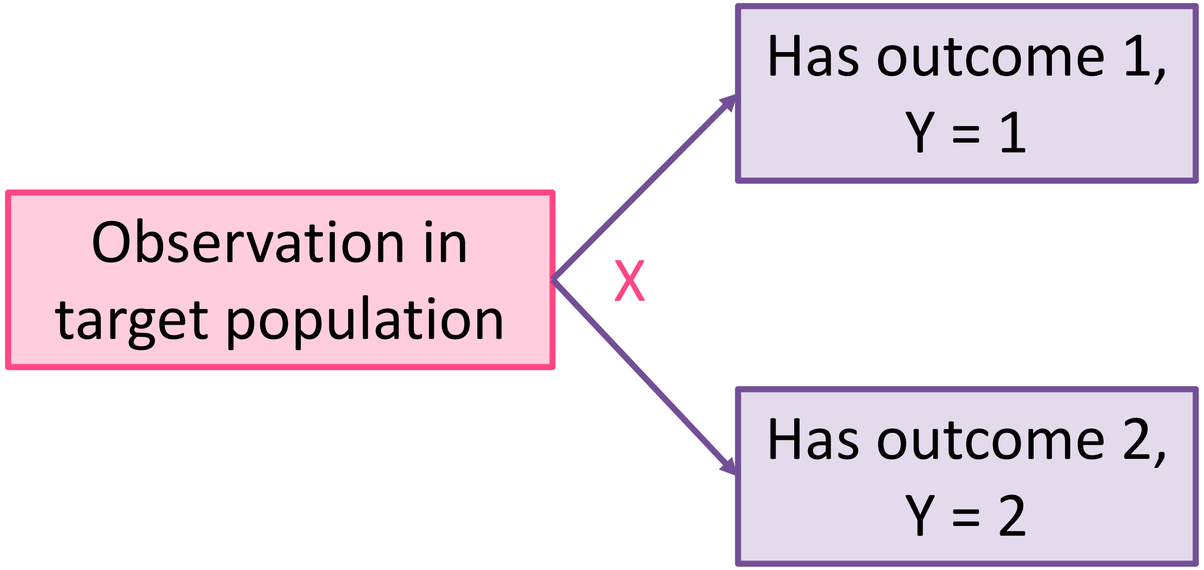
Statistical inference for association studies in the presence of binary outcome misclassification

Kimberly A. Hochstedler and Martin T. Wells

Cornell University

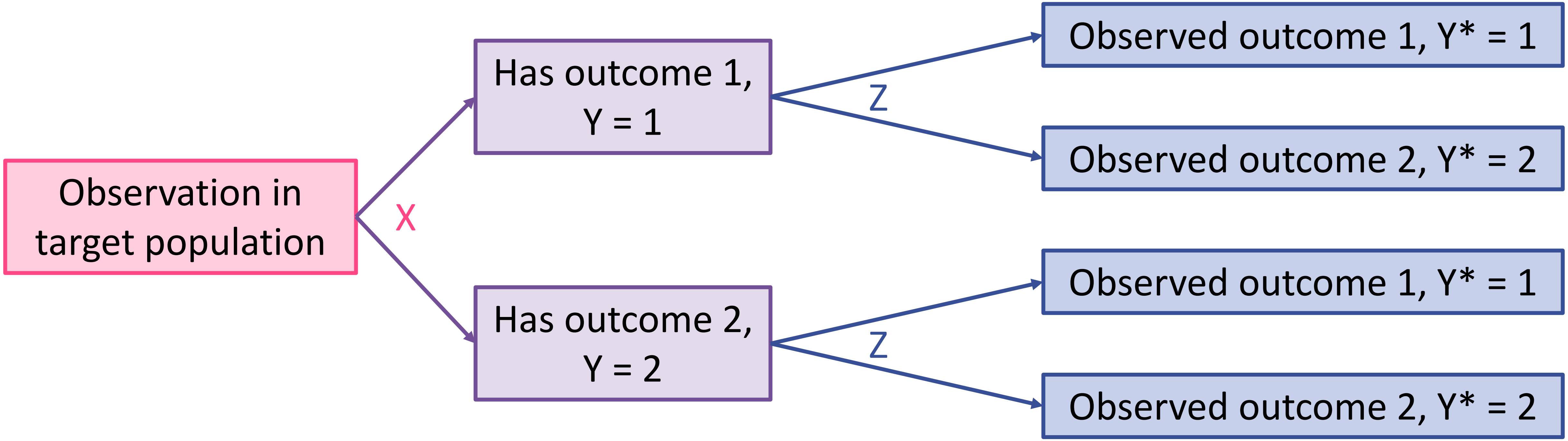
Problem setting

- Interested in the association between X and the **binary variable** Y.



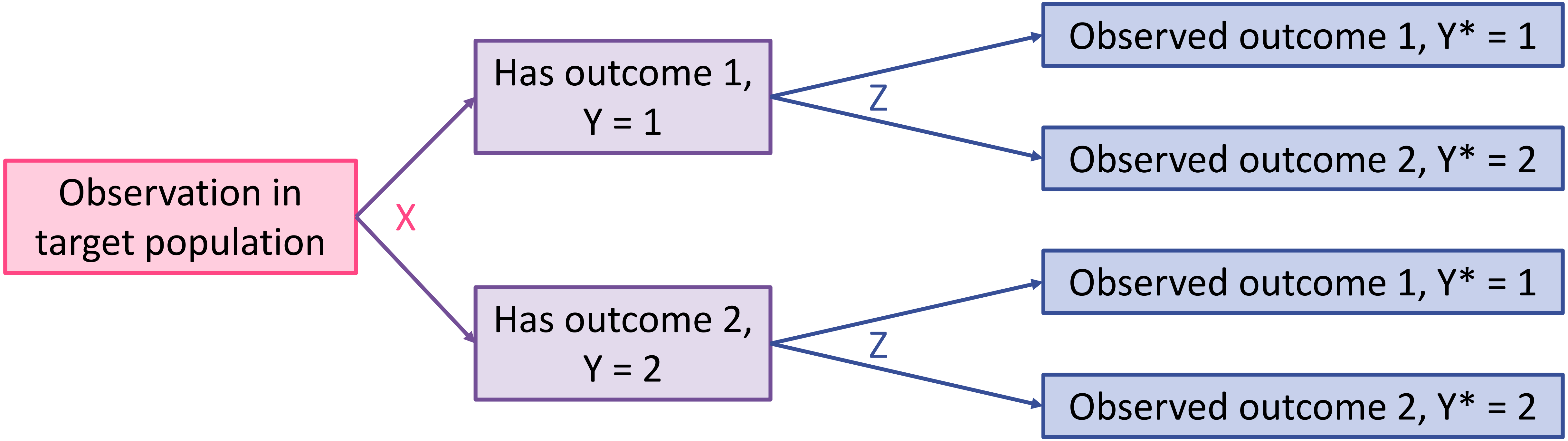
Problem setting

- Interested in the association between X and the **binary variable** Y.
- Measure Y using an instrument that is **not always accurate**, and obtain Y*.



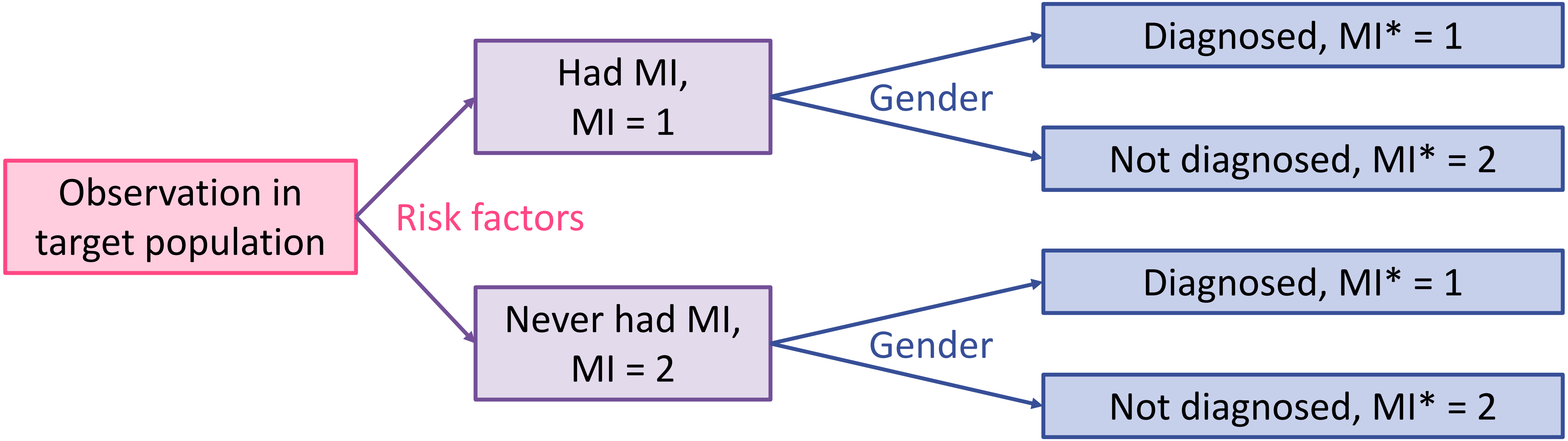
Problem setting

- Interested in the association between X and the **binary variable** Y.
- Measure Y using an instrument that is **not always accurate**, and obtain Y*.
- A third variable, Z, is related to the **misclassification mechanism**.

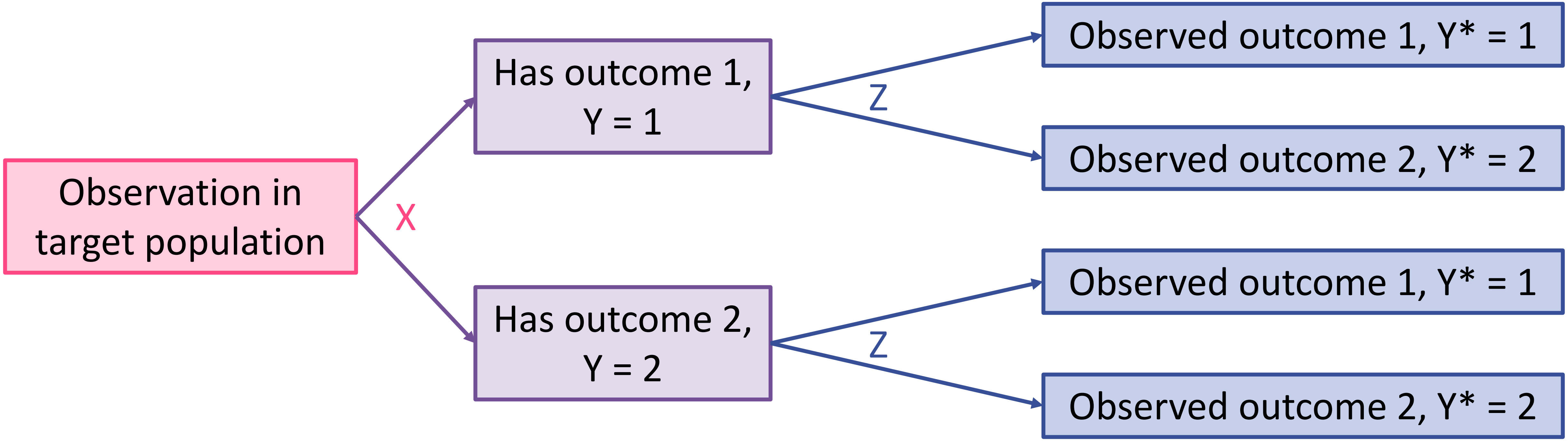


Example

- Interested in the association between Risk Factors and the **binary variable** MI.
- Measure MI using **self-reported medical diagnoses**, and obtain MI*.
- A third variable, gender, is related to the **misclassification mechanism**.



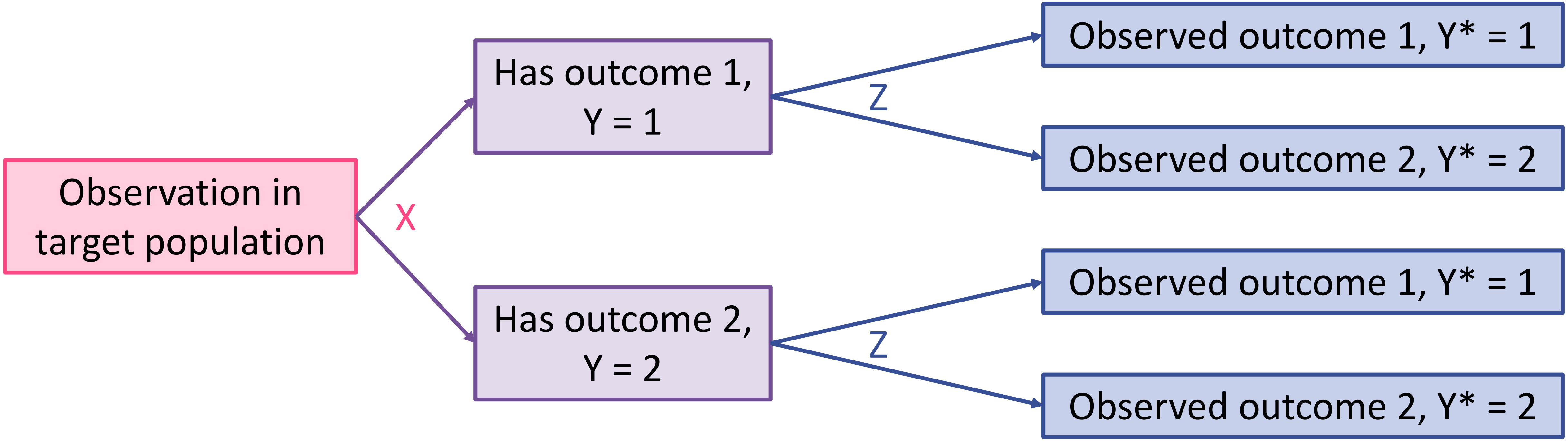
Misclassification model



True outcome mechanism: $\text{logit}\{P(Y = j|X;\beta)\} = \beta_{j0} + \beta_{jX} X$

Observation mechanism: $\text{logit}\{P(Y^* = k|Y = j, Z;\gamma)\} = \gamma_{kj0} + \gamma_{kjZ} Z$

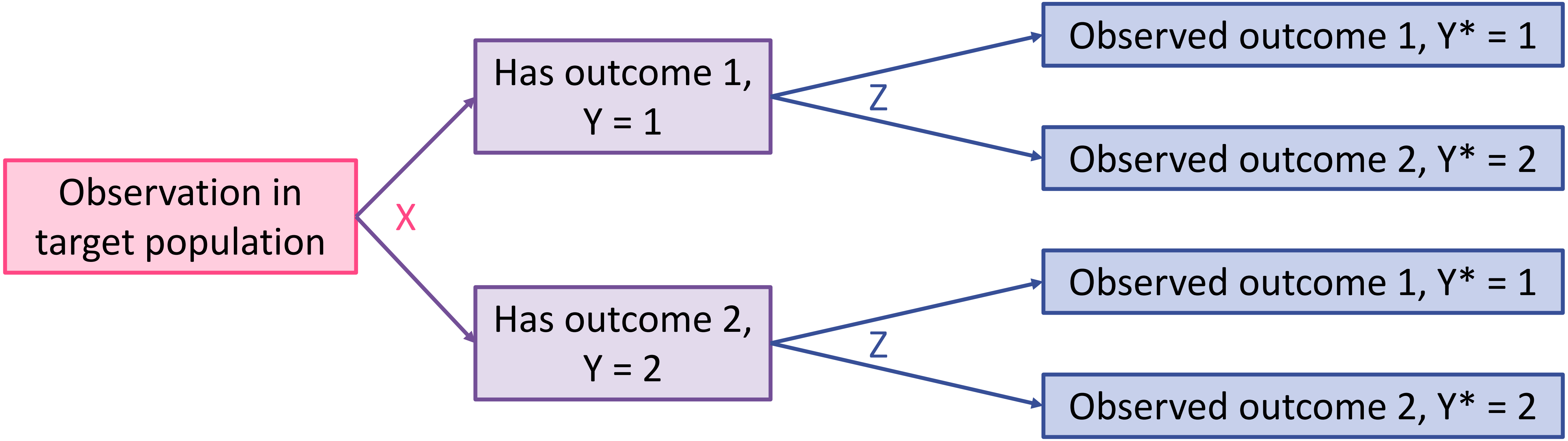
Misclassification model



$$P(Y_i = j | X; \beta) = \pi_{ij} = \frac{\exp\{\beta_{j0} + \beta_{jX} X_i\}}{1 + \exp\{\beta_{j0} + \beta_{jX} X_i\}}$$

$$P(Y_i^* = k | Y_i = j, Z; \gamma) = \pi^*_{ikj} = \frac{\exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}{1 + \exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}$$

Misclassification model

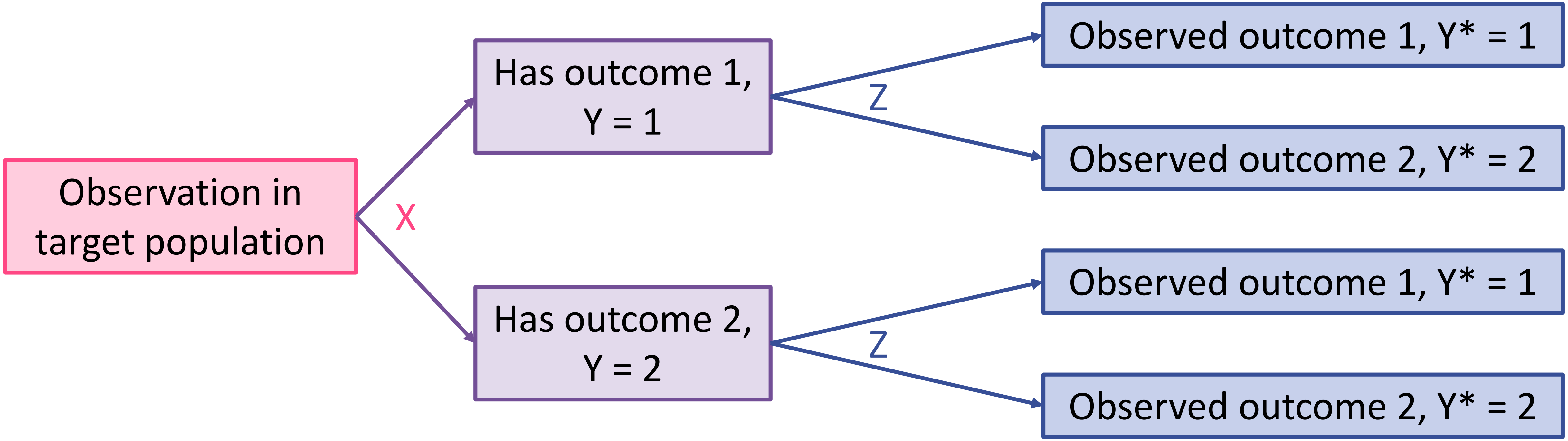


$$P(Y_i = j | X; \beta) = \pi_{ij} = \frac{\exp\{\beta_{j0} + \beta_{jX} X_i\}}{1 + \exp\{\beta_{j0} + \beta_{jX} X_i\}}$$

Primary interest: Estimating β

$$P(Y_i^* = k | Y_i = j, Z; \gamma) = \pi_{ikj}^* = \frac{\exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}{1 + \exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}$$

Misclassification model



$$P(Y_i = j | X; \beta) = \pi_{ij} = \frac{\exp\{\beta_{j0} + \beta_{jX} X_i\}}{1 + \exp\{\beta_{j0} + \beta_{jX} X_i\}}$$

Primary interest: Estimating β

$$P(Y_i^* = k | Y_i = j, Z; \gamma) = \pi^*_{ikj} = \frac{\exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}{1 + \exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}$$

Secondary interest: Estimating γ

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{P(Y_i^{*} = k | Y_i = j, Z_i)\} \right]$$

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j|X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{P(Y_i^{*} = k|Y_i = j, Z_i)\} \right]$$

$$y_{ij} = \mathbb{I}\{Y_i = j\}$$

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{P(Y_i^{*} = k | Y_i = j, Z_i)\} \right]$$

$$y_{ij} = \mathbb{I}\{Y_i = j\}$$

True outcome portion

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{P(Y_i^{*} = k | Y_i = j, Z_i)\} \right]$$

$$y_{ij} = \mathbb{I}\{Y_i = j\}$$

True outcome portion

Observed outcome, given true outcome portion

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^* \log\{P(Y_i^* = k | Y_i = j, Z_i)\} \right]$$

$$y_{ij} = \mathbb{I}\{Y_i = j\}$$

True outcome portion

Observed outcome, given true outcome portion

$$= \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{\pi_{ij}\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^* \log\{\pi_{ikj}^*\} \right]$$

Complete data log-likelihood

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

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{P(Y_i = j | X_i)\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{P(Y_i^{*} = k | Y_i = j, Z_i)\} \right]$$

$$y_{ij} = \mathbb{I}\{Y_i = j\}$$

True outcome portion

Observed outcome, given true outcome portion

$$= \sum_{i=1}^{N} \left[\sum_{j=1}^{2} y_{ij} \log\{\pi_{ij}\} + \sum_{j=1}^{2} \sum_{k=1}^{2} y_{ij} y_{ik}^{*} \log\{\pi_{ikj}^{*}\} \right]$$

Estimation with the EM Algorithm

Expectation Step ⟷ Maximization Step

Estimation with the EM Algorithm

Expectation Step ⟷ **Maximization Step**

$$w_{ij} = P(Y_i = j | Y_i^*, X, Z) = \sum_{k=1}^{2} \frac{y_{ik}^* \pi_{ikj}^* \pi_{ij}}{\sum_{\ell=1}^{2} \pi_{ik\ell}^* \pi_{i\ell}}$$

Estimation with the EM Algorithm

Expectation Step ⟷ **Maximization Step**

$$w_{ij} = P(Y_i = j | Y_i^*, X, Z) = \sum_{k=1}^{2} \frac{y_{ik}^* \pi_{ikj}^* \pi_{ij}}{\sum_{\ell=1}^{2} \pi_{ik\ell}^* \pi_{i\ell}}$$

$$Q = \sum_{i=1}^{N} \Big[\sum_{j=1}^{2} w_{ij} \log\{\pi_{ij}\} + \sum_{j=1}^{2} \sum_{k=1}^{2} w_{ij} y_{ik}^* \log\{\pi_{ikj}^*\} \Big]$$

Estimation with the EM Algorithm

Expectation Step ⟷ Maximization Step

$$w_{ij} = P(Y_i = j | Y_i^*, X, Z) = \sum_{k=1}^{2} \frac{y_{ik}^* \pi_{ikj}^* \pi_{ij}}{\sum_{\ell=1}^{2} \pi_{ik\ell}^* \pi_{i\ell}}$$

$$Q = \sum_{i=1}^{N} \Big[\sum_{j=1}^{2} w_{ij} \log\{\pi_{ij}\} + \sum_{j=1}^{2} \sum_{k=1}^{2} w_{ij} y_{ik}^* \log\{\pi_{ikj}^*\} \Big]$$

$$Q_\beta = \sum_{i=1}^{N} \Big[\sum_{j=1}^{2} w_{ij} \log\{\pi_{ij}\} \Big]$$

$$Q_{\gamma_{k1}} = \sum_{i=1}^{N} \Big[\sum_{k=1}^{2} w_{i1} y_{ik}^* \log\{\pi_{ik1}^*\} \Big]$$

$$Q_{\gamma_{k2}} = \sum_{i=1}^{N} \Big[\sum_{k=1}^{2} w_{i2} y_{ik}^* \log\{\pi_{ik2}^*\} \Big]$$

Label switching

- **Label switching:** When a mixture model likelihood is **invariant under relabeling** of the mixture components, resulting in **multimodal likelihood functions**.

Label switching

- **Label switching:** When a mixture model likelihood is **invariant under relabeling** of the mixture components, resulting in **multimodal likelihood functions**.

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \Big[y_{i1} \log\{\pi_{i1}\} + y_{i2} \log\{\pi_{i2}\}$$

$$+ y_{i1} y_{i1}^{*} \log\{\pi_{i11}^{*}\} + y_{i1} y_{i2}^{*} \log\{\pi_{i21}^{*}\} + y_{i2} y_{i1}^{*} \log\{\pi_{i12}^{*}\} + y_{i2} y_{i2}^{*} \log\{\pi_{i22}^{*}\} \Big]$$

$$\ell_{complete}(\beta, \gamma; X, Z) = \sum_{i=1}^{N} \Big[y_{i2} \log\{\pi_{i2}\} + y_{i1} \log\{\pi_{i1}\}$$

$$+ y_{i2} y_{i1}^{*} \log\{\pi_{i12}^{*}\} + y_{i2} y_{i2}^{*} \log\{\pi_{i22}^{*}\} + y_{i1} y_{i1}^{*} \log\{\pi_{i11}^{*}\} + y_{i1} y_{i2}^{*} \log\{\pi_{i21}^{*}\} \Big]$$

Label switching

- Suppose we have a single predictor X and a single predictor Z:

$$\sum_{i=1}^{N}\Big[y_{i1}\beta_0 + y_{i1}x_i\beta_X - (y_{i1}+y_{i2})\log\{1+\exp\{\beta_0 + x_i\beta_X\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i1}\log\{1+\exp\{\gamma_{110}+z_i\gamma_{11Z}\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i2}\log\{1+\exp\{\gamma_{120}+z_i\gamma_{12Z}\}\}\Big]$$

$$= \sum_{i=1}^{N}\Big[y_{i2}(-\beta_0) + y_{i2}x_i(-\beta_X) - (y_{i1}+y_{i2})\log\{1+\exp\{-\beta_0 + x_i(-\beta_X)\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i2}\log\{1+\exp\{\gamma_{120}+z_i\gamma_{12Z}\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i1}\log\{1+\exp\{\gamma_{110}+z_i\gamma_{11Z}\}\}\Big]$$

Label switching

- Suppose we have a single predictor X and a single predictor Z:

$$\sum_{i=1}^{N}\Big[y_{i1}\beta_0 + y_{i1}x_i\beta_X - (y_{i1}+y_{i2})\log\{1+\exp\{\beta_0 + x_i\beta_X\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i1}\log\{1+\exp\{\gamma_{110}+z_i\gamma_{11Z}\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i2}\log\{1+\exp\{\gamma_{120}+z_i\gamma_{12Z}\}\}\Big]$$

$$= \sum_{i=1}^{N}\Big[y_{i2}(-\beta_0) + y_{i2}x_i(-\beta_X) - (y_{i1}+y_{i2})\log\{1+\exp\{-\beta_0 + x_i(-\beta_X)\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i2}\log\{1+\exp\{\gamma_{120}+z_i\gamma_{12Z}\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*}+y_{i2}^{*})y_{i1}\log\{1+\exp\{\gamma_{110}+z_i\gamma_{11Z}\}\}\Big]$$

Label switching

- Suppose we have a single predictor X and a single predictor Z:

$$\sum_{i=1}^{N} \Big[y_{i1}\beta_0 + y_{i1}x_i\beta_X - (y_{i1} + y_{i2}) \log\{1 + \exp\{\beta_0 + x_i\beta_X\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*} + y_{i2}^{*})y_{i1} \log\{1 + \exp\{\gamma_{110} + z_i\gamma_{11Z}\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*} + y_{i2}^{*})y_{i2} \log\{1 + \exp\{\gamma_{120} + z_i\gamma_{12Z}\}\} \Big]$$

$$= \sum_{i=1}^{N} \Big[y_{i2}(-\beta_0) + y_{i2}x_i(-\beta_X) - (y_{i1} + y_{i2}) \log\{1 + \exp\{-\beta_0 + x_i(-\beta_X)\}\}$$

$$+ y_{i2}y_{i1}^{*}\gamma_{120} + y_{i2}y_{i1}^{*}z_i\gamma_{12Z} - (y_{i1}^{*} + y_{i2}^{*})y_{i2} \log\{1 + \exp\{\gamma_{120} + z_i\gamma_{12Z}\}\}$$

$$+ y_{i1}y_{i1}^{*}\gamma_{110} + y_{i1}y_{i1}^{*}z_i\gamma_{11Z} - (y_{i1}^{*} + y_{i2}^{*})y_{i1} \log\{1 + \exp\{\gamma_{110} + z_i\gamma_{11Z}\}\} \Big]$$

Label switching

- There are two sets of parameters that yield the **exact same likelihood value**.

$$\beta_0, \beta_X, \gamma_{110}, \gamma_{11Z}, \gamma_{120}, \gamma_{12Z}$$

Label switching

- There are two sets of parameters that yield the **exact same likelihood value**.

$$\beta_0, \beta_X, \gamma_{110}, \gamma_{11Z}, \gamma_{120}, \gamma_{12Z}$$

$$\Updownarrow$$

$$-\beta_0, -\beta_X, \gamma_{120}, \gamma_{12Z}, \gamma_{110}, \gamma_{11Z}$$

Correcting label switching

- There is a quantity that has different values when each parameter set is used to compute it:

$$P(Y_i^* = k | Y_i = j, Z; \gamma) = \pi^*_{ikj} = \frac{\exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}{1 + \exp\{\gamma_{kj0} + \gamma_{kjZ} Z_i\}}$$

Correcting label switching

Algorithm 1 Correcting label switching in binary outcome misclassification models

Compute average π^*_{jj} for all $j \in \{1, 2\}$ using $\hat{\beta}$ and $\hat{\gamma}$.
if $\pi^*_{jj} > 0.50$ for all $j \in \{1, 2\}$ **then**
  $\hat{\beta}_{corrected} \leftarrow \hat{\beta}$
  $\hat{\gamma}_{corrected} \leftarrow \hat{\gamma}$
else
  $\hat{\beta}_{corrected} \leftarrow -\hat{\beta}$
  $\hat{\gamma}_{corrected,k1} \leftarrow \hat{\gamma_{k2}}$
  $\hat{\gamma}_{corrected,k2} \leftarrow \hat{\gamma_{k1}}$
end if

Correcting label switching

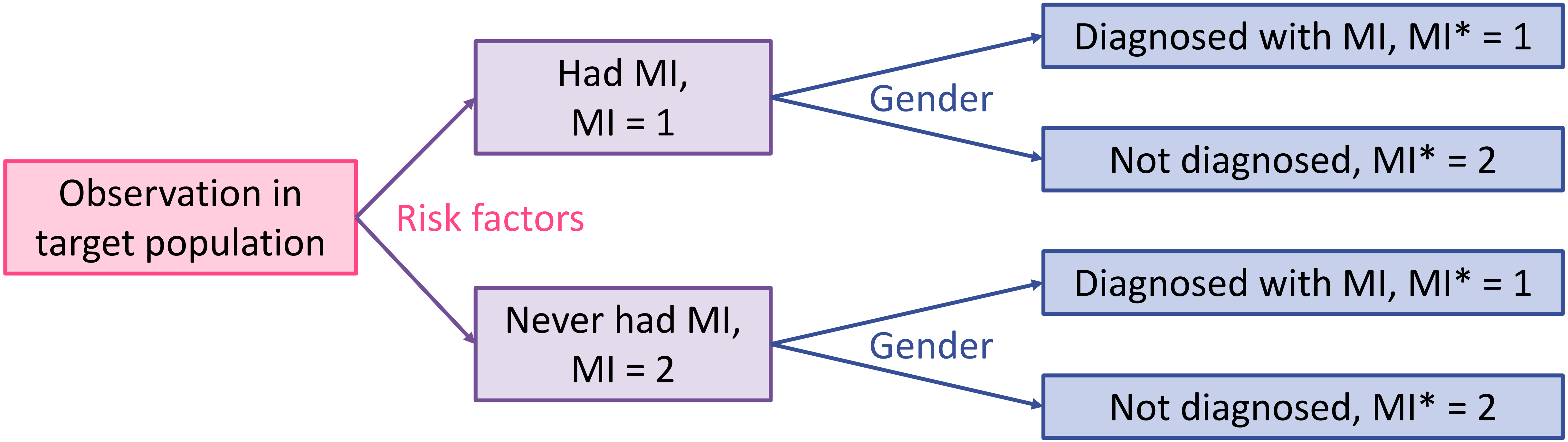
Algorithm 1 Correcting label switching in binary outcome misclassification models

Compute average π^*_{jj} for all $j \in \{1, 2\}$ using $\hat{\beta}$ and $\hat{\gamma}$.
if $\pi^*_{jj} > 0.50$ for all $j \in \{1, 2\}$ **then**
$\quad \hat{\beta}_{corrected} \leftarrow \hat{\beta}$
$\quad \hat{\gamma}_{corrected} \leftarrow \hat{\gamma}$
else
$\quad \hat{\beta}_{corrected} \leftarrow -\hat{\beta}$
$\quad \hat{\gamma}_{corrected,k1} \leftarrow \hat{\gamma_{k2}}$
$\quad \hat{\gamma}_{corrected,k2} \leftarrow \hat{\gamma_{k1}}$
end if

Assumption:
Outcome categories are correctly classified at least 50% of the time.

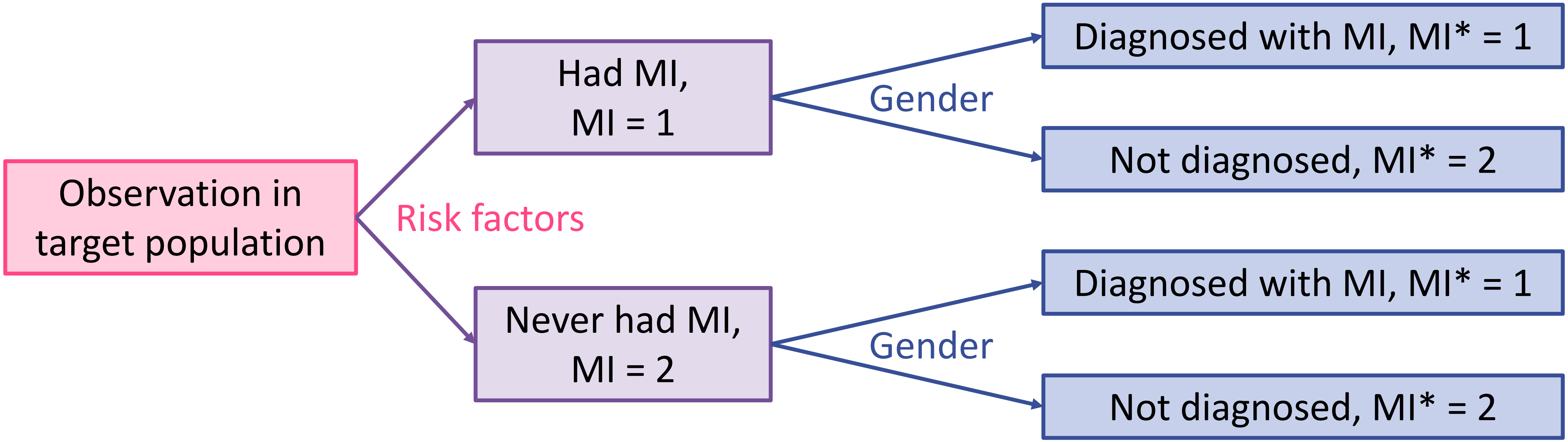
- Apply to EM estimates.

Applied Example



- **Goal:** Understand the risk factors for MI.
  - MI is suspected to be misdiagnosed differentially based on patient age and gender.
  - Data from 2020 MEPS survey.

Applied Example



- **Model for true MI:** MI ~ Smoking Status + Exercise Habits + Age
- **Model observed MI given true MI:** MI* | MI ~ Age + Gender

Applied Example

- **Model for true MI:** MI ~ Smoking Status + Exercise Habits + Age
- **Model observed MI given true MI:** MI* | MI ~ Age + Gender

	EM		Naive Analysis	
	Est.	SE	Est.	SE
β_0	-4.374	0.065	-3.576	0.078
β_{smoke}	1.544	0.107	0.635	0.109
$\beta_{exercise}$	0.303	0.126	0.184	0.084
β_{age}	0.094	0.010	0.059	0.003
γ_{110}	2.969	0.100	-	-
$\gamma_{11,gender}$	-1.766	0.036	-	-
$\gamma_{11,age}$	-0.198	0.005	-	-
γ_{120}	-3.580	0.112	-	-
$\gamma_{12,gender}$	-0.818	0.108	-	-
$\gamma_{12,age}$	0.084	0.005	-	-

Effects are attenuated when we do not account for misclassification of MI

Applied Example

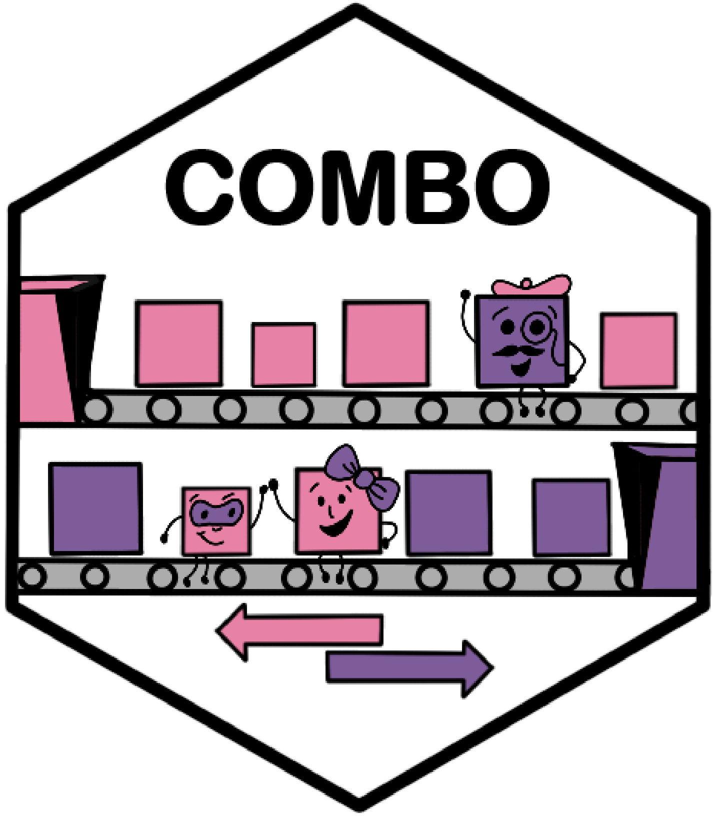
- **Model for true MI:** MI ~ Smoking Status + Exercise Habits + Age
- **Model observed MI given true MI:** MI* | MI ~ Age + Gender

	Estimated Specificity P(no MI* | no MI)	**Estimated Sensitivity** P(MI* | MI)
Men	94.4%	76.3%
Women	97.1%	59.1%

Want to use this method?

- You’re in luck!

- *COMBO* R Package (coming soon to a CRAN repository near you).
  - **Co**rrecting **M**isclassified **B**inary **O**utcomes

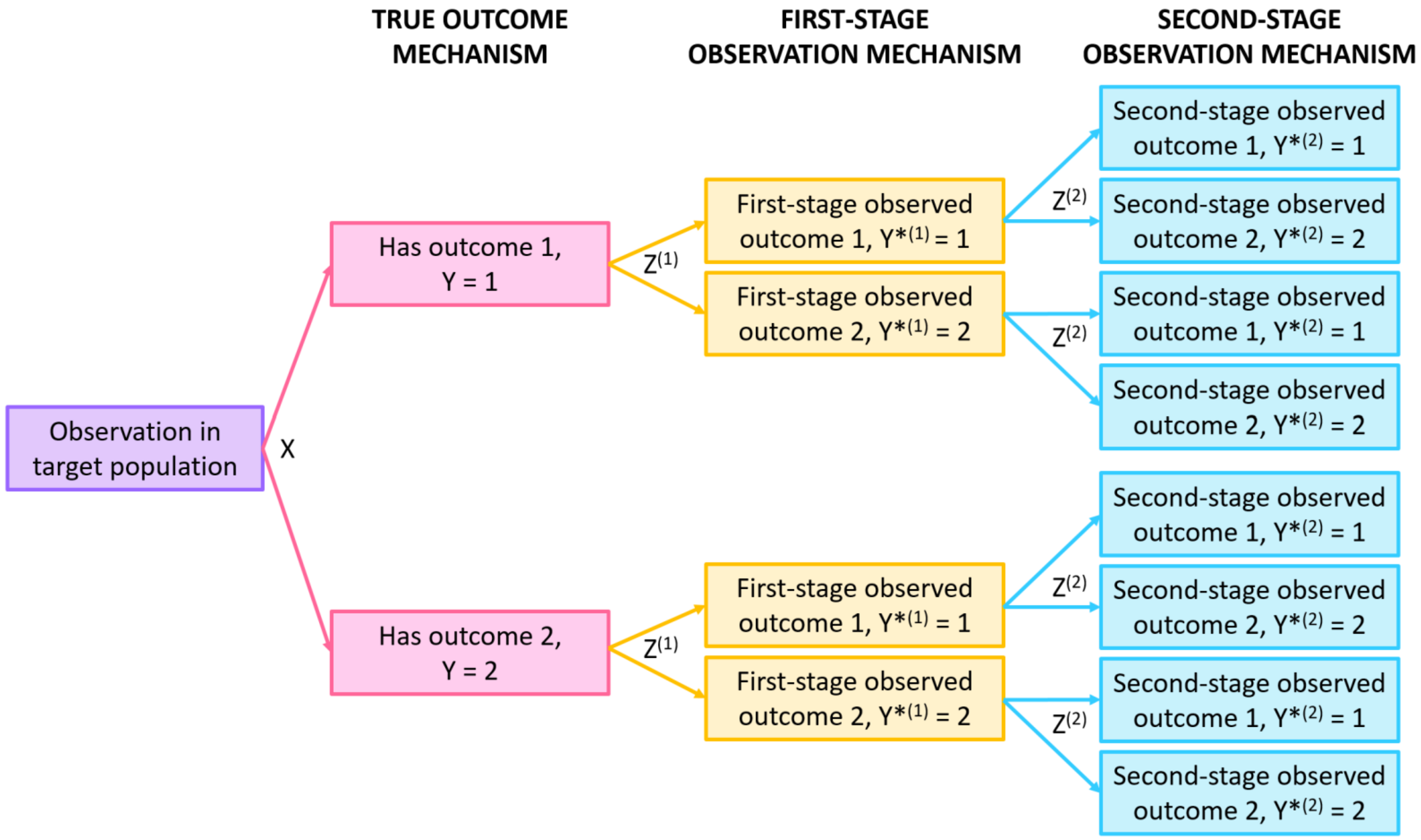


Conclusions and Next Steps

- We can use the proposed EM algorithm to **estimate associations** when a **binary outcome is potentially misclassified**.

Conclusions and Next Steps

- **Extensions:**
  - Outcomes with more than 2 categories.
  - More than one outcome "stage".



Thank you!

Kimberly A. Hochstedler - kah343@cornell.edu

Cornell Bowers CIS
Statistics and Data Science

Notation cheat sheet and more info on "COMBO" available at: **bit.ly/R_COMBO**