WASHINGTON STATE UNIVERSITY
College of Pharmacy and Pharmaceutical Sciences

2026 USTUR Scientific Advisory Committee Meeting
Hampton Inn, Richland, Washington; April 16–17, 2026

Correcting Outcome Misclassification: a Simulation Study

Stacey L. McComish, *Associate in Research*

United States Transuranium and Uranium Registries

1845 Terminal Drive, Suite 201, Richland, WA 99354

ustur.wsu.edu | s.mccomish@wsu.edu

"Learning from Plutonium and Uranium Workers"

Correcting for Misclassification

- Correction method adapted from a paper by Rogan and Gladen, 1978
- Corrects for probability

$$P_{corr} = \frac{(P_{mis} - over)}{(1 - over - under)}$$

P_{mis} = misclassified probability

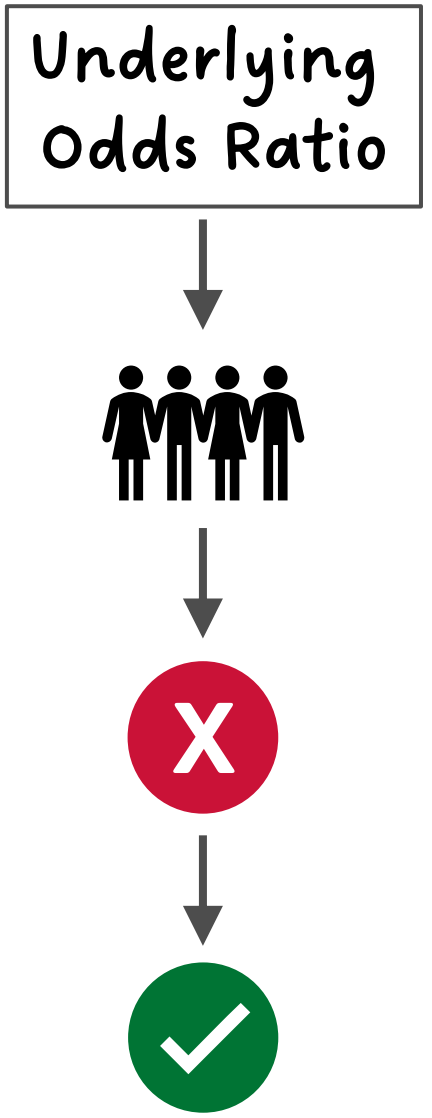
P_{corr} = corrected probability
over = over-misclass rate
under = under-misclass rate

Note: Can generate probabilities that are NOT between 0 and 1:
If $P_{corr}<0$, set P_{corr} to 10^{-5}
If $P_{corr}>1$, set P_{corr} to $1-10^{-5}$

Negative probabilities can happen when the baseline disease rate is less than the over-misclassification rate.

A Simulation Study

- Step 1: Define the **underlying** association between dose and disease (OR=1, 2, or 4.5)
- Step 2: Randomly generate possible **population** level datasets
- Step 3: Randomly simulate possible **misclassified** datasets
- Step 4: Randomly simulate possible **corrected** datasets



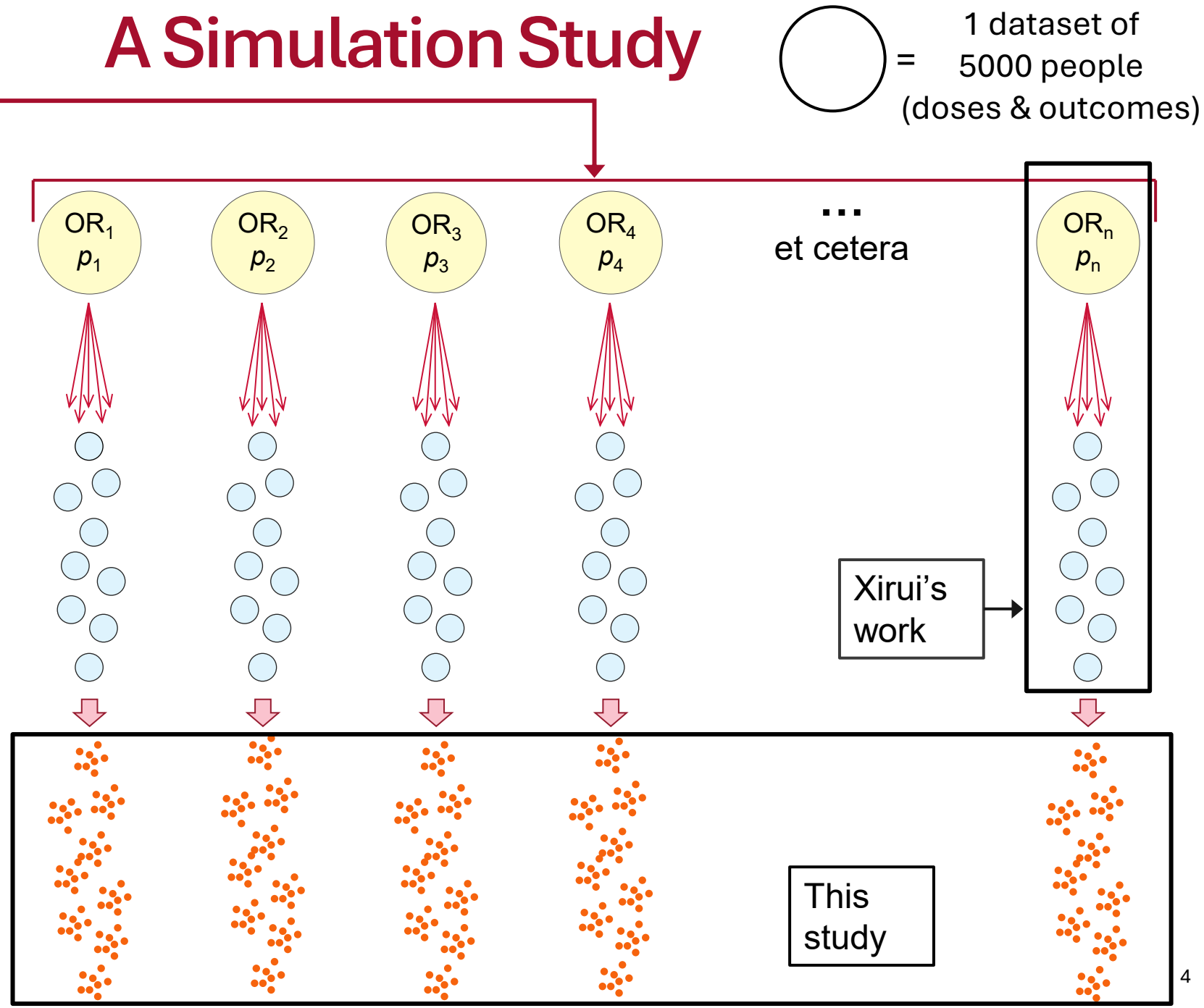
A Simulation Study

Underlying Association
what an epidemiological study wants to know

Population *what people really died from*

Misclassified
death certificate errors

Corrected



A Simulation Study

Underlying Association
what an epidemiological study wants to know

Population level *what people really died from*

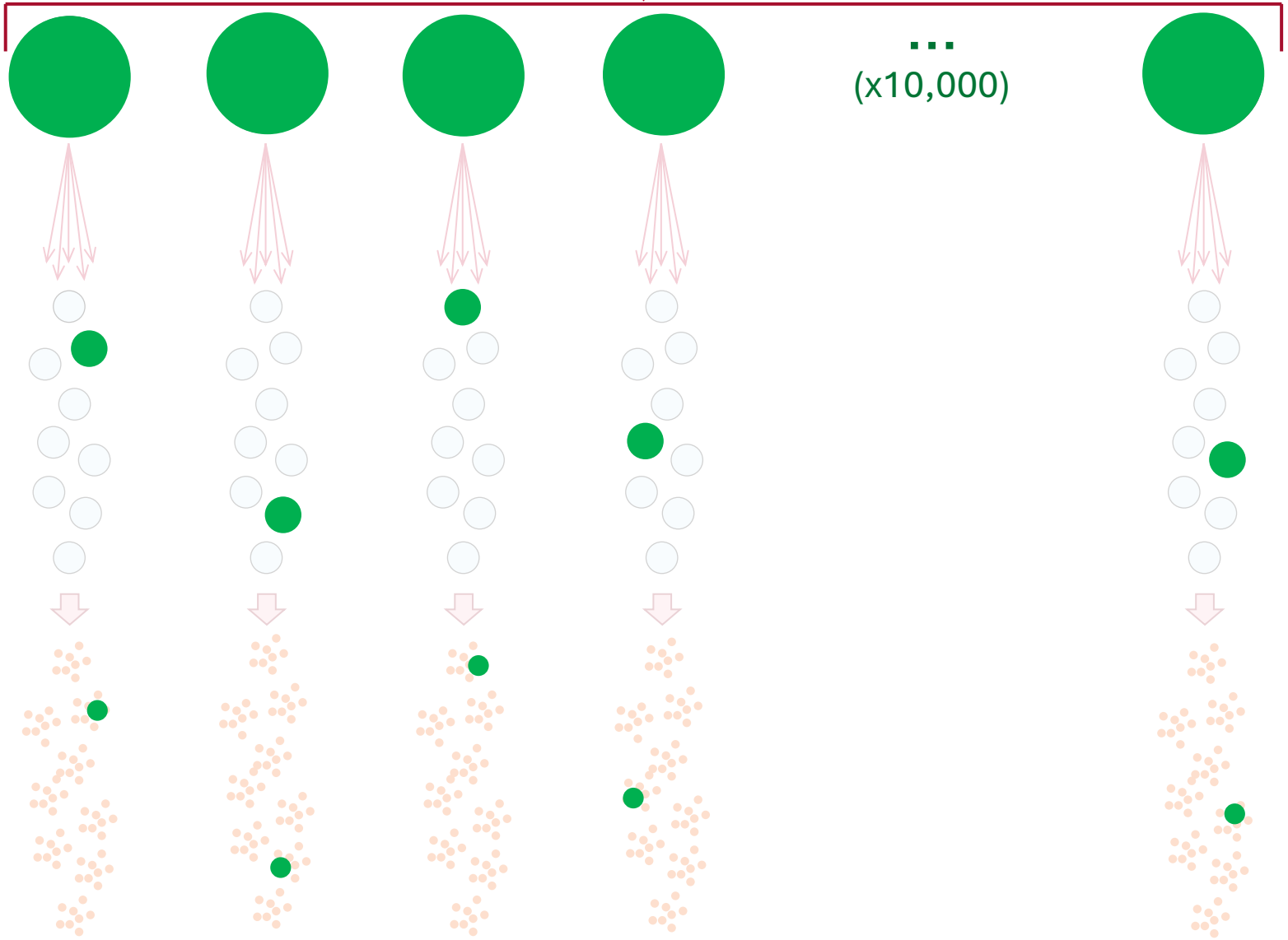
baseline cancer = 20%

Misclassified
death certificate errors

under-misclass = 20%
over-misclass =20%

Corrected

Dataset used in simulation



How Does the Correction Work?

INPUT
1 Misclassified dataset

1 dataset = doses and outcomes for 5,000 people

Logistic Fit
Fit the dataset using a logistic regression

The logistic regression results in **two regression coefficients**

- ✓ β_0: related to baseline disease rate
- ✓ β_1: related to the odds ratio, OR = exp(β_1)

OR=1: no association between dose and disease

OR>1: causative association between dose and disease (if statistically significant)

Regression coefficients can be used to calculate probability of disease as a function of dose.

How Does the Correction Work?

INPUT
1 Misclassified dataset

1 dataset = doses and outcomes for 5,000 people

Logistic Fit
Fit the dataset using a logistic regression

Misclassified Outcome Probabilities
Use fit to calculate outcome probabilities for all 5,000 doses

Logistic probability function used to calculate probability of disease for each of the 5,000 doses:

$$p(dose) = 1/\left[1 + e^{-(\beta_0+\beta_1 \cdot dose)}\right]$$

How Does the Correction Work?

INPUT
1 Misclassified dataset

1 dataset = doses and outcomes for 5,000 people

Logistic Fit
Fit the dataset using a logistic regression

Misclassified Outcome Probabilities
Use fit to calculate outcome probabilities for all 5,000 doses

Corrected Outcome Probabilities
Correct each misclassified probability using a correction factor

$$P_{corr} = \frac{(P_{mis} - over)}{(1 - over - under)}$$

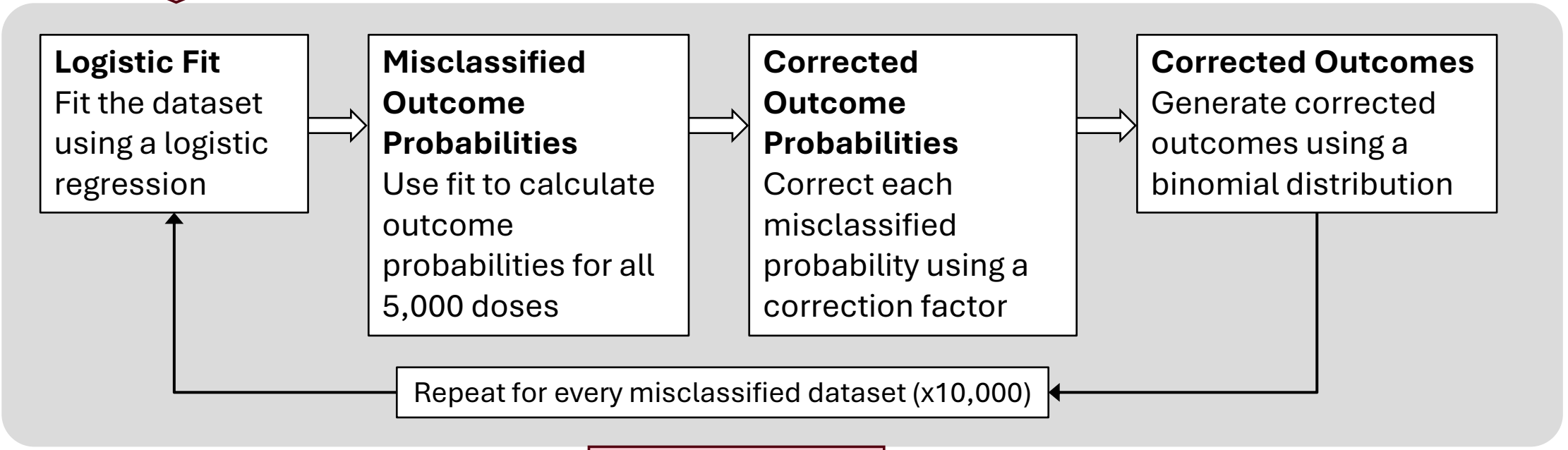
P_{mis} = misclassified probability

P_{corr} = corrected probability
over = over-misclass rate
under = under-misclass rate

How Does the Correction Work?

INPUT
1 Misclassified dataset

1 dataset = doses and outcomes for 5,000 people



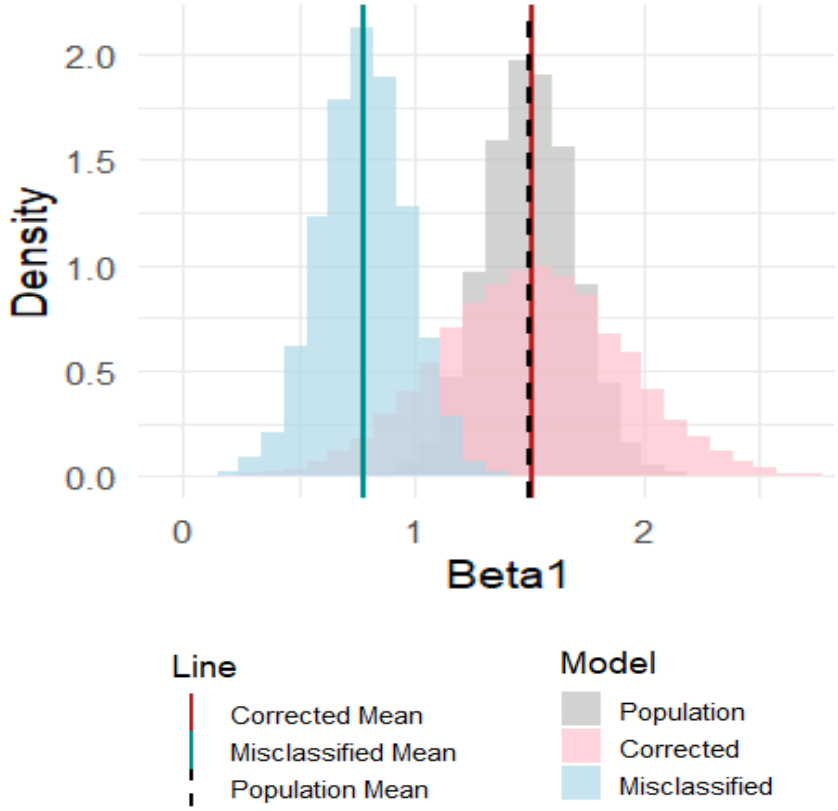
OUTPUT
10,000 Corrected datasets

Other outputs:

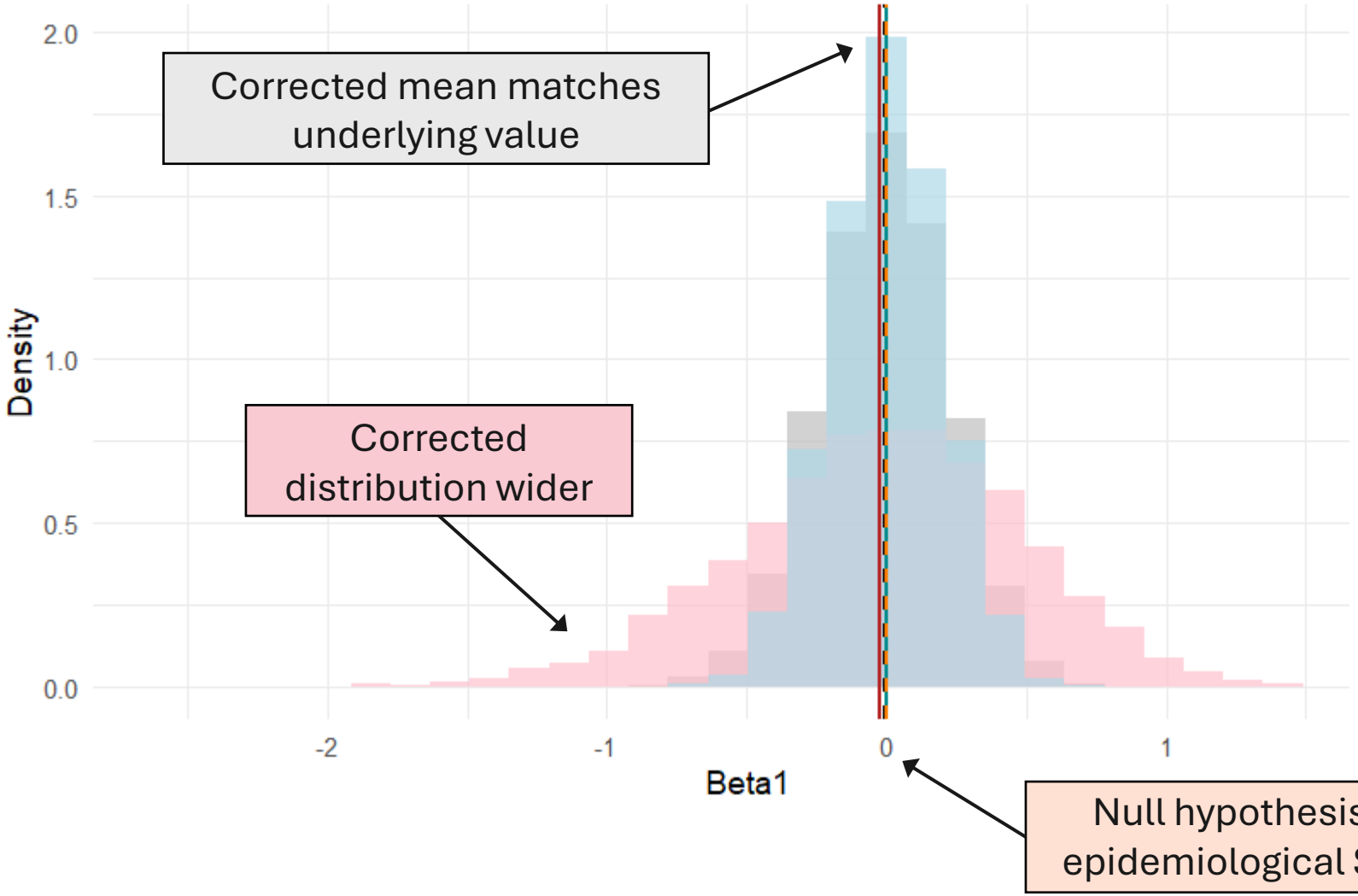
- ✓ Underlying β_1/OR
- ✓ 10,000 Population datasets
- ✓ 10,000 Misclassified datasets

Correction Results

- For each dataset,
  - ✓ β_1
  - ✓ 95% confidence interval for β_1
- For each distribution of 10,000 datasets,
  - ✓ Mean β_1
  - ✓ Standard deviation of β_1
- Could plot histograms for OR, but β_1 is better suited to show distortions in the histograms



No Dose Response ($OR_{Underlying}=1/\beta_1=0$)



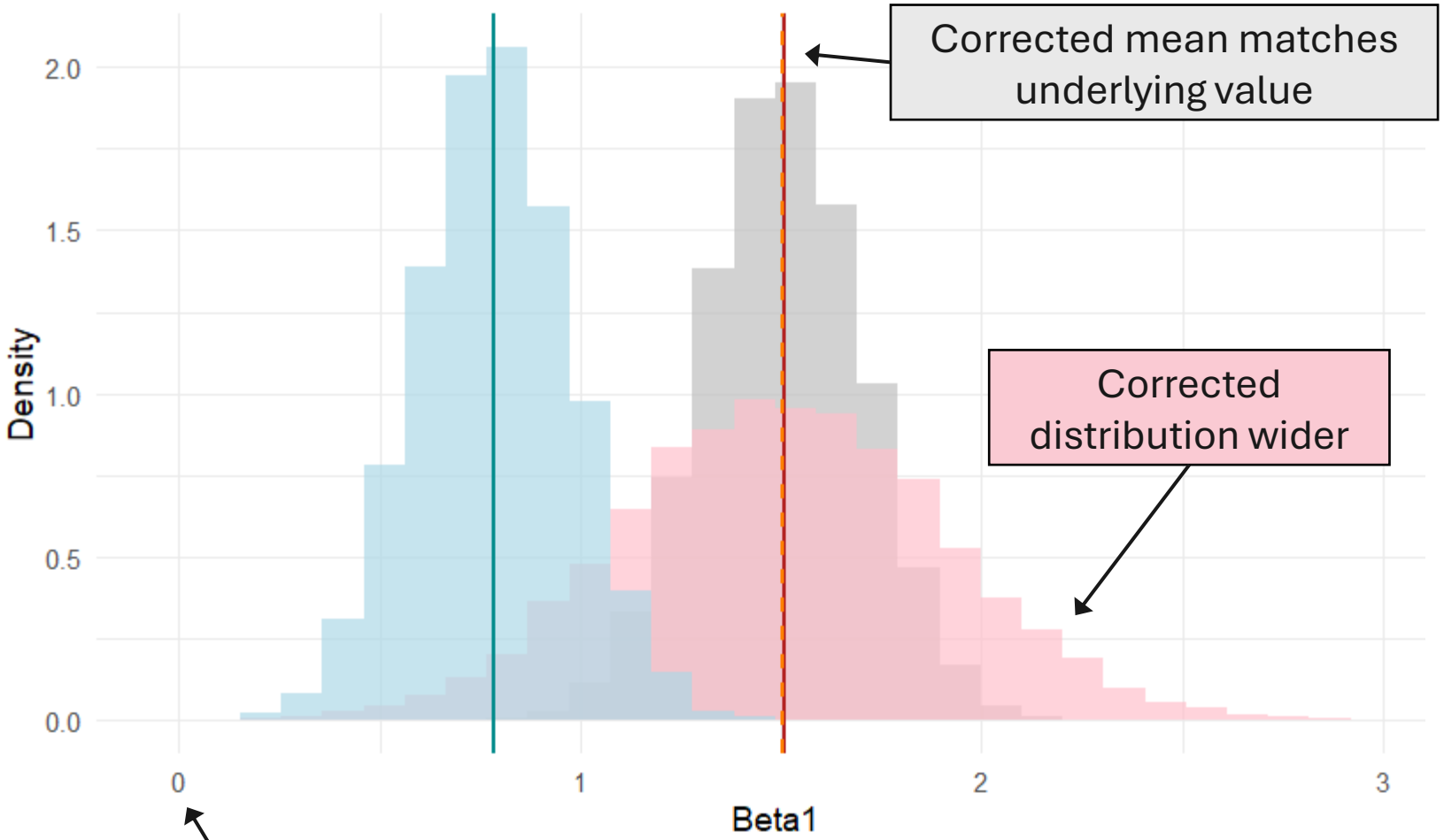
- All three distributions are centered around the underlying β_1 value

Line
- Corrected Mean
- Misclassified Mean
- True Value
- Population Mean

Model
- Population
- Corrected
- Misclassified

Mean ± SD:	**β_1**
Population	-0.007±0.23
Misclassified	0.001±0.20
Corrected	-0.024±0.52

Strong Dose Response ($OR_{Underlying}$=4.5/β_1=1.5)



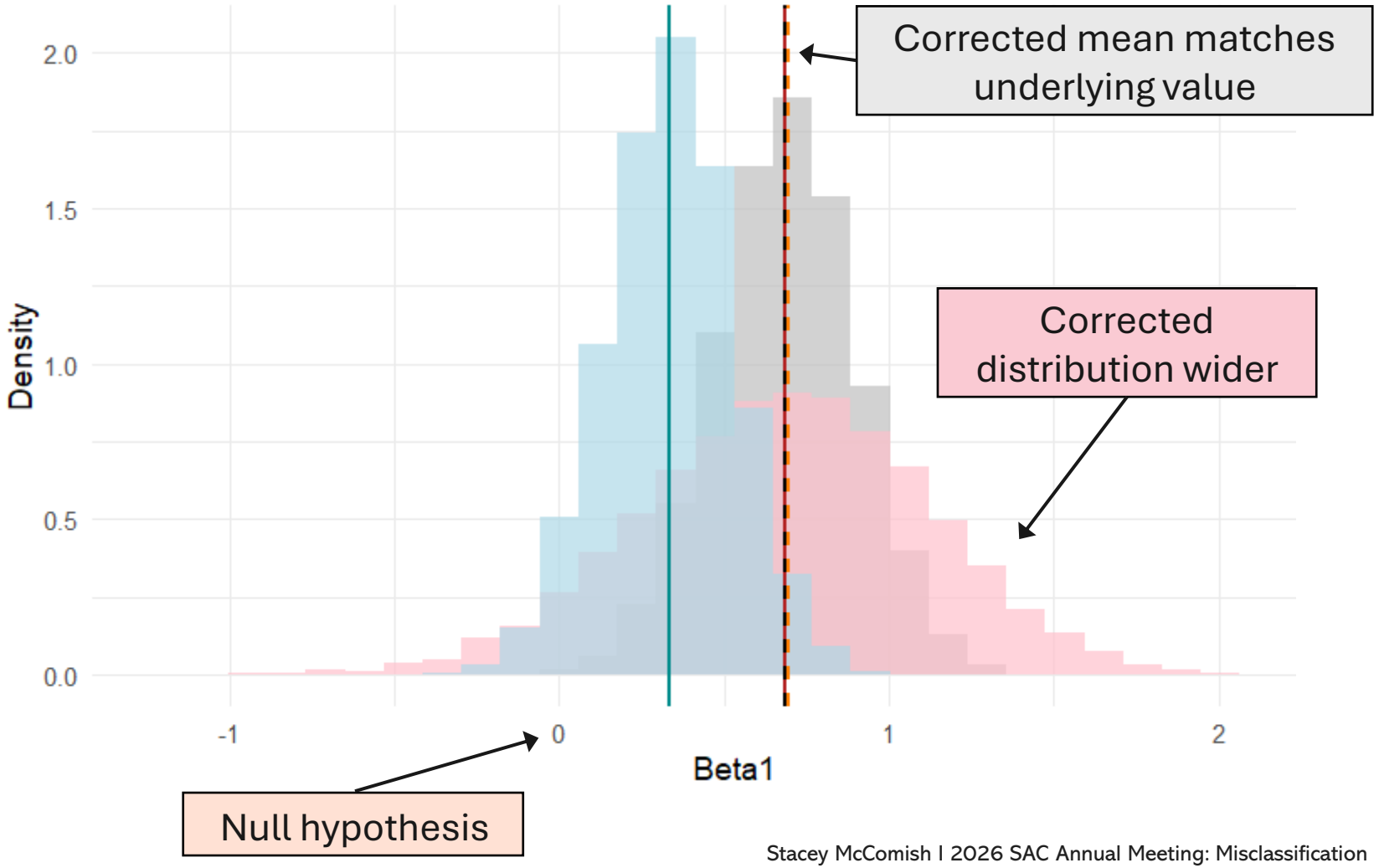
- Misclassification shifts the distribution toward the null, and correction recenters it around the underlying β_1 value

Line
- Corrected Mean
- Misclassified Mean
- True Value
- Population Mean

Model
- Population
- Corrected
- Misclassified

Mean ± SD:	**β_1**
Population	1.51±0.20
Misclassified	0.78±0.19
Corrected	1.51±0.40

Moderate Dose Response ($OR_{Underlying}=2/\beta_1=0.7$)



- Misclassification shifts the distribution toward the null, and correction recenters it around the underlying β_1 value

Line
- Corrected Mean
- Misclassified Mean
- True Value
- Population Mean

Model
- Population
- Corrected
- Misclassified

Mean ± SD:	β_1
Population	0.67±0.21
Misclassified	0.33±0.19
Corrected	0.68±0.44

The Next Step: Study Significance

- So far, we have discussed how well corrected data matches the underlying β_1 value
- But what does this mean for the findings of epidemiological studies?

False Positive	False Negative
Reality No dose response	**Reality** Dose response
Study conclusion Dose response	**Study conclusion** No dose response

_p_-values determine significance

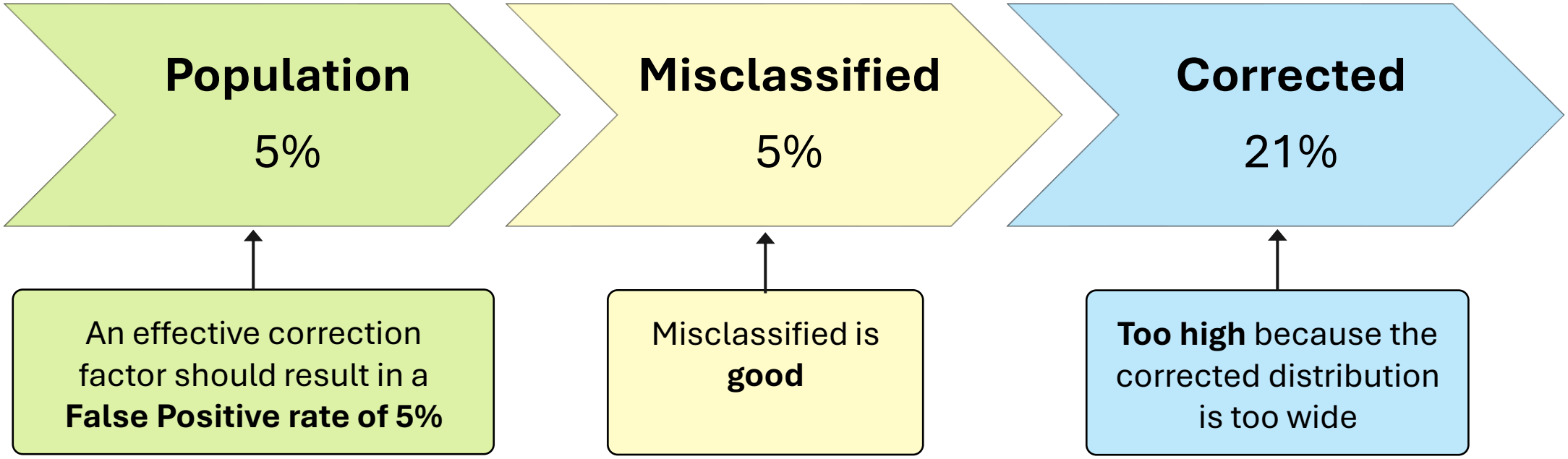
$p \leq 0.05$: significant (dose response)

$p > 0.05$: non-significant (no dose response)

False Positive Rate

Step 1: Run the simulation using $OR_{Underlying}=1$ (no effect)

Step 2: Calculate the **false positive rate** for each distribution

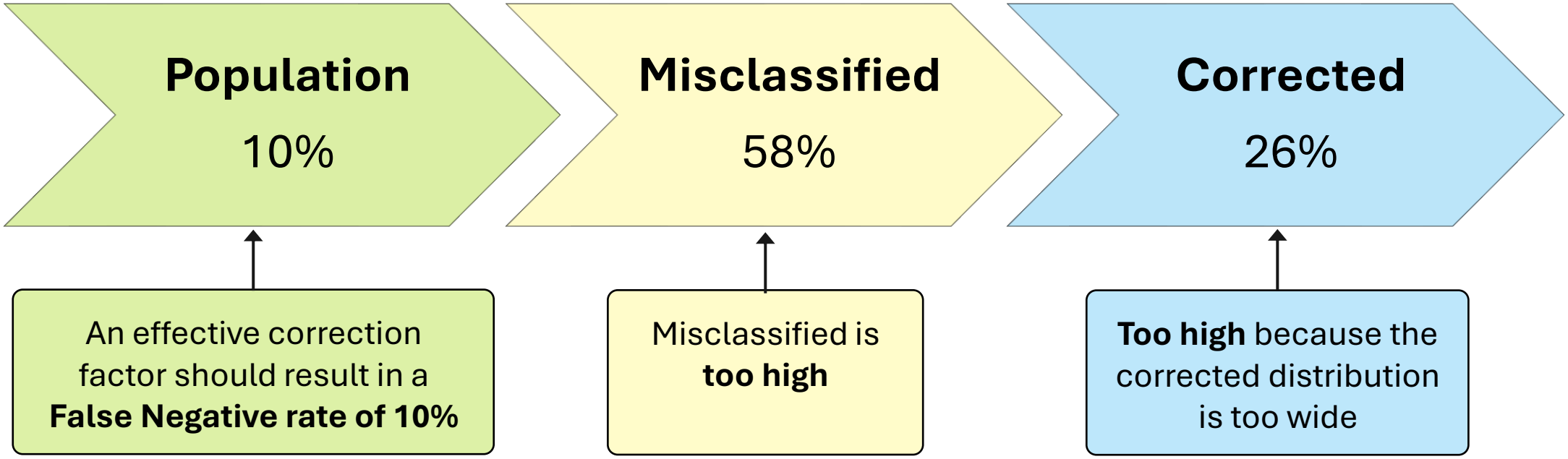


There is no need to correct to correct data if $OR_{Underlying}=1$

False Negative Rate

Step 1: Run the simulation using $OR_{Underlying}=2$ (causative effect)

Step 2: Calculate the **false negative rate** for each distribution

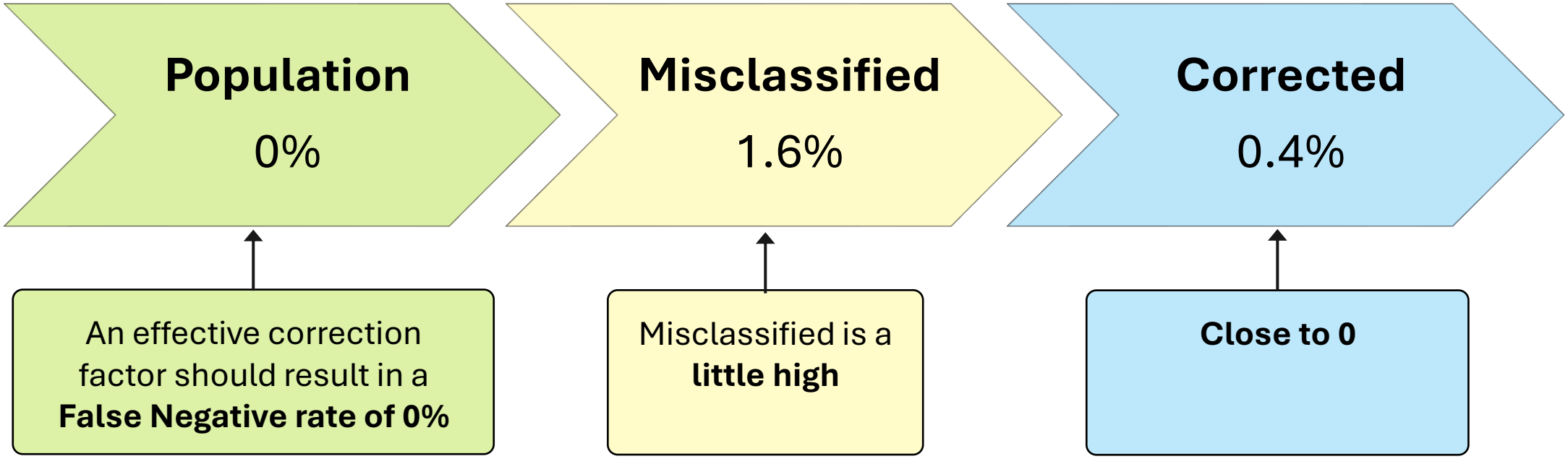


The correction method not useful at $OR_{Underlying}=2$

False Negative Rate

Step 1: Run the simulation using $OR_{Underlying}$=4.5 (causative effect)

Step 2: Calculate the **false negative rate** for each distribution



The correction method possibly useful for $OR_{Underlying}$=4.5

Summary of Significance Results

Dataset Type	False Positive Rate	False Negative Rate			
	OR_{Under}=1	OR_{Under}=1.5	OR_{Under}=2†	OR_{Under}=3	OR_{Under}=4.5
Population	**5.0%**	**54%**	**11%**	**0.05%**	**0%**
Misclassified	5.0%	83%	58%	18%	1.7%
Corrected	21.3%	52%‡	26%	5.5%	0.4%

†approximate, based on only 10,000 datasets

‡Includes both false negatives and corrected odds ratios that were less than 1

- After correction
  - ✓ False positive rate is too high
  - ✓ False negative rate is improved compared to misclassified, but can be quite large
- Clear trend where false negative rates decrease as $OR_{Underlying}$ increases

Summary

Results

- **Bias:** mean values were successfully corrected
- **Variability:** corrected distributions were too wide

Results

- **False positives** made worse
- **False negatives** improved for large odds ratios

Conclusions

- OR=1: correction method not useful
- OR>1: correction method may be useful if OR is large

Note: These findings are valid if baseline disease rate is greater than the over-misclassification rate.

Current Work

- Explore impact of other factors on false negative and false positive rates: number of study subjects, misclassification rates, and baseline disease rate
- Improve the correction method (reduce the width of the corrected distribution)

Questions? Thoughts?