Learning Models from Data with Measurement Error: Tackling Underreporting

Roy Adams, Yuelong Ji, Xiaobin Wang, and Suchi Saria

Johns Hopkins University

Introduction

Goal: Estimate the distribution of **outcome Y** given **exposure A** and **covariates X** from non-experimental data.

Introduction

Goal: Estimate the distribution of **outcome Y** given **exposure A** and **covariates X** from non-experimental data.

Measurement error is common source of bias when using non-experimental data.

Introduction

Goal: Estimate the distribution of **outcome Y** given **exposure A** and **covariates X** from non-experimental data.

Measurement error is common source of bias when using non-experimental data.

- We focus on **underreporting error.**

Introduction

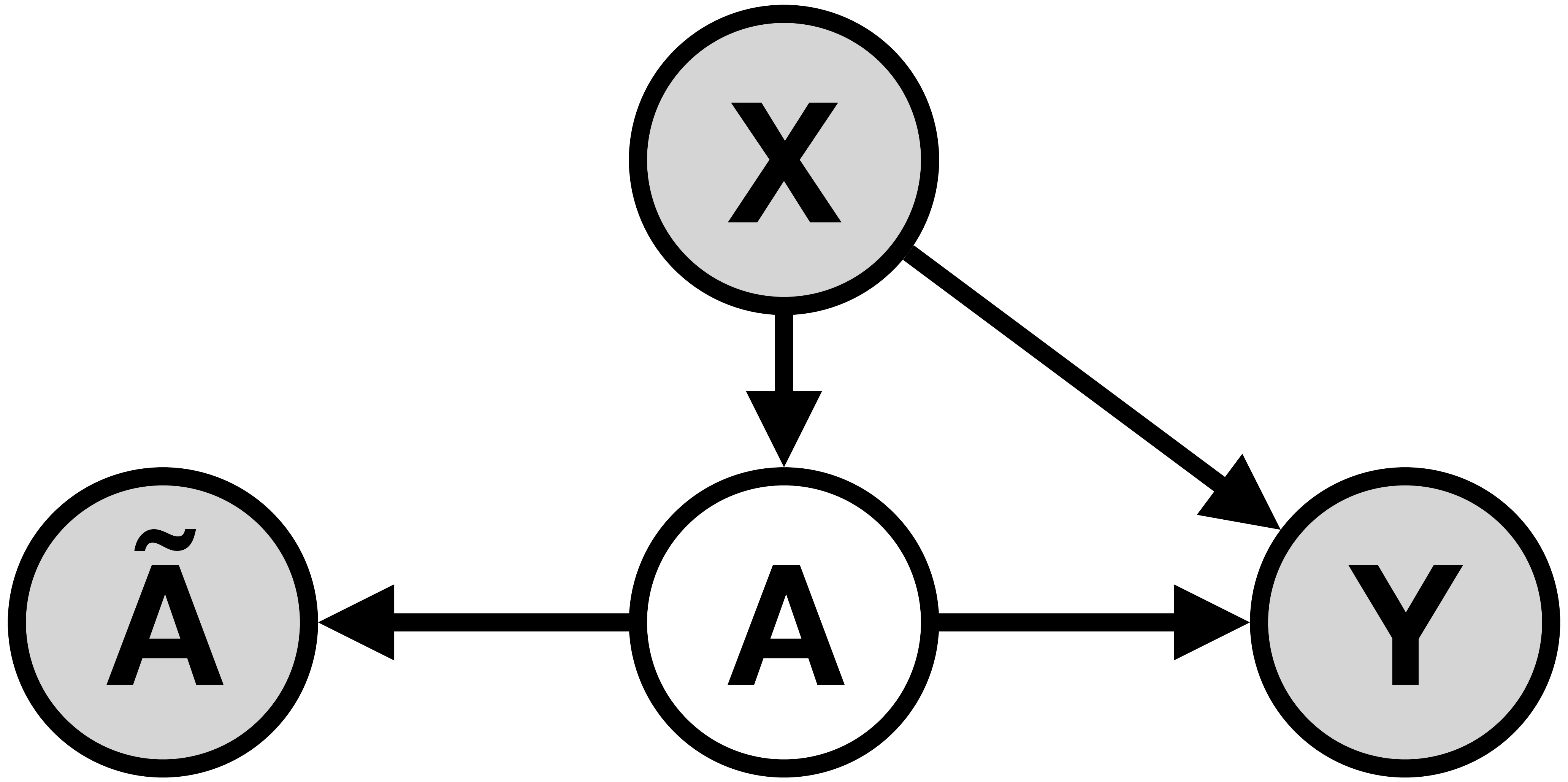
Goal: Estimate the distribution of **outcome Y** given **exposure A** and **covariates X** from non-experimental data.

Measurement error is common source of bias when using non-experimental data.

- We focus on **underreporting error.**
  - E.g. survey data of sensitive variables such as drug use.

Model

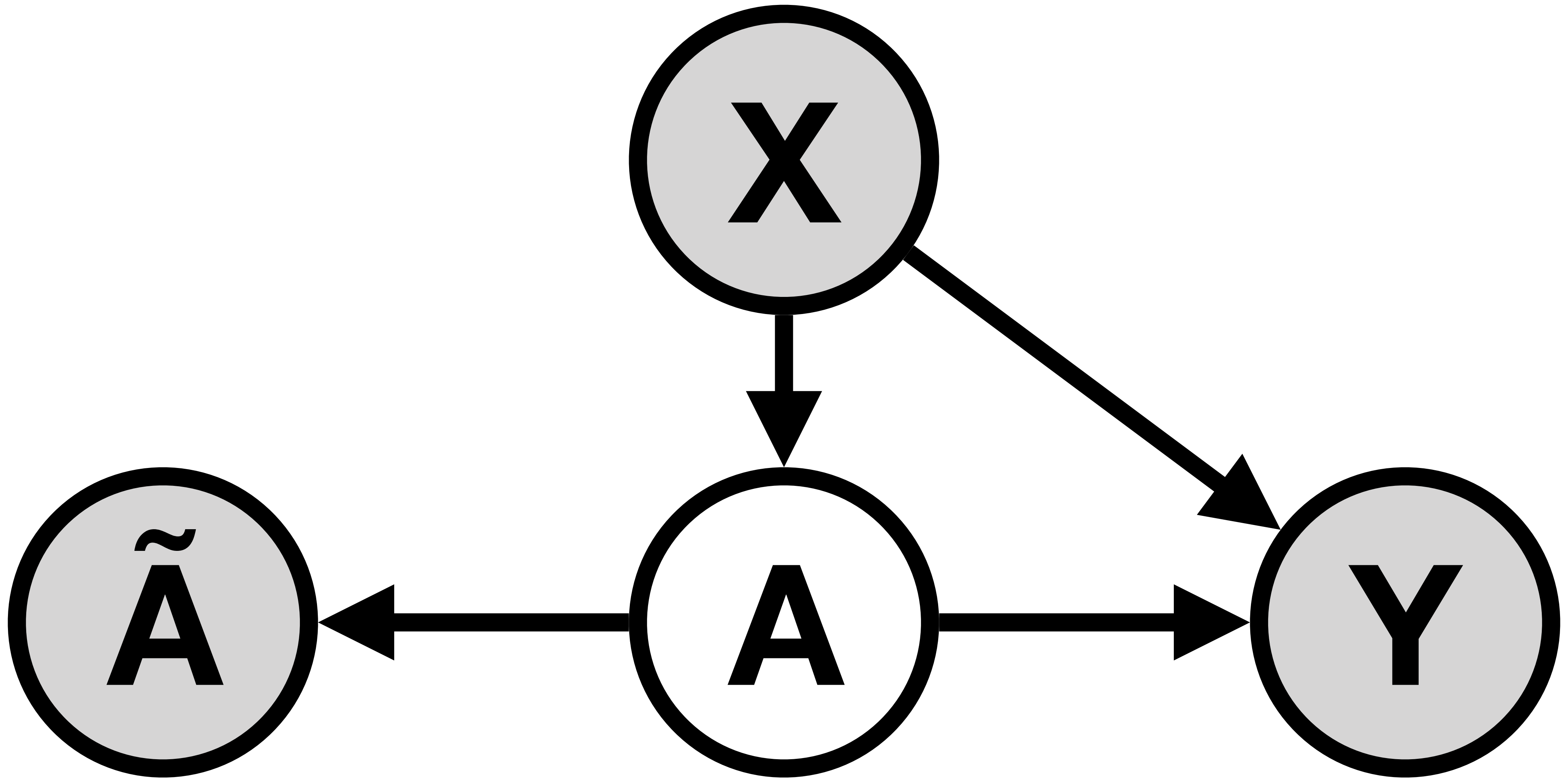
Updated goal: Estimate the distribution of outcome Y given exposure A and covariates X when **exposure observations Ã are subject to underreporting errors**.



Model

Updated goal: Estimate the distribution of outcome Y given exposure A and covariates X when **exposure observations Ã are subject to underreporting errors**.

Assumptions:

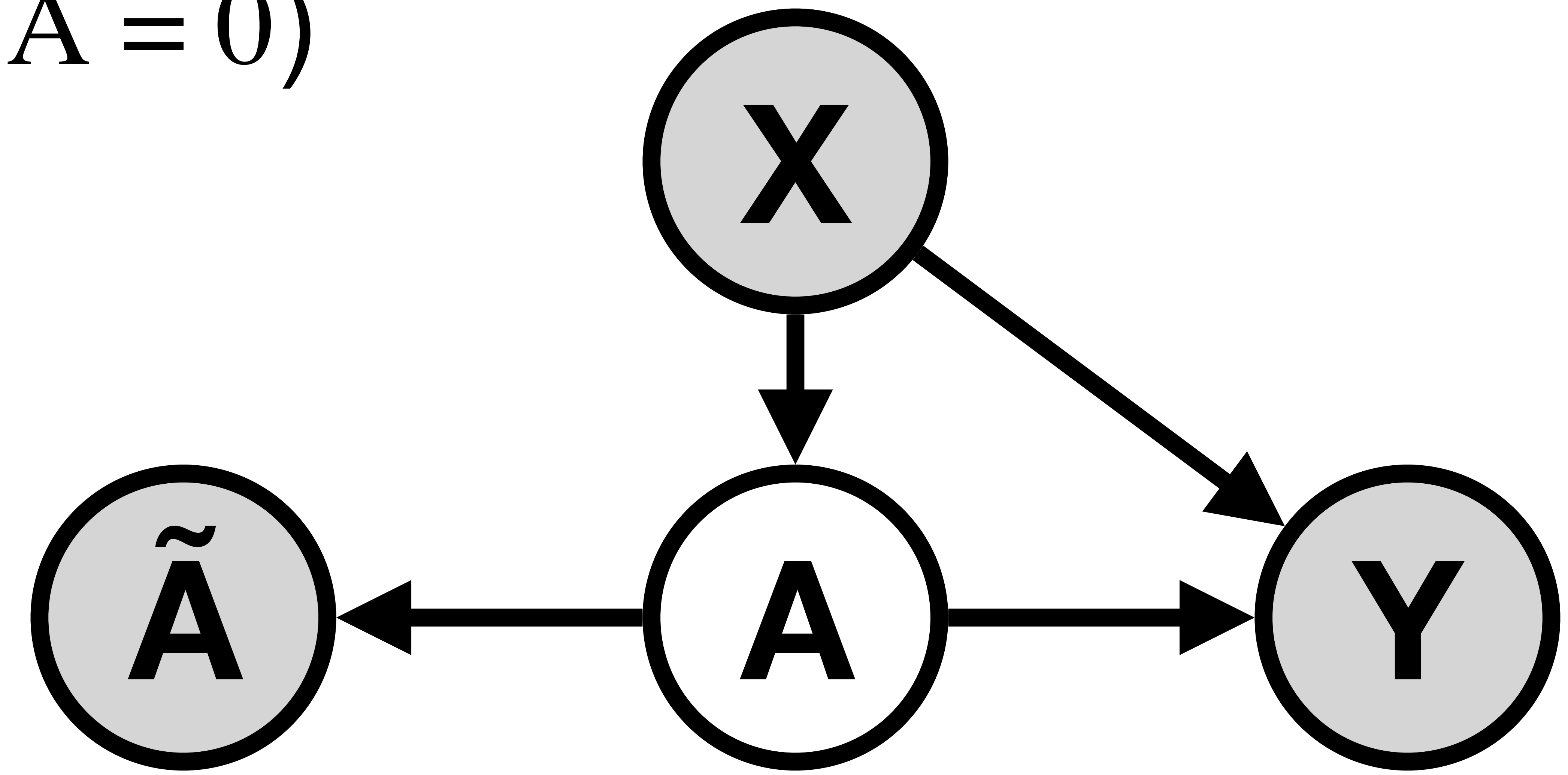


Model

Updated goal: Estimate the distribution of outcome Y given exposure A and covariates X when **exposure observations Ã are subject to underreporting errors**.

Assumptions:

1. Strict underreporting ($A = 0 \implies \tilde{A} = 0$)

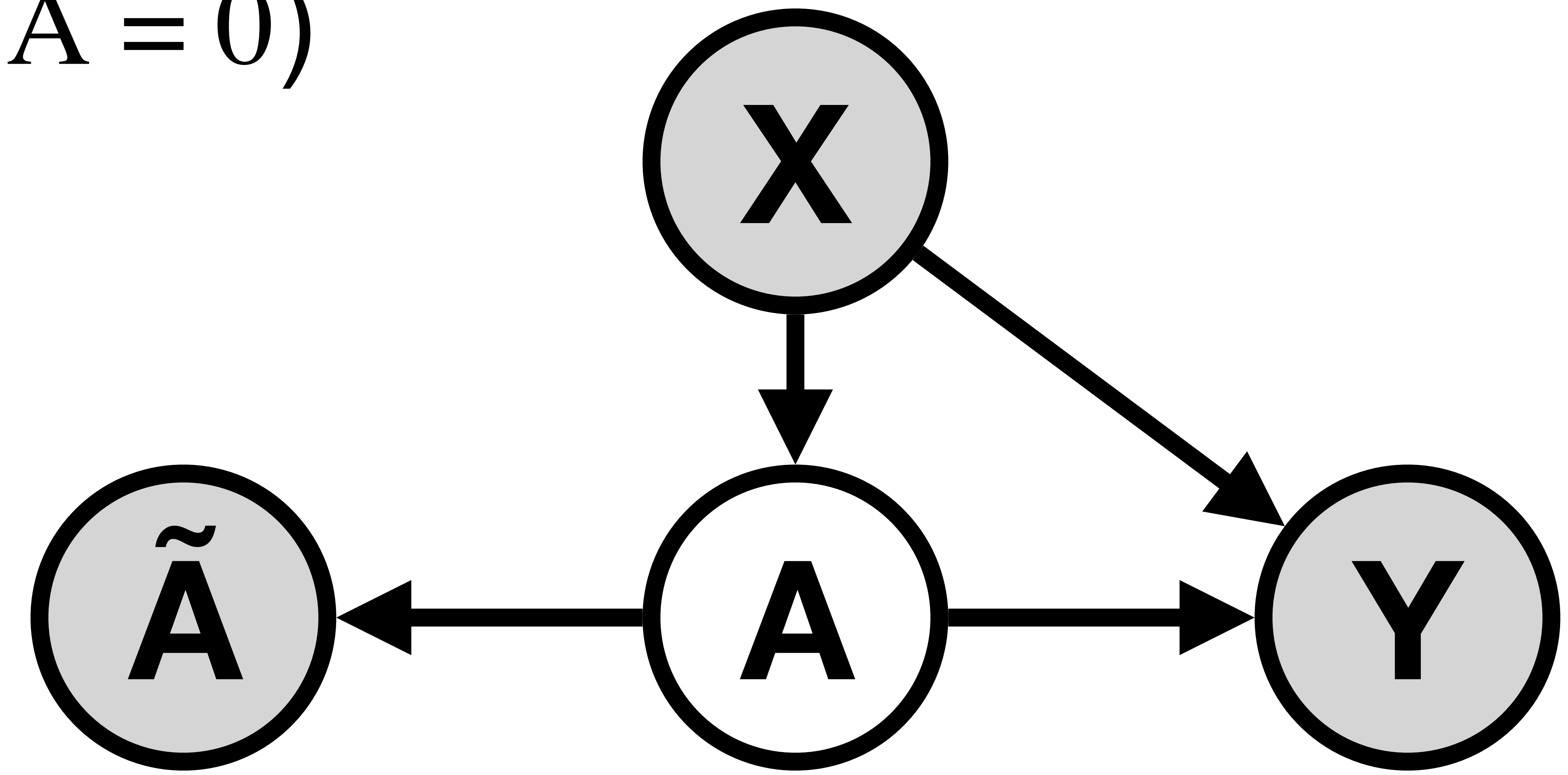


Model

Updated goal: Estimate the distribution of outcome Y given exposure A and covariates X when **exposure observations Ã are subject to underreporting errors**.

Assumptions:

1. Strict underreporting ($A = 0 \implies \tilde{A} = 0$)
2. $\tilde{A}$ is independent of X given A

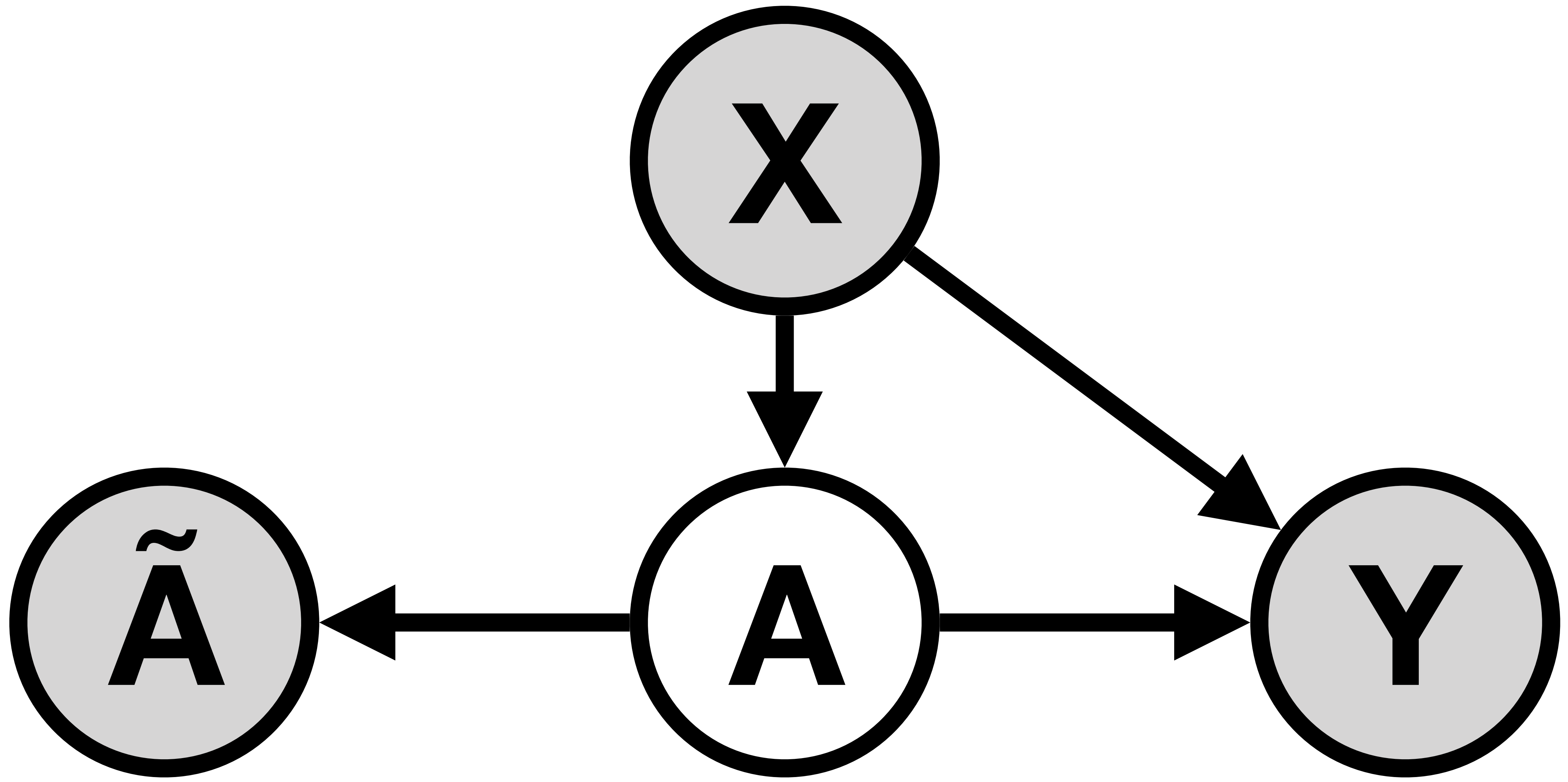


Model

Outcome model … $p_\theta(Y \mid A, X)$

Exposure model … $p_\phi(A \mid X)$

Error model ……… $p_\tau(\tilde{A} \mid A)$

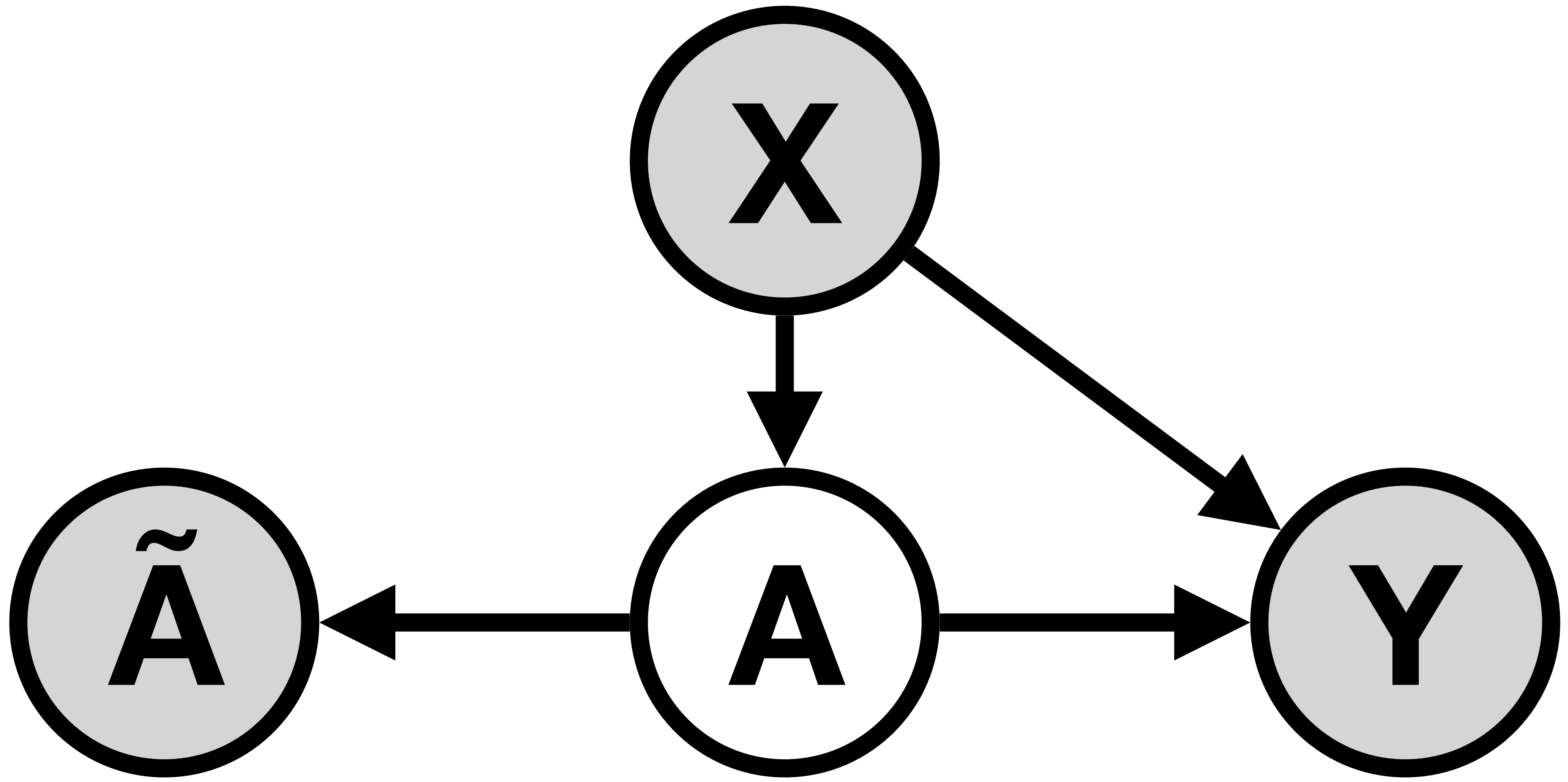


Model

Outcome model … $p_\theta(Y \mid A, X)$

Exposure model … $p_\phi(A \mid X)$

Error model ……… $p_\tau(\tilde{A} \mid A)$



Maximize the **log marginal likelihood**:

$$\max_{\theta,\phi,\tau} \sum_i \log \sum_a p_\theta(y_i \mid a, x_i) p_\tau(\tilde{a}_i \mid a) p_\phi(a \mid x_i)$$

Identifiability

Identifiability

We prove three separate identifiability conditions:

Identifiability

We prove three separate identifiability conditions:

1. The error distribution is known

Identifiability

We prove three separate identifiability conditions:

1. The error distribution is known
2. We have a second error-prone exposure observation

Identifiability

We prove three separate identifiability conditions:

1. The error distribution is known
2. We have a second error-prone exposure observation
3. Under assumptions about the form of the exposure distribution (see paper/poster for details)

Identifiability

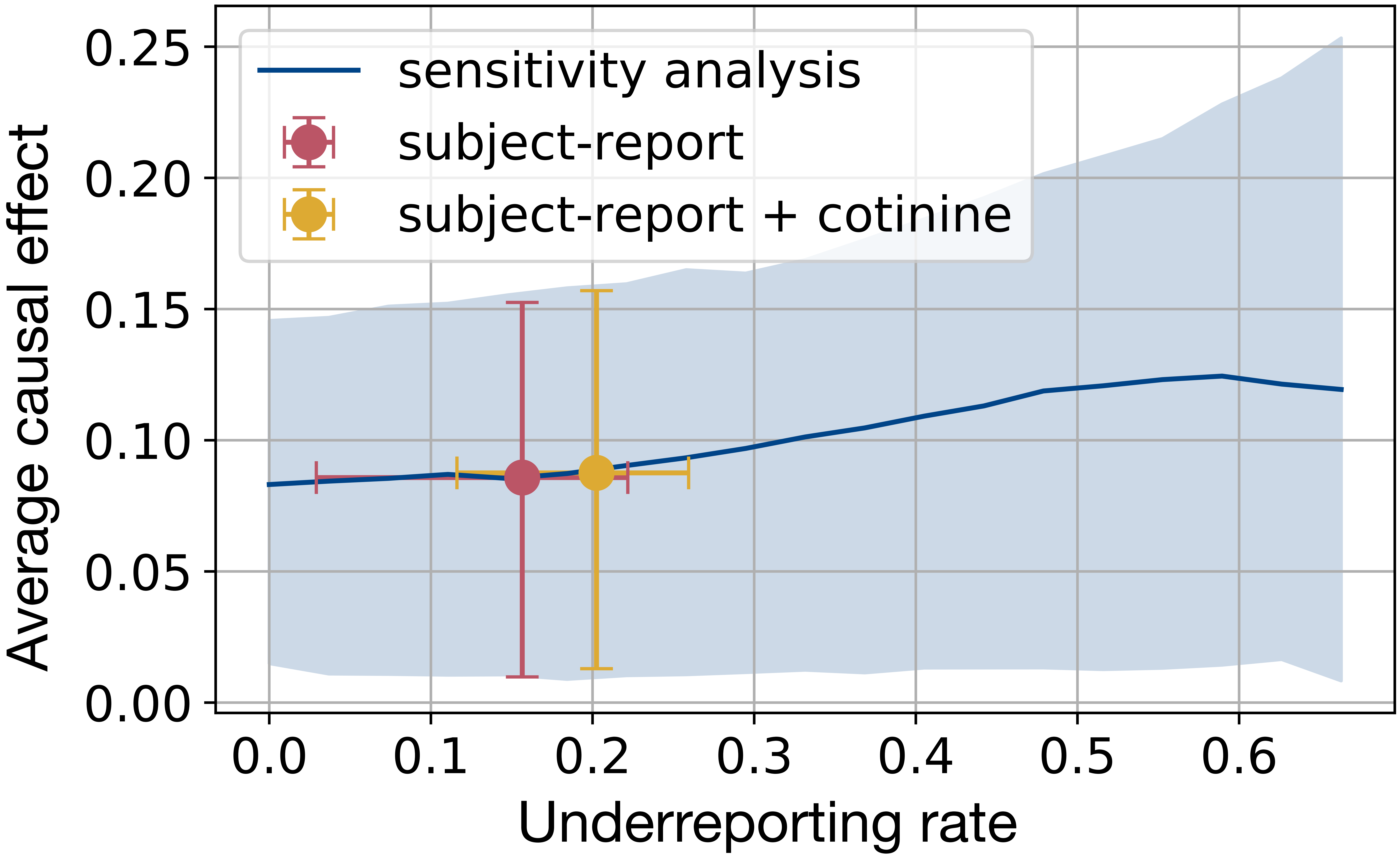
We prove three separate identifiability conditions:

1. The error distribution is known
2. We have a second error-prone exposure observation
3. Under assumptions about the form of the exposure distribution (see paper/poster for details)

In particular:

If $\mathbf{X}$ **is not independent of** $\mathbf{A}$ and $p(A \mid X)$ is a **logit, probit, or cloglog regression model**, then $p(Y, \tilde{A} \mid X)$ is identifiable.

Maternal drug use and childhood obesity



Thanks!

Come see poster #75