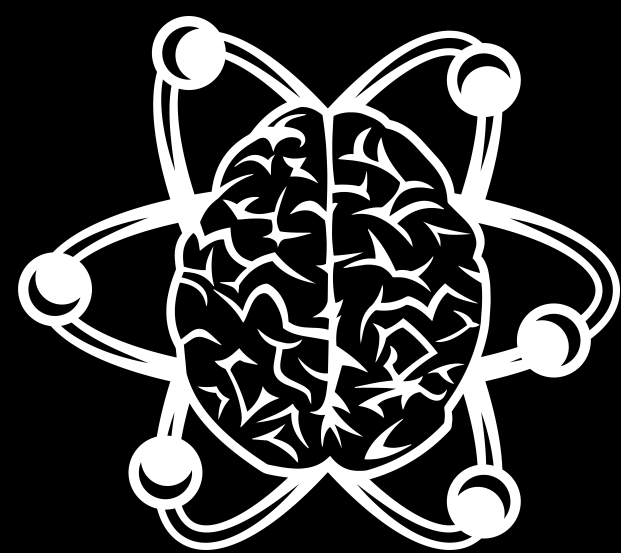
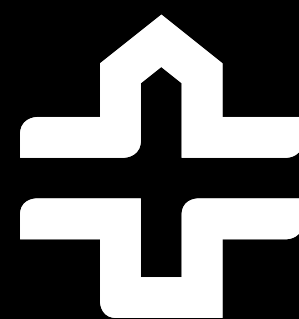




# CHOOSING PREDICTORS TO LINEAR REGRESSION MODELS

Turku PET Centre Brain Imaging Course 2026

Tuulia Malen, Turku PET Centre  
[tukama@utu.fi](mailto:tukama@utu.fi)



**‘Blindly tossing variables into the causal salad is never a good idea.’**

**- R. McElreath**



# Reference

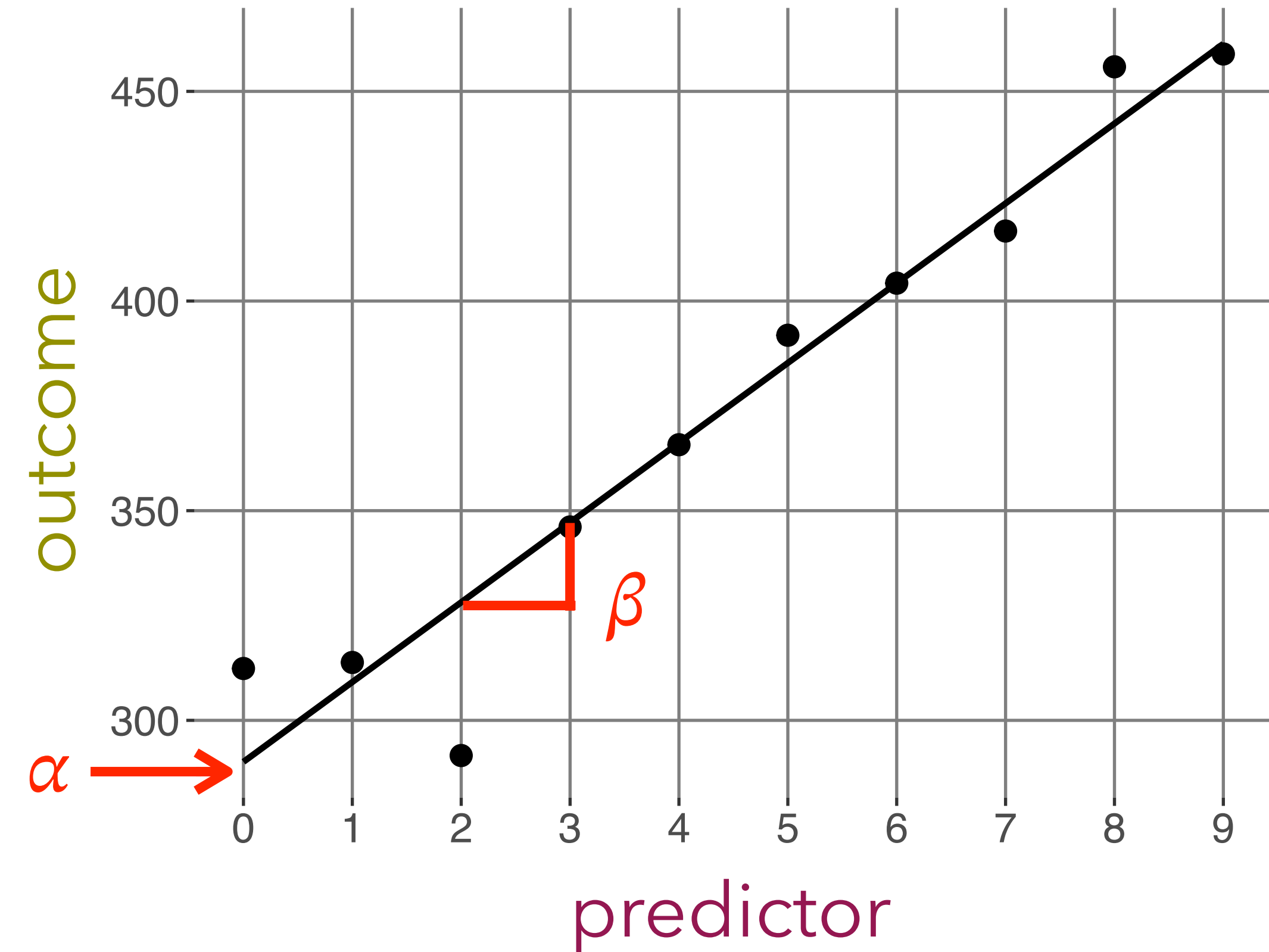
- Richard McElreath, an American professor of Anthropology, Max Planck Institute, Germany
- Statistical Rethinking. *A Bayesian Course with Examples in R and Stan* (2<sup>nd</sup> edition, 2020). Chapman and Hall/ CRC.
  - Chapters 5-7 (mainly)
  - The 2024 edition of the course, including slides, lectures and exercise material:
    - [https://github.com/rmcelreath/stat\\_rethinking\\_2024](https://github.com/rmcelreath/stat_rethinking_2024)
    - <https://www.youtube.com/watch?v=mBEA7PKDmiY&list=PLDcUM9US4XdPz-KxHM4XHt7uUVGWWVSus&index=5>
- Visualization
  - Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL <http://www.posit.co/>.
  - OpenAI's DALL·E via ChatGPT

# Introduction: Linear regression modeling

- Estimate associations between one or more predictors (independent variables) and an outcome (dependent variable)
- Predictor effect on outcome
  - How age, sex, and chronic pain affect the opioid receptor availability?
  - How dopamine synthesis capacity affects the dopamine receptor availability?

# Introduction: General linear regression with main effects

- If one predictor:
  - $\text{outcome}_{\text{obs}} = \alpha + \beta * \text{predictor}_{\text{obs}} + \text{residual}_{\text{obs}} \text{ (error)}$
- $\alpha$  = intercept ( $\text{outcome}$  when  $\text{predictor} = 0$ )
- $\beta$  = **the effect** = regression coefficient = the change in  $\text{outcome}$  with one-unit increase in  $\text{predictor}$
- $\alpha$  and  $\beta$  are estimated by the regression model based on our data so that the residuals (distances between observations and the fit= line) are minimized
- Syntax (e.g. in R) often something like:  $\text{outcome} \sim \text{predictor}$





# Introduction: General linear regression with main effects

- If one predictor:

- $\text{outcome}_{\text{obs}} = \alpha + \beta^* \text{predictor}_{\text{obs}} + \text{residual}_{\text{obs}}$

- If two (or more) predictors:

- $\text{outcome}_{\text{obs}} = \alpha + \beta^* \text{predictor1}_{\text{obs}} + \beta_2^* \text{predictor2}_{\text{obs}} + \dots + \text{residual}_{\text{obs}}$

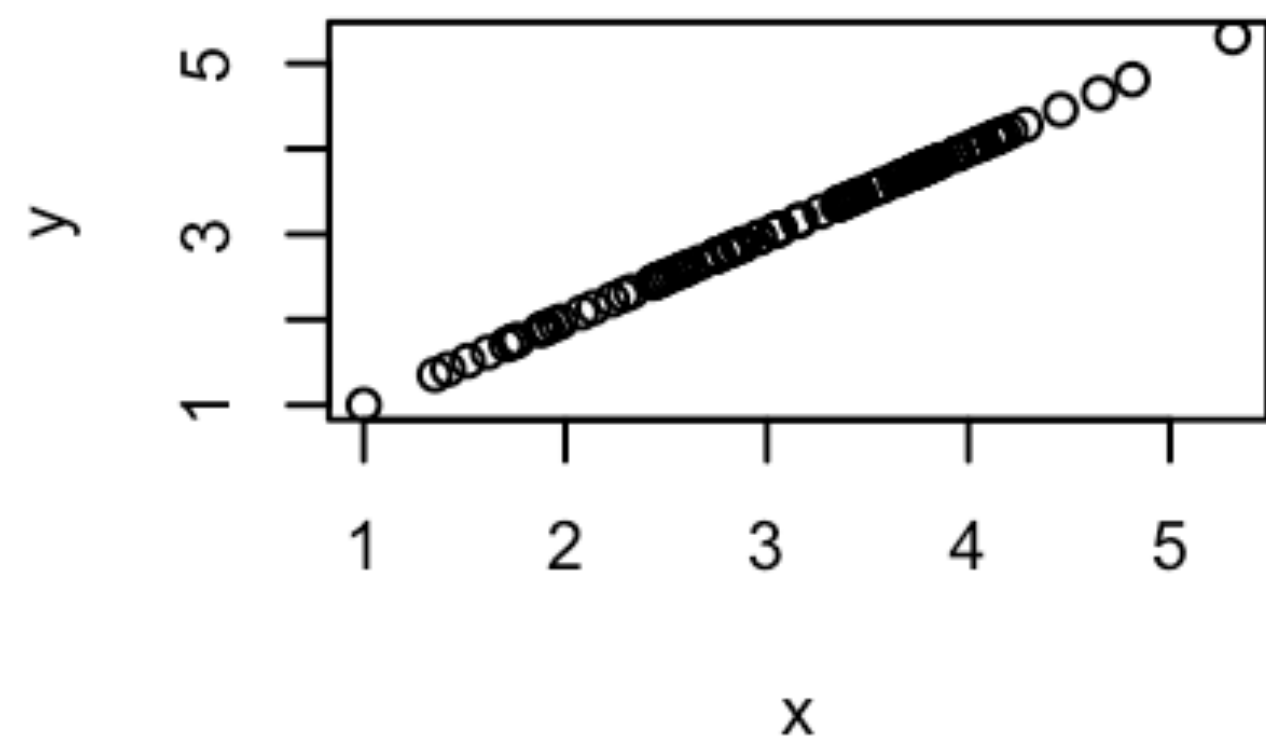
- The effect of  $\text{predictor1}$ , when  $\text{predictor2}$  is adjusted ('controlled') for

# Correlation & Causality

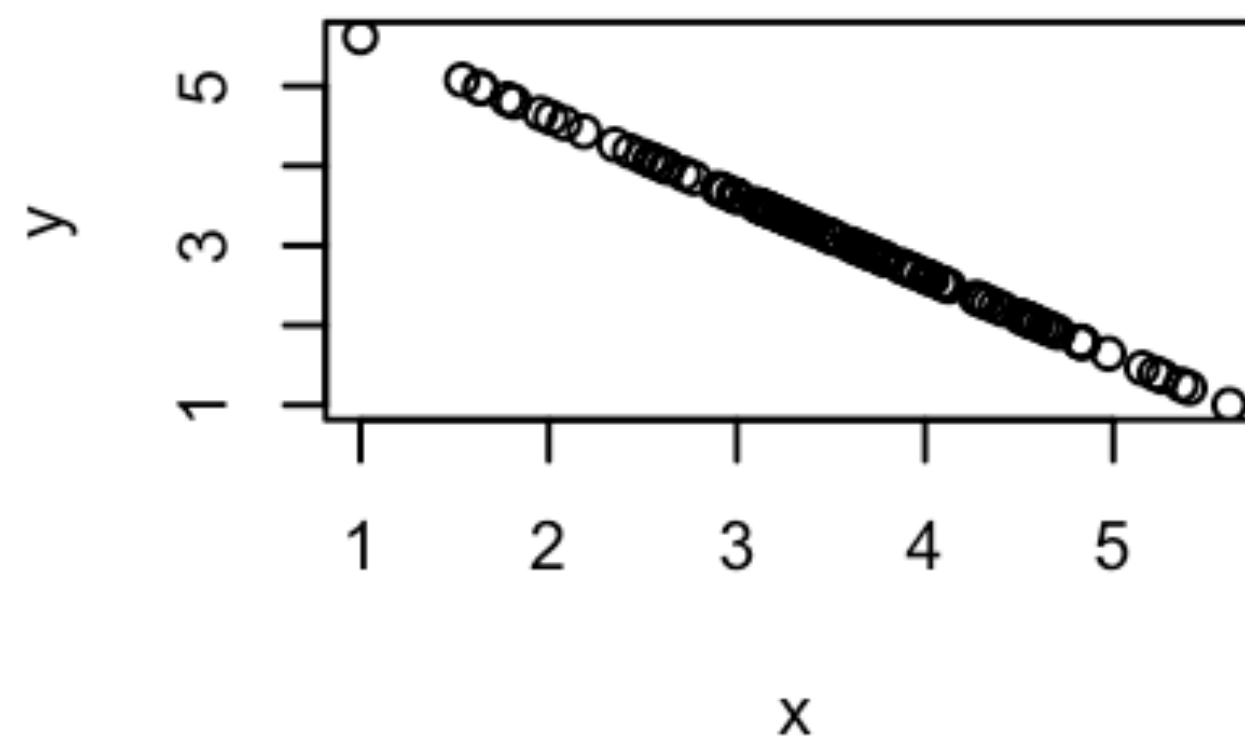
- Linear regression is based on correlation between the variables (**predictors** and **outcome**)
- **Correlation is common and it does not necessarily reveal causality, but only association**
  - Thus, the *effect* is only an association, although it sounds causal!
  - <https://www.tylervigen.com/spurious-correlations>
    - High correlation of butter consumption and ticket prices at North American movie theaters from 2001 to 2021.

# Correlation & Causality

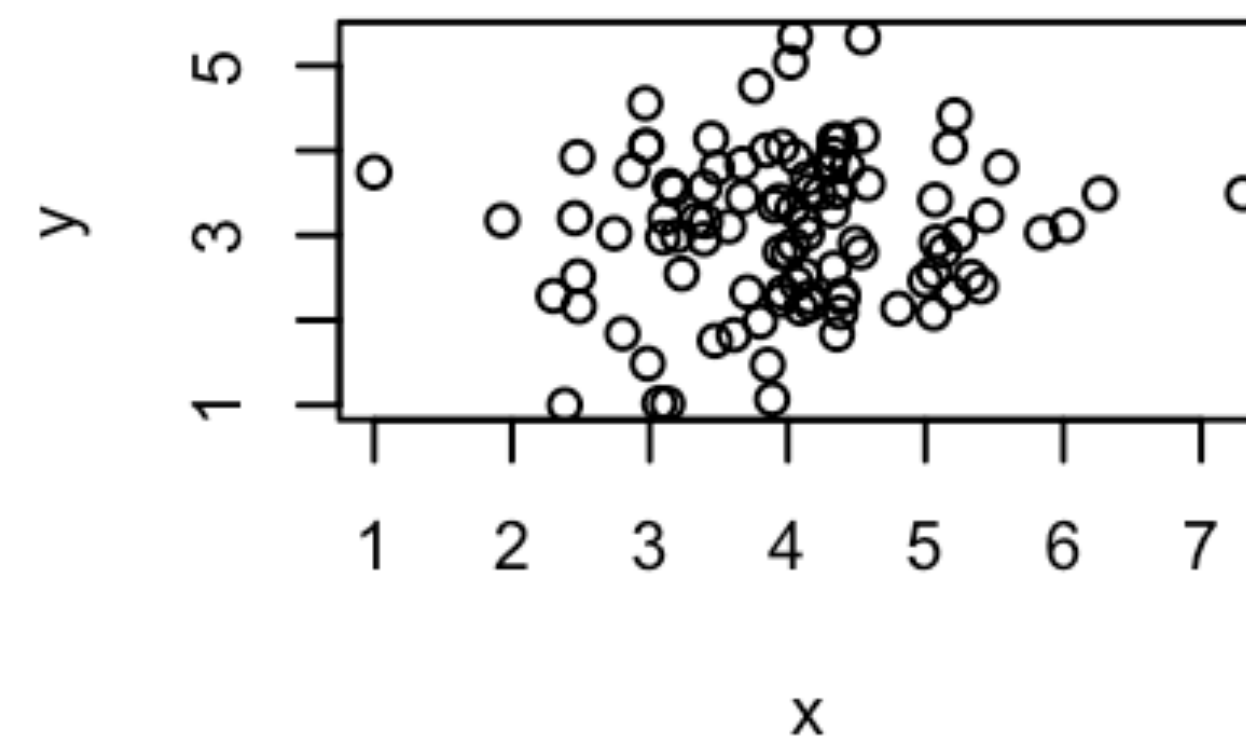
**Perfect Positive Correlation = 1**



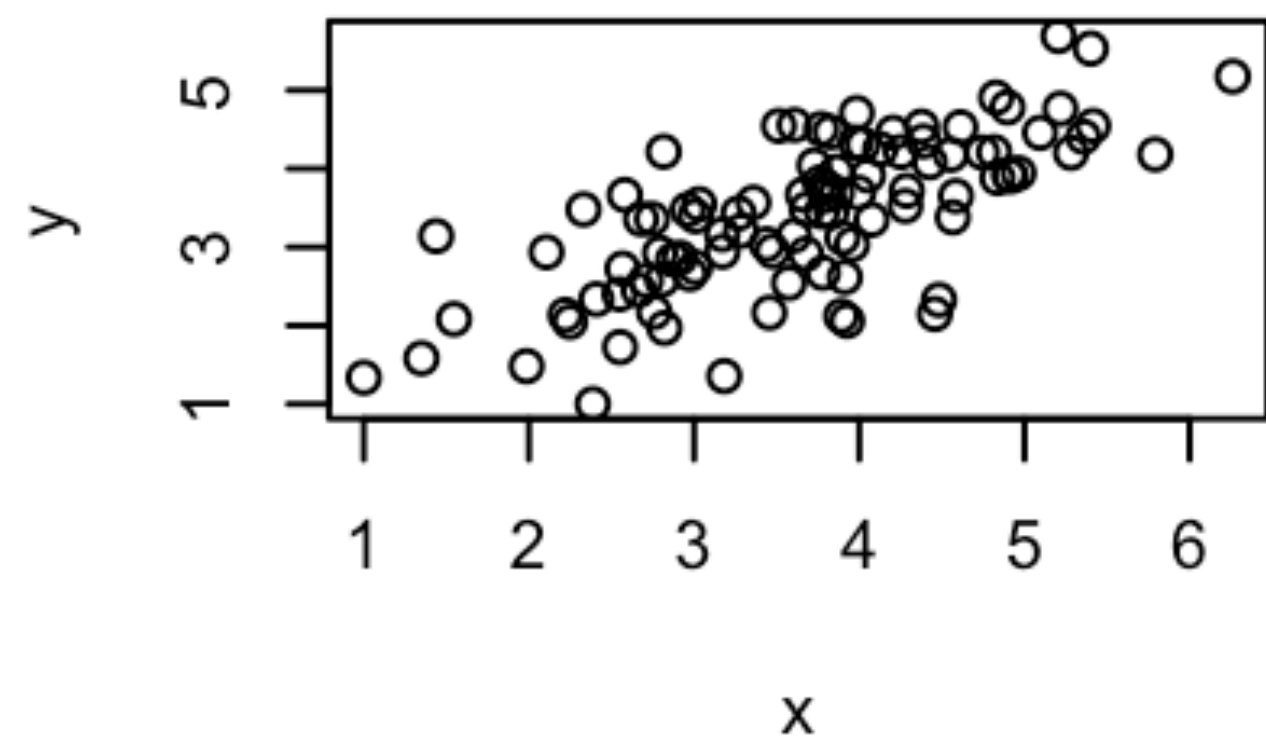
**Perfect Negative Correlation = -1**



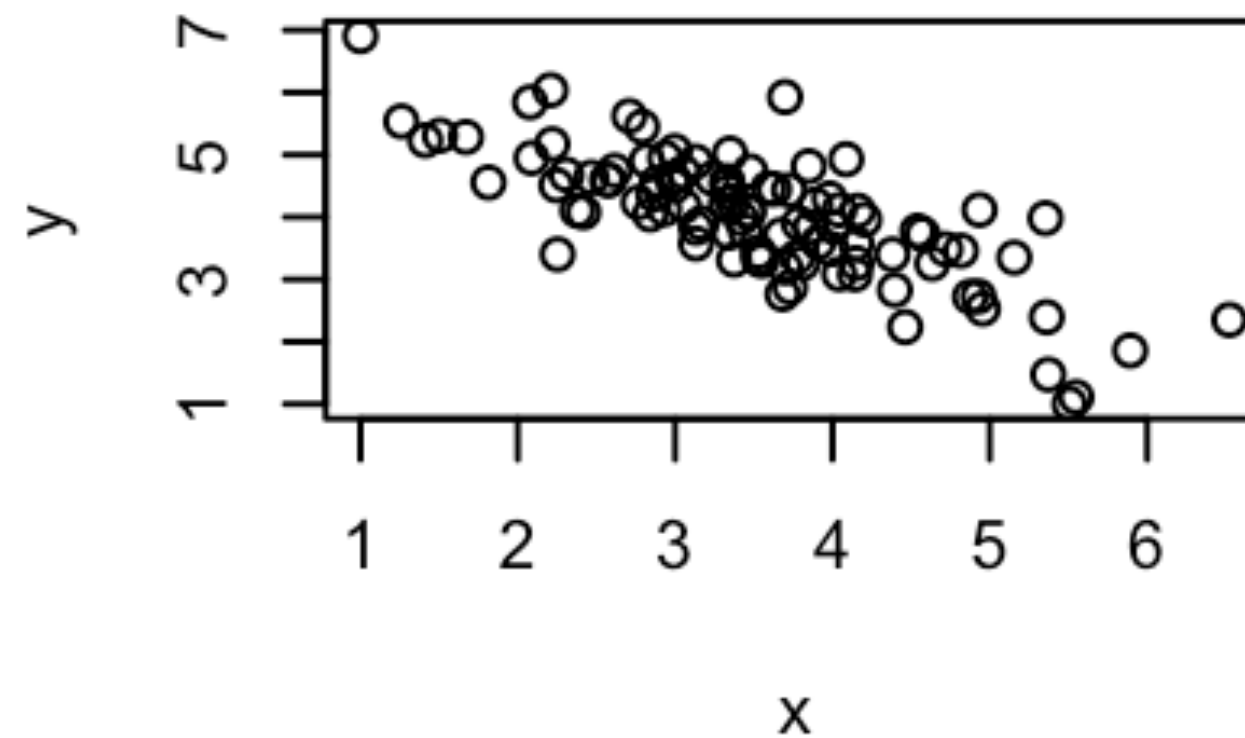
**Zero-Correlation = 0**



**Positive Correlation = 0.7**



**Negative Correlation = -0.7**





# Correlation & Causality

- Many times causality (the effect!) is the key interest in our studies
  - Which factors cause the disease?
  - **Predictor**= the factor that ***predicts*** the **outcome**
  - Some settings allow interpretations of causality: controlled trials where pre and post treatment measures, sometimes longitudinal data

**Which predictors to include in the model?**



# Which **predictors** to include in the model?

- Just the ones that are interesting (research question)?
- Everything that we could possibly get?
- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + uninteresting\* ones
  - Interest in the disorder effect that is independent of the effects of age and sex:  
disorder, age and sex

\*Many times called covariates. Btw the model treats them just like any other predictors

# Which **predictors** to include in the model?

- Just the ones that are interesting (research question)
- Everything that we can get? *'Blindly tossing variables into the causal salad is never a good idea.'*
- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + uninteresting\* ones
  - Interest in the disorder effect that is independent of the effects of age and sex:  
disorder, age and sex

\*Many times called covariates. Btw the model treats them just like any other predictors

# Which **predictors** to include in the model?

- Just the ones that are interesting (research question)?
- Everything that we could possibly get?
- Variables that are known to systematically explain variance in the outcome
  - Including the interesting + uninteresting\* ones
  - Interest in the disorder effect that is independent of the effects of age and sex:  
disorder, age and sex

\*Many times called covariates. Btw the model treats them just like any other predictors



# Which **predictors** to include in the model?

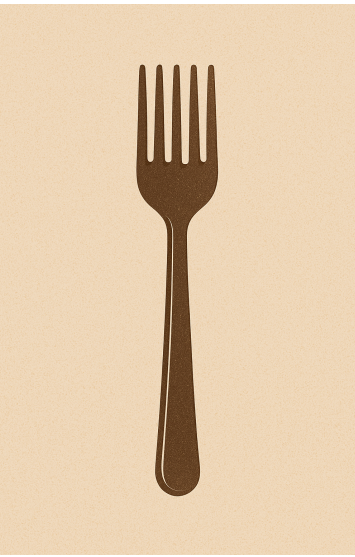
- Variables that are known to systematically explain variance in the outcome
- But with caution...
- Let's think of causality between our modeling variables **to avoid confounds**
  - Although many times difficult, we can try to interpret (or rule out) causality between our variables
    - Longitudinal data: Happiness today cannot cause happiness yesterday
    - Receptor availability cannot cause age

# Confounds

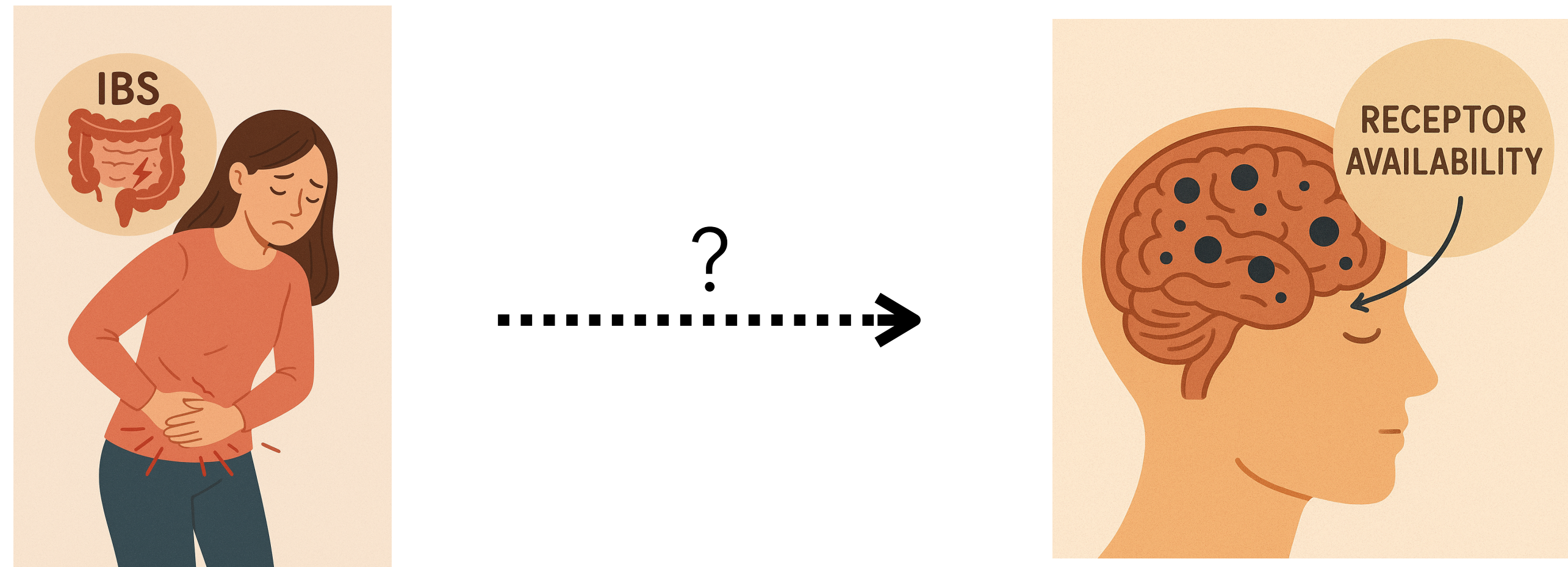
- I said: Let's think of causality between our modeling variables **to avoid confounds**. What are confounds?
- Features of our data and model that will **mislead us about the effects**
  - Produce false effects
  - Hide existing effects
- Basic confound types: Fork, pipe, collider, descendant



# Confounds: The Fork



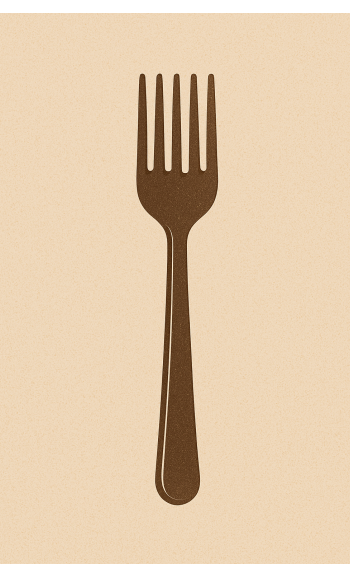
- What is the effect of IBS severity on receptor availability?



IBS= Irritated bowel syndrome



# Confounds: The Fork

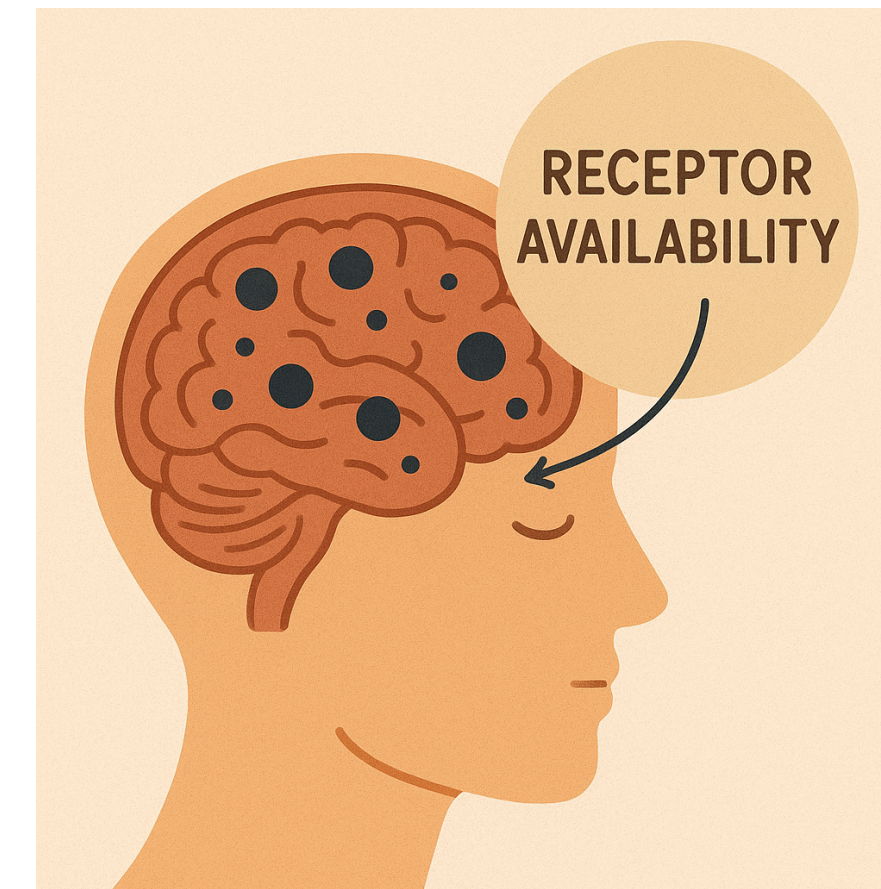
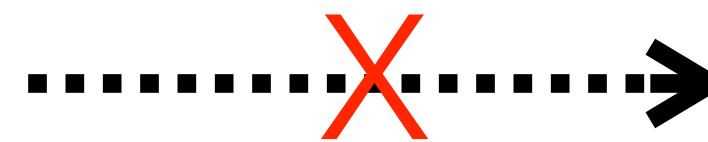


- Let's assume...

- Many vegan proteins, e.g. beans induce IBS symptoms in vulnerable individuals



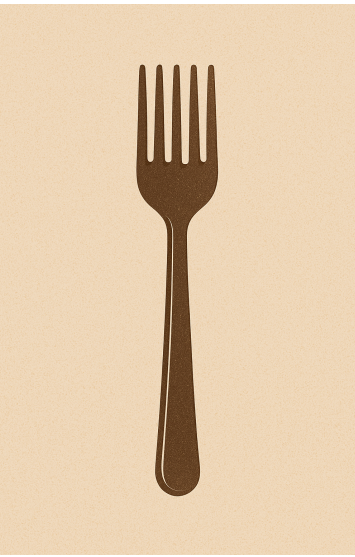
- Many vegan products are healthy for the brain



Directed acyclic graph (DAG), arrows point the direction of the causality



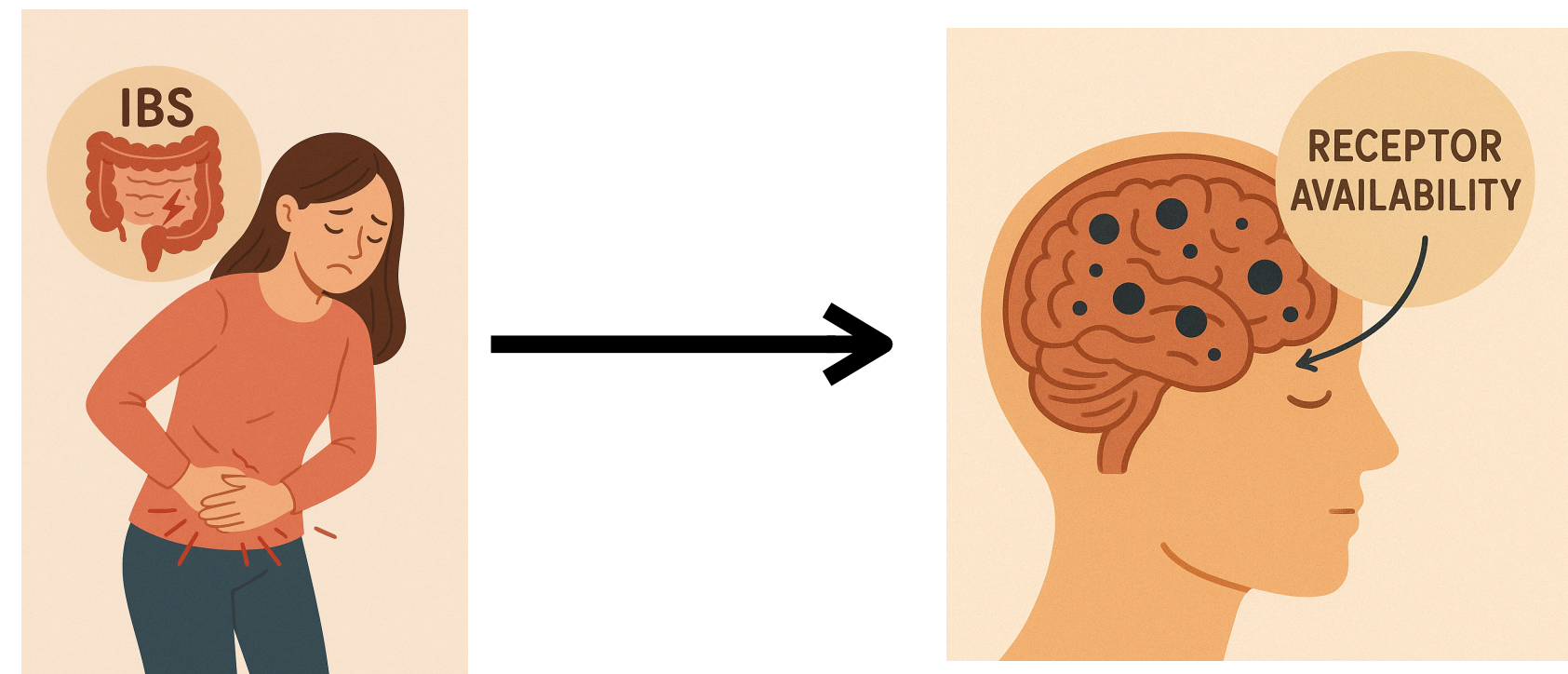
# Confounds: The Fork



- What is the effect of IBS severity on receptor availability?

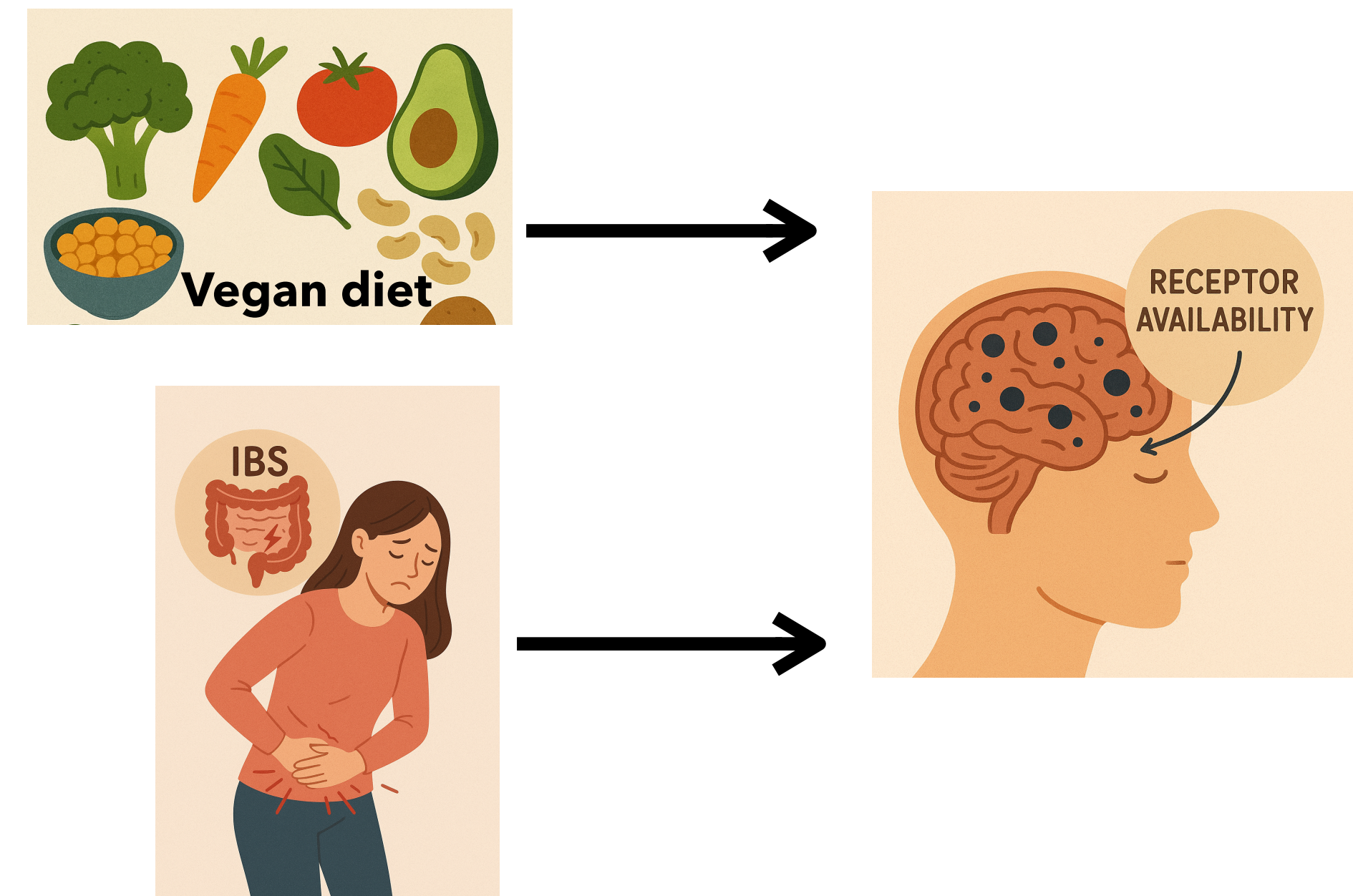
## MODEL 1

Receptor availability  $\sim$  IBS severity



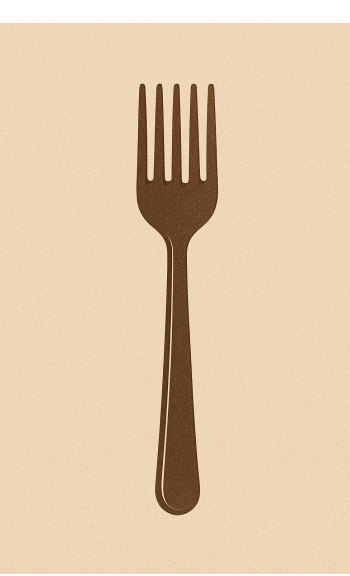
## MODEL 2

Receptor availability  $\sim$  IBS severity + vegan diet





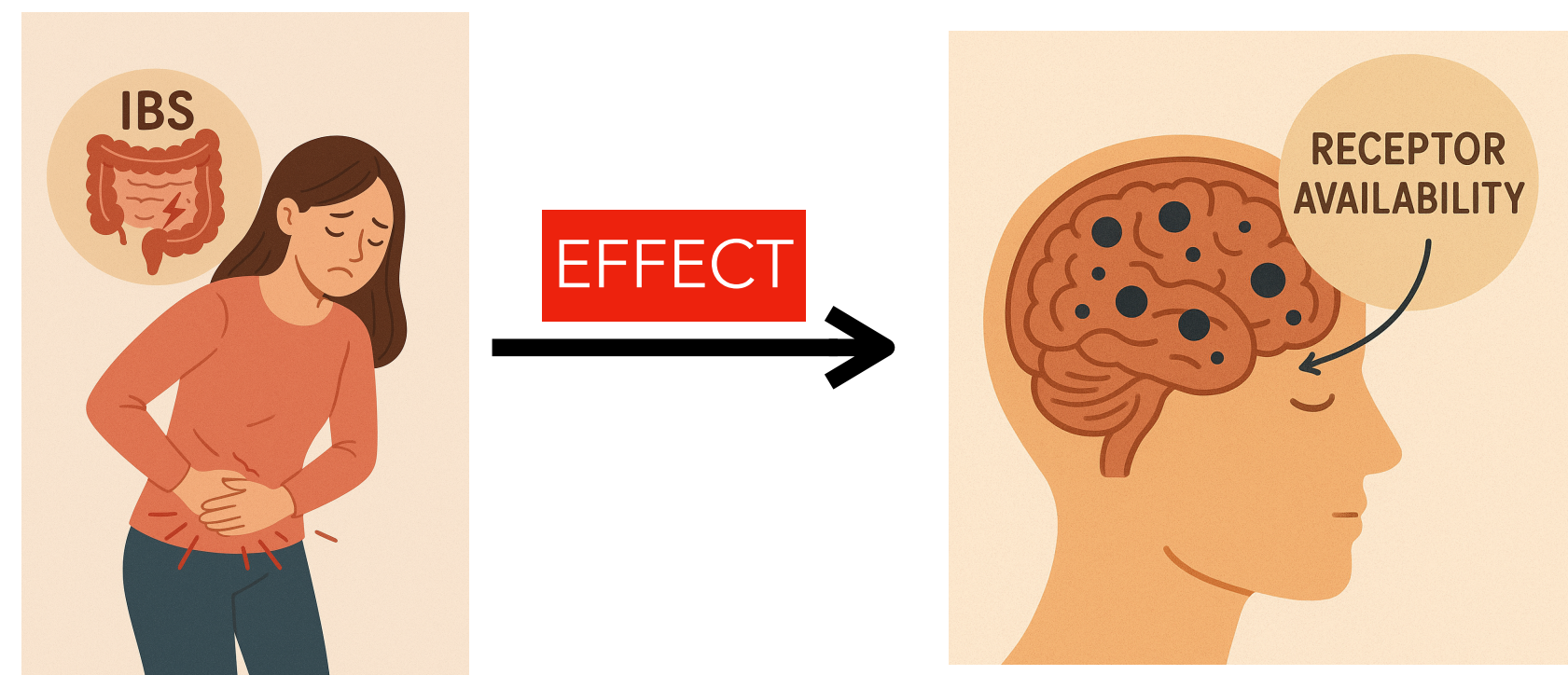
# Confounds: The Fork



- What is the effect of IBS severity on receptor availability?

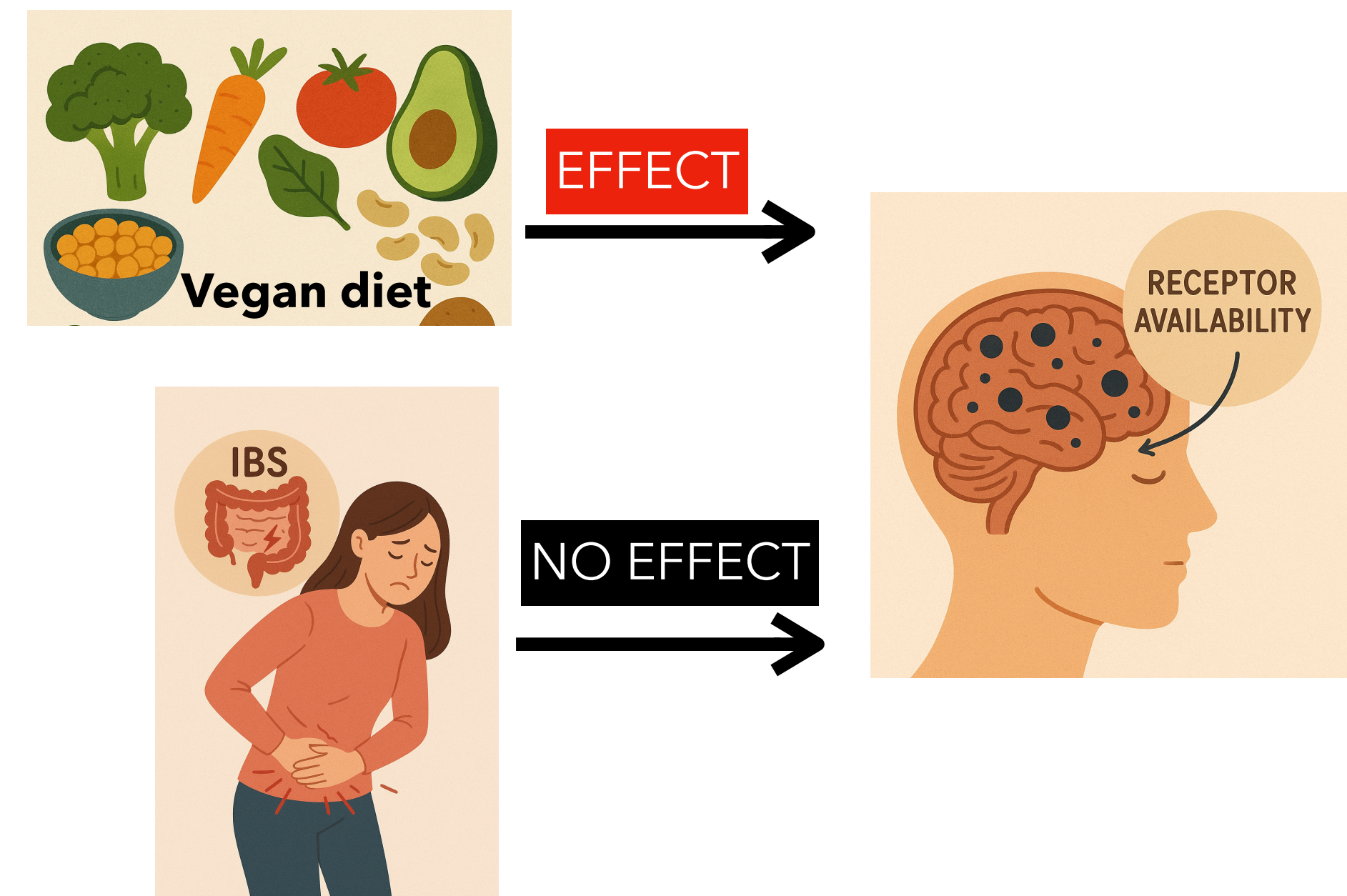
## MODEL 1

Receptor availability  $\sim$  IBS severity



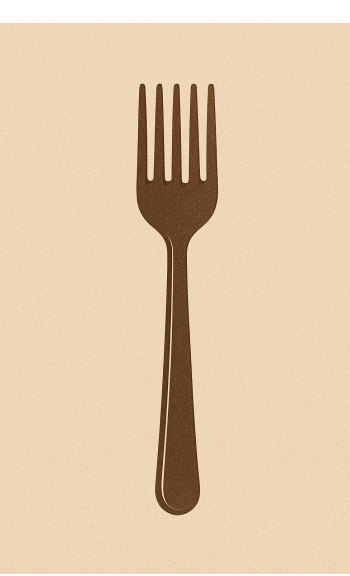
## MODEL 2

Receptor availability  $\sim$  IBS severity + vegan diet





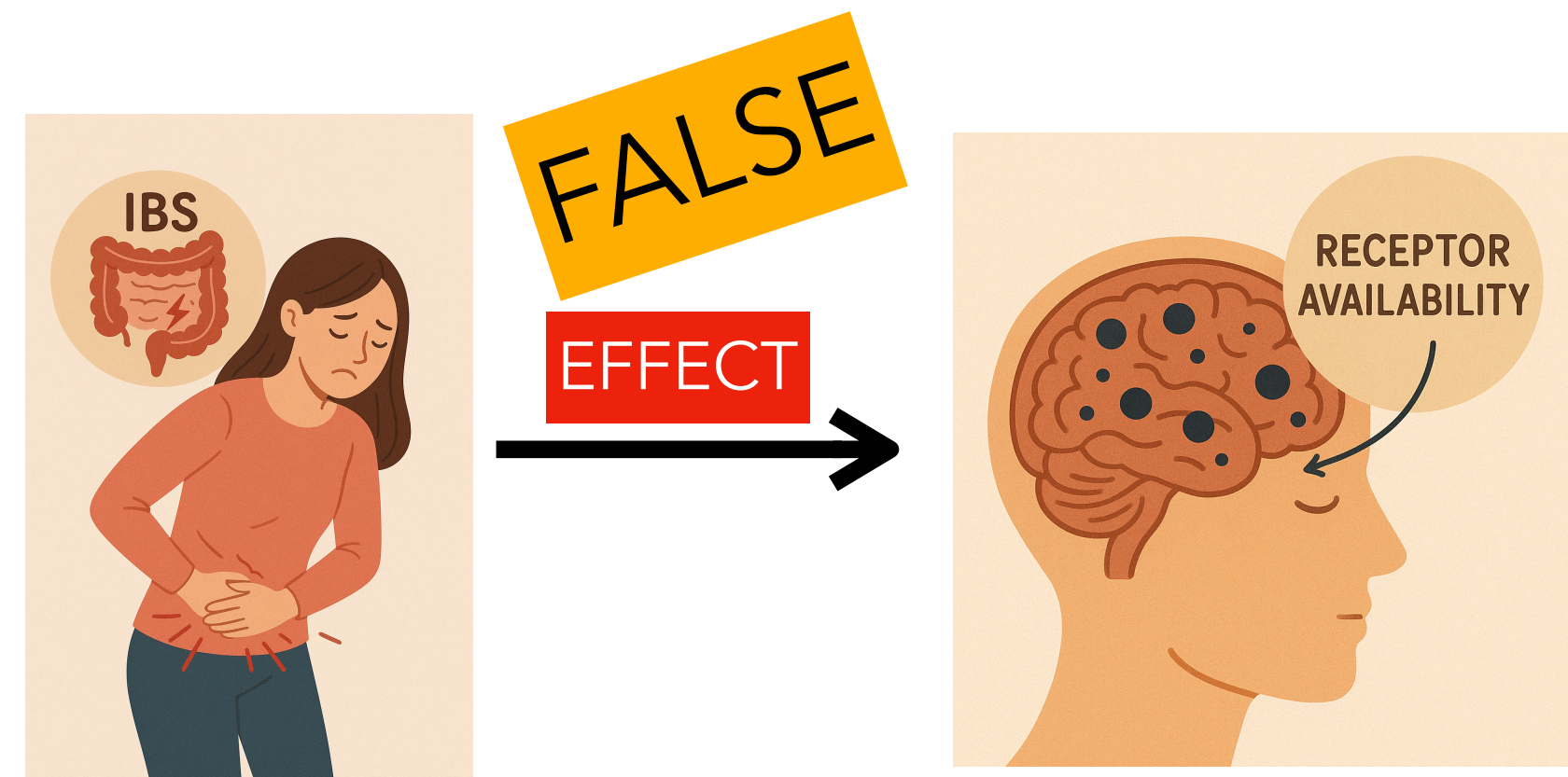
# Confounds: The Fork



- What is the effect of IBS severity on receptor availability?

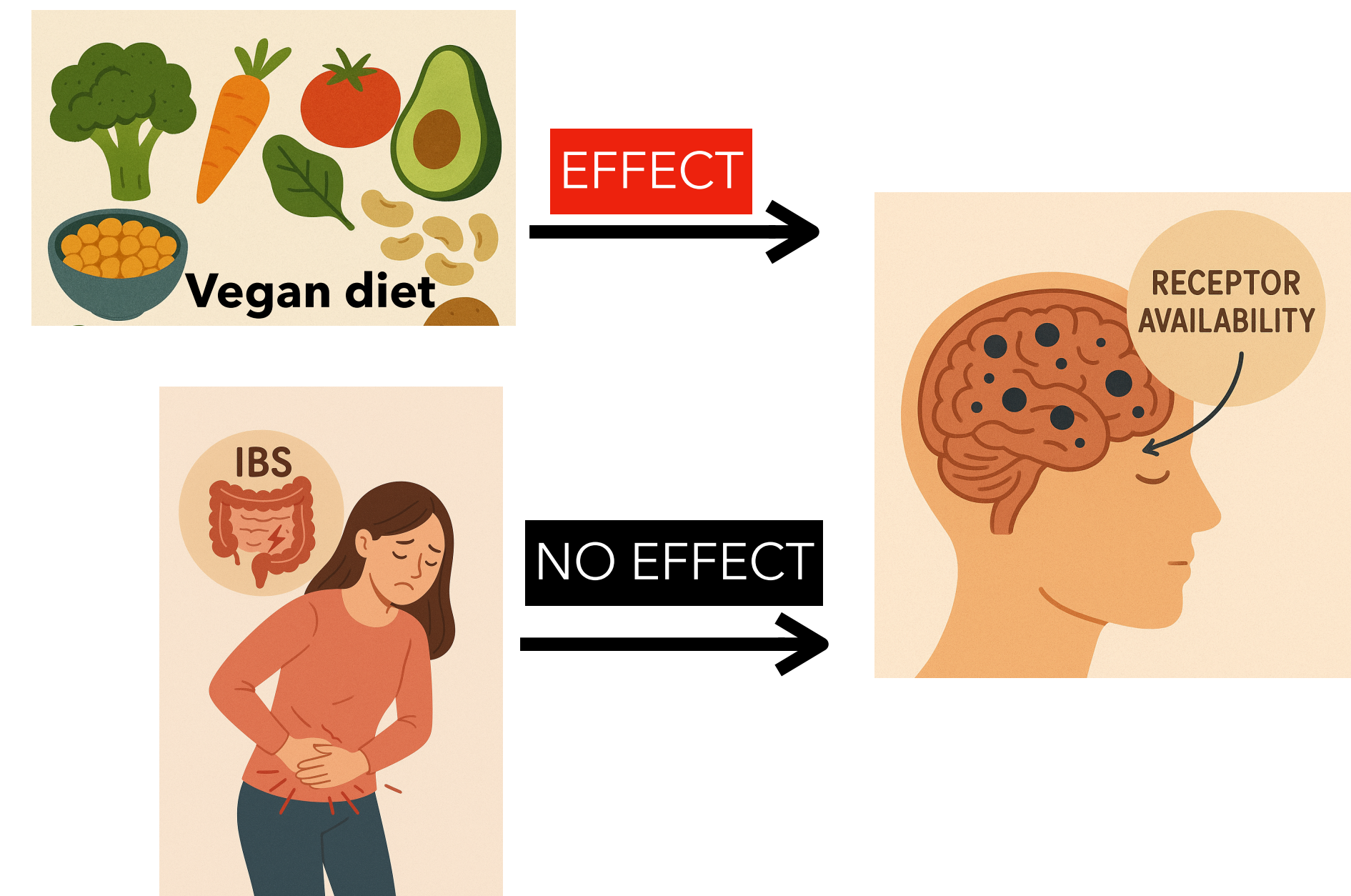
## MODEL 1

Receptor availability  $\sim$  IBS severity



## MODEL 2

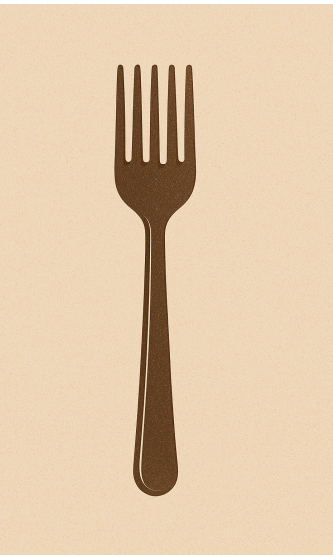
Receptor availability  $\sim$  IBS severity + vegan diet



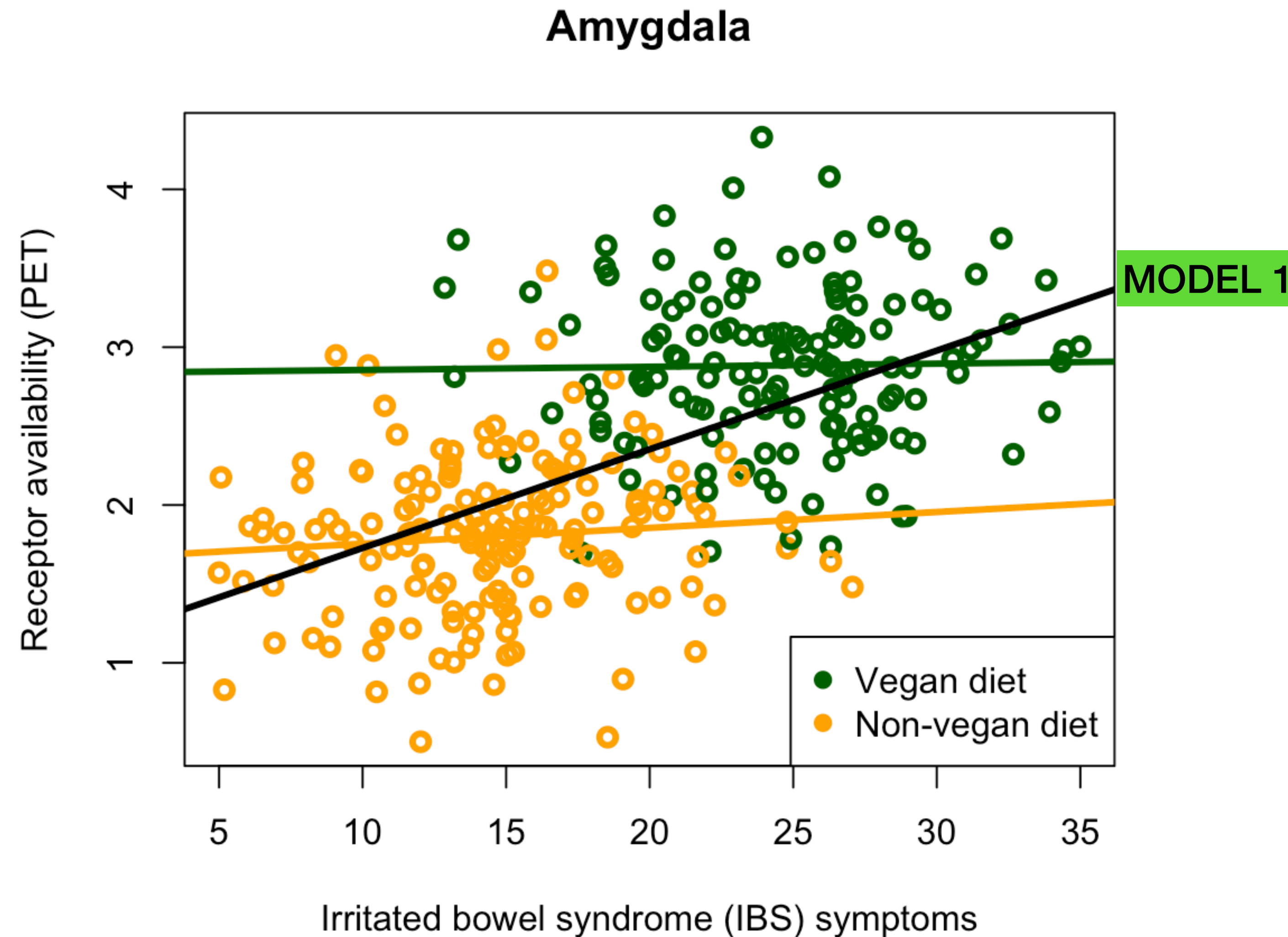
- The vegan effect misinterpreted as the IBS effect



# Confounds: The Fork



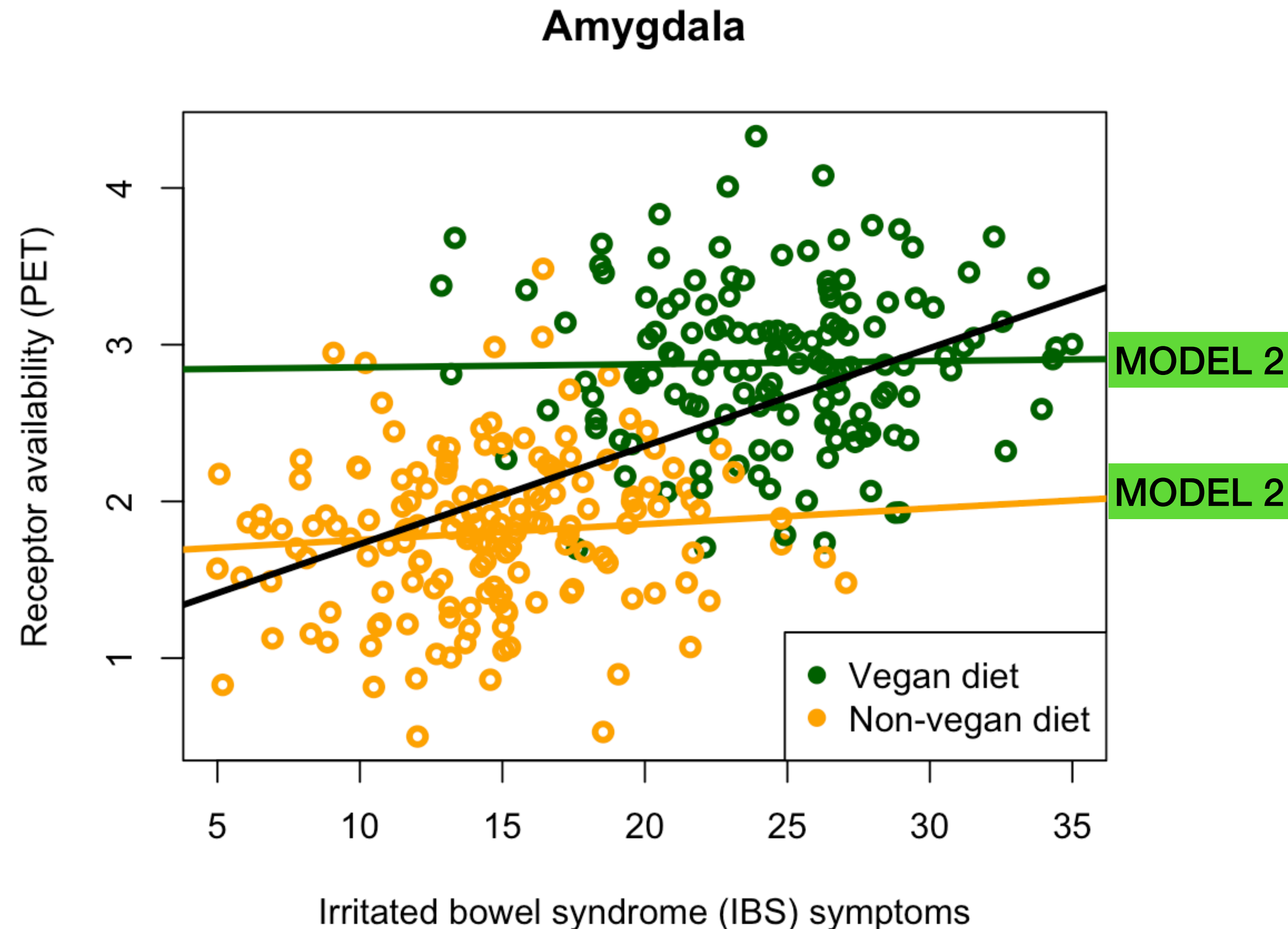
- What is the effect of IBS severity on receptor availability?



# Confounds: The Fork

Disclaimer: These are not the regression lines from the model (we do not get IBS effect separately for vegans and non-vegans from main effects model), but the model is estimating the effect of IBS when diet is constant. So this is describing the characteristics of the data that induces the confound! (Considers also the upcoming examples)

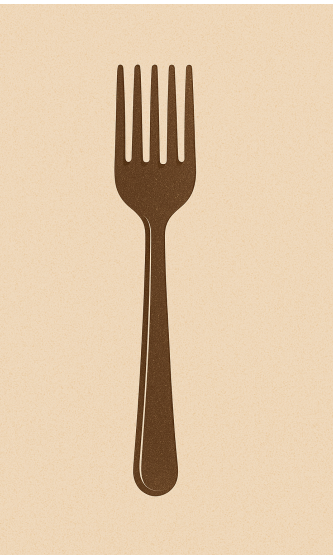
- What is the effect of IBS severity on receptor availability?





# Confounds: The Fork

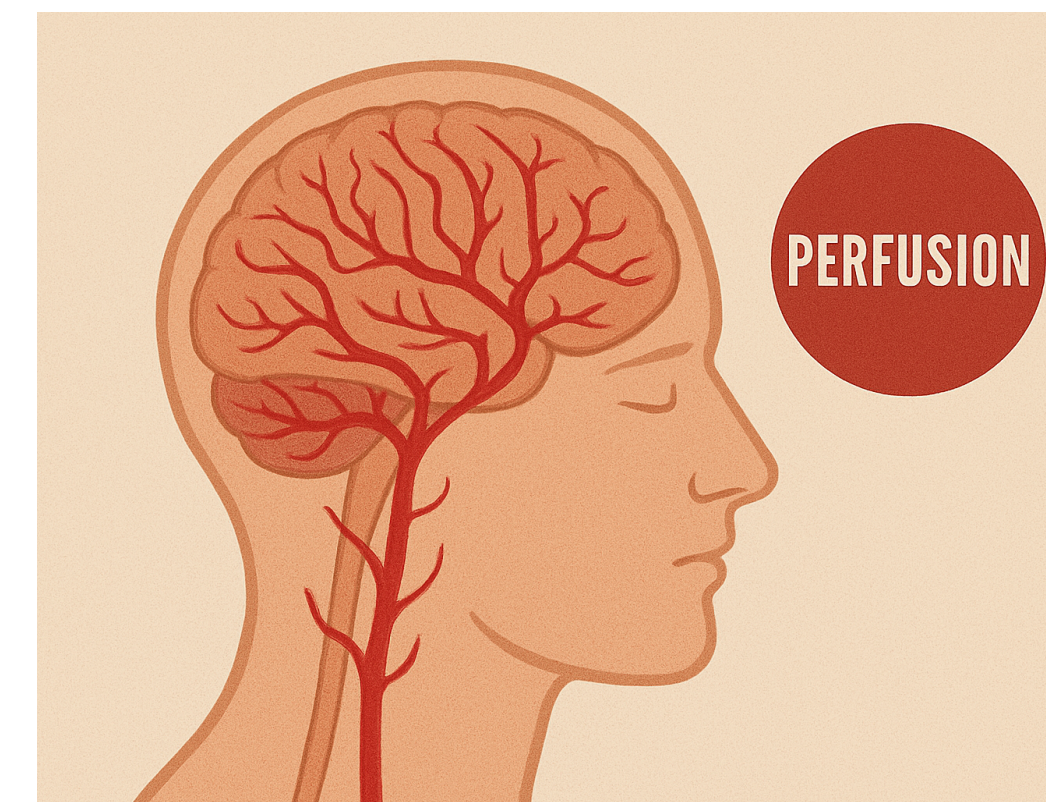
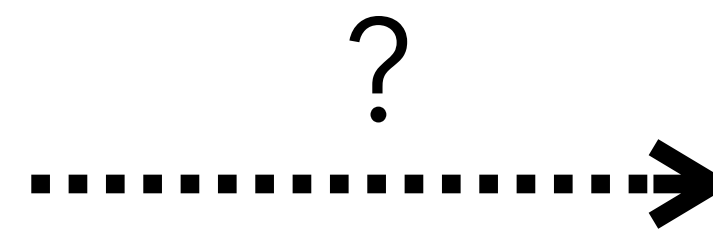
## The general rule of thumb



- If a certain variable affects both the interesting predictor and the outcome (vegan diet), its effect should be 'cleaned' from the data by adding it as a predictor

# Confounds: The Pipe

- What is the effect of cold exposure on brain perfusion?

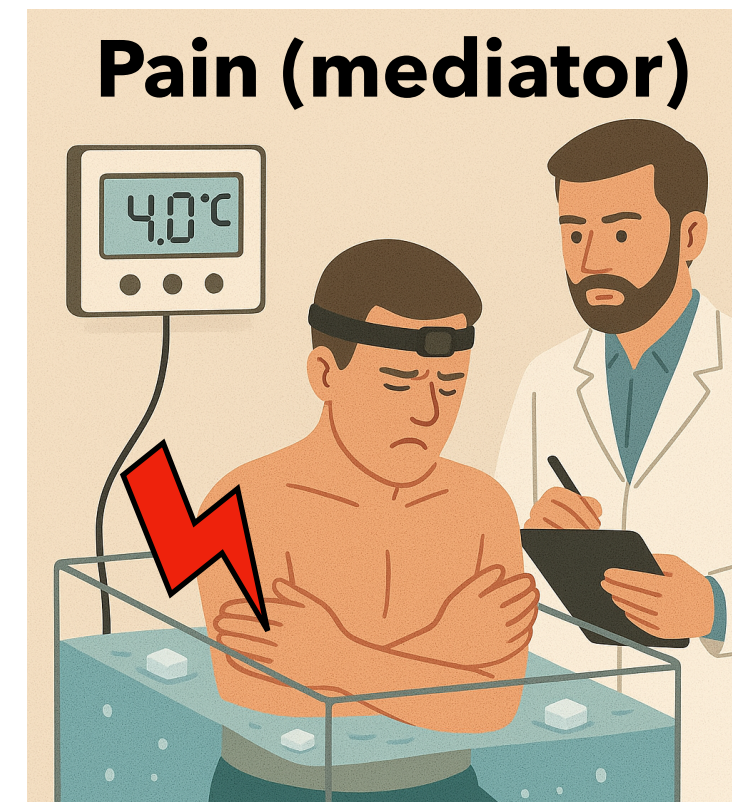




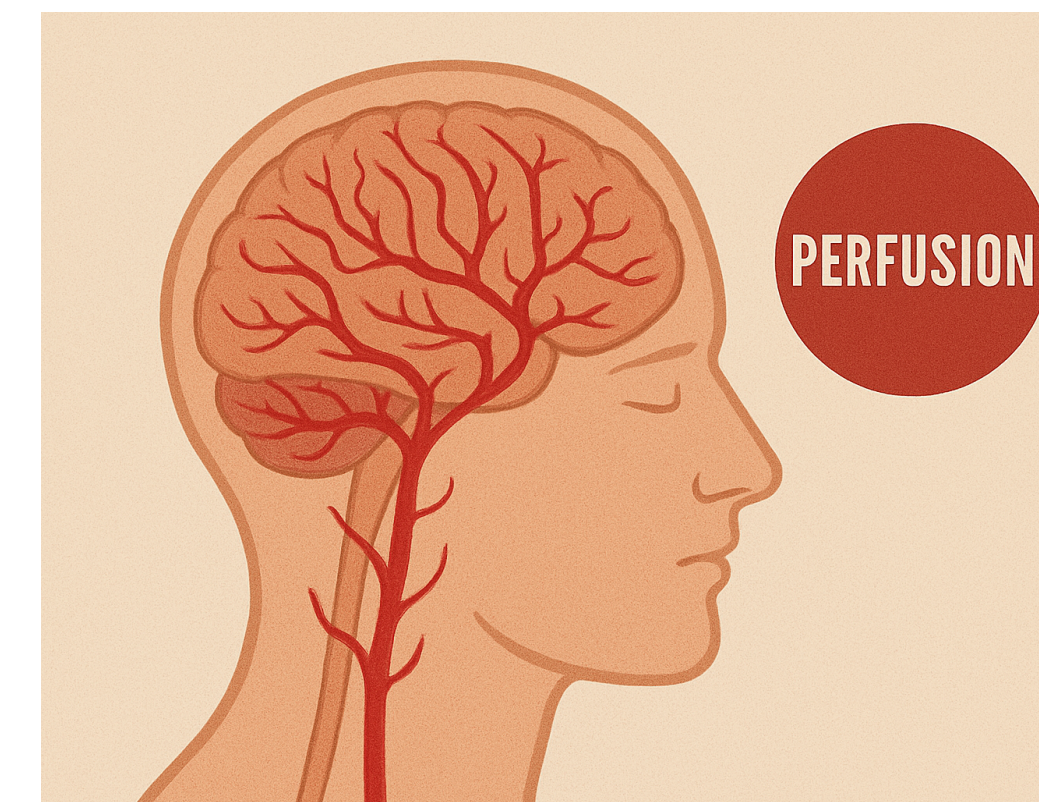
# Confounds: The Pipe



- Let's assume...
- The colder the water, the more likely the individual experiences pain



- Pain increases perfusion





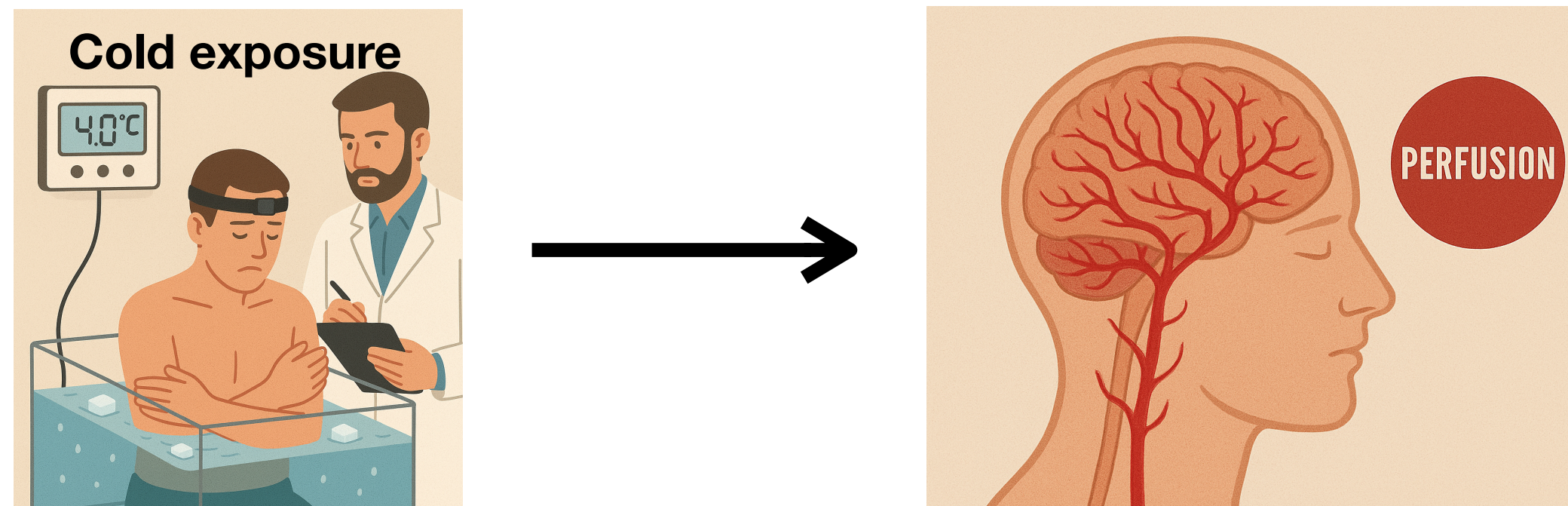
# Confounds: The Pipe

- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?



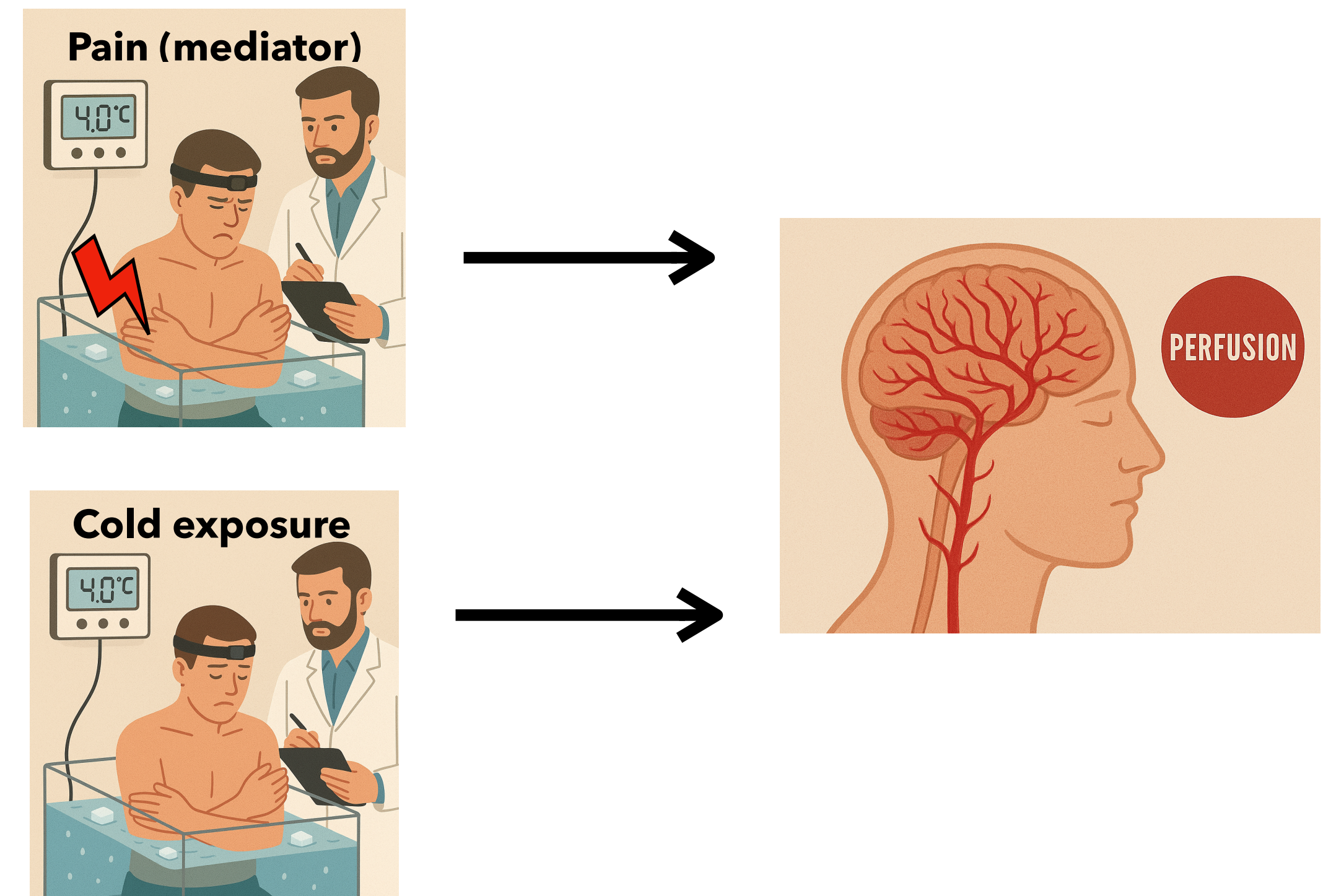
## MODEL 1

Perfusion ~ cold exposure



## MODEL 2

Perfusion ~ cold exposure + pain





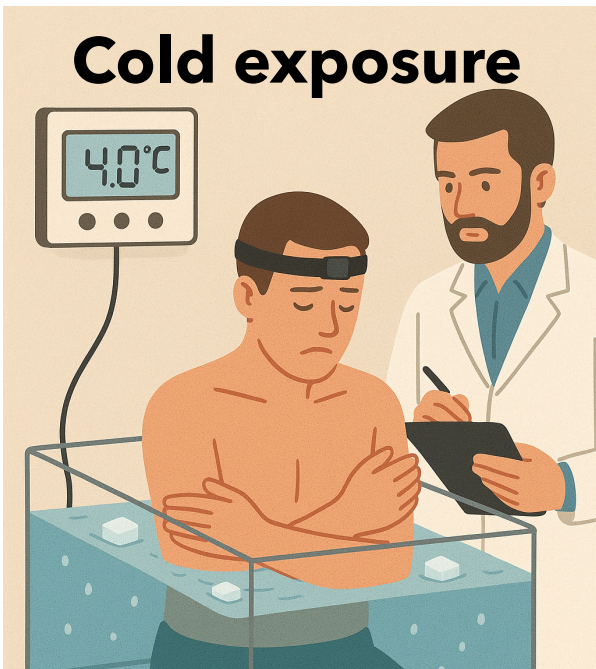
# Confounds: The Pipe

- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?

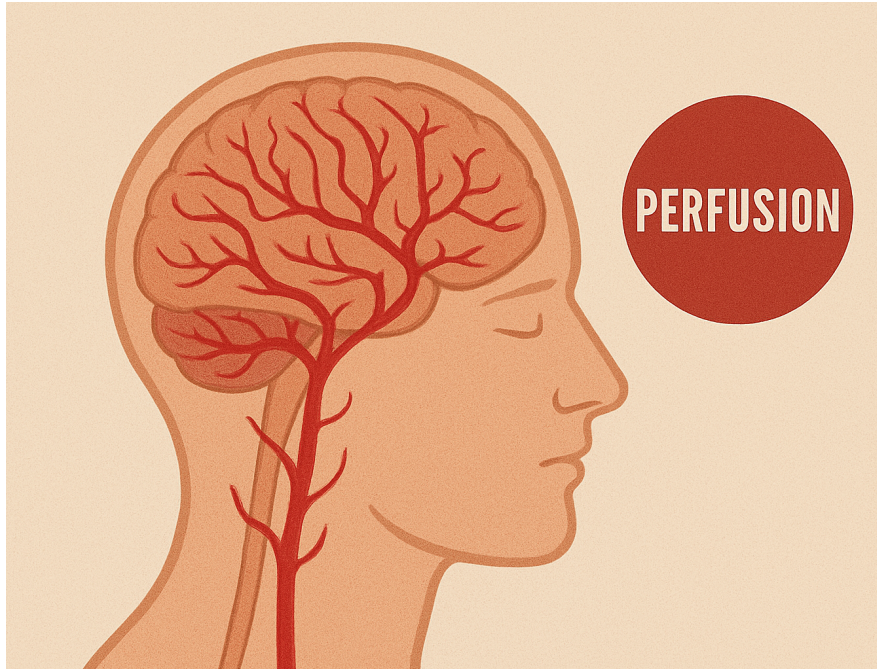


## MODEL 1

Perfusion ~ cold exposure



EFFECT

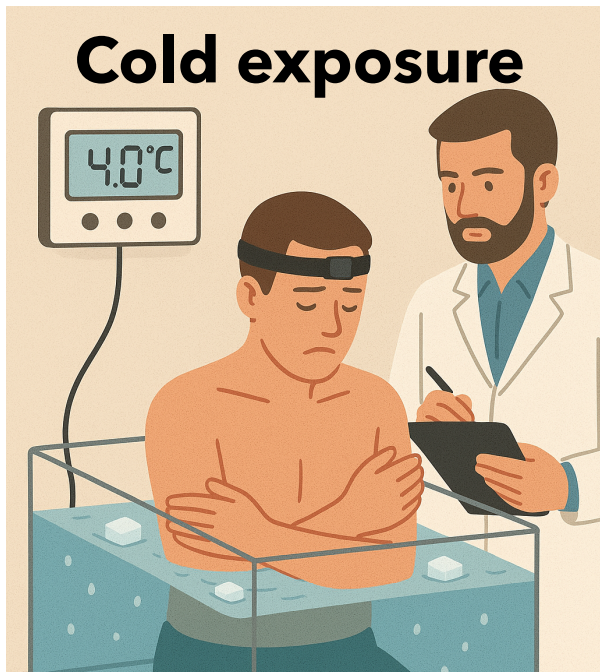
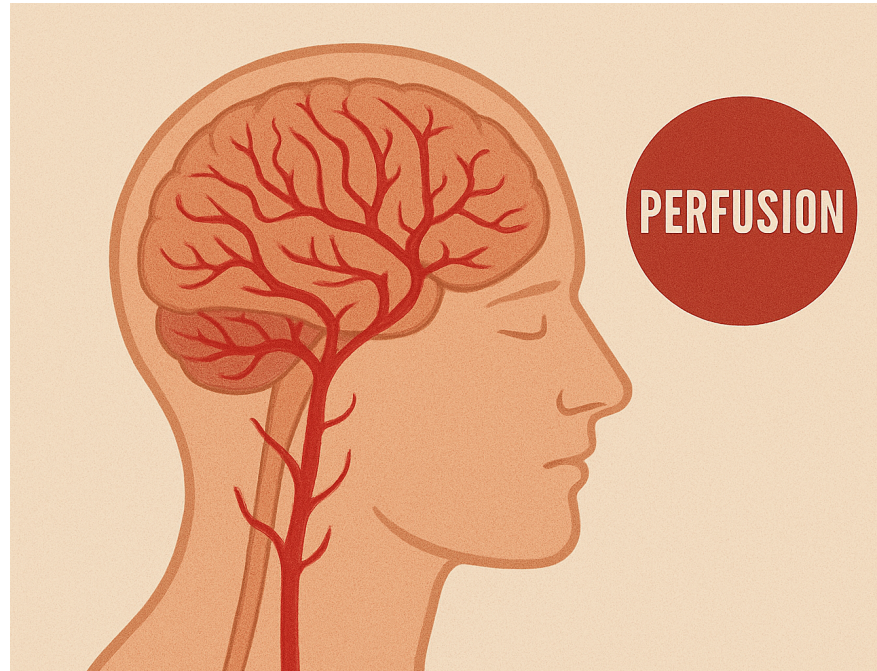


## MODEL 2

Perfusion ~ cold exposure + pain



EFFECT



NO EFFECT



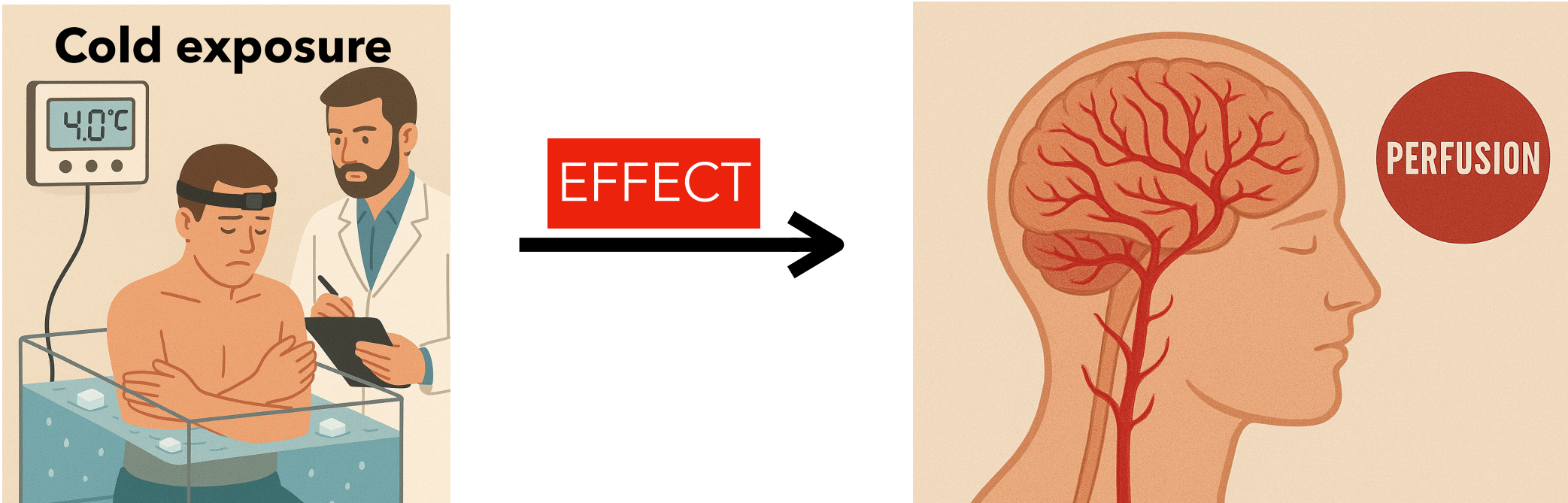
# Confounds: The Pipe



- What is the **total effect of cold exposure** (including the pain-mediated effect) on brain perfusion?

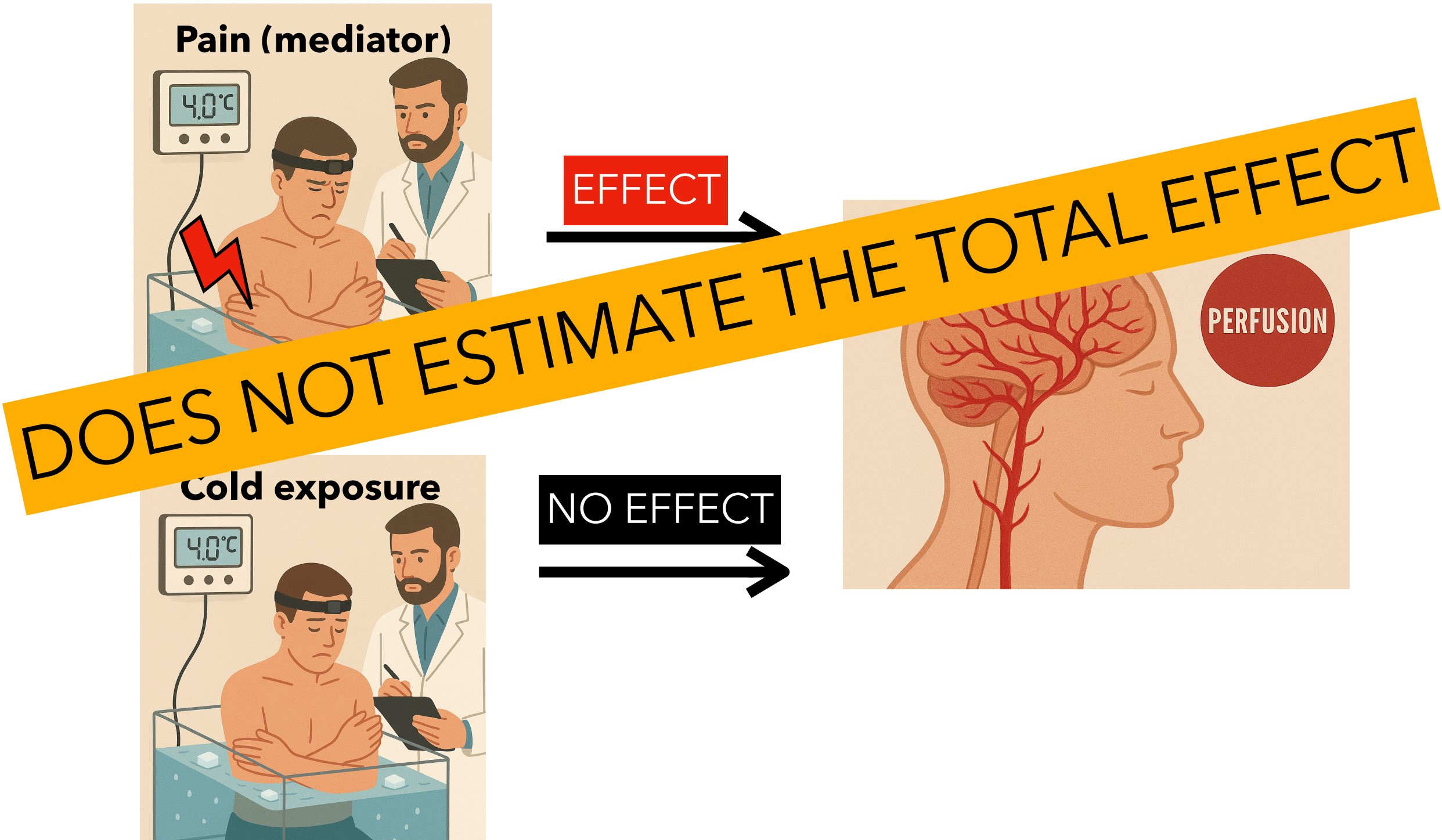
## MODEL 1

Perfusion ~ cold exposure



## MODEL 2

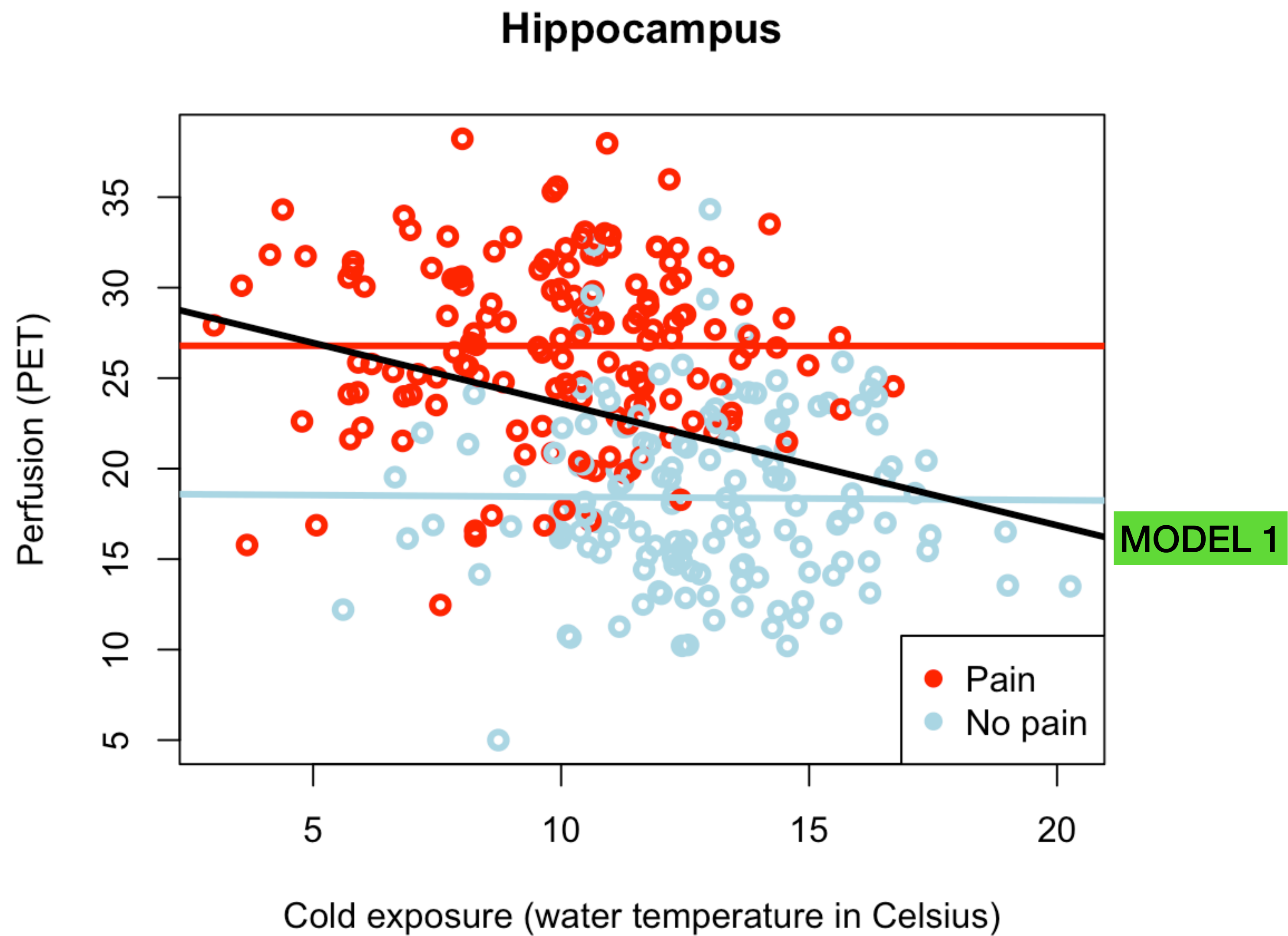
Perfusion ~ cold exposure + pain



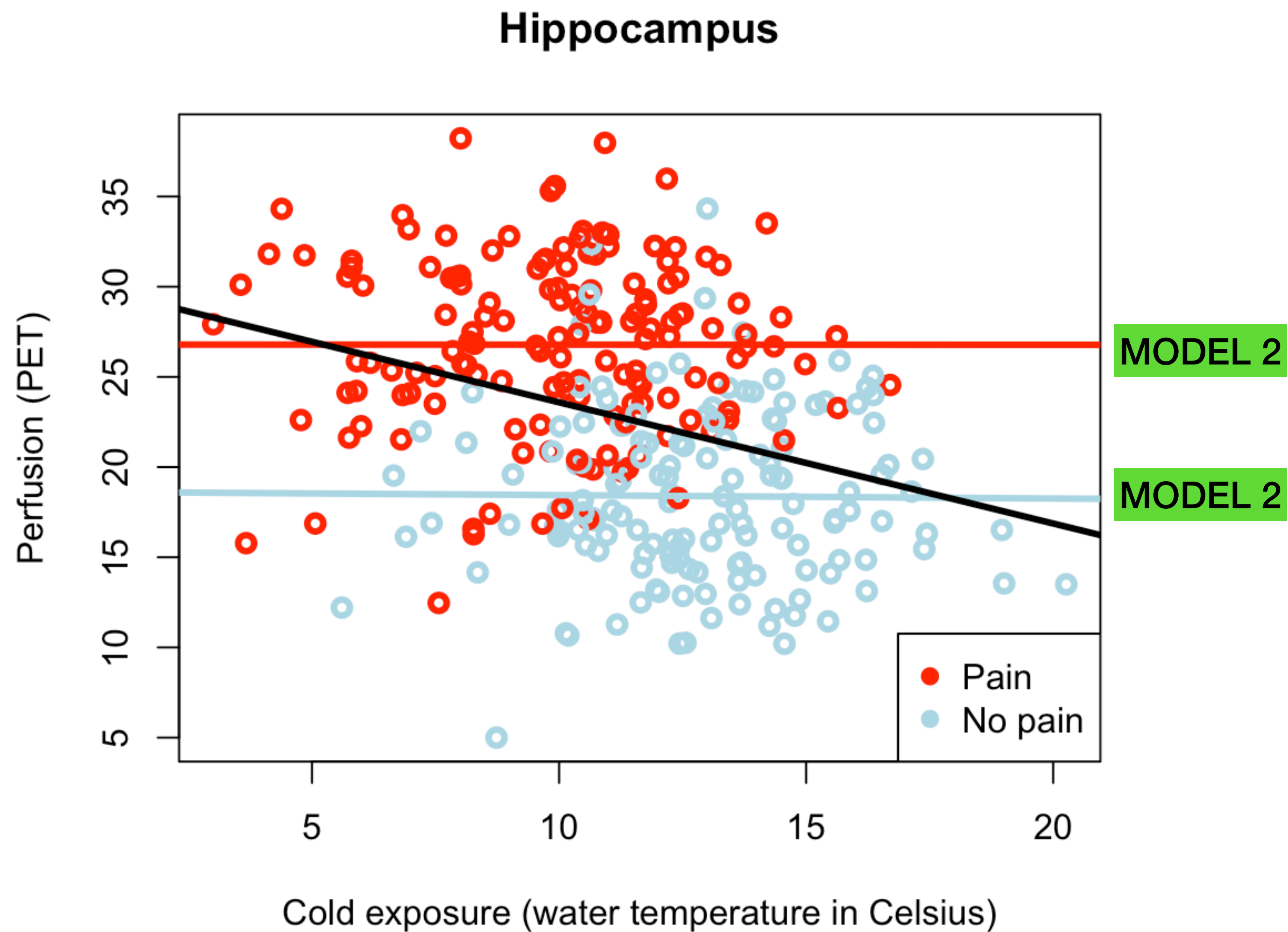
- Include pain if interested in the pain-independent effect of cold-exposure



# Confounds: The Pipe



# Confounds: The Pipe





# Confounds: The Pipe

## The general rule of thumb

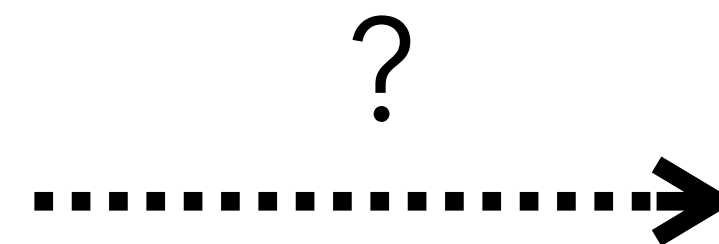
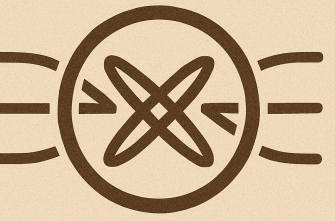


- **Let's not include treatment consequences** as predictors if interested in the total treatment effect
- Drug reduces heart rate that lowers anxiety
  - Anxiety  $\sim$  drug + heart rate
    - The drug doesn't work (bad conclusion, post-treatment bias)



# Confounds: The Collider

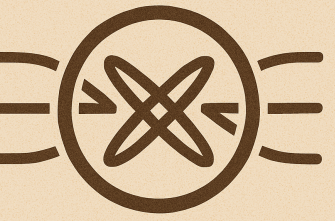
- Diathesis-stress model: genetic vulnerability + stressors = disorder onset
- What is the effect of genetic vulnerability on environmental stressors?



PG = pathological gambling



# Confounds: The Collider



- Let's assume...



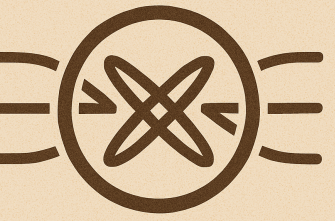
- Both genetic vulnerability and environmental stressors increase the likelihood of pathological gambling



PG = pathological gambling



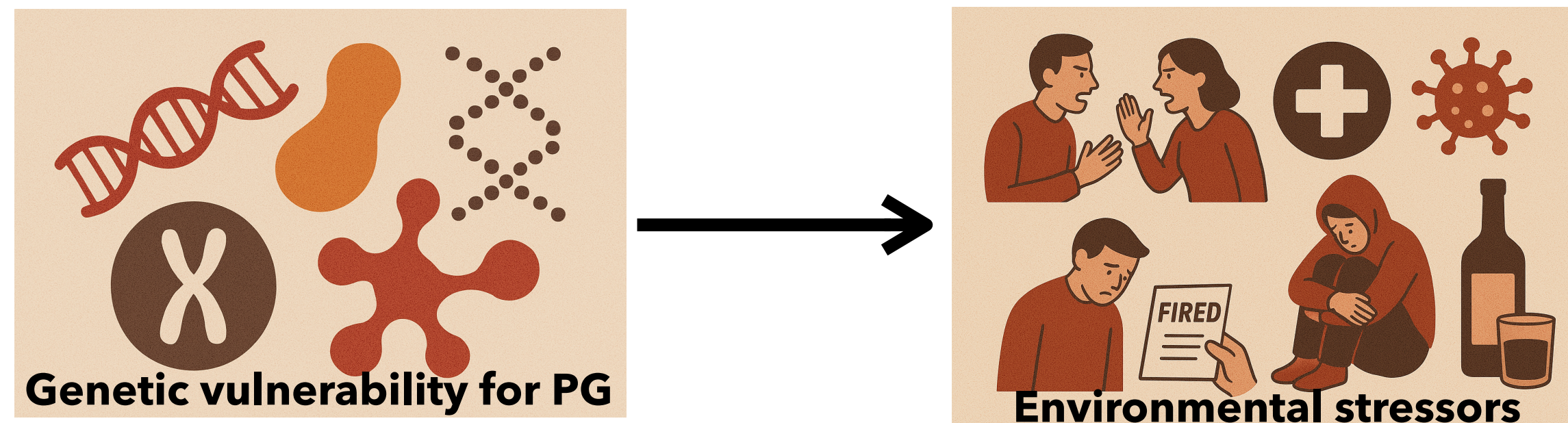
# Confounds: The Collider



- What is the effect of genetic vulnerability on environmental stressors?

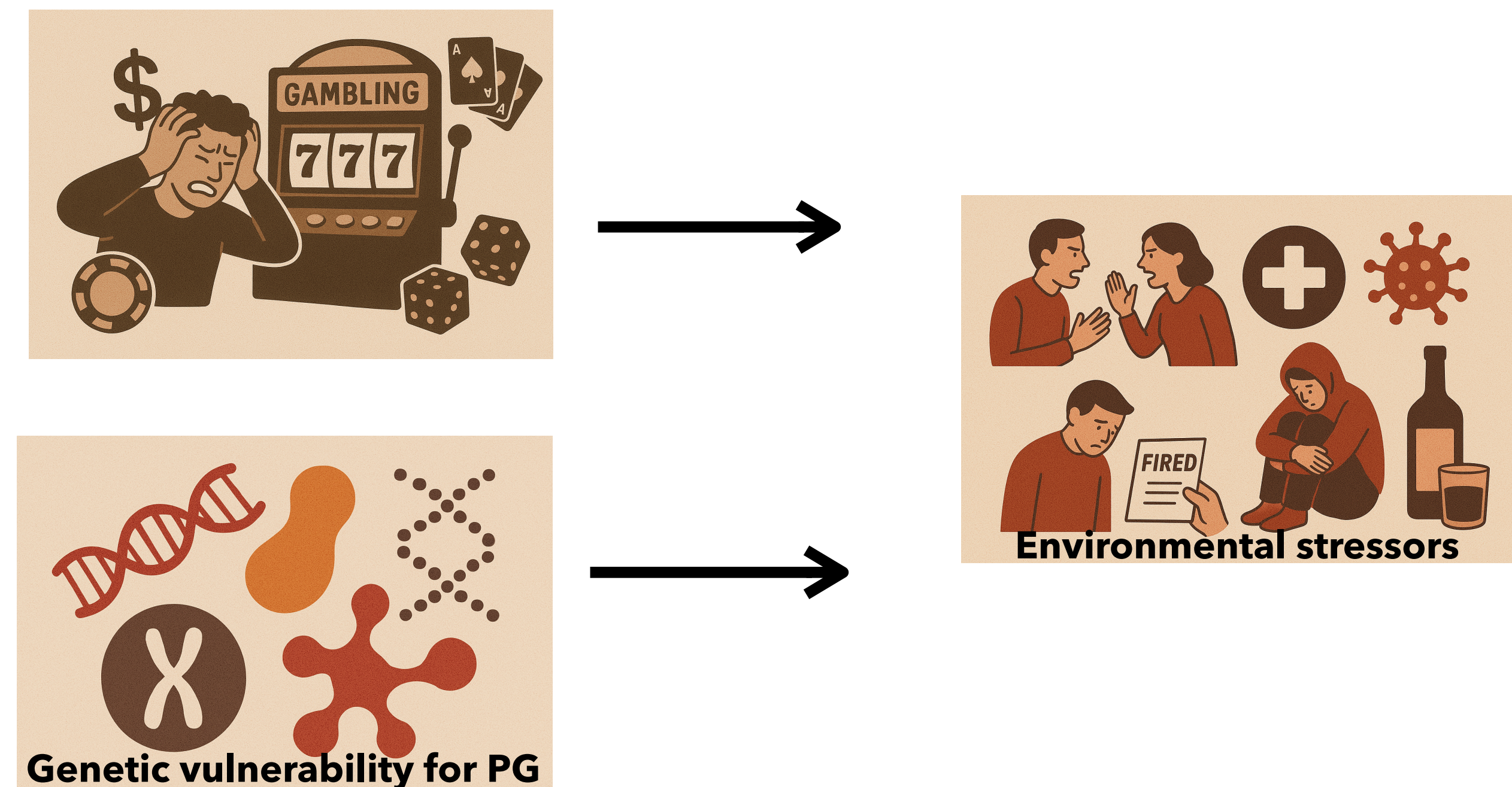
## MODEL 1

Environmental stressors ~ genetic vulnerability



## MODEL 2

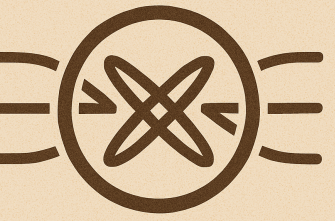
Environmental stressors ~ genetic vulnerability + clinical status



PG = pathological gambling



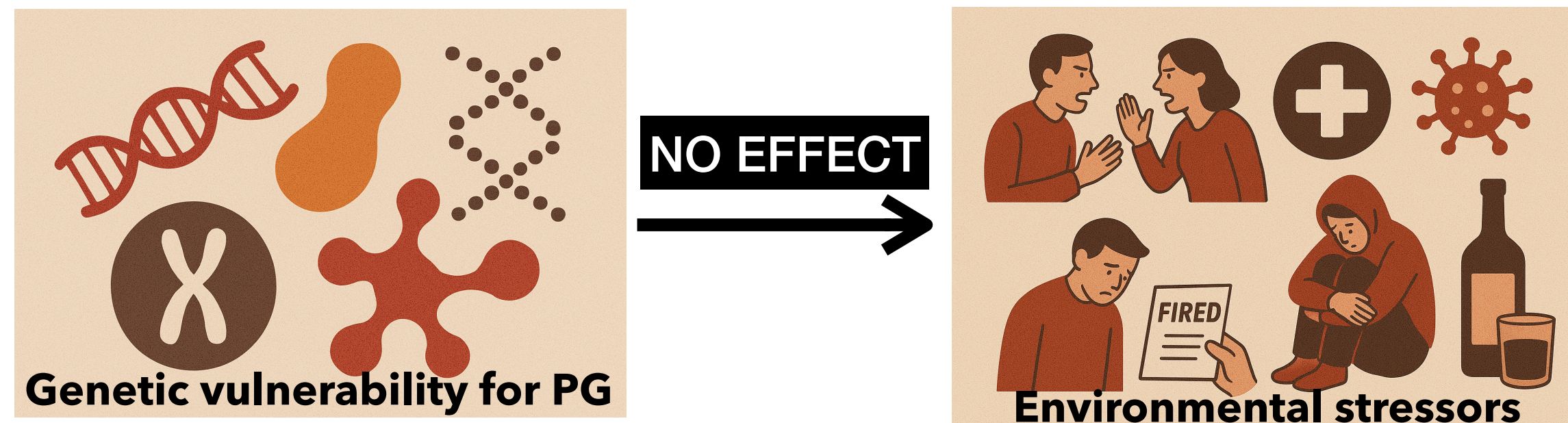
# Confounds: The Collider



- What is the effect of genetic vulnerability on environmental stressors?

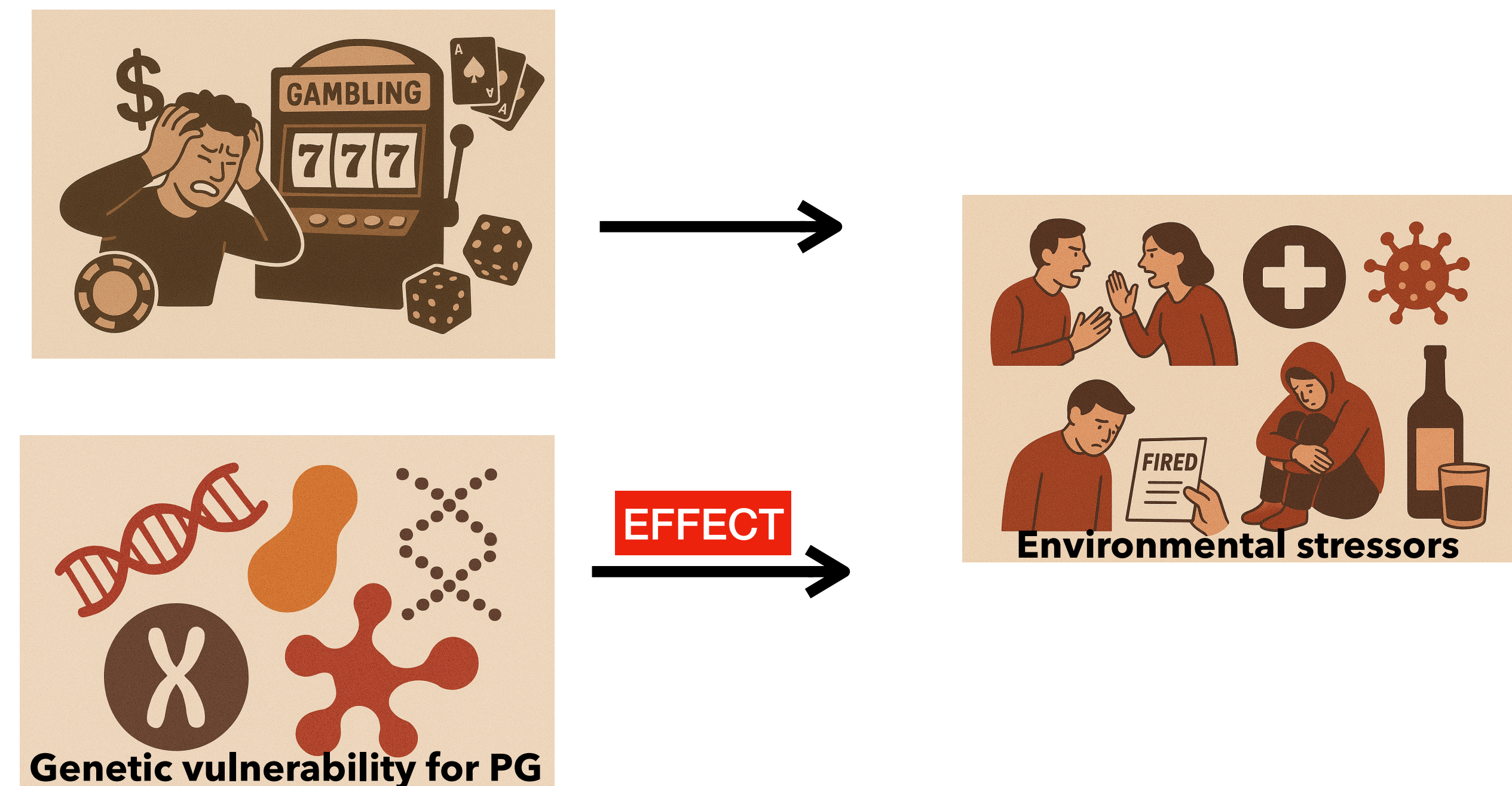
## MODEL 1

Environmental stressors ~ genetic vulnerability



## MODEL 2

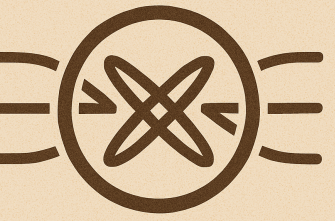
Environmental stressors ~ genetic vulnerability + clinical status



PG = pathological gambling



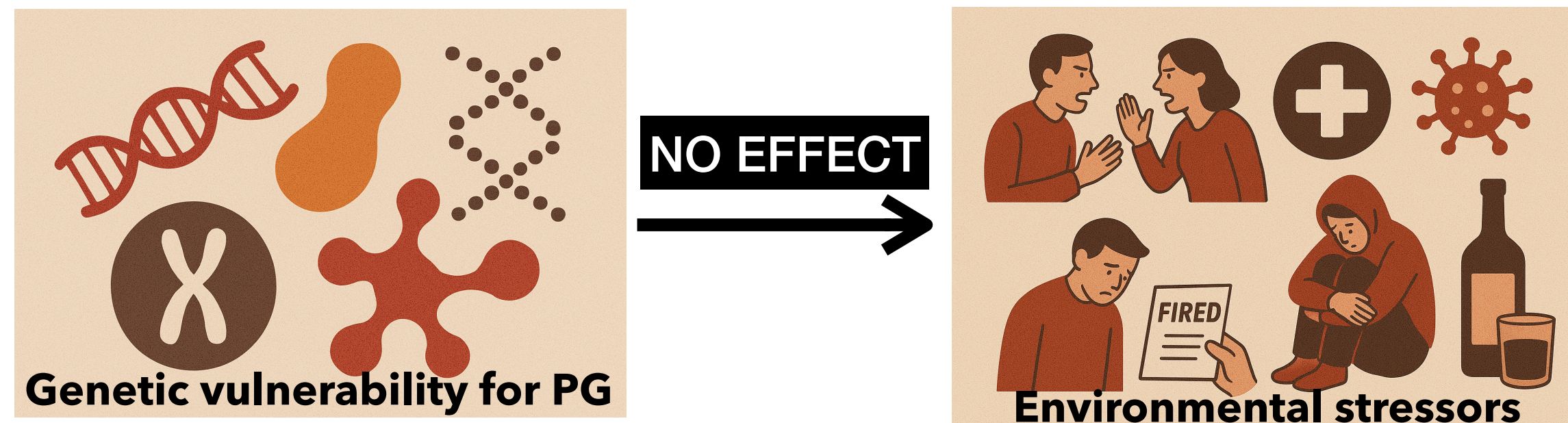
# Confounds: The Collider



- What is the effect of genetic vulnerability on environmental stressors?

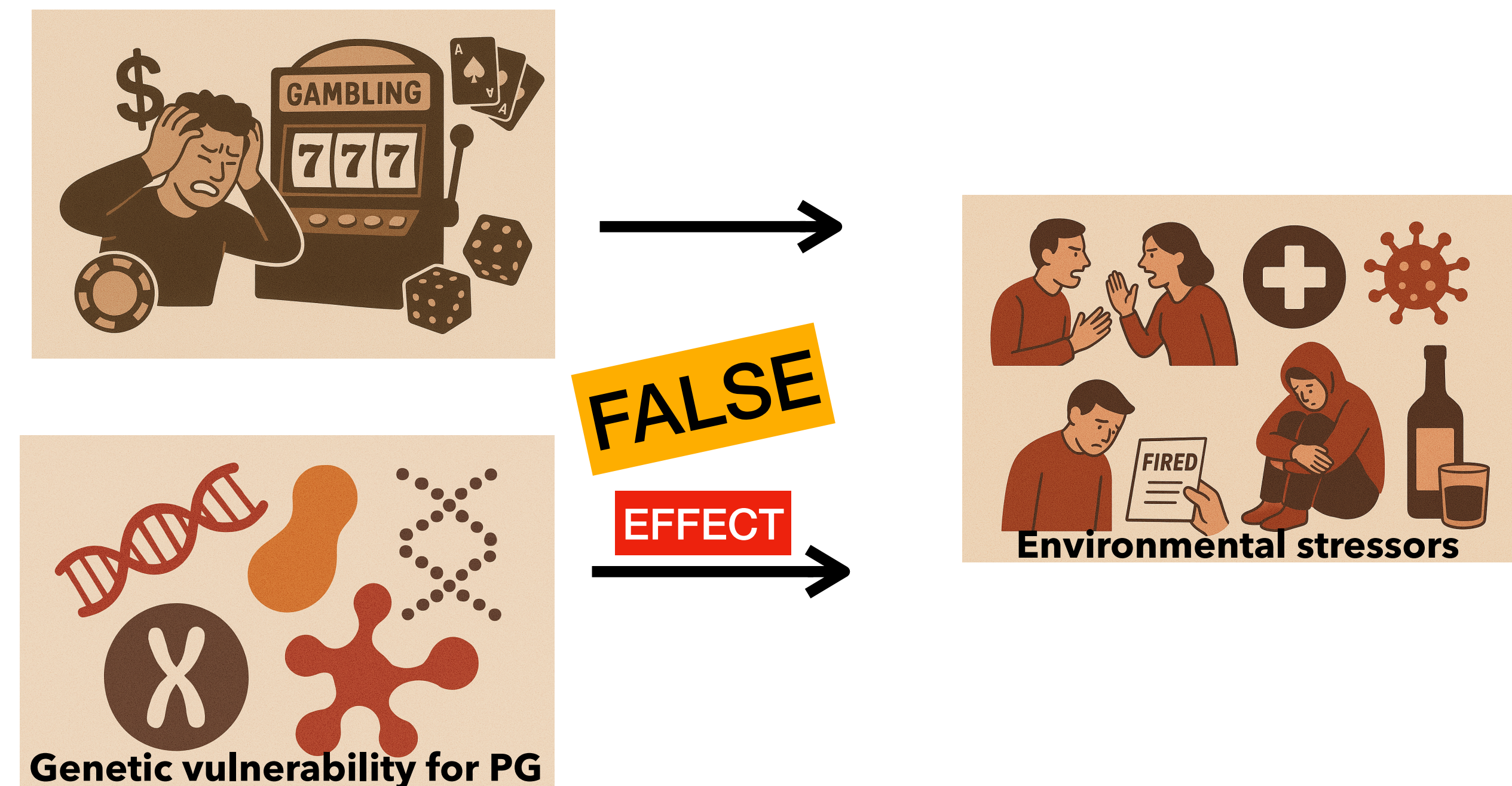
## MODEL 1

Environmental stressors ~ genetic vulnerability



## MODEL 2

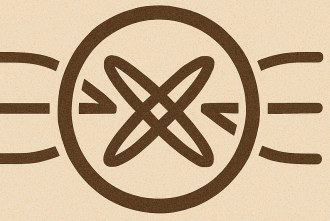
Environmental stressors ~ genetic vulnerability + clinical status



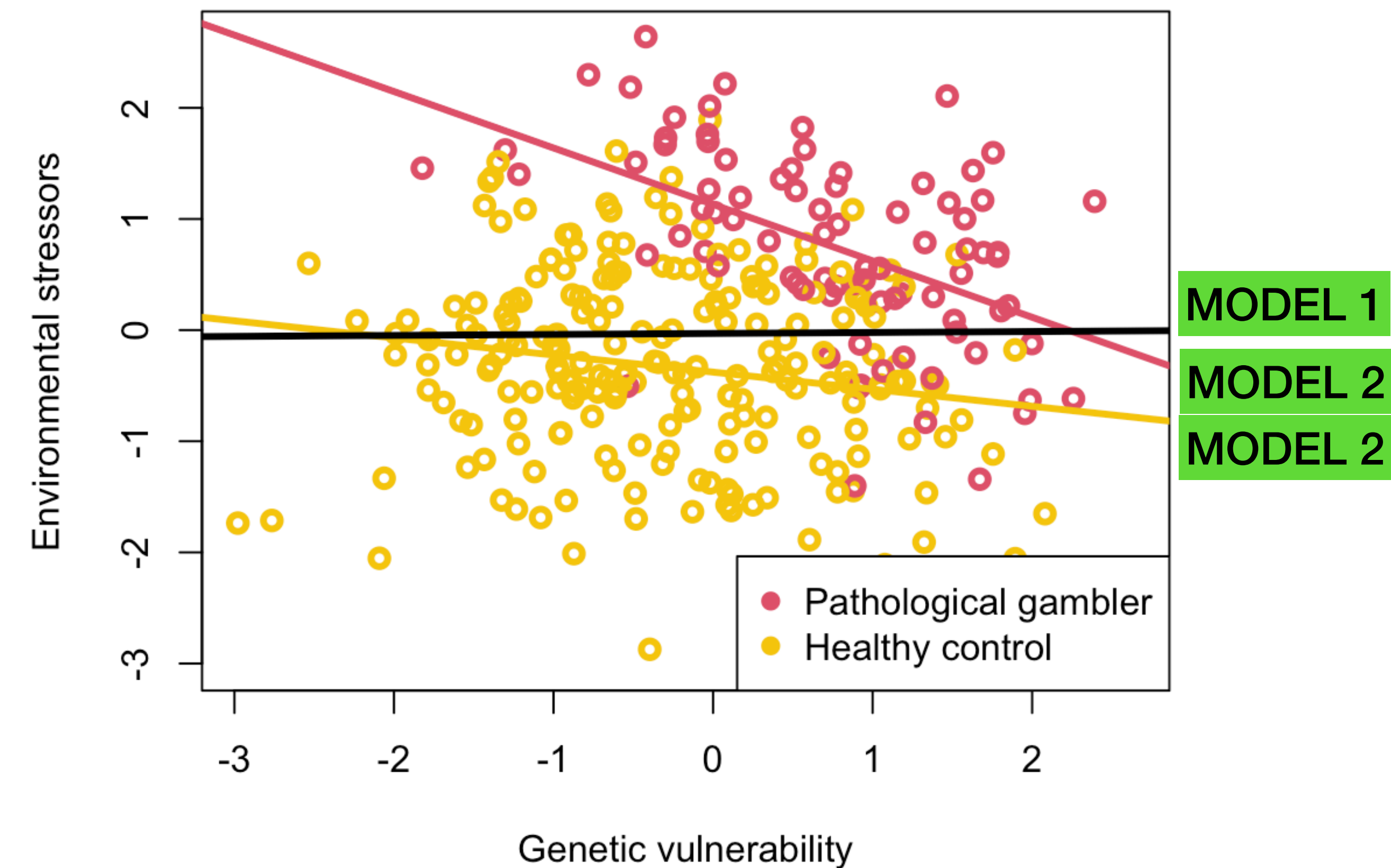
PG = pathological gambling



# Confounds: The Collider



- What is the effect of genetic vulnerability on environmental stressors?

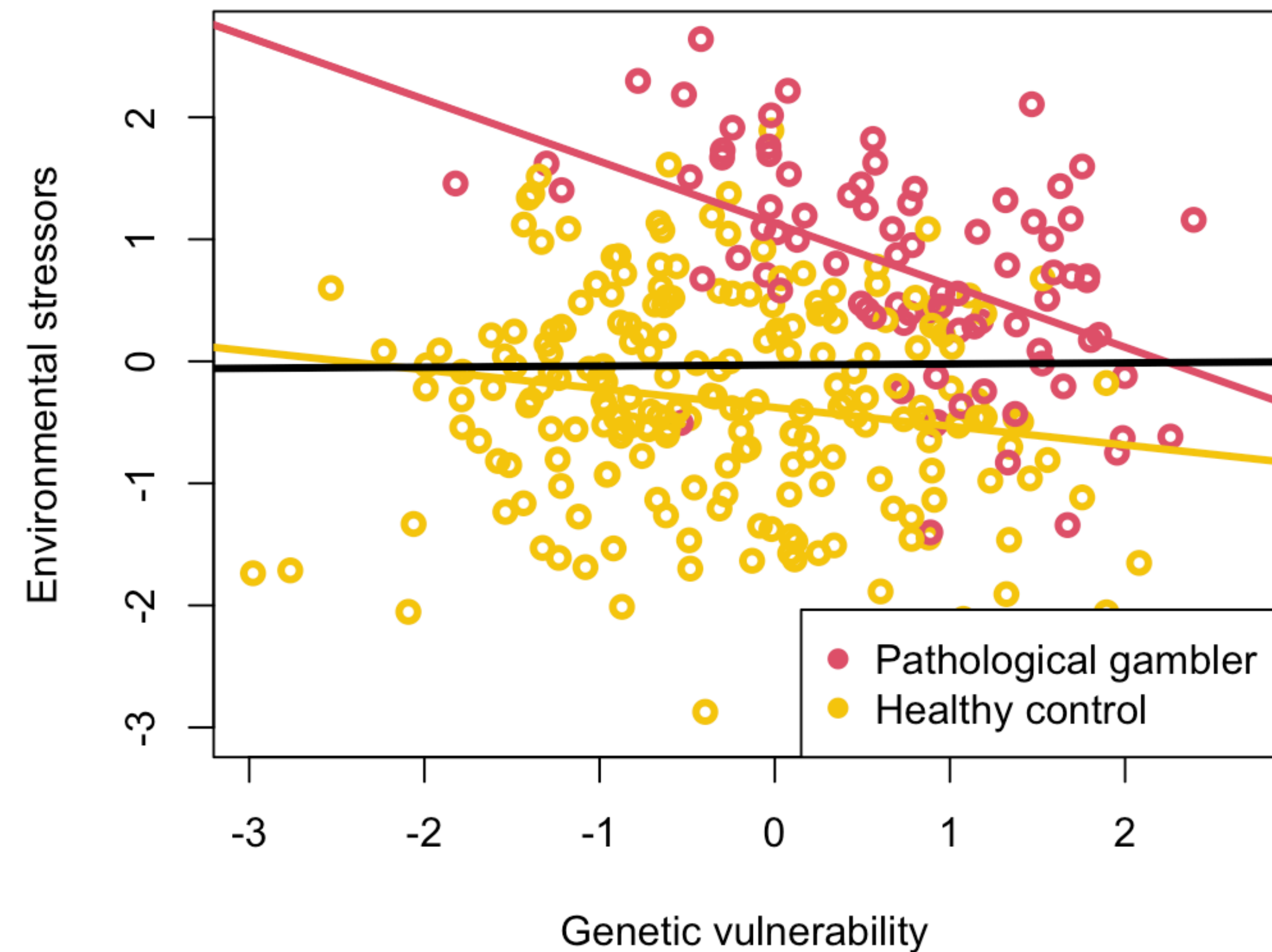
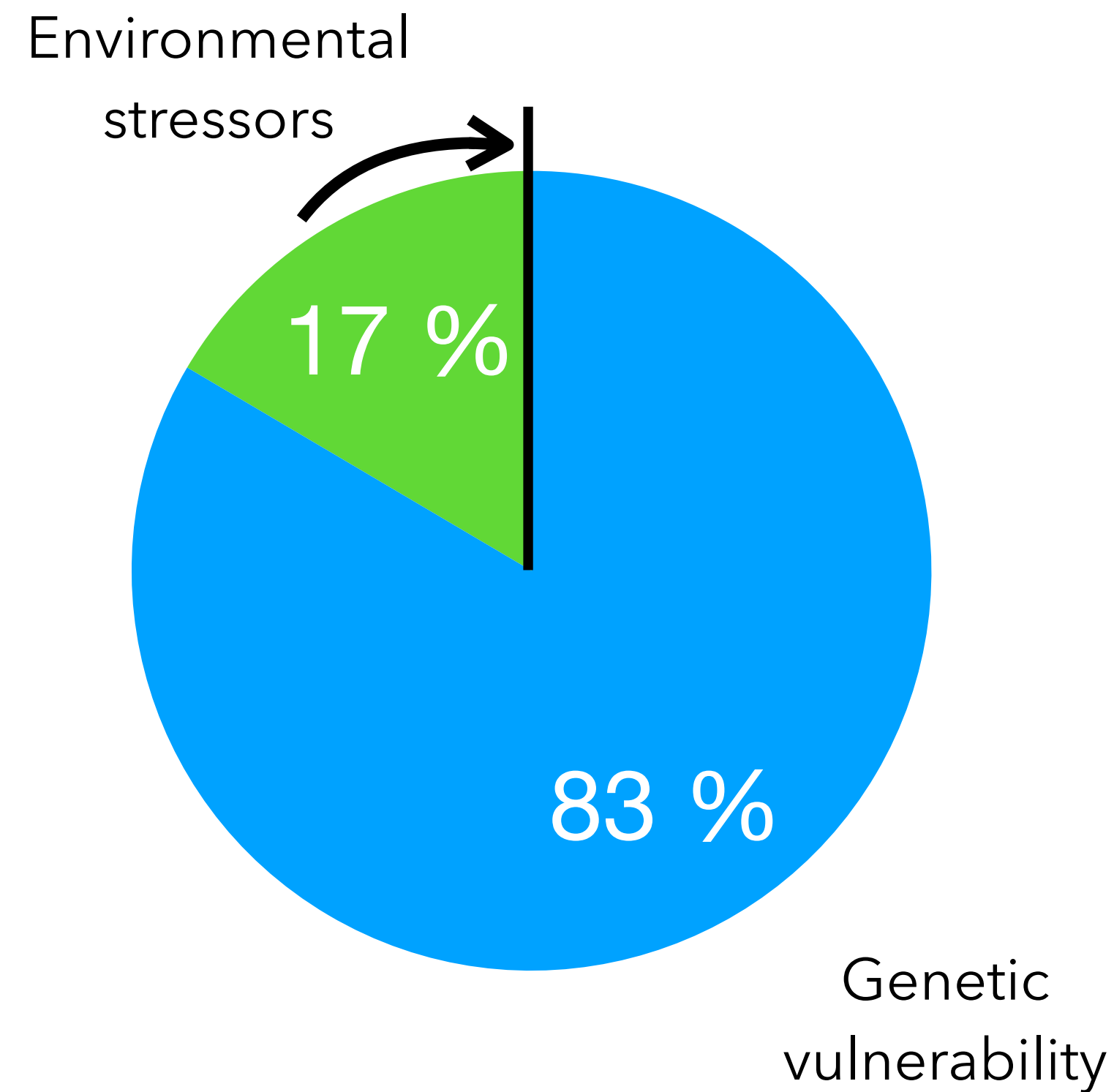
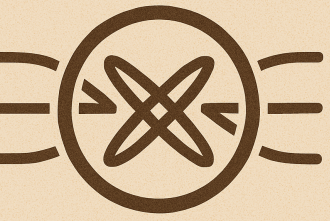


- Clinical status is a problematic collider
- Clinical status (pathological gambler, healthy control) as a predictor
- Creates a **fake association between genetic vulnerability and environmental stressors**



# Confounds: The Collider

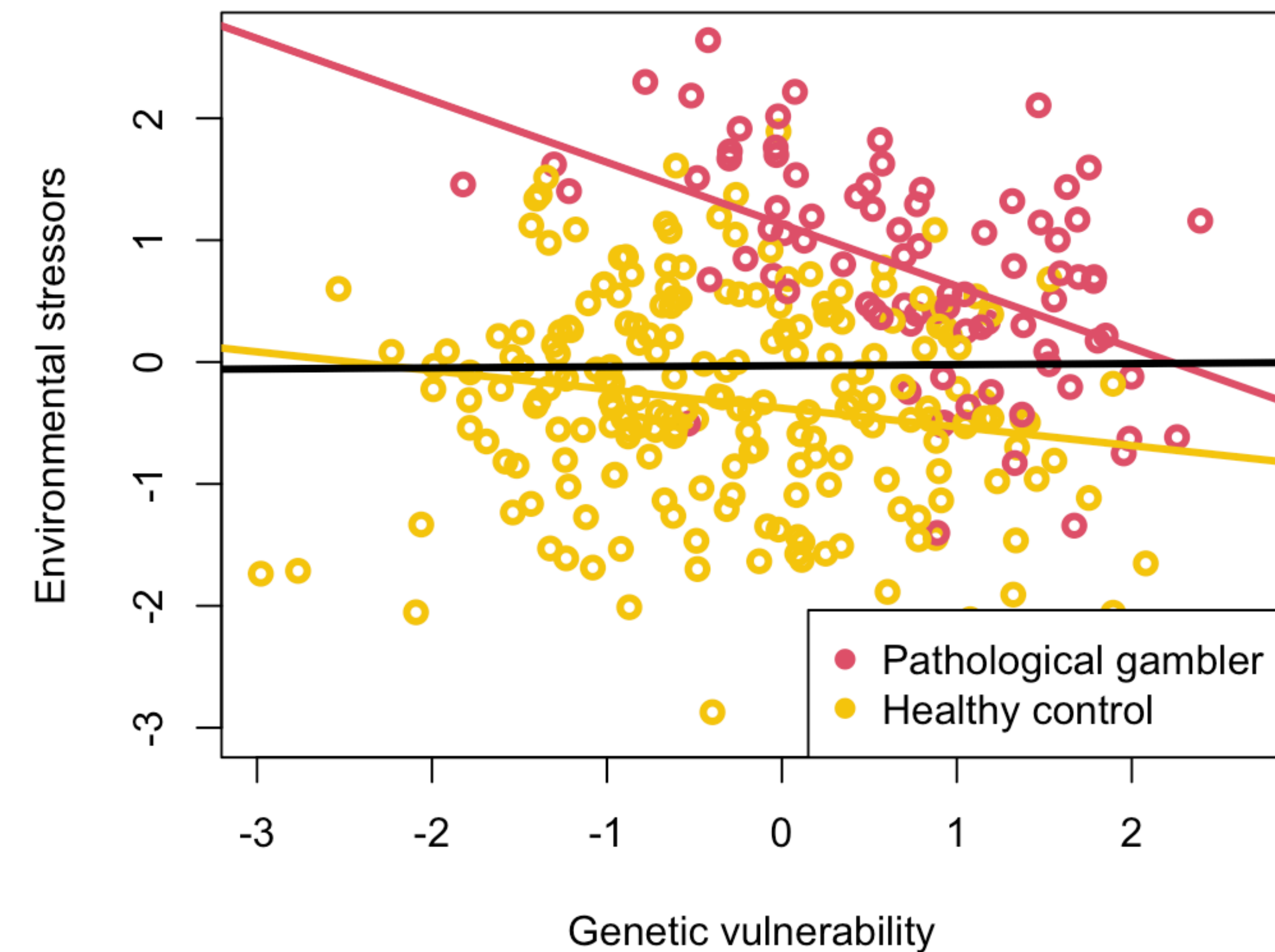
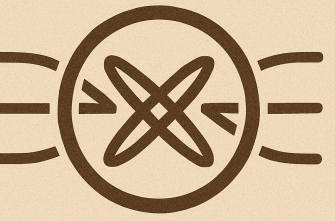
## The thresholding effect





# Confounds: The Collider

## The thresholding effect

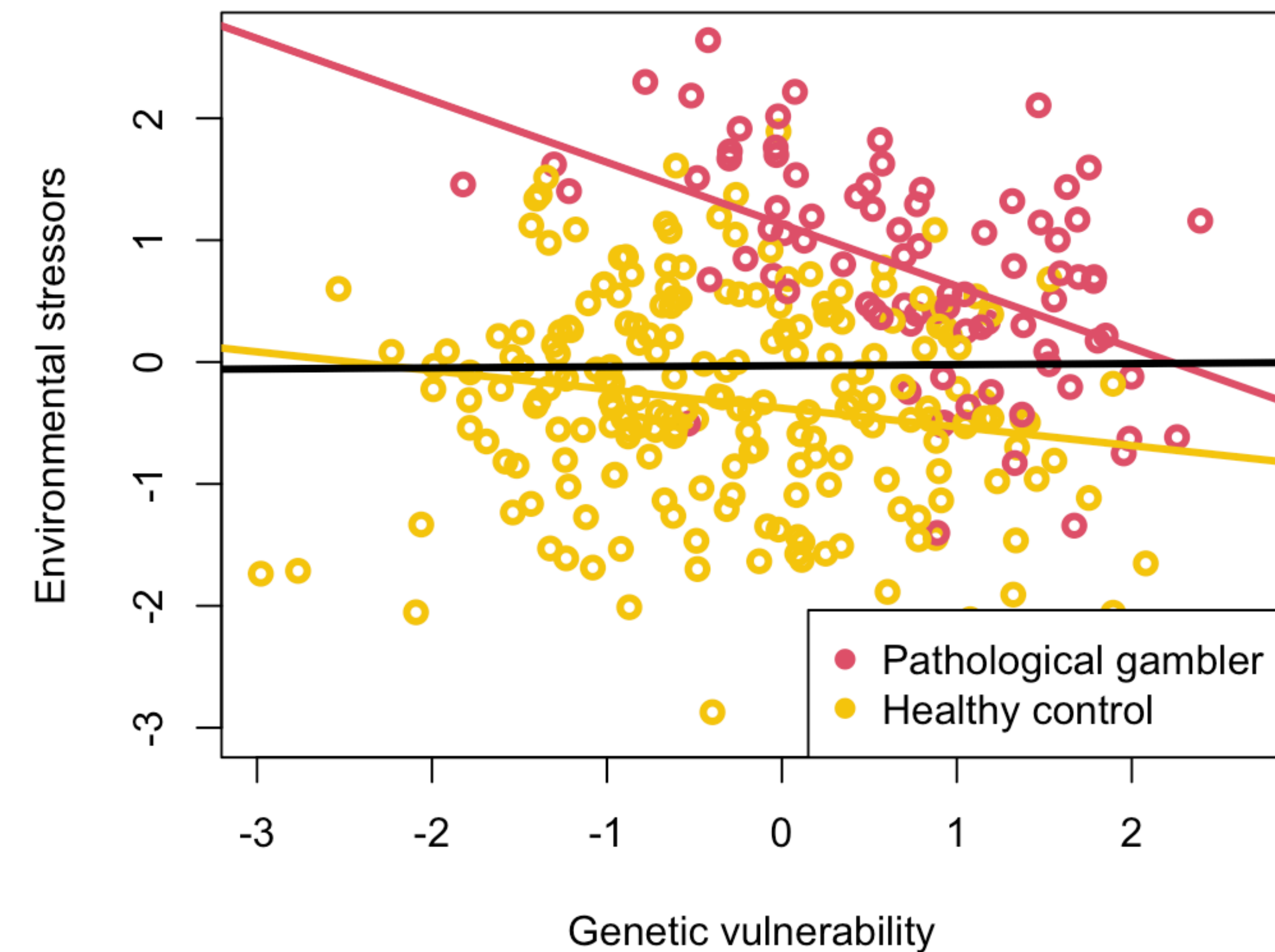
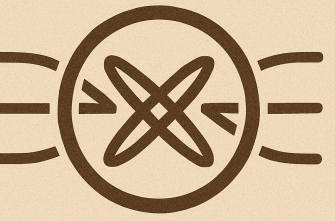


- If either genetic vulnerability or environmental stressors big enough, then onset of pathological gambling
- The higher the one, the lower the other needs to be for onset
- Either one must be high, we are missing the low-low corner
- Looking at the genetic-environmental association at each level of clinical status, there is an arbitrarily strong association



# Confounds: The Collider

## The thresholding effect

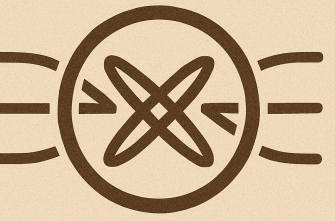


- So it can well be true, that subjects with PG show this 'higher the one, lower the other' -tendency
- Yet, it does not mean that genetic vulnerability for PG causes less environmental stressors (higher the vulnerability, lower the stressors)



# Confounds: The Collider

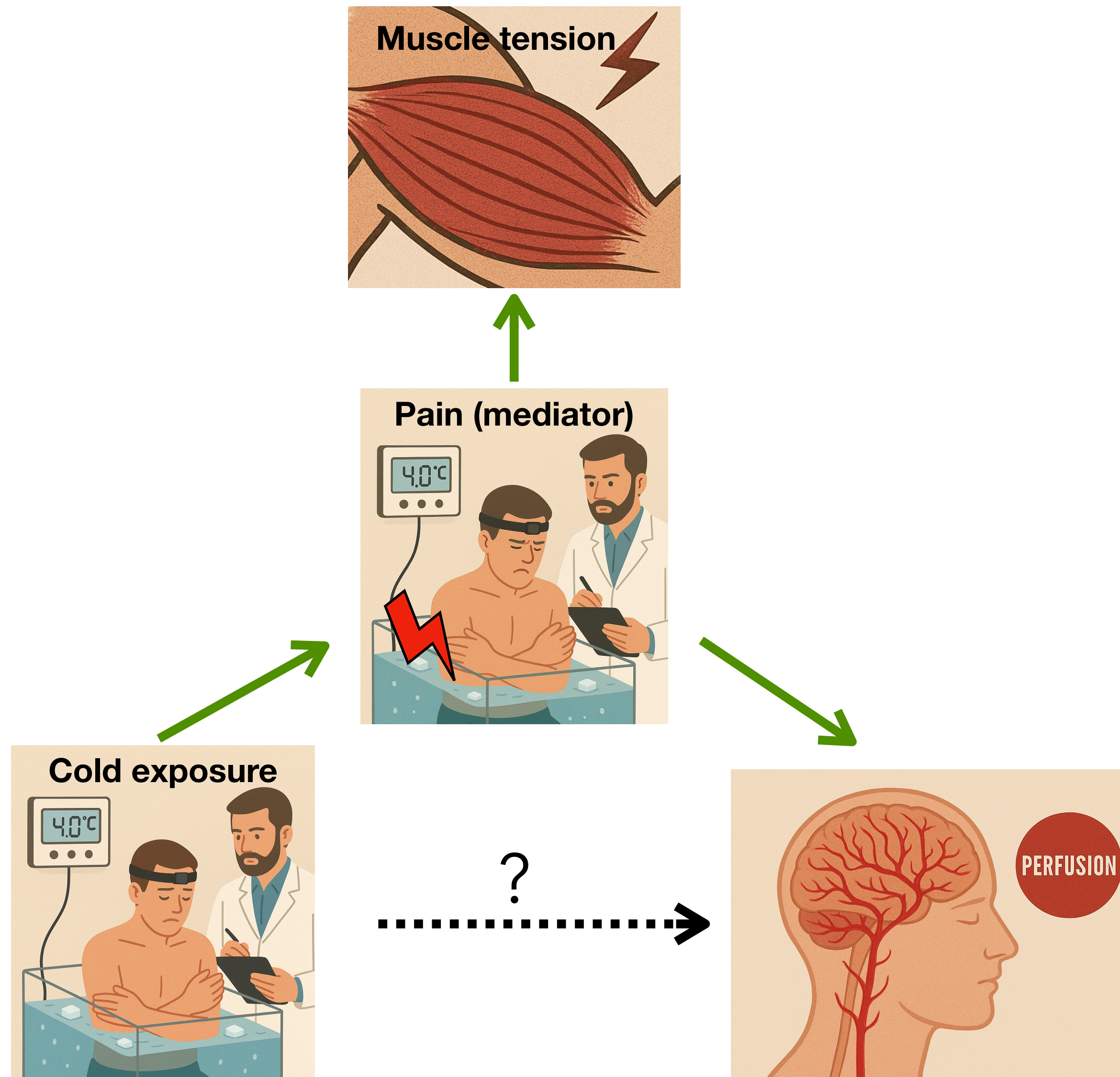
## The general rule of thumb



- Let's consider
  - The subpopulations we have in the data
  - In which population we want to assess the interesting effect
    - Visualizing data is always useful



# Confounds: The Descendant



- Behavior depends on where the descendant is attached to (here pipe)
- Muscle tension is a descendant (child) for pain (parent)
- Including muscle tension like including its 'parent' pain, but a bit weaker - not a clone but includes some of the same information
- If interested in the total effect, leave out both the parent and the child
- If interested in the mediator-independent effect, may be better to include one but not both



# Theory vs practice

- The DAGs make sense in theory. In practice, they may become extremely large and cryptic as the number of variables increases
- Often, we do not know the causal paths between the variables
  - If we did, would we study them?
- Awareness of the potential pitfalls in the interpretation of our findings!



# Useful practices

- **Assess correlation of the predictors, multicollinearity metrics (e.g. VIF)**
  - High correlations: Are the predictors measures of the same thing, is one a proxy of the other?
    - Yes: Can we choose only one of them or combine several variables into one umbrella variable (metabolic strain from body mass index, waist measurement, cholesterol...)?
- **Model comparison**
  - Modify the set of the predictors, then compare models: Essential changes in the findings?
    - No: No signs of major problems with the predictor combination, might be good to report the model comparison in publications
    - Yes: Analyze more in detail and consider the causal paths



# Underfitting & Overfitting

- Not too few, not too many predictors
- **Underfitting: Too few**
  - Too general
  - Is missing meaningful predictors that in real life influence the outcome
    - Missing the age effect in brain data, while neural functions and structures are clearly affected by age

# Underfitting & Overfitting

- Not too few, not too many predictors
- **Overfitting: Too many**
  - Difficult to interpret: The causal paths become difficult to handle
  - Fits the current sample 'perfectly' but does not generalize well in the population



**'Blindly tossing variables into the causal salad is never a good idea.'**

**- R. McElreath**