Noisy-Or Risk Allocation: A Probabilistic Model for Attributable Risk Estimation

Amelia J. Averitt, MPH MA PhD[1,2]; Adler Perotte, MD MA[2]

1. Regeneron Genetics Center, Clinical Informatics. Tarrytown, NY.
2. Columbia University, Department of Biomedical Informatics. New York, NY.

Columbia University Department of Biomedical Informatics

Regeneron Science to Medicine®

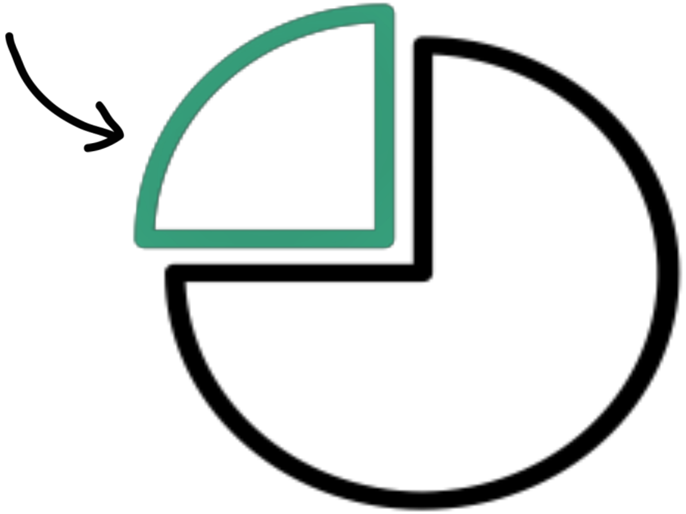
Rothman, 1976

risks are an important tool for
the understanding and communication of
exposure-outcome relationships

Leviton, 1973; Miettinen, 1974; Markush, 1977

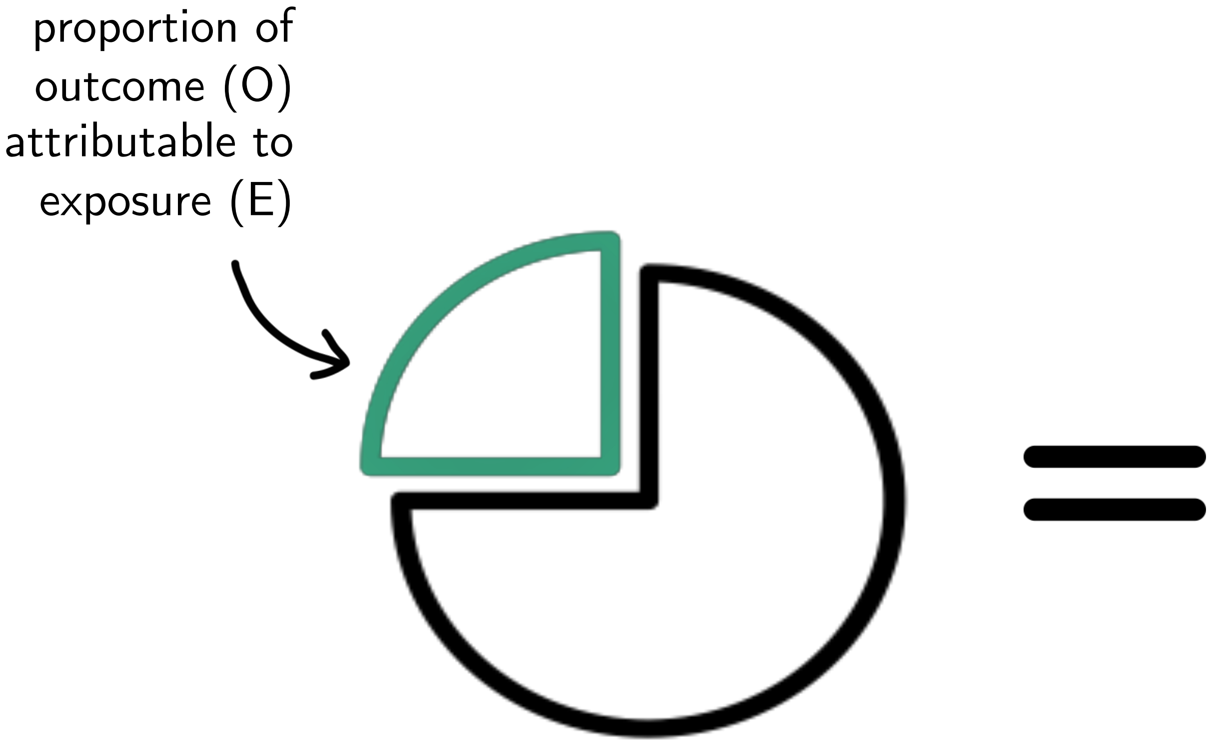
attributable risk (AR) estimation

proportion of outcome (O) attributable to exposure (E)



Leviton, 1973; Miettinen, 1974; Markush, 1977

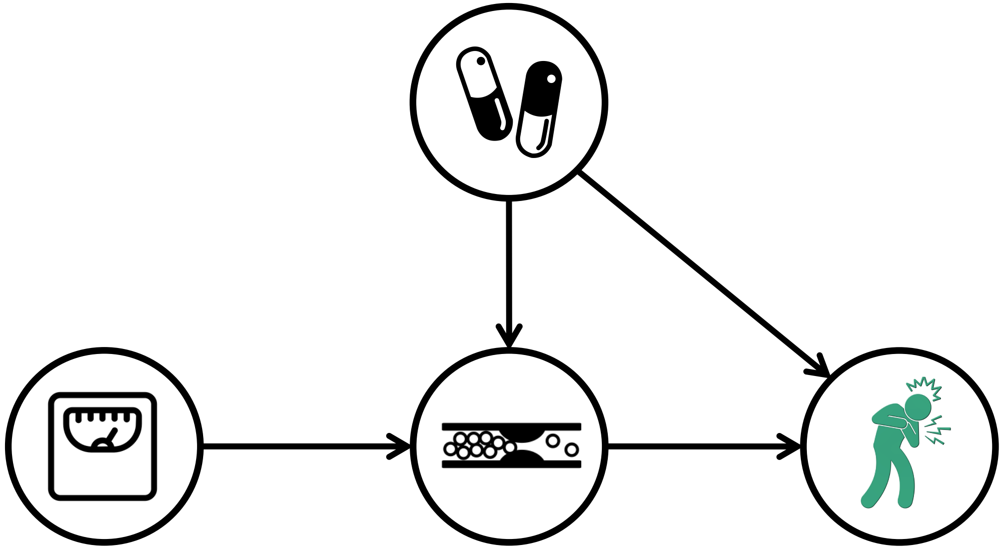
attributable risk (AR) estimation



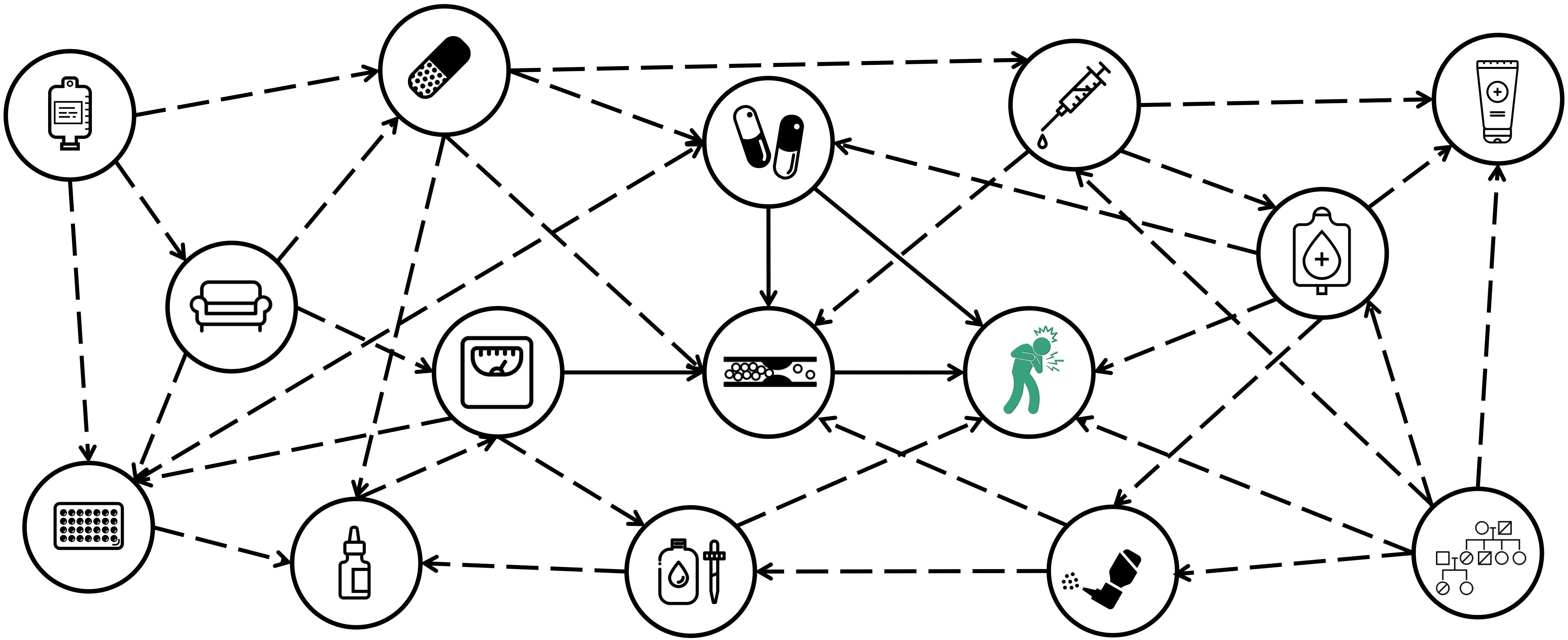
"excess risk" Levin, 1953

$$AR = \frac{P(O) - P(O|\neg E)}{P(O)}$$

attributable risk (AR) estimation



attributable risk (AR) estimation



attributable risk (AR) estimation

the setting of binary exposures and outcomes

attributable risk (AR) estimation

the setting of binary exposures and outcomes

calculation of excess risk Levin, 1953

$$AR = \frac{P(D) - P(D|\neg E)}{P(D)}$$

approximation using disproportionality methods Bate, 2002

- multi-gamma Poisson shrinker (MGPS)
- risk ratios (RR)

$$AR = \frac{RR - 1}{RR}$$

regression-based methods Hayashi, 2018; Caster, 2007

- penalized logistic regression

attributable risk (AR) estimation

the setting of binary exposures and outcomes

calculation of excess risk Levin, 1953

$$AR = \frac{P(D) - P(D|\neg E)}{P(D)}$$

approximation using disproportionality methods Bate, 2002

- multi-gamma Poisson shrinker (MGPS)
- risk ratios (RR)

$$AR = \frac{RR - 1}{RR}$$

regression-based methods Hayashi, 2018; Caster, 2007

- penalized logistic regression

1 global inference

2 local inference

3 account for confounding

attributable risk (AR) estimation

the setting of binary exposures and outcomes

calculation of excess risk Levin, 1953

$$AR = \frac{P(D) - P(D|\neg E)}{P(D)}$$

approximation using disproportionality methods Bate, 2002

- multi-gamma Poisson shrinker (MGPS)
- risk ratios (RR)

$$AR = \frac{RR - 1}{RR}$$

regression-based methods Hayashi, 2018; Caster, 2007

- penalized logistic regression

1 global inference

2 local inference

3 account for confounding

attributable risk (AR) estimation

the setting of binary exposures and outcomes

calculation of excess risk Levin, 1953

$$AR = \frac{P(D) - P(D|\neg E)}{P(D)}$$

approximation using disproportionality methods Bate, 2002

- multi-gamma Poisson shrinker (MGPS)
- risk ratios (RR)

$$AR = \frac{RR - 1}{RR}$$

regression-based methods Hayashi, 2018; Caster, 2007

- penalized logistic regression

1 global inference

2 local inference

3 account for confounding

a model that meets all of these criteria is preferred

attributable risk (AR) estimation

the setting of binary exposures and outcomes

calculation of excess risk Levin, 1953

$$AR = \frac{P(D) - P(D|\neg E)}{P(D)}$$

approximation using disproportionality methods Bate, 2002

- multi-gamma Poisson shrinker (MGPS)
- risk ratios (RR)

$$AR = \frac{RR - 1}{RR}$$

regression-based methods Hayashi, 2018; Caster, 2007

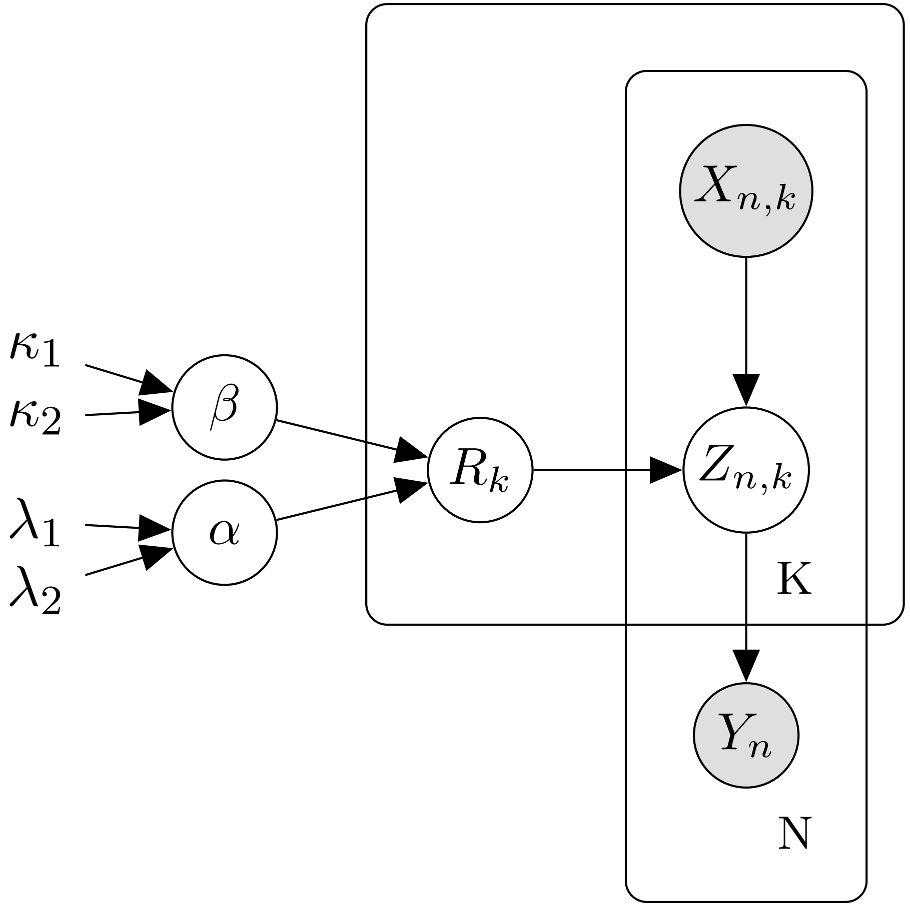
- penalized logistic regression

1 global inference

2 local inference

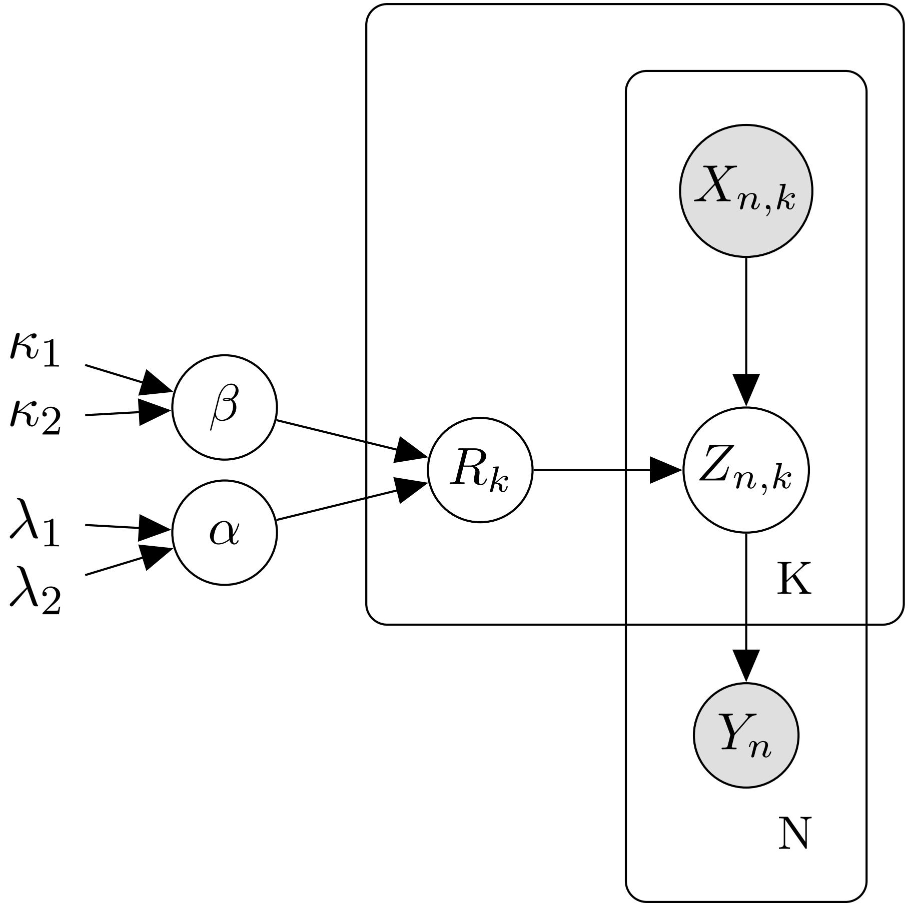
3 account for confounding

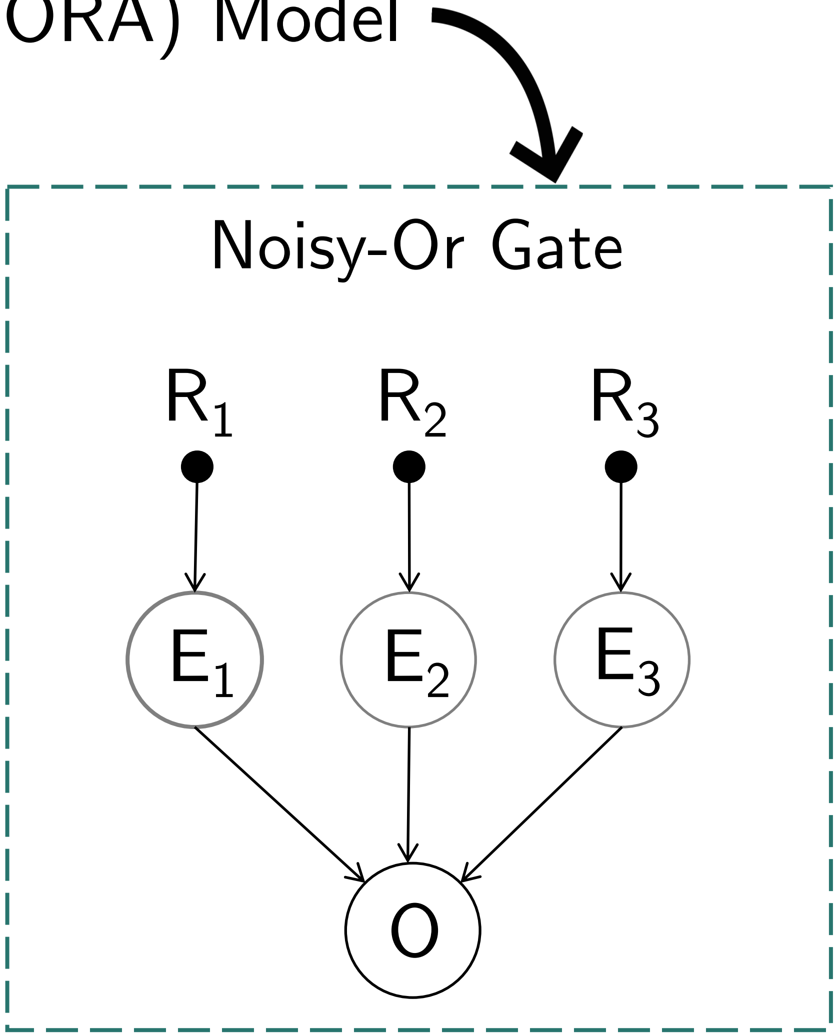
Noisy-Or Risk Allocation (NORA) Model



Good, 1959; Pearl, 1988; Srinivas, 1993; Antonucci, 2011; Parsons, n.d.

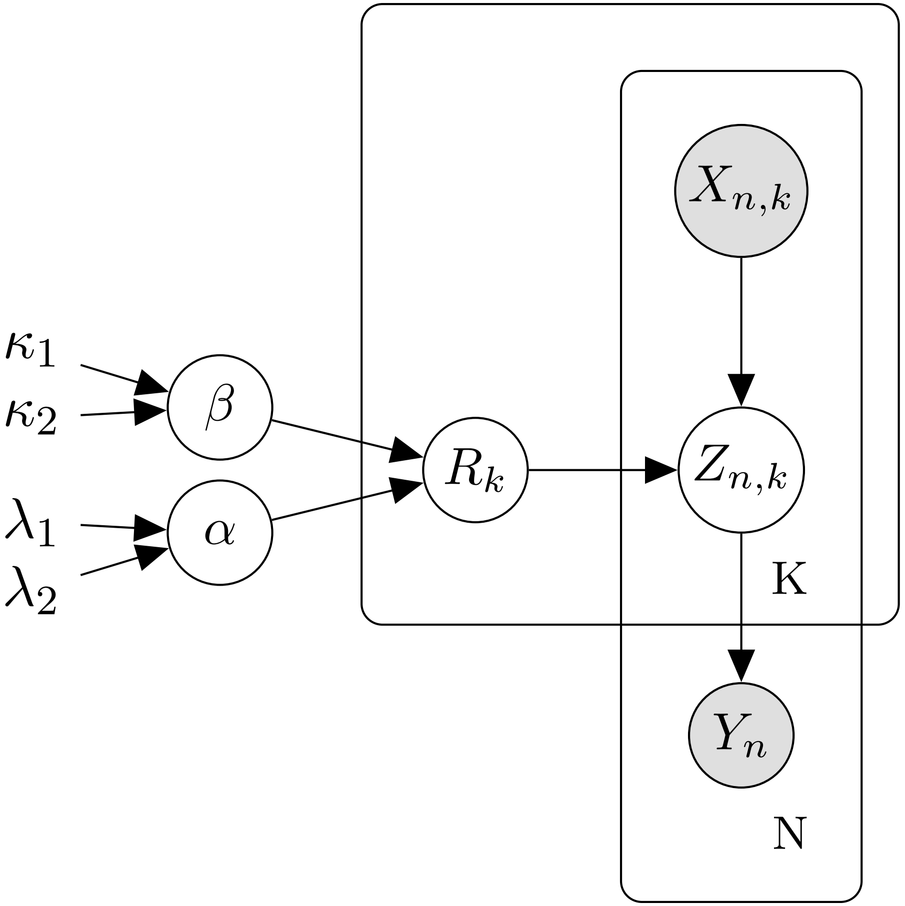
Noisy-Or Risk Allocation (NORA) Model

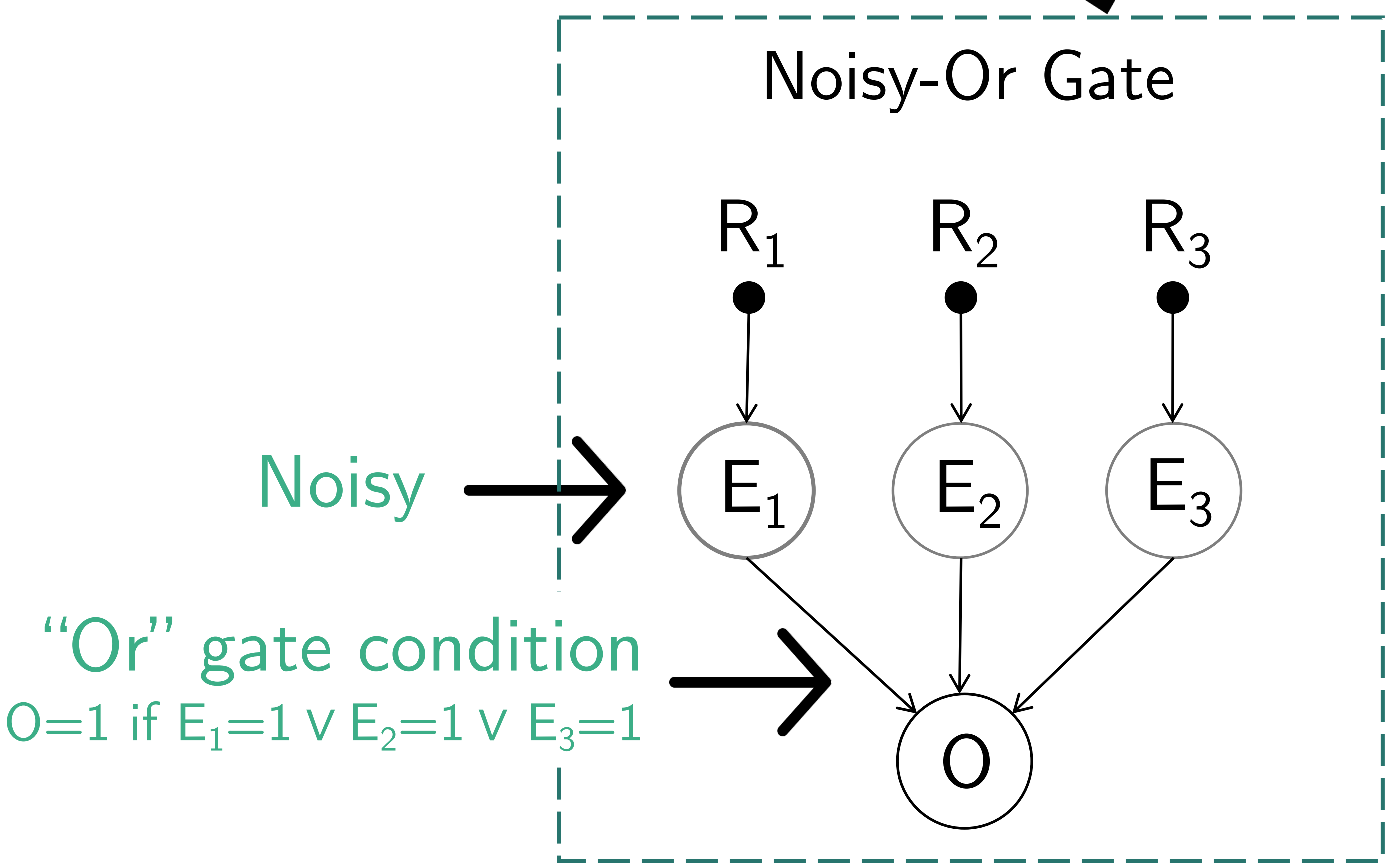




Good, 1959; Pearl, 1988; Srinivas, 1993; Antonucci, 2011; Parsons, n.d.

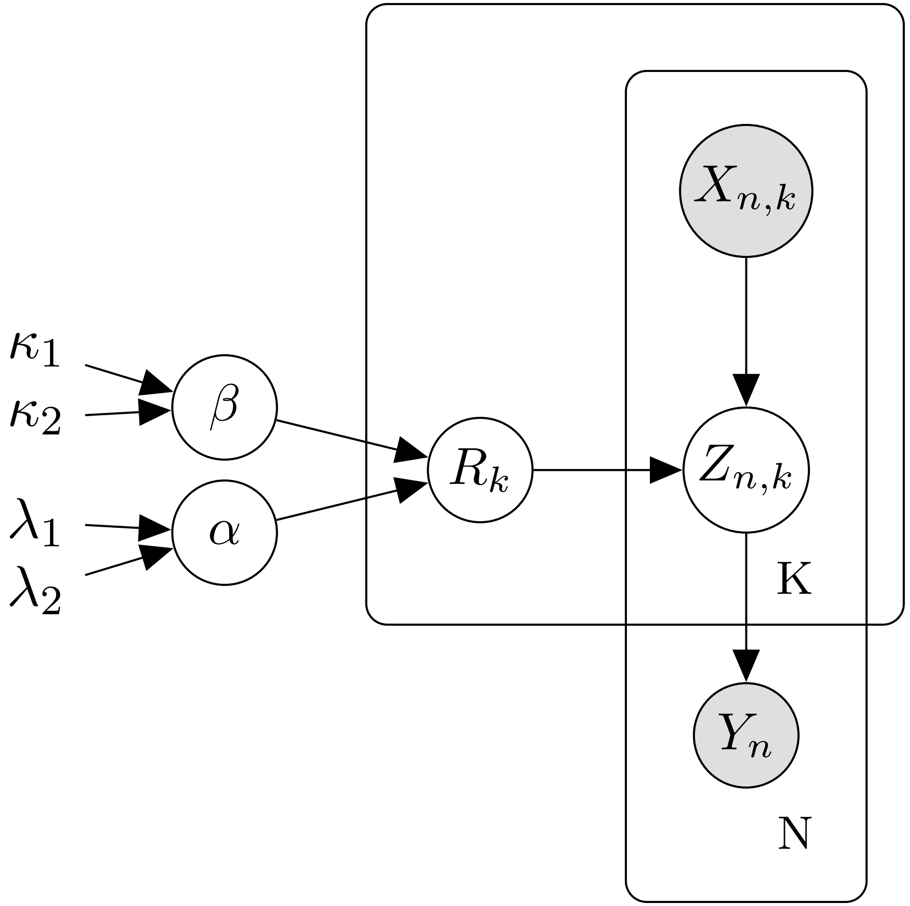
Noisy-Or Risk Allocation (NORA) Model





Good, 1959; Pearl, 1988; Srinivas, 1993; Antonucci, 2011; Parsons, n.d.

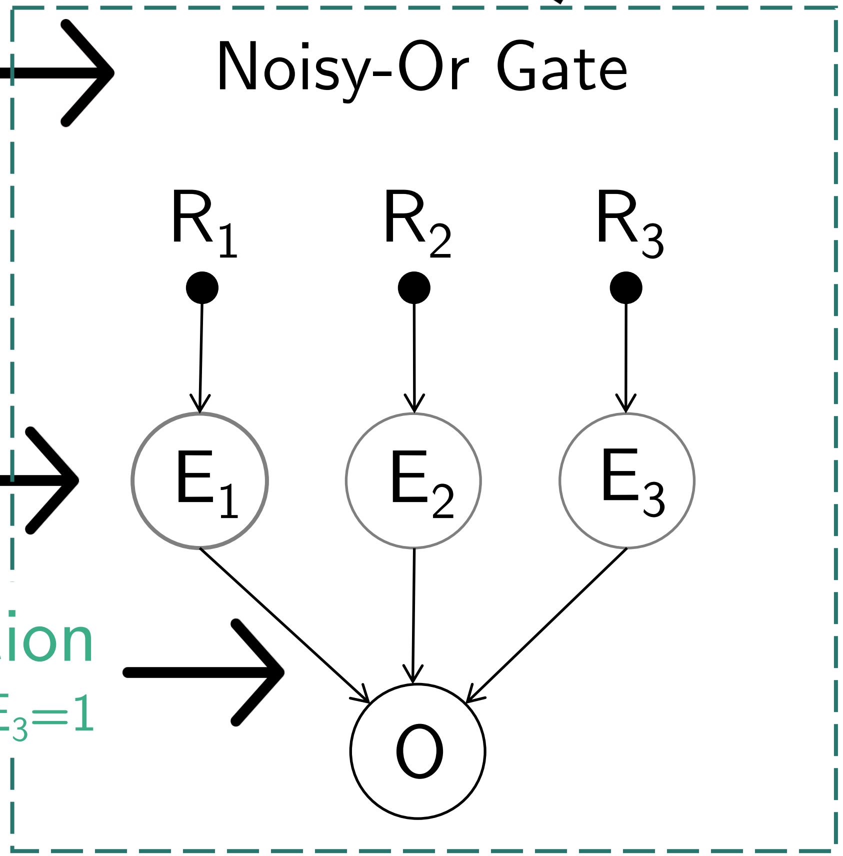
Noisy-Or Risk Allocation (NORA) Model



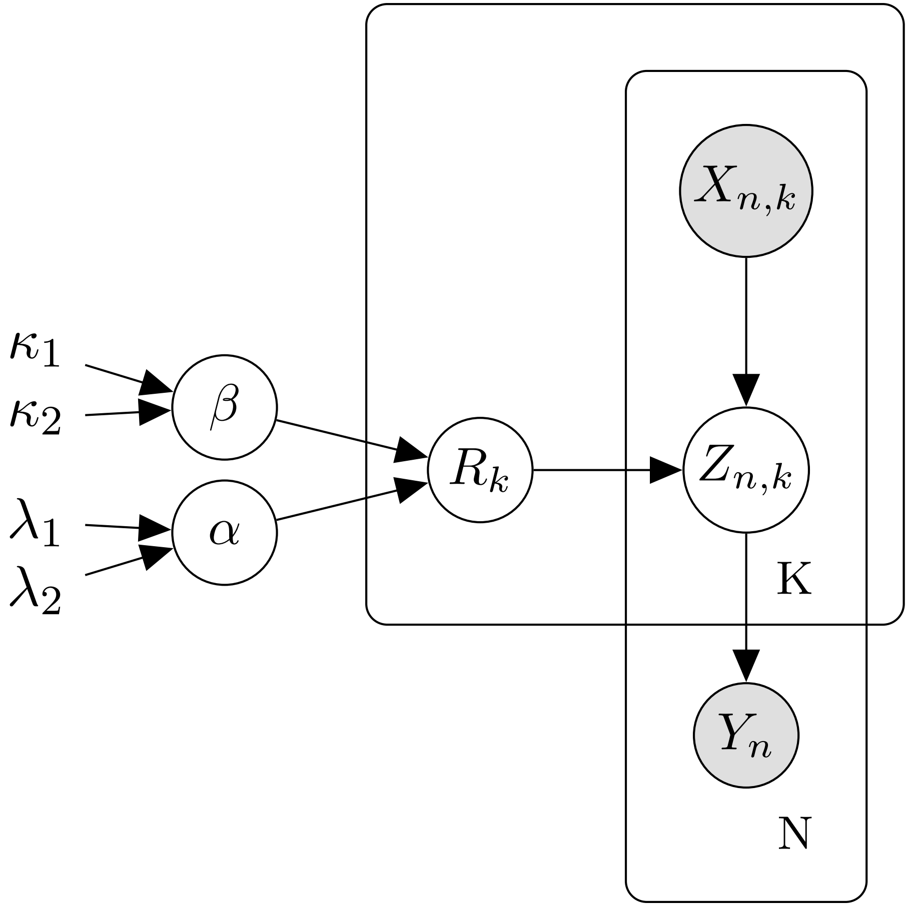
embeds the notion of *causal independence*

Noisy

"Or" gate condition
O=1 if E_1=1 ∨ E_2=1 ∨ E_3=1



Noisy-Or Risk Allocation (NORA) Model



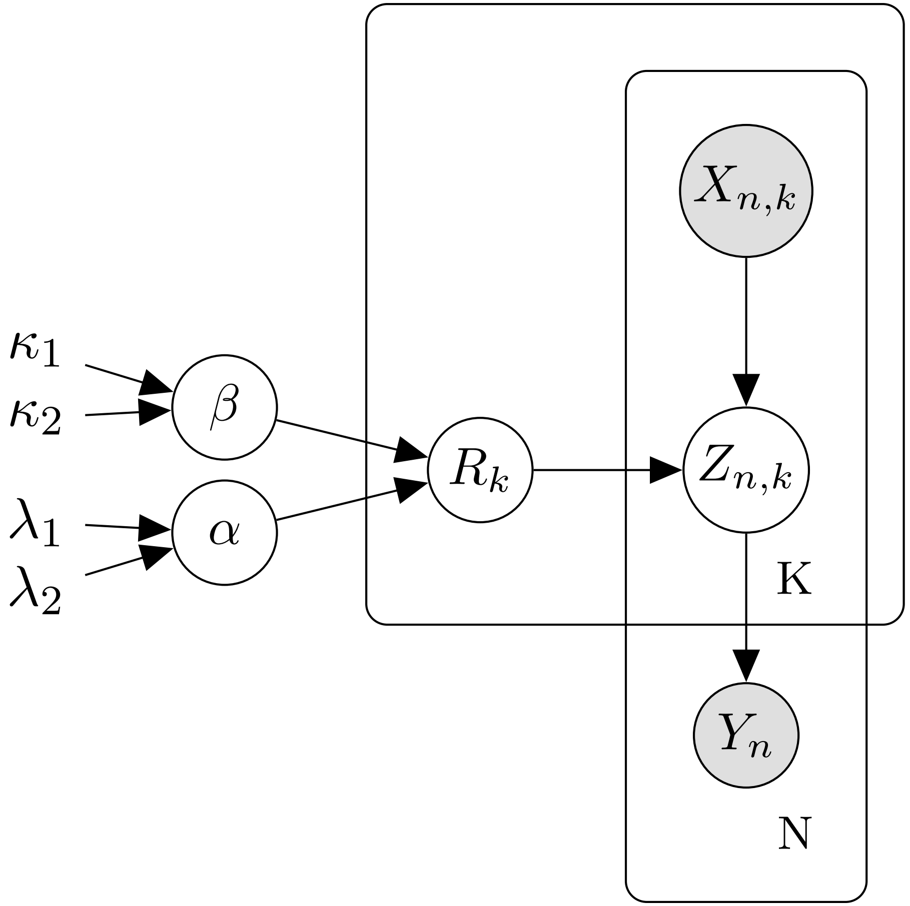
Nodes → variables
Edges → relationships between variables

(grey circle) → observed

(white circle) → unobserved

N = number of people
K = number of exposures

Noisy-Or Risk Allocation (NORA) Model

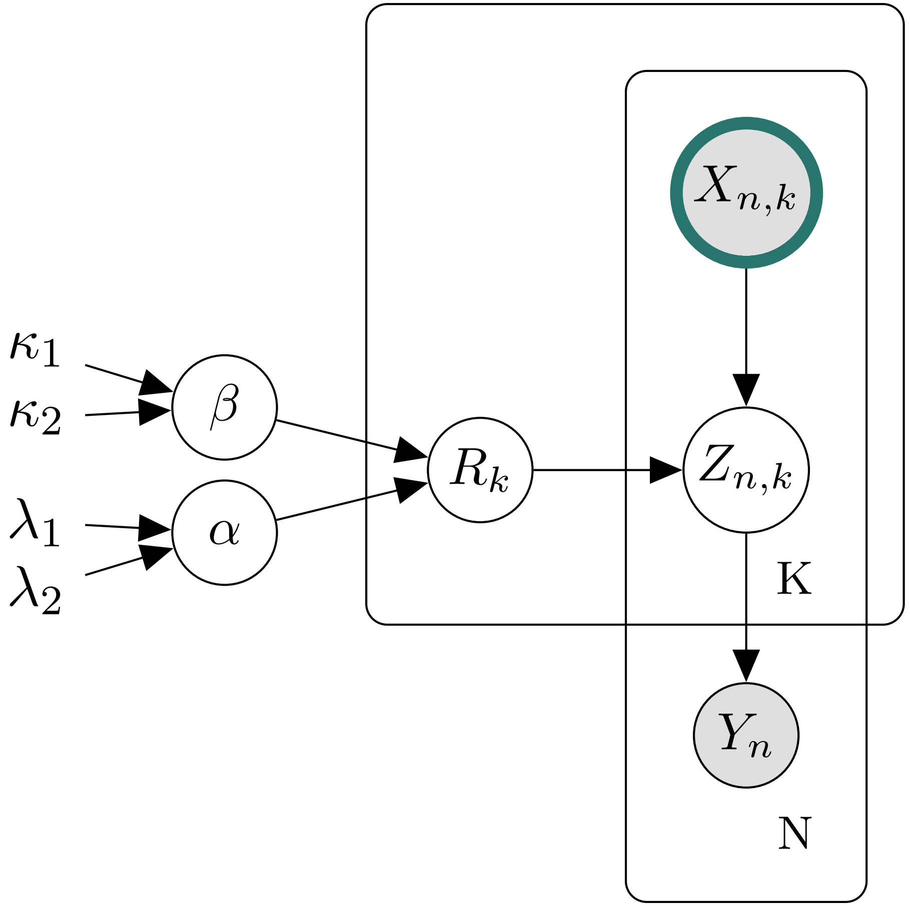


Nodes → variables
Edges → relationships between variables

● → observed
○ → unobserved

N = number of people
K = number of exposures ← "Noisy"

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables

● → observed

○ → unobserved

N = number of people
K = number of exposures

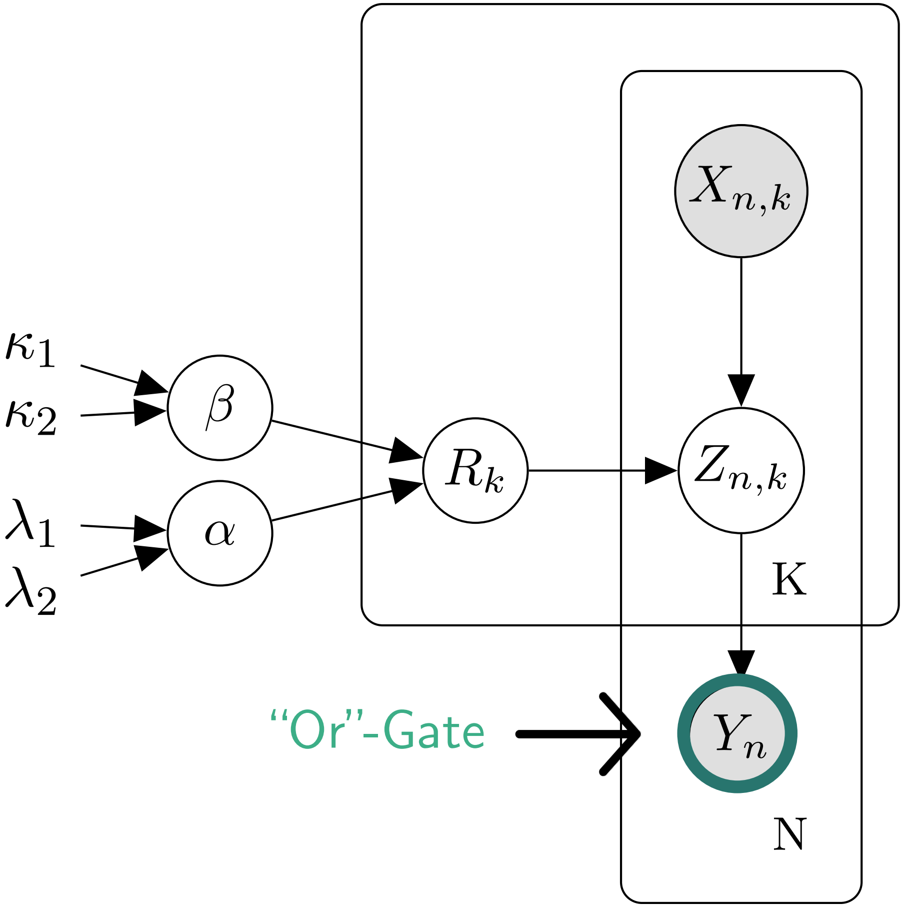
$X_{n,k} \sim \text{Bernoulli}(\epsilon)$ — *exposure k for person n*

$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$ — *outcome for person n*

$R_k \sim \text{Beta}(\alpha, \beta)$ — *risk for exposure k*

$Z_{n,k} \sim \text{Bernoulli}(X_{n,k} R_k)$ — *activation of exposure k for person n*

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables

● → observed
○ → unobserved

N = number of people
K = number of exposures

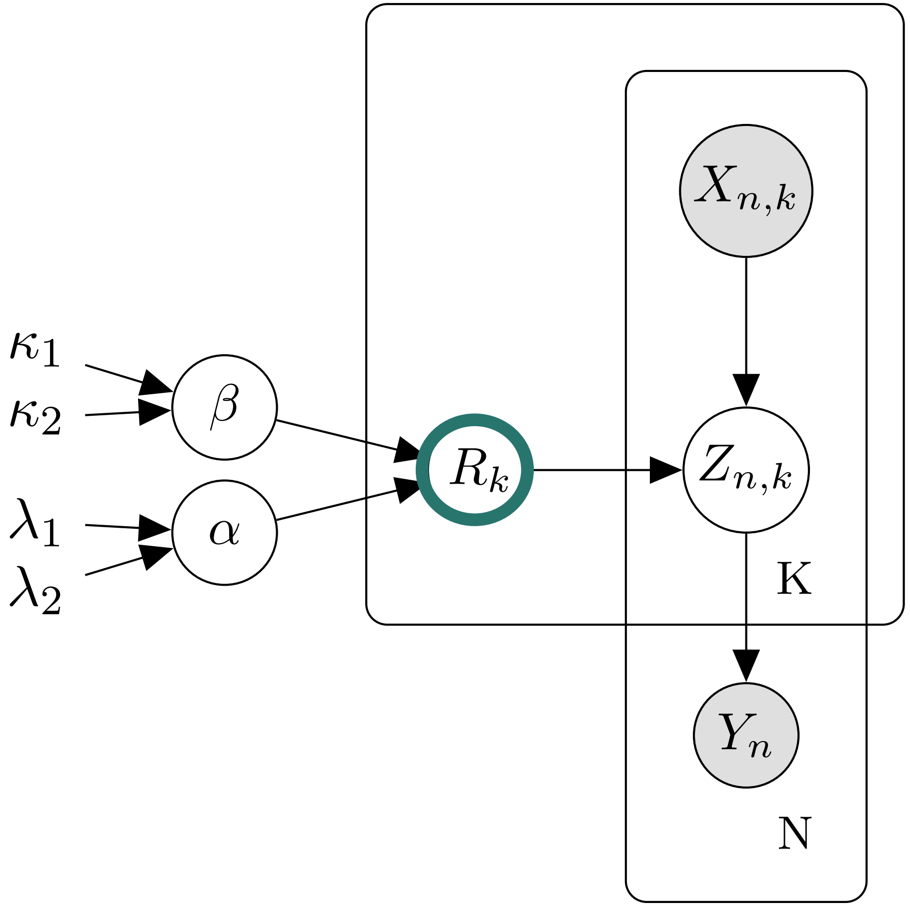
$X_{n,k} \sim \text{Bernoulli}(\epsilon)$ *exposure k for person n*

$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$ *outcome for person n*

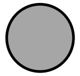
$R_k \sim \text{Beta}(\alpha, \beta)$ *risk for exposure k*

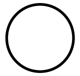
$Z_{n,k} \sim \text{Bernoulli}(X_{n,k} R_k)$ *activation of exposure k for person n*

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables

 → observed

 → unobserved

N = number of people
K = number of exposures

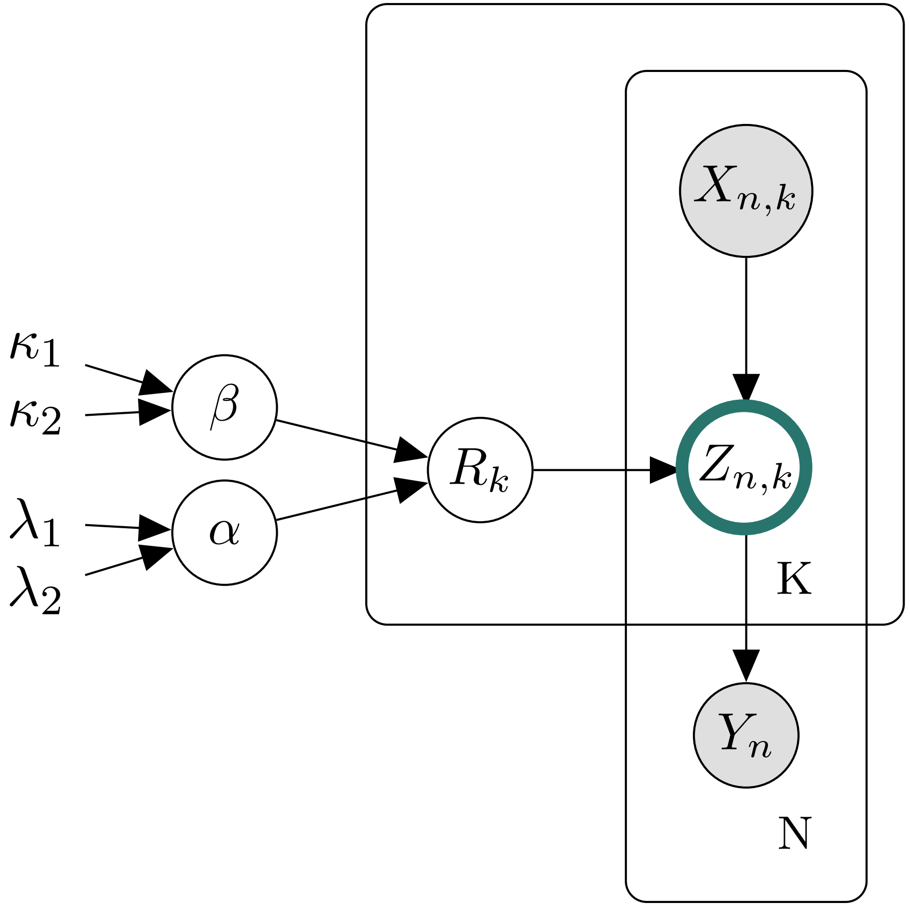
$X_{n,k} \sim \text{Bernoulli}(\epsilon)$ *exposure k for person n*

$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$ *outcome for person n*

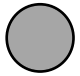
$R_k \sim \text{Beta}(\alpha, \beta)$ *risk for exposure k*

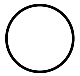
$Z_{n,k} \sim \text{Bernoulli}(X_{n,k} R_k)$ *activation of exposure k for person n*

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables


→ observed


→ unobserved

N = number of people
K = number of exposures

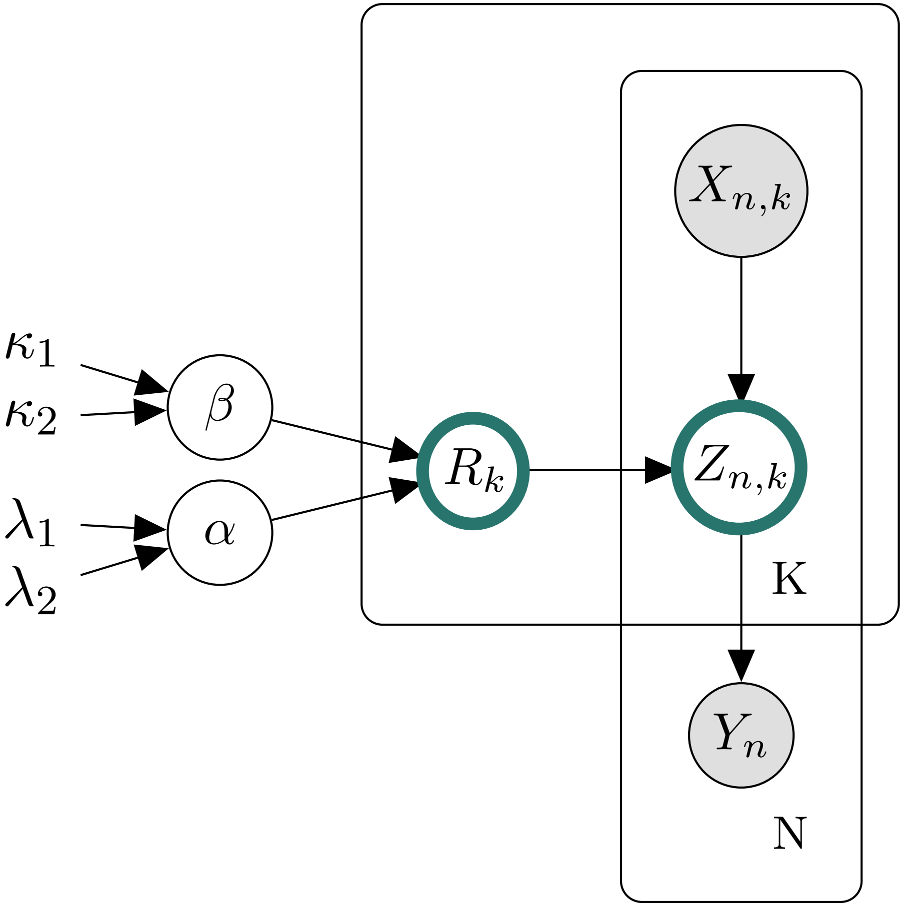
$X_{n,k} \sim \text{Bernoulli}(\epsilon)$ — *exposure k for person n*

$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$ — *outcome for person n*

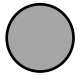
$R_k \sim \text{Beta}(\alpha, \beta)$ — *risk for exposure k*

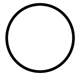
$Z_{n,k} \sim \text{Bernoulli}(X_{n,k}R_k)$ — *activation of exposure k for person n*

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables

 → observed

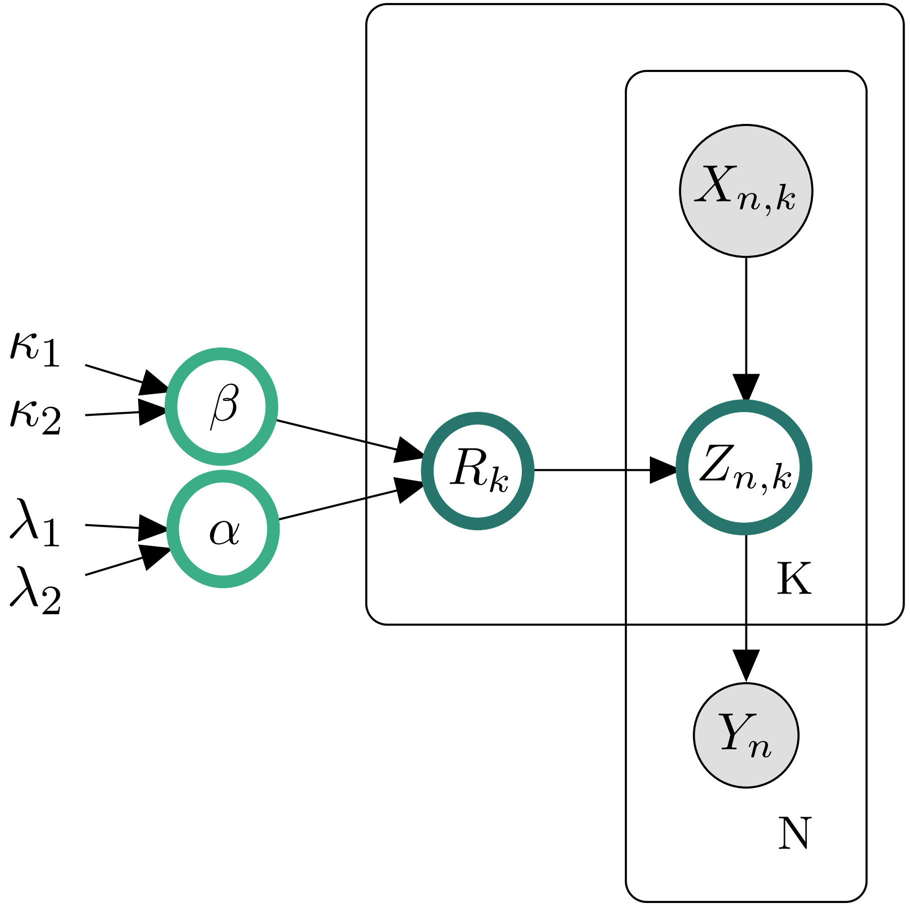
 → unobserved

- R & Z learned by Gibbs sampling

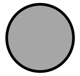
N = number of people
K = number of exposures

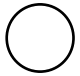
$X_{n,k} \sim \text{Bernoulli}(\epsilon)$	*exposure k for person n*
$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$	*outcome for person n*
$R_k \sim \text{Beta}(\alpha, \beta)$	*risk for exposure k*
$Z_{n,k} \sim \text{Bernoulli}(X_{n,k} R_k)$	*activation of exposure k for person n*

Noisy-Or Risk Allocation (NORA) Model



Nodes → variables
Edges → relationships between variables


→ observed

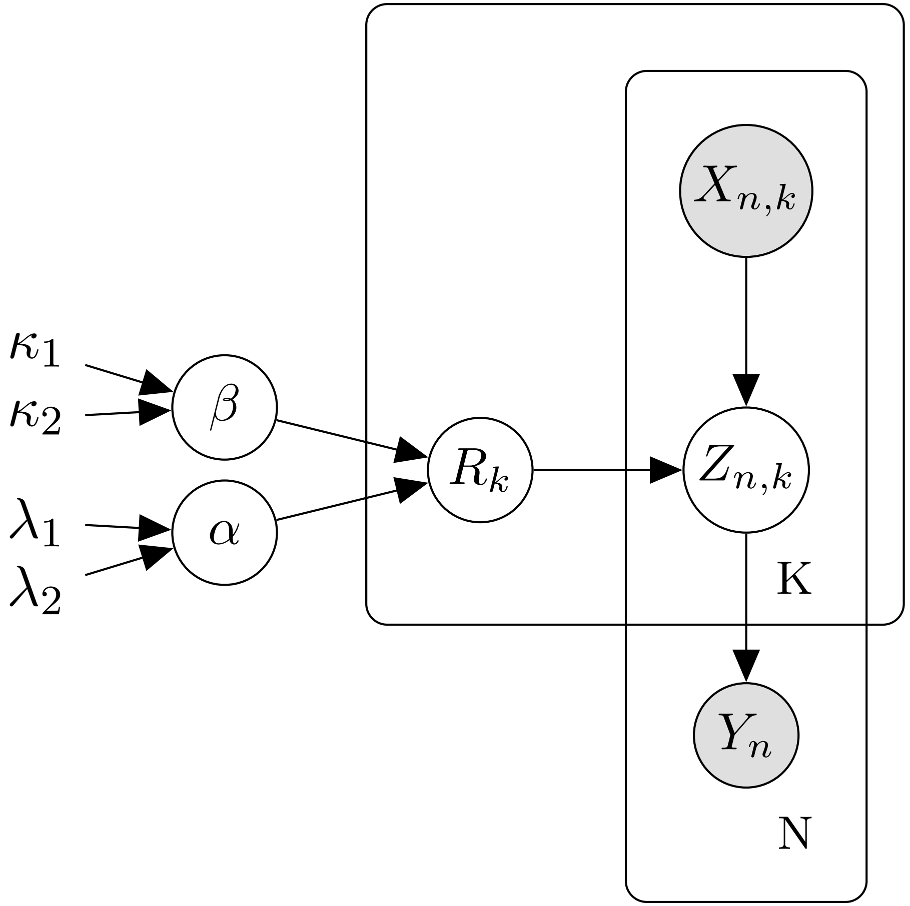

→ unobserved

- R & Z learned by Gibbs sampling
- α & β learned by Metropolis Hastings

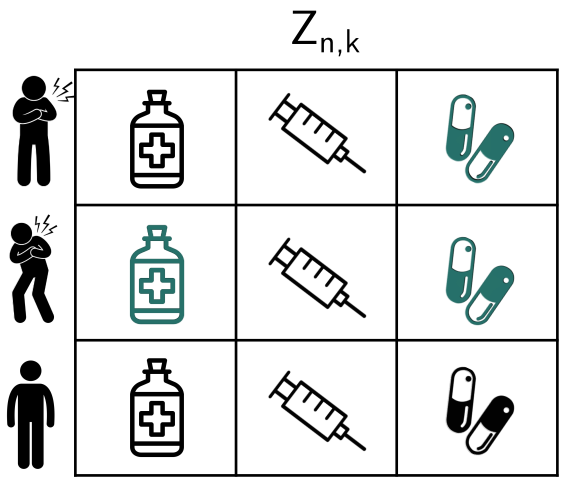
N = number of people
K = number of exposures

$X_{n,k} \sim \text{Bernoulli}(\epsilon)$	*exposure k for person n*
$Y_n = \begin{cases} 1 \text{ if any } Z_k = 1 \\ 0 \text{ if all } Z_k = 0 \end{cases}$	*outcome for person n*
$R_k \sim \text{Beta}(\alpha, \beta)$	*risk for exposure k*
$Z_{n,k} \sim \text{Bernoulli}(X_{n,k} R_k)$	*activation of exposure k for person n*

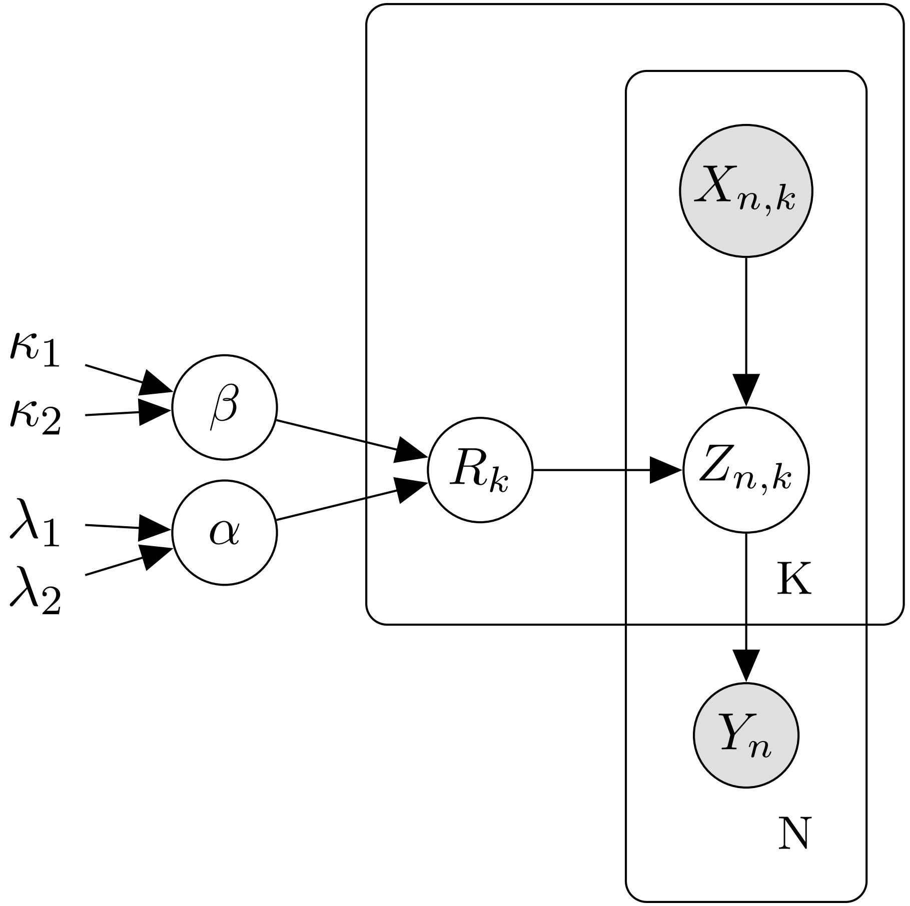
Noisy-Or Risk Allocation (NORA) Model



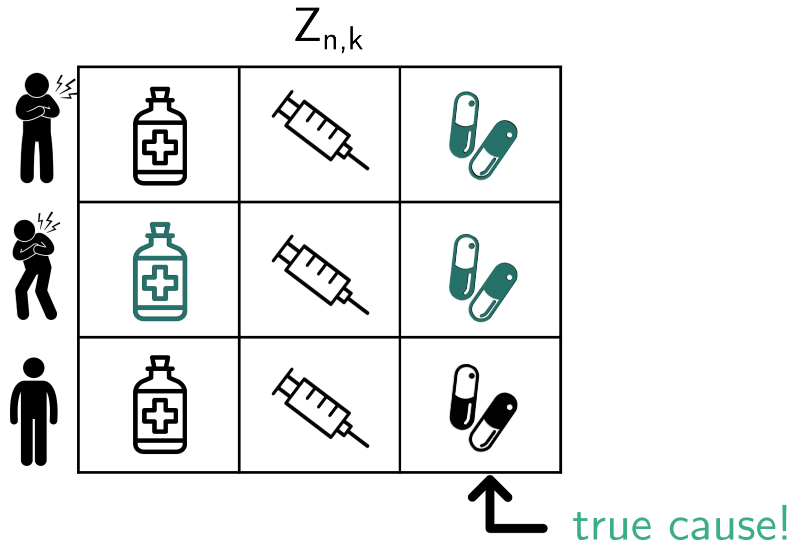
- through activations (Z) we can infer which exposures are causal for one patient and a one outcome (local inferences)



Noisy-Or Risk Allocation (NORA) Model

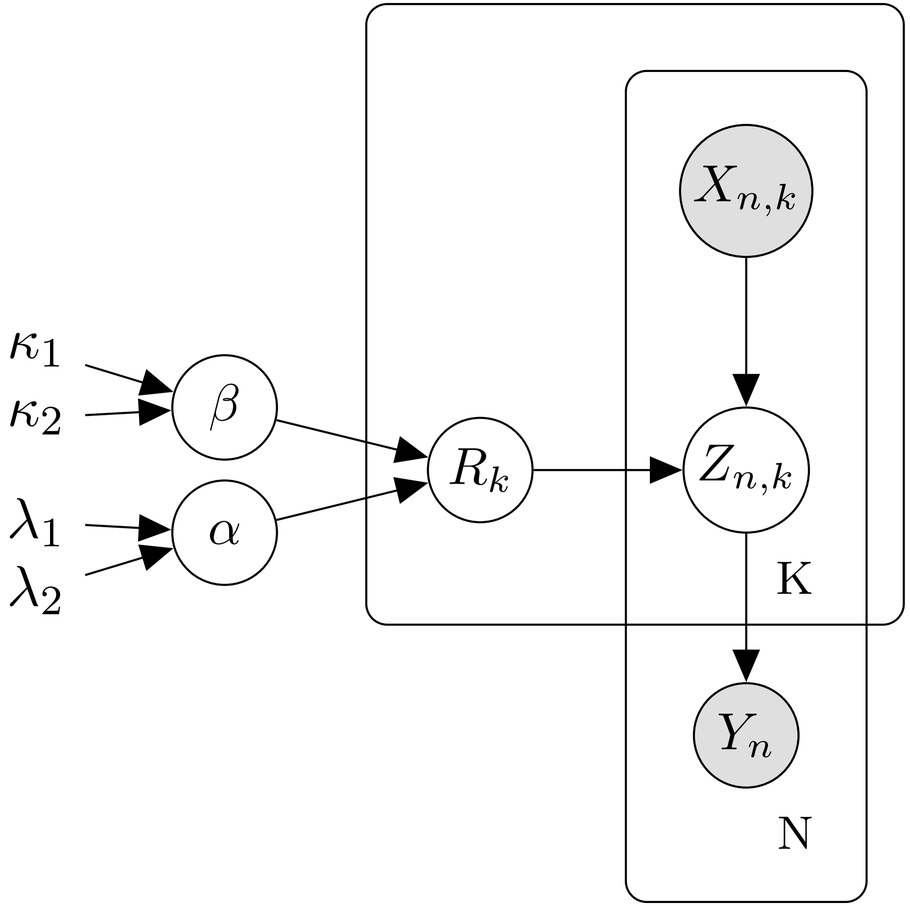


- through activations (Z) we can infer which exposures are causal for one patient and a one outcome (local inferences)

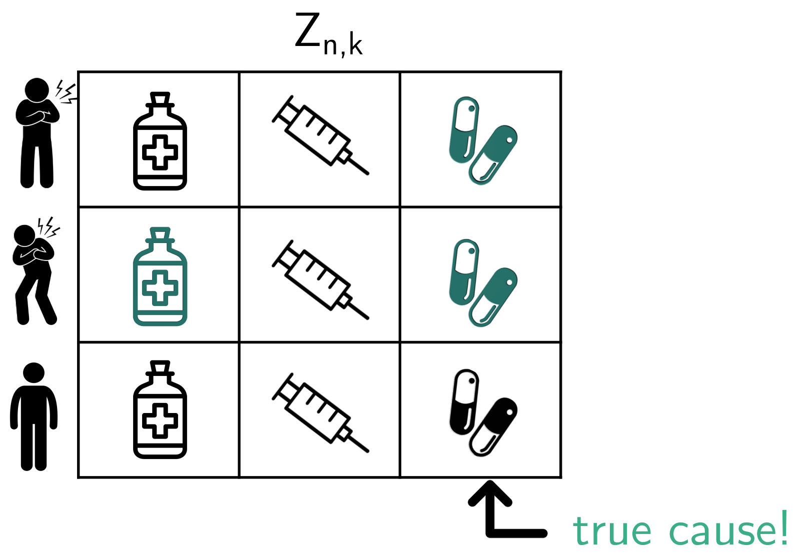


- use Z's to infer the attributable risks for an exposure causing an outcome (global inferences)

Noisy-Or Risk Allocation (NORA) Model



- through activations (Z) we can infer which exposures are causal for one patient and a one outcome (local inferences)

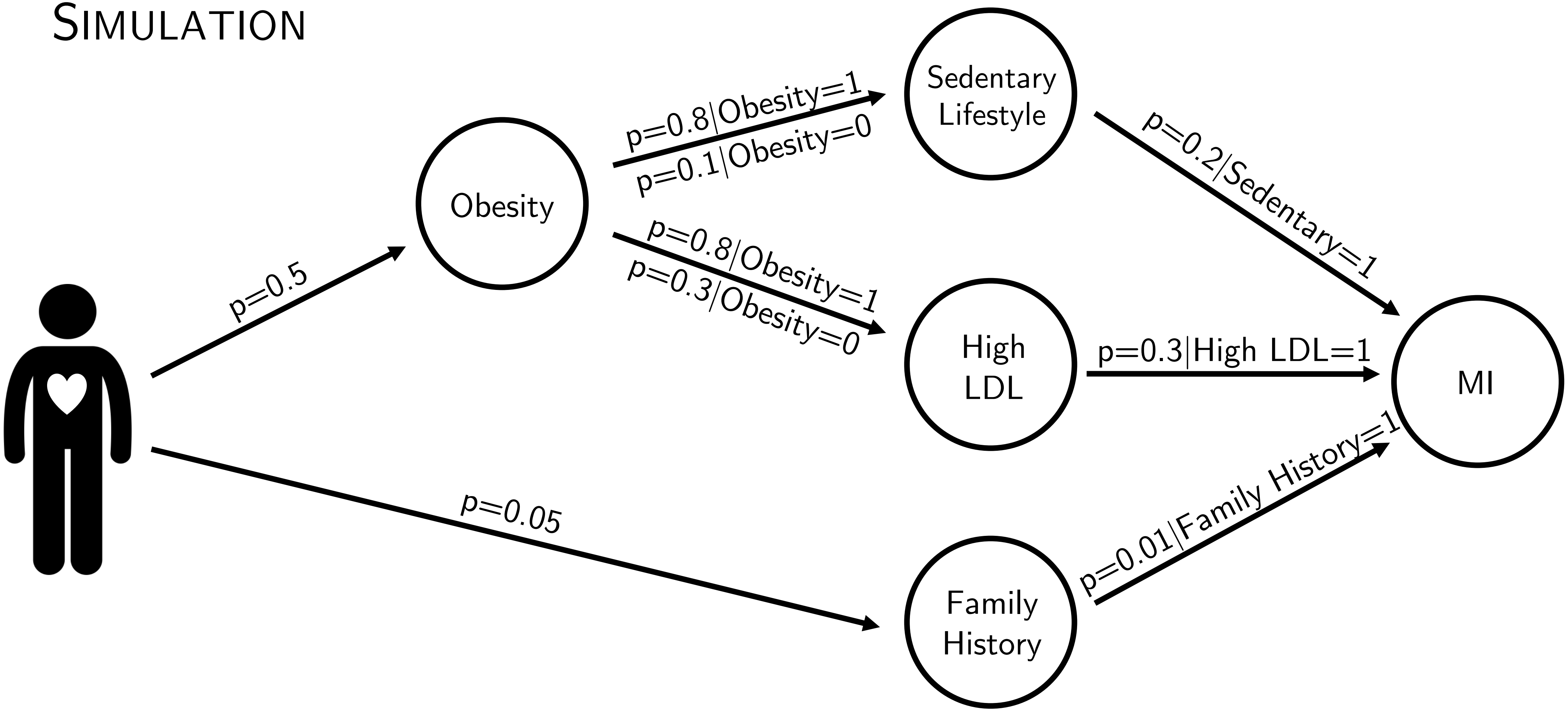


- use Z's to infer the attributable risks for an exposure causing an outcome (global inferences)
- multivariate setting accounts for confounding!

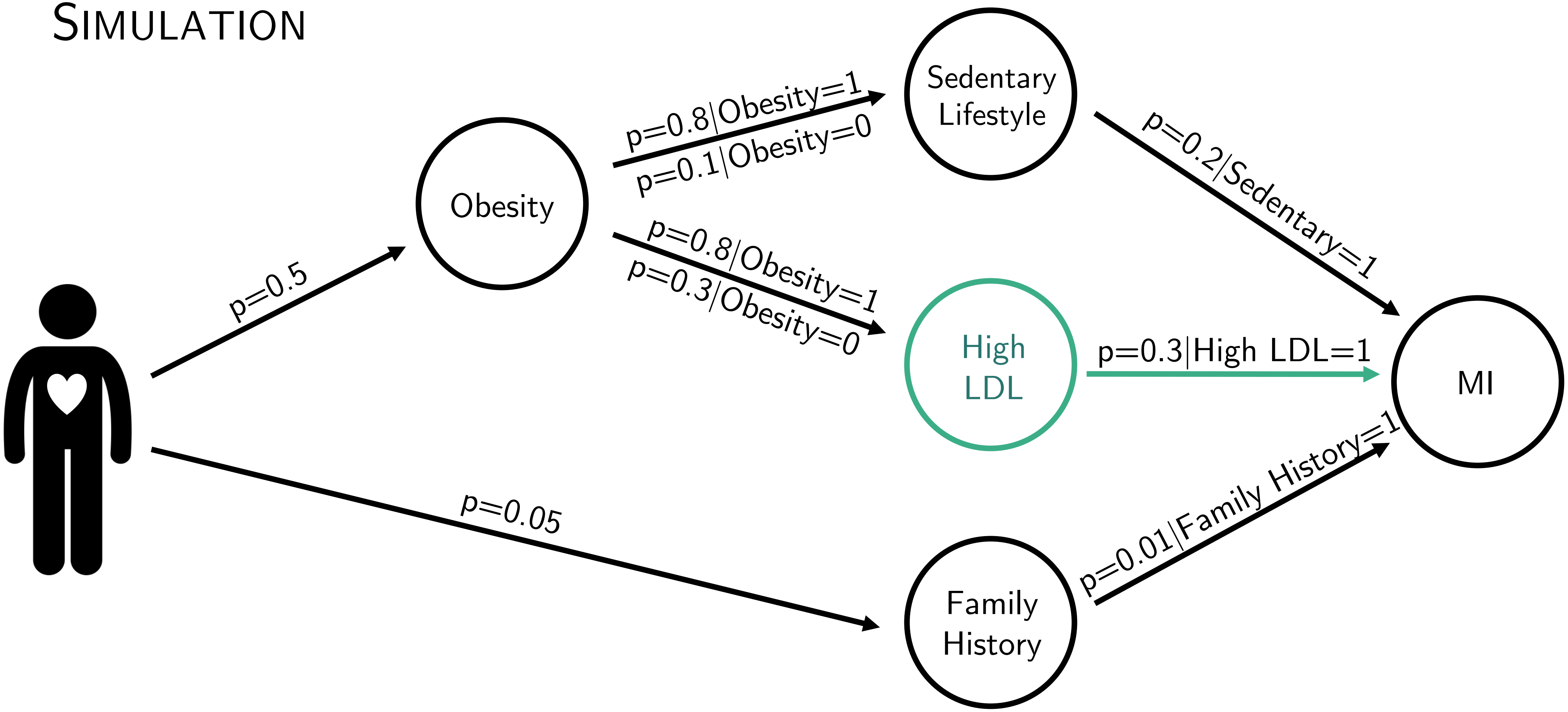
1. a simulation in which the ground truth is known
2. application to real-world clinical data

1. a simulation in which the ground truth is known
2. application to real-world clinical data

SIMULATION

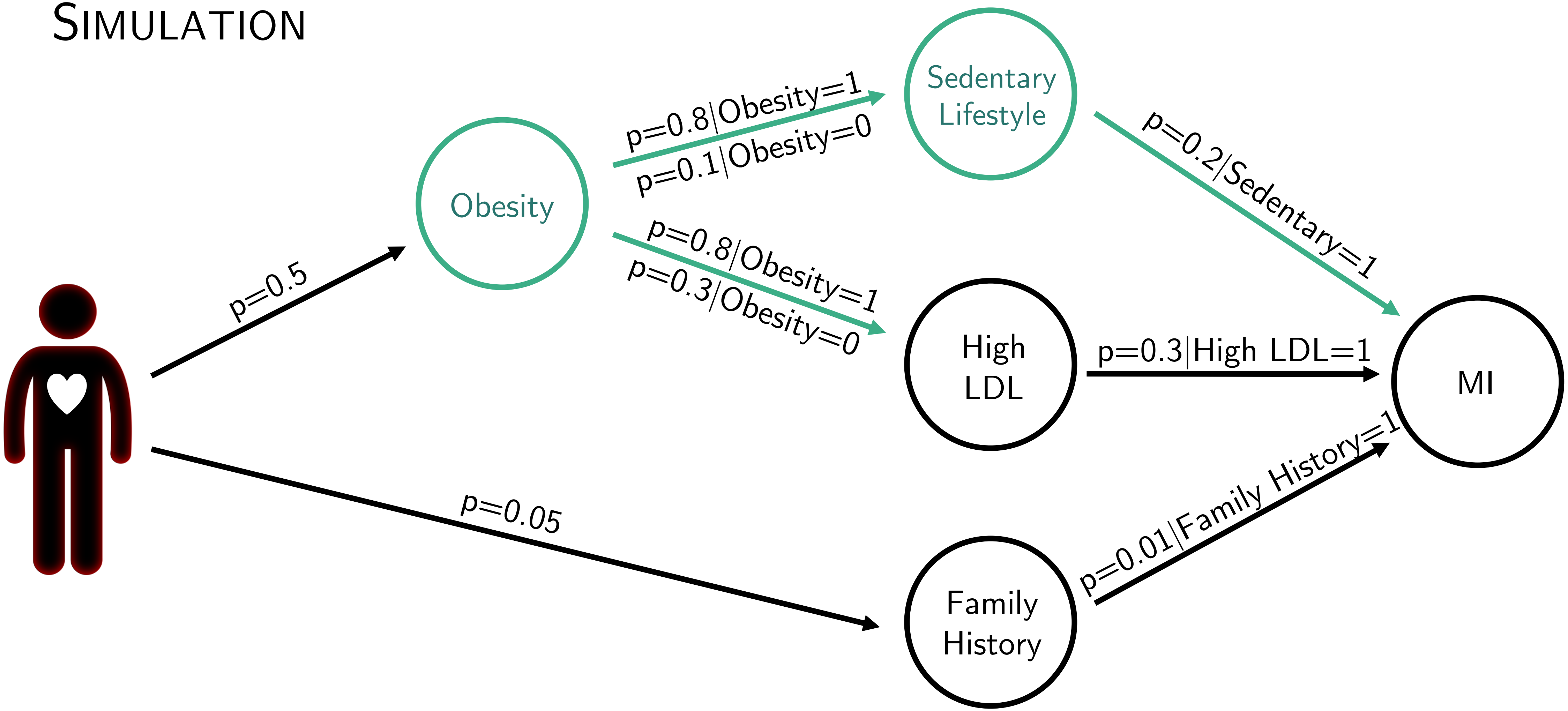

p=0.5
Obesity
p=0.8|Obesity=1
p=0.1|Obesity=0
p=0.8|Obesity=1
p=0.3|Obesity=0
Sedentary Lifestyle
High LDL
Family History
p=0.05
p=0.2|Sedentary=1
p=0.3|High LDL=1
p=0.01|Family History=1
MI

SIMULATION



Obesity
Sedentary Lifestyle
High LDL
Family History
MI
p=0.5
p=0.05
p=0.8|Obesity=1
p=0.1|Obesity=0
p=0.8|Obesity=1
p=0.3|Obesity=0
p=0.2|Sedentary=1
p=0.3|High LDL=1
p=0.01|Family History=1

SIMULATION



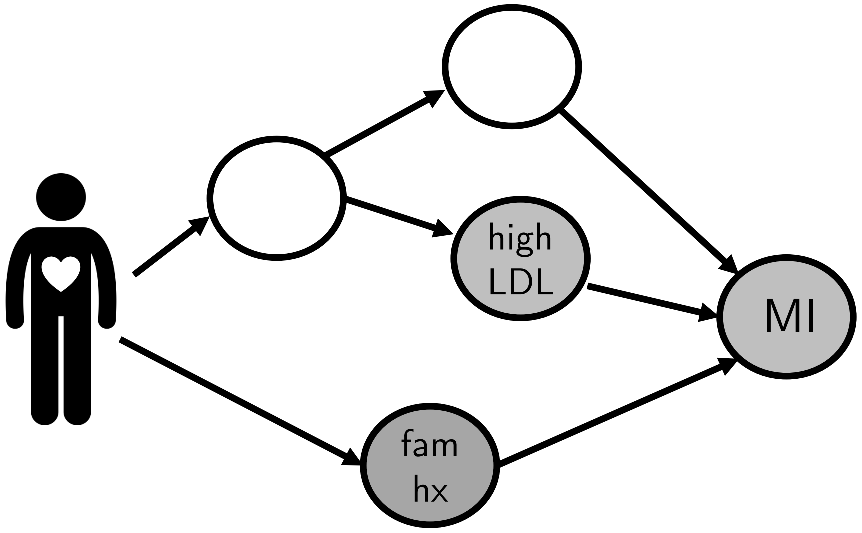
Obesity
Sedentary Lifestyle
High LDL
Family History
MI
p=0.5
p=0.05
p=0.8|Obesity=1
p=0.1|Obesity=0
p=0.8|Obesity=1
p=0.3|Obesity=0
p=0.2|Sedentary=1
p=0.3|High LDL=1
p=0.01|Family History=1

Experimentation

Simulation → Evaluation

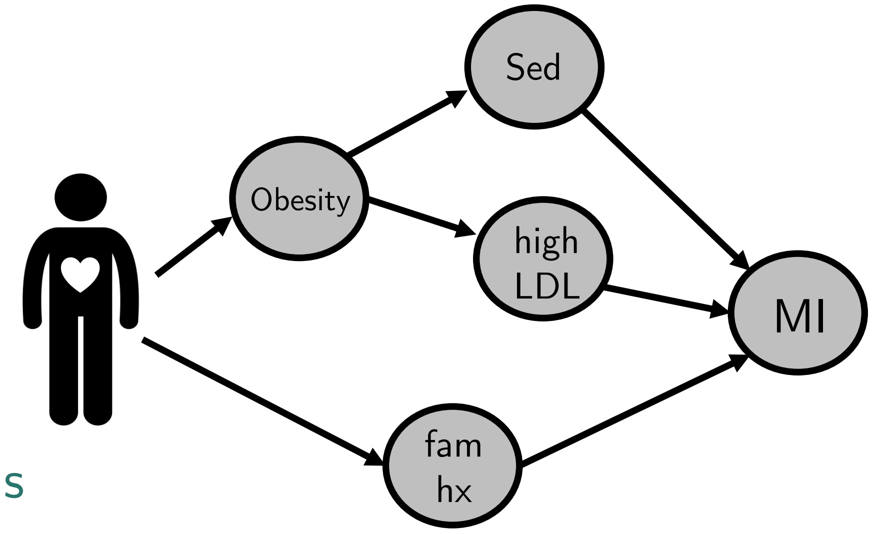
Censored Model

p(MI | high LDL, family hx)



Uncensored Model

p(MI | high LDL, obesity, sedentary, family hx)



1 correctness of implementation

2 robustness to bias

SIMULATION → RESULTS

1. correctness of implementation → recover the true attributable risk of High LDL (p=0.30)

Risk of MI from high LDL

	NORA	LR
Censored Model	0.21	0.44
Uncensored Model	0.30	0.50

← ground truth = 0.30

SIMULATION → RESULTS

1. correctness of implementation → recover the true attributable risk of High LDL (p=0.30)

Risk of MI from high LDL

	NORA	LR
Censored Model	0.21	0.44
Uncensored Model	0.30	0.50

← ground truth = 0.30

2. robustness to bias → how close are estimates of High LDL attributable risks to the ground truth when backdoor paths are not observed?

Bias in high LDL estimate

	NORA	LR
% bias Censored vs Truth	30.5%	46.6%
% bias Uncensored vs Truth	1.3%	66.7%

← less biased when backdoor paths are *unobserved*

← less biased when backdoor paths are *observed*

1. a simulation in which the ground truth is known
2. application to real-world clinical data

APPLICATION TO CLINICAL DATA

1. create noisy, observational cohorts to investigate a single outcome
2. apply NORA and comparators to learn ARs of exposures for outcome
   - Levin 1953 calculation
   - disproportionality analyses (MGPS-EGBM & RR)
   - penalized logistic regression (L1)

exposure-outcome relationship	outcome
condition – condition	heart failure kidney disease renal impairment disorder of the spleen Kaposi sarcoma glaucoma
procedure – condition	mucositis hypothyroidism
drug – adverse drug reaction	hearing loss mucositis

Application to Clinical Data → Evaluation

APPLICATION TO CLINICAL DATA → EVALUATION

demonstrate global inference of risk over high AR exposures from NORA versus the high AR exposures from the comparators

- compare AR estimates from literature vs estimates from NORA and comparators ⟶ Kaposi Sarcoma

APPLICATION TO CLINICAL DATA → EVALUATION

demonstrate global inference of risk over high AR exposures from NORA versus the high AR exposures from the comparators

- compare AR estimates from literature vs estimates from NORA and comparators → Kaposi Sarcoma

demonstrate local inference of outcome over AR estimates from NORA and comparators

- predictive performance (AUROC) of NORA vs L1 and comparators on a held-out dataset

demonstrate local inference of risk over high AR exposures from NORA for an individual.

- choose a patient with outcome, simulate activations using exposures and risks learned from model → Heart Failure

Application to Clinical Data → Results → global inference of risk

Kaposi sarcoma - a rare type of cancer that is most commonly seen in the HIV/AIDs patients, but is occasionally seen in HIV-negative populations of severely immunocompromised patients. Angeletti, 2008

	NORA	L1	Levin 1953	RR	MGPS-EBGM
Human immunodeficiency virus infection	0.0070	0.7872	0.2022	0.9566	0.9603

APPLICATION TO CLINICAL DATA → RESULTS → global inference of risk

KAPOSI SARCOMA - a rare type of cancer that is most commonly seen in the HIV/AIDs patients, but is occasionally seen in HIV-negative populations of severely immunocompromised patients. Angeletti, 2008

	NORA	L1	Levin 1953	RR	MGPS-EBGM	Gold-Standard Liu, 2015
Human immunodeficiency virus infection	0.0070	0.7872	0.2022	0.9566	0.9603	0.0048

Application to Clinical Data → Results → global inference of risk

Kaposi sarcoma - a rare type of cancer that is most commonly seen in the HIV/AIDs patients, but is occasionally seen in HIV-negative populations of severely immunocompromised patients. Angeletti, 2008

	NORA	L1	Levin 1953	RR	MGPS-EBGM	Gold-Standard Liu, 2015
Human immunodeficiency virus infection	0.0070	0.7872	0.2022	0.9566	0.9603	0.0048

APPLICATION TO CLINICAL DATA → RESULTS → local inference of outcome

Outcome
disseminated intravascular coagulation
glaucoma
hearing loss
heart failure
Kaposi sarcoma
mucositis (exposures drugs)
renal impairment
splenomegaly
hypothyroidism
mucositis (exposures procedures)

APPLICATION TO CLINICAL DATA → RESULTS → local inference of outcome

Outcome	AUROC NORA	AUROC L1
disseminated intravascular coagulation	0.8878	0.7773
glaucoma	0.7017	0.6999
hearing loss	0.5056	0.6329
heart failure	0.8030	0.7953
Kaposi sarcoma	0.8011	0.5624
mucositis (exposures drugs)	0.5291	0.6560
renal impairment	0.8170	0.7965
splenomegaly	0.6248	0.5000
hypothyroidism	0.5643	0.6349
mucositis (exposures procedures)	0.5829	0.6134

Application to Clinical Data → Results → local inference of outcome

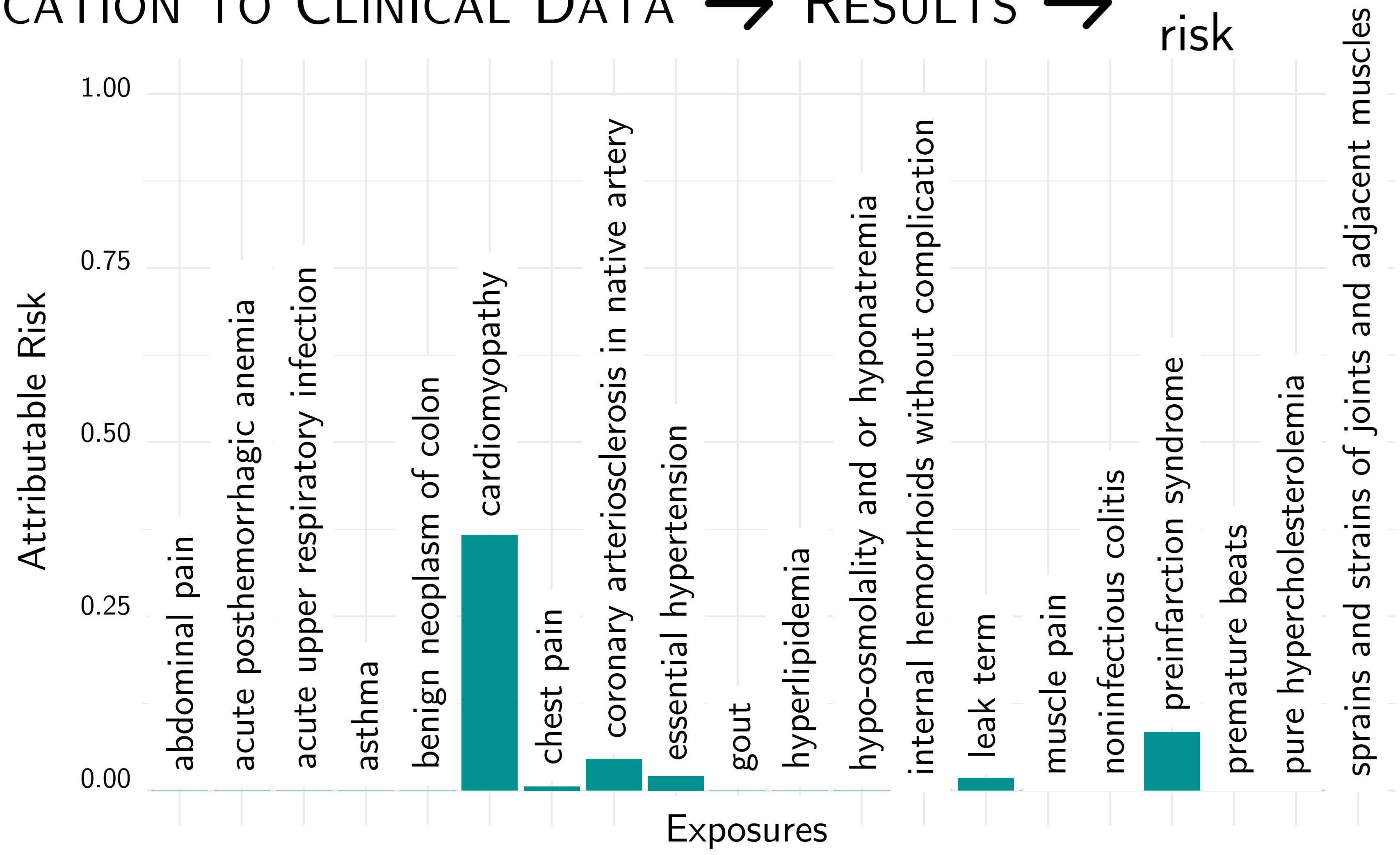
Outcome	AUROC NORA	AUROC L1
disseminated intravascular coagulation	0.8878	0.7773
glaucoma	0.7017	0.6999
hearing loss	0.5056	0.6329
heart failure	0.8030	0.7953
Kaposi sarcoma	0.8011	0.5624
mucositis (exposures drugs)	0.5291	0.6560
renal impairment	0.8170	0.7965
splenomegaly	0.6248	0.5000
hypothyroidism	0.5643	0.6349
mucositis (exposures procedures)	0.5829	0.6134

*exposures for a single patient with heart failure

Application to Clinical Data → Results → local inference of risk

*exposures for a single patient with heart failure

Application to Clinical Data → Results → local inference of risk



Significance

- NORA identifies a cohesive set of high-AR risk factors that have reasonable estimates of risk.
- Simulations suggest that the model may be more robust to confounding than logistic regression.
- NORA support global and local inferences, which helps us rectify care of a single patient with public-health.

Significance

- NORA identifies a cohesive set of high-AR risk factors that have reasonable estimates of risk.
- Simulations suggest that the model may be more robust to confounding than logistic regression.
- NORA support global and local inferences, which helps us rectify care of a single patient with public-health.

Limitations

- violation of causal independence
- we make certain distributional choices
- may learn patterns in EHR documentation rather than true causes
    - complicated by assumption that timestamps are correct
- unseen confounders

Funding

This research is supported by grants R01LM009886-10 and T15LM007079 from The National Library of Medicine.

Thank you. Questions?

amelia.averitt@regeneron.com

http://ajaveritt.com

https://www.linkedin.com/in/ameliajaveritt

@AJAveritt