

Rで統計学の勉強をするためのデータ生成

調整が回帰分析の係数に与える影響を、
数式とDAGを見ながら実験してみる



自己紹介

本で行うことのイメージ

医学（疫学）研究のよくある形/本で行うこと
95%信頼区間の実験

DAGの基本

DAGとは/DAGは研究と関係ある？

先ほどの重回帰分析をDAGで表してみる/DAGのパス

Colliderの特徴/Colliderの特徴の例

他のDAGのパスの特徴：中間変数

他のDAGのパスの特徴：交絡因子

他のDAGのパスの特徴：Colliderの調整2

課題：Rで次の事象が発生するかを確認してみてください

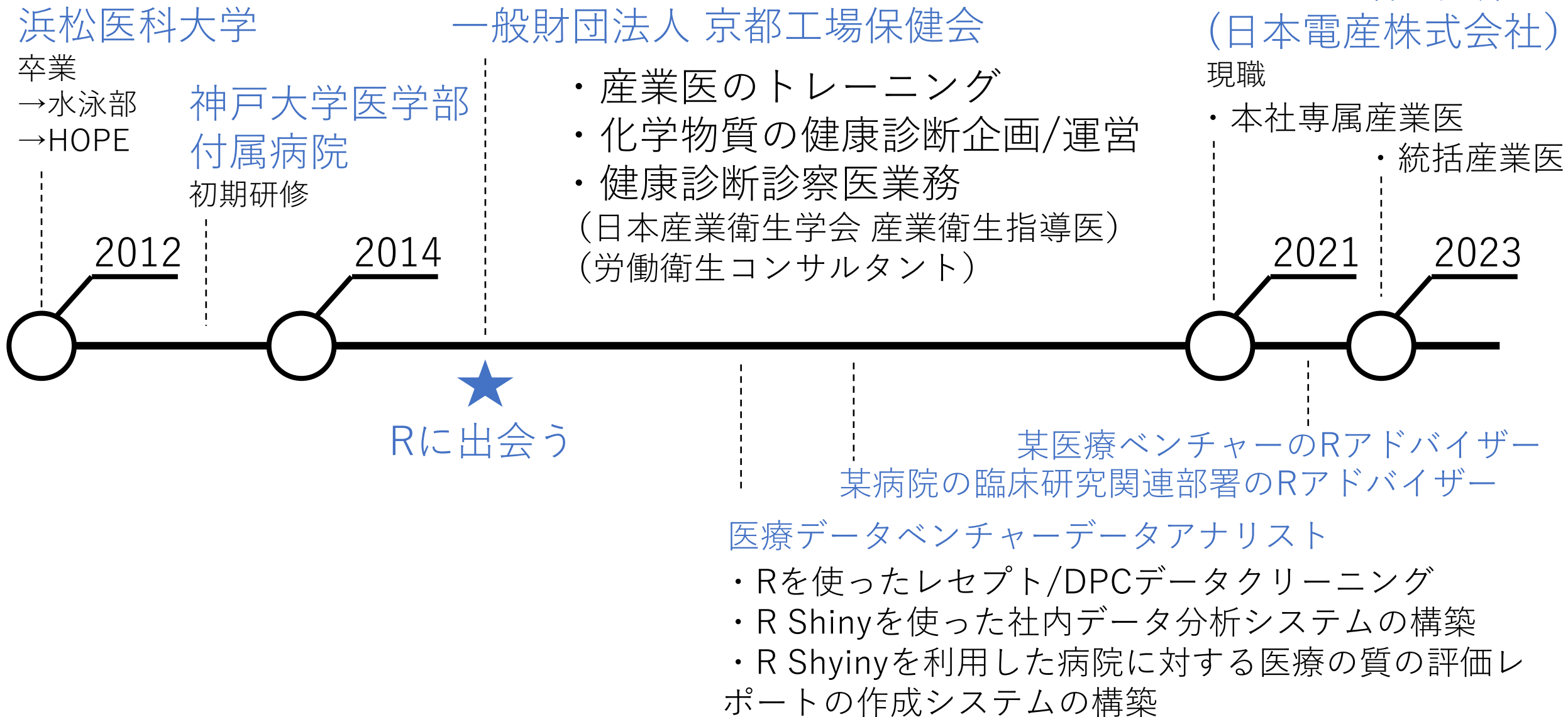
バックドアパス

DAGは「どの変数を調整するか」という議論で有用

時間が余れば

喫煙と出生体重のパラドックス

医師/産業医としての経歴

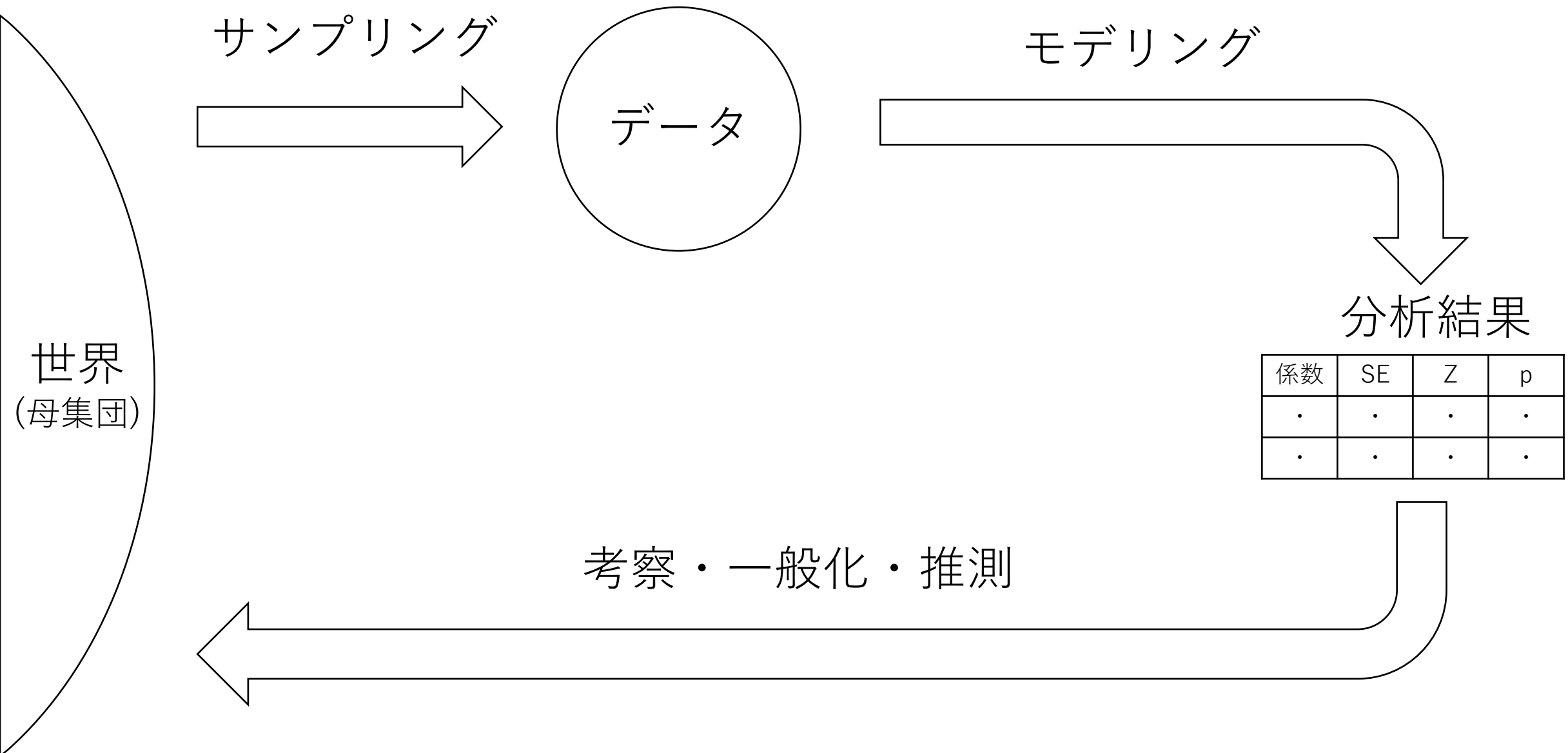


Rユーザーとしての経歴 全て兼業

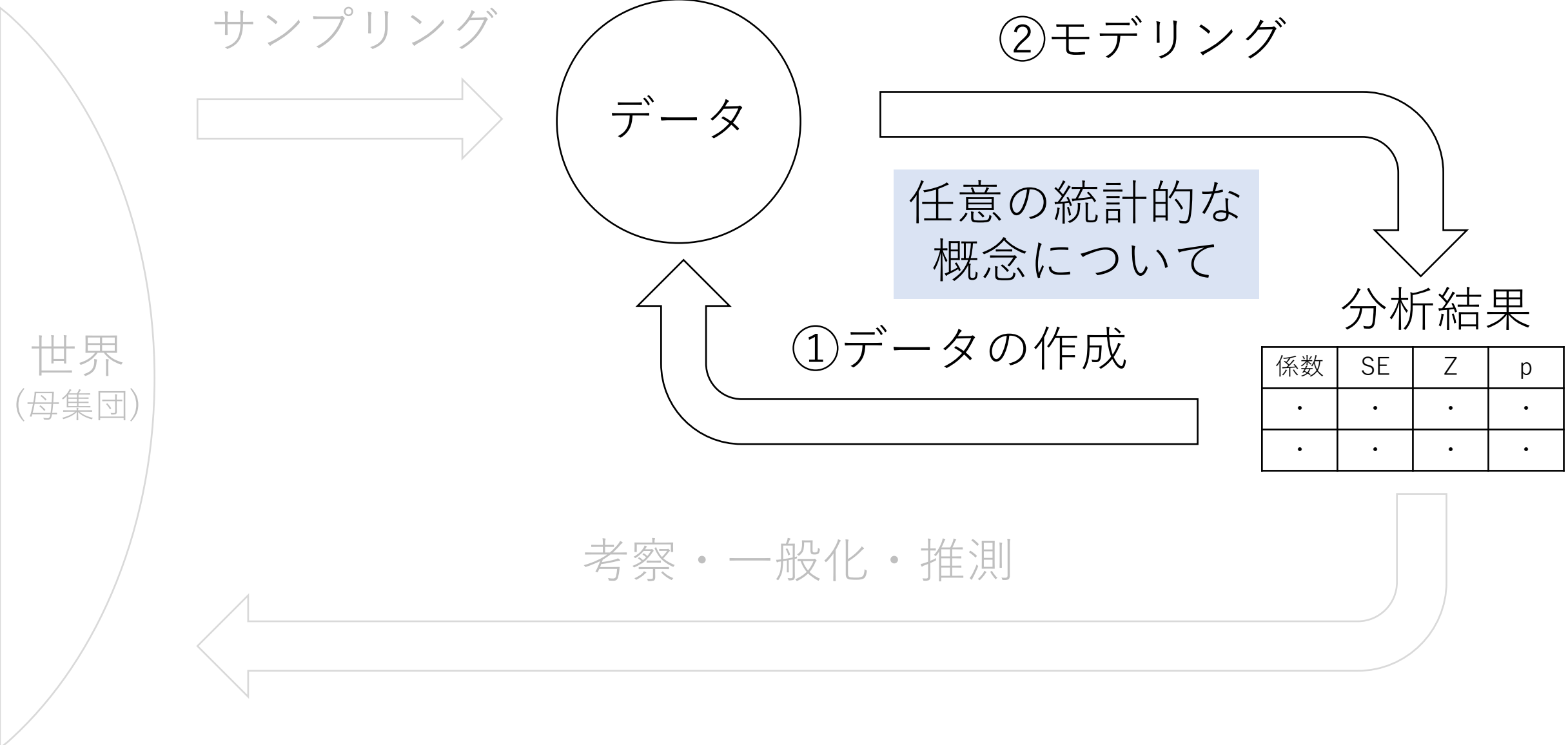
他、Rに関する書籍の出版など

本で行うことのイメージ

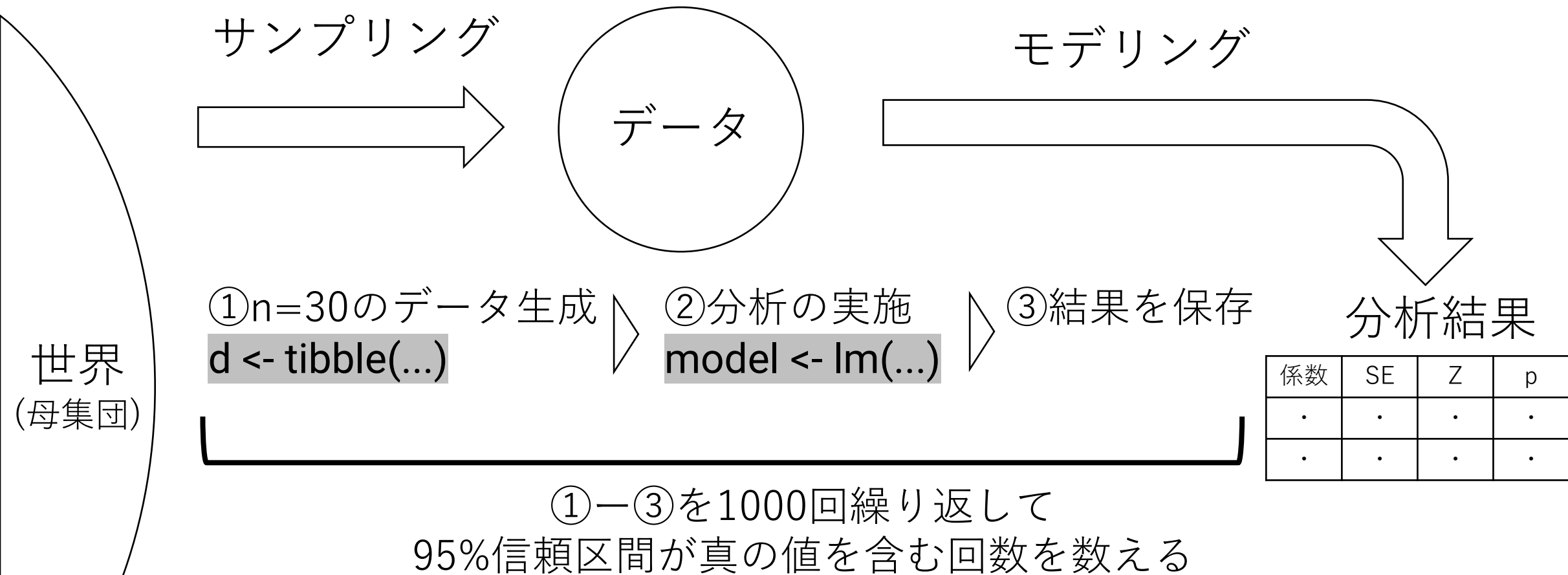
医学（疫学）研究のよくある形



本で行うこと



95%信頼区間の実験

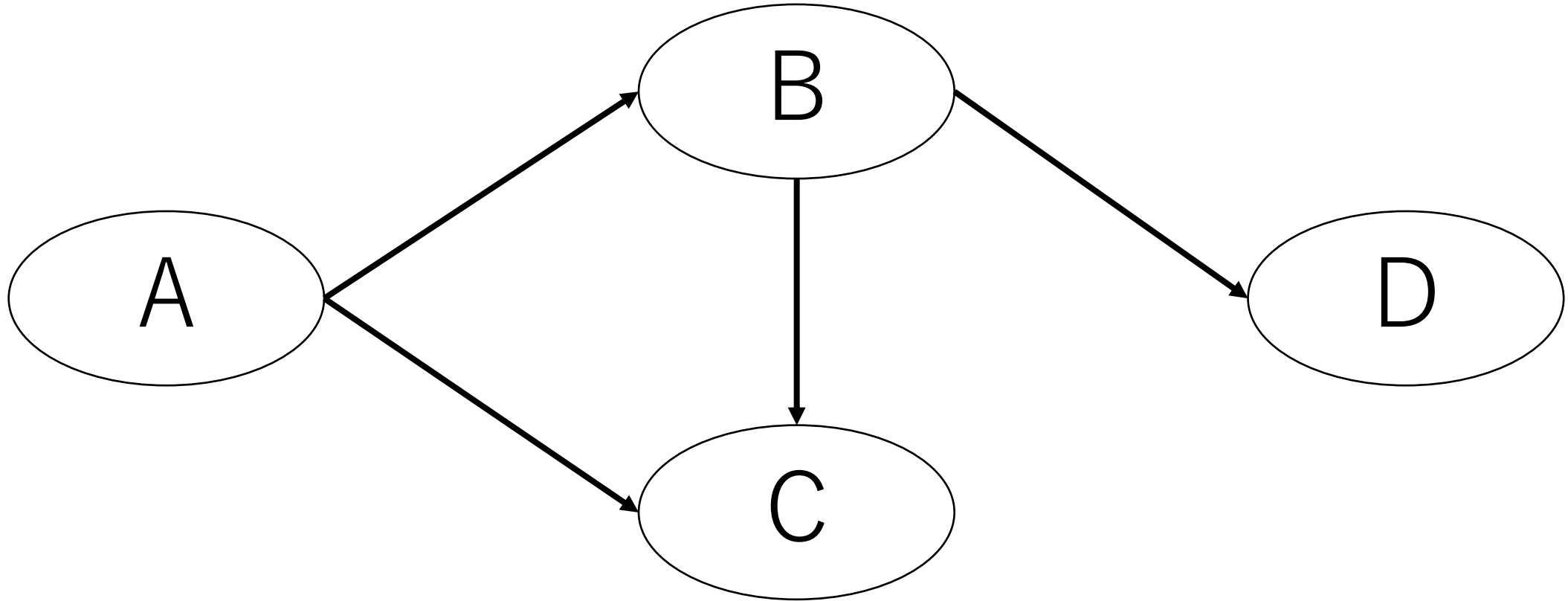


真のモデル： $Y = 20 + X_1 + 2X_2 + 3X_3 + \epsilon$

DAGの基本

DAGとは

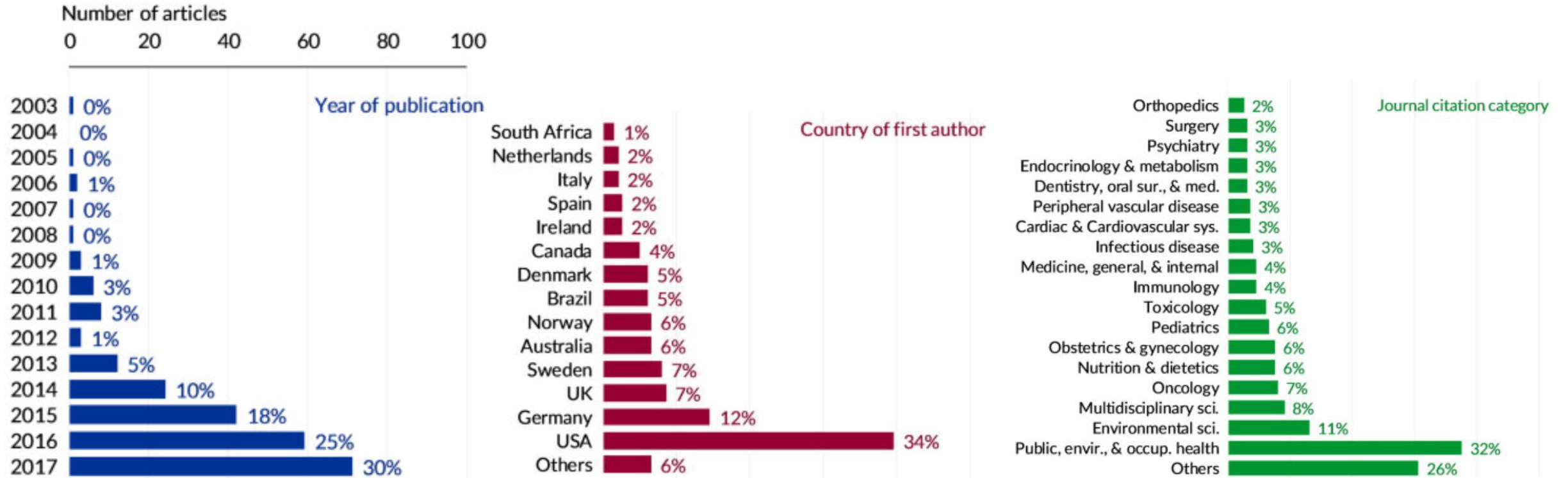
DAG : Directed Acyclic Graph (有向非巡回グラフ)



有向：ノード (○) からエッジ (→)

非巡回：同じノードにエッジをたどって戻ってこない

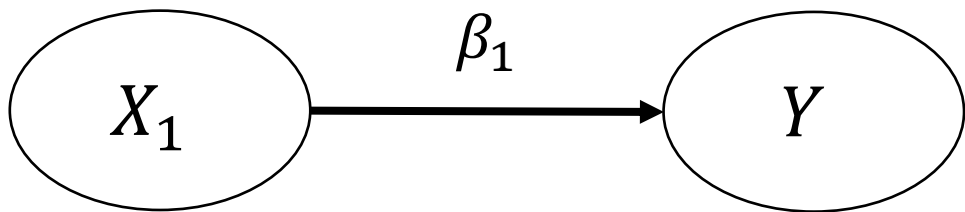
DAGは研究と関係ある？



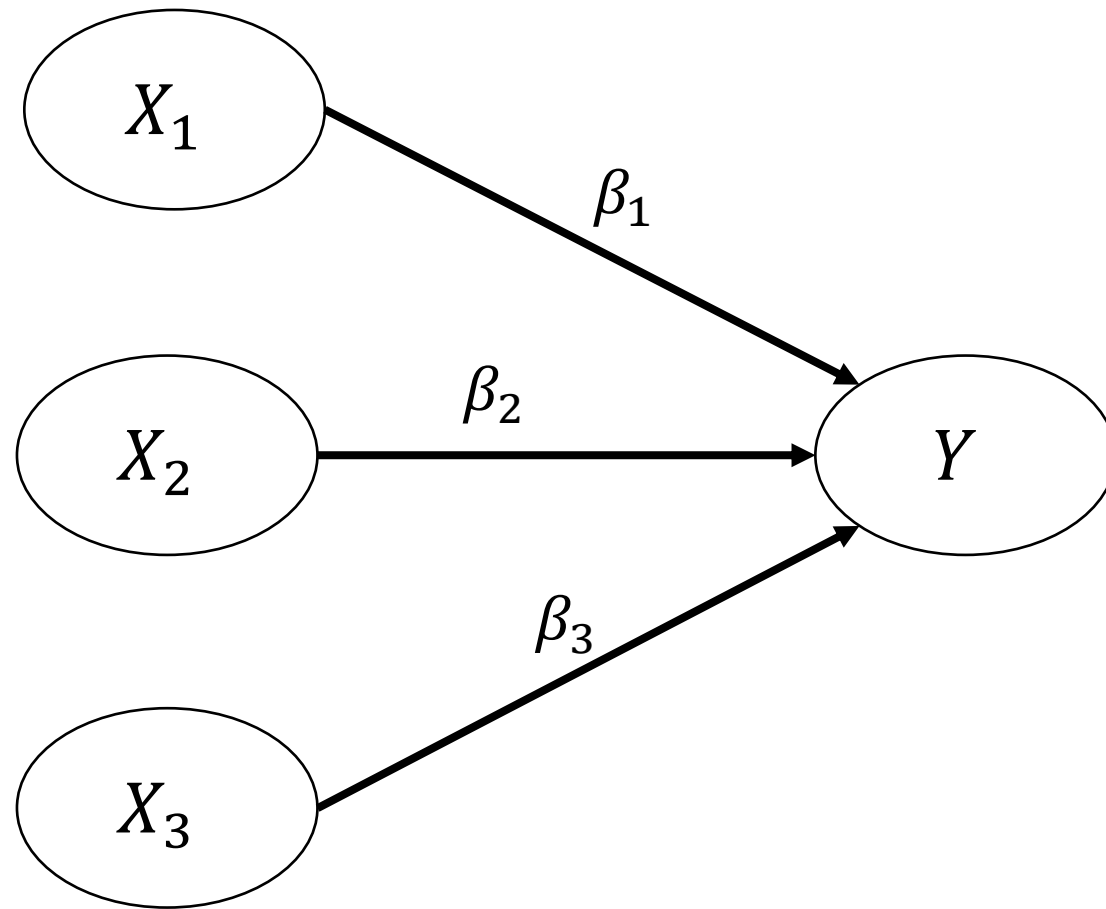
観察研究で因果関係について言及する場合に利用されることがある
DAGに基づいて分析の仮説などを明瞭に共有することができる

先ほどの重回帰分析をDAGで表してみる

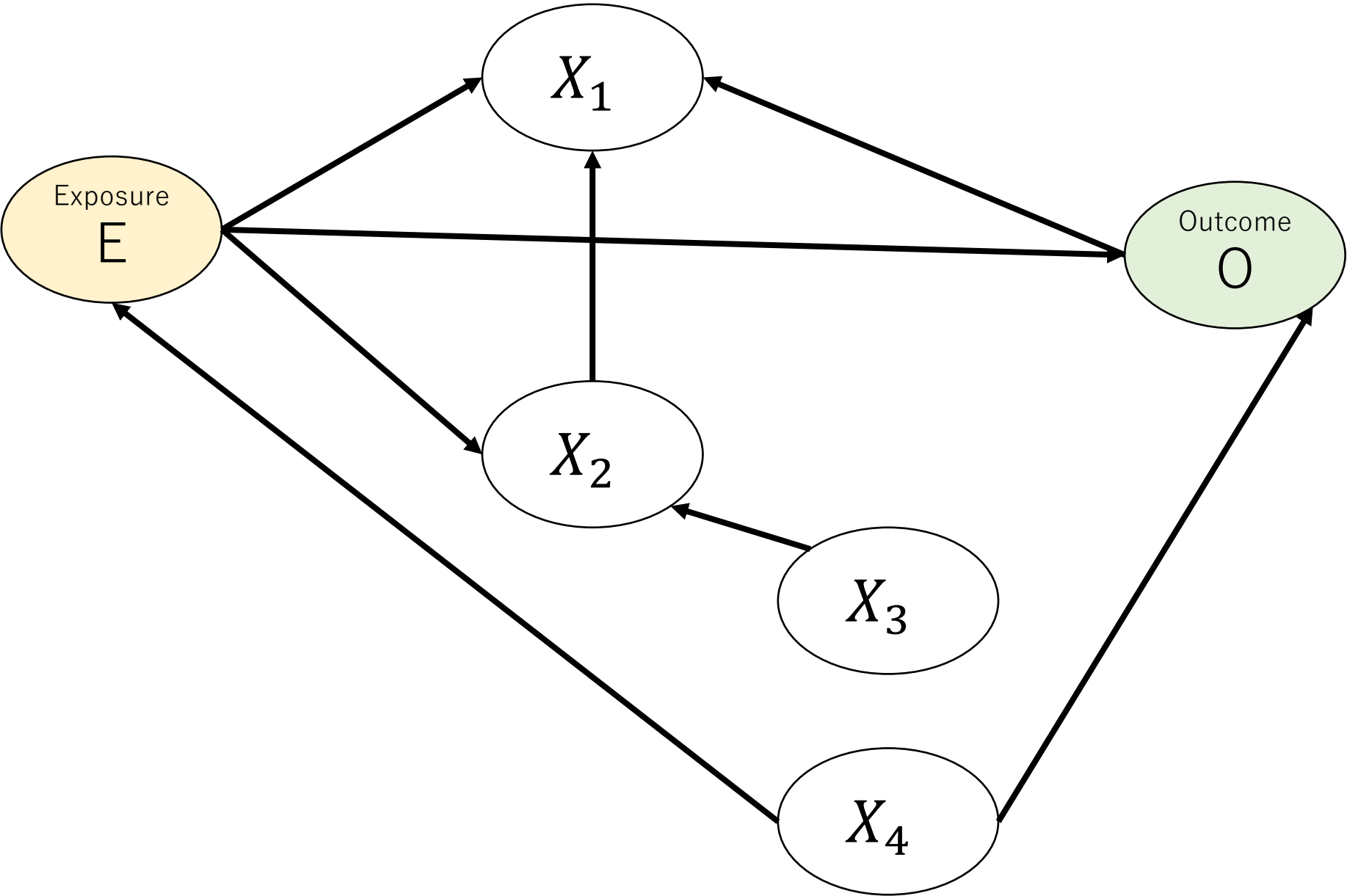
モデル 1 : $Y = \beta_0 + \beta_1 X_1 + \epsilon$



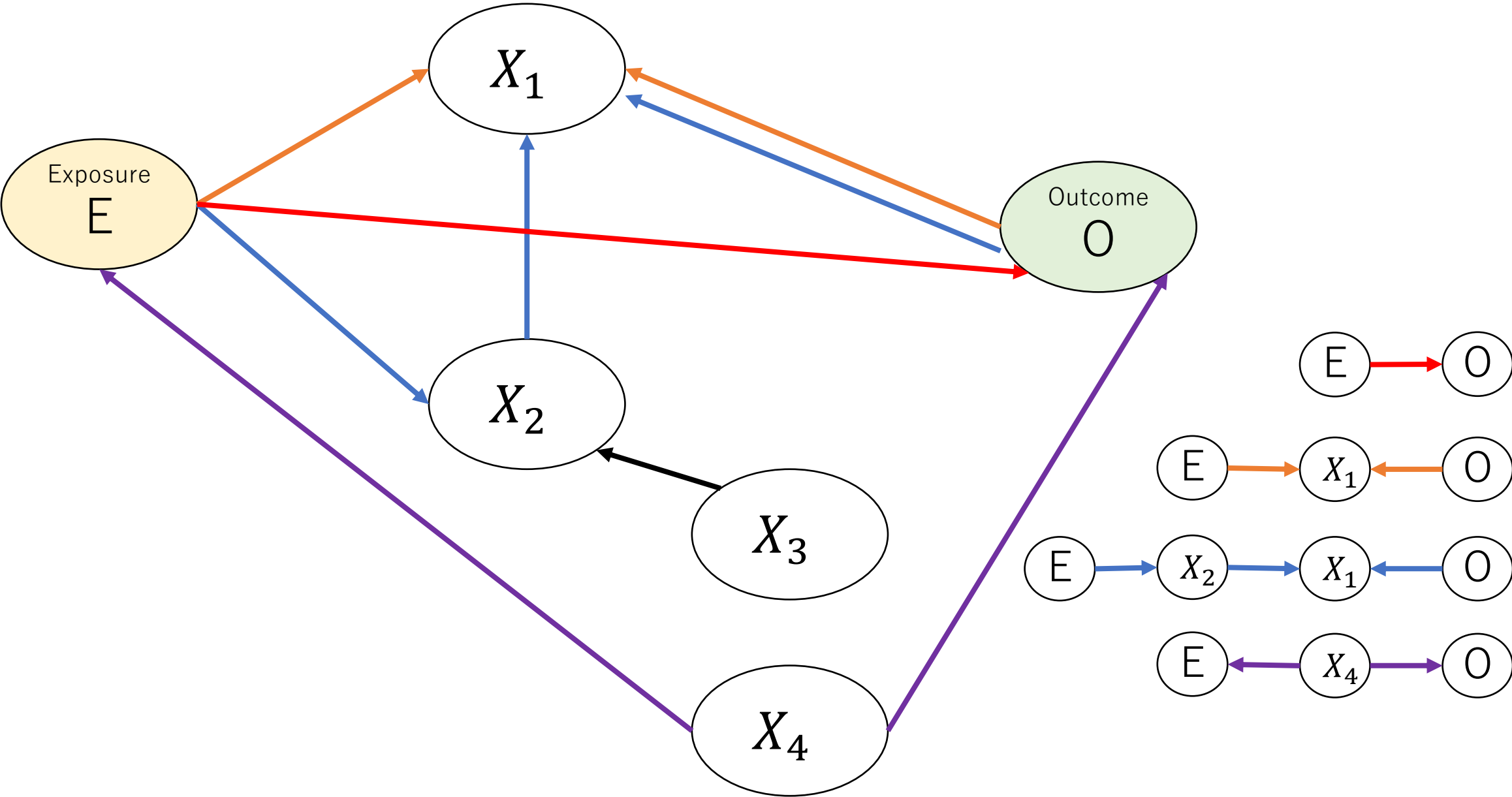
モデル 2 : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$



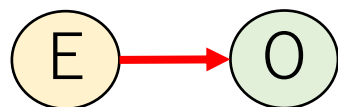
DAGのパス



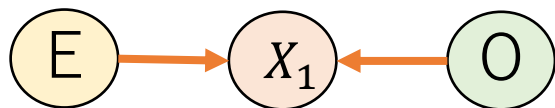
DAGのパス



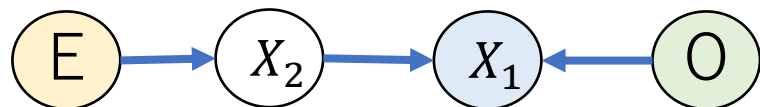
DAGのパス



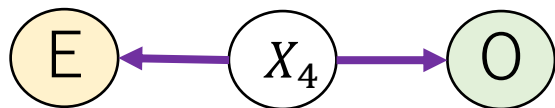
EからOのパスは開いている



$E \rightarrow X_1 \leftarrow O$ はColliderとなっており、EからOのパスは閉じている



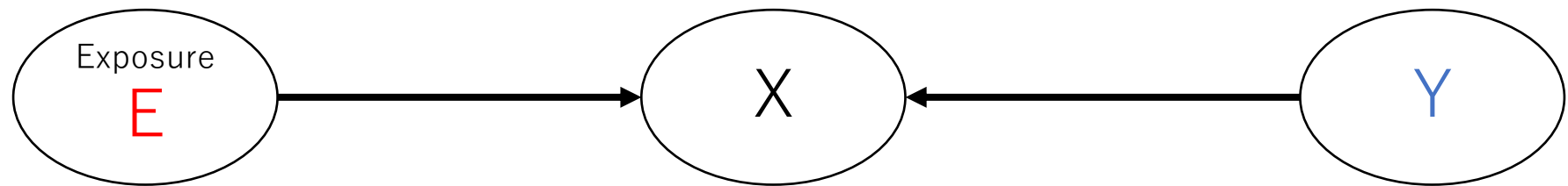
$X_2 \rightarrow X_1 \leftarrow O$ がColliderとなっており、EからOのパスは閉じている



$E \leftarrow X_4 \rightarrow O$ は開いたパスであるため、EからOのパスは開いている

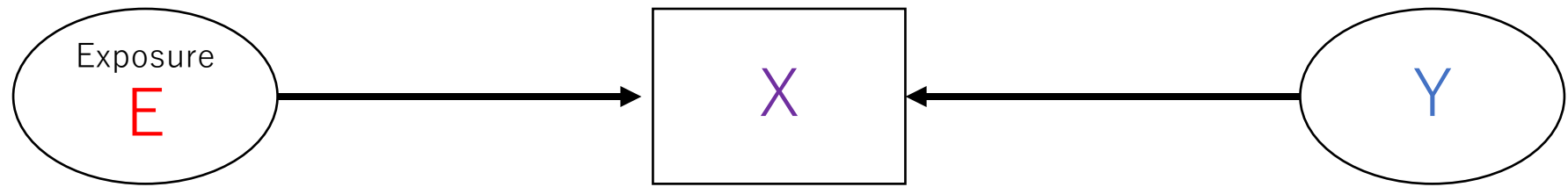
Colliderの特徴

EとYの間にColliderがあれば、EとYのパスが閉じている



$$Y = \beta_0 + \beta_1 E + \epsilon \qquad \rightarrow \beta_1 \text{ は } 0$$

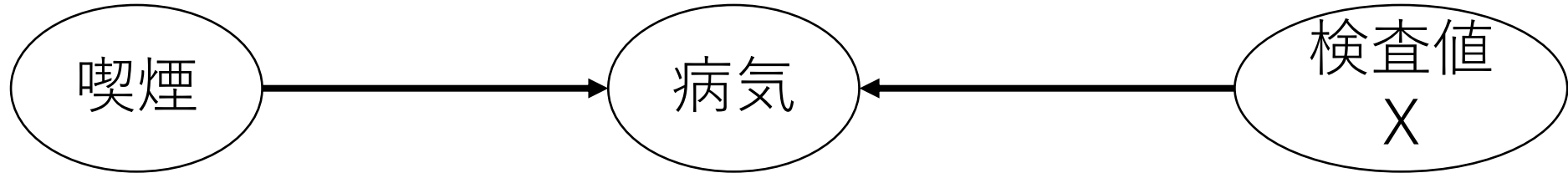
Colliderを調整（□で表現）すれば、パスが開く



$$Y = \beta_0 + \beta_1 E + \beta_2 X + \epsilon \qquad \rightarrow \beta_1 \text{ が } 0 \text{ 以外}$$

Colliderの特徴の例

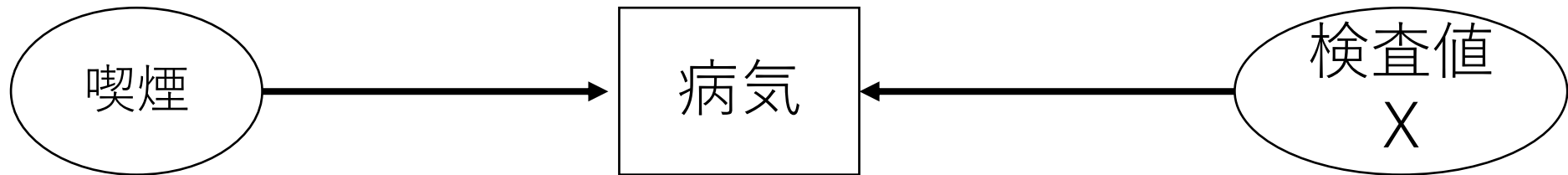
喫煙をするとある病気が発症する。検査値Xが病気を引き起こす指標



$$\text{検査値}X = \beta_0 + \beta_1 \text{喫煙} + \epsilon \quad \rightarrow \beta_1 \text{は} 0$$

正しい結論：喫煙と検査値には関係がない

ただ、もし病気をモデルに入れて（調整して）分析すると・・・



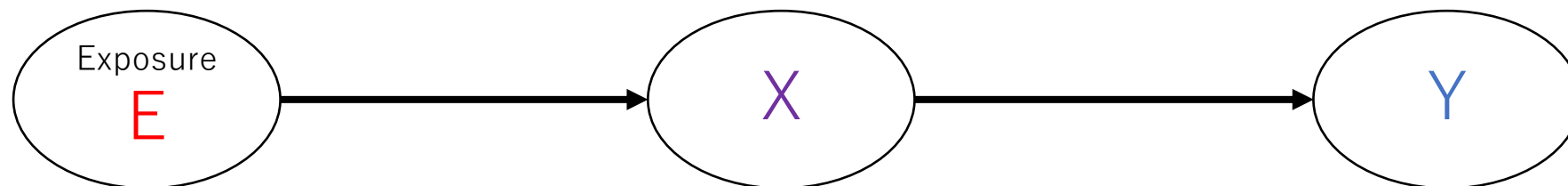
$$\text{検査値}X = \beta_0 + \beta_1 \text{喫煙} + \beta_2 \text{病気} + \epsilon \quad \rightarrow \beta_1 \text{が} 0 \text{以外}$$

誤った結論：喫煙すると検査値が変化する

Rで見てみましょう！

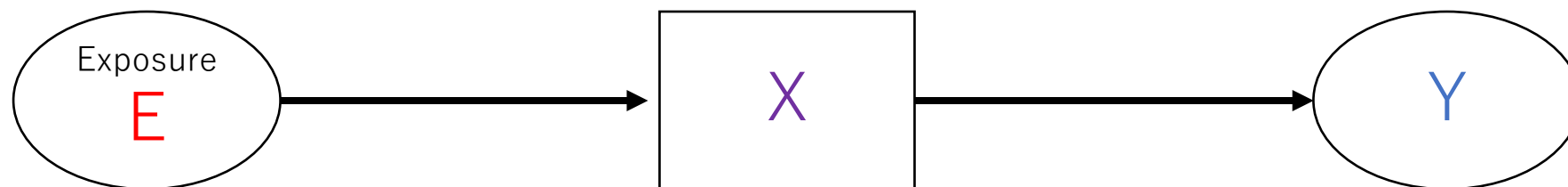
他のDAGのパスの特徴：中間変数

次のように矢印が伸びている場合は、EからYへのパスは開いている
Xは中間変数



$$Y = \beta_0 + \beta_1 E + \epsilon \quad \rightarrow \beta_1 \text{ は } 0 \text{ 以外}$$

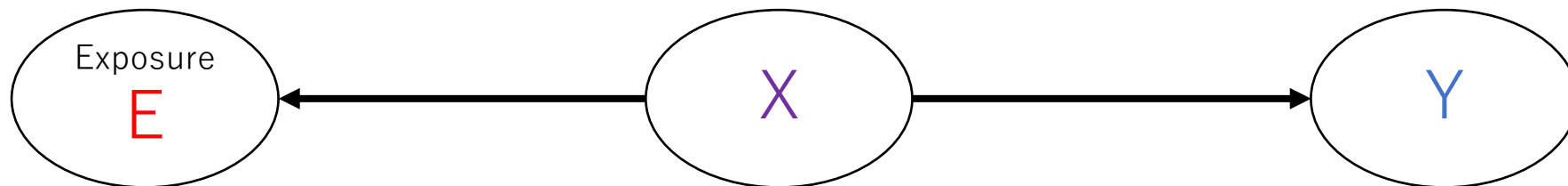
中間変数を調整すれば、パスが閉じる



$$Y = \beta_0 + \beta_1 E + \beta_2 X + \epsilon \quad \rightarrow \beta_1 \text{ が } 0$$

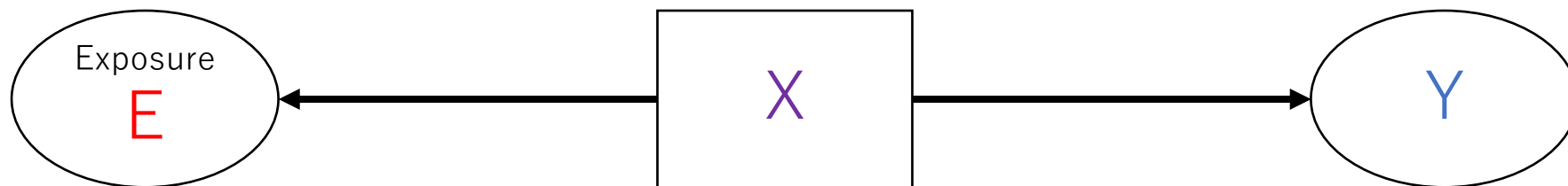
他のDAGのパスの特徴：交絡因子

次のように矢印が伸びている場合は、EからYへのパスは開いている
Xは交絡因子



$$Y = \beta_0 + \beta_1 E + \epsilon \quad \rightarrow \beta_1 \text{ は } 0 \text{ 以外}$$

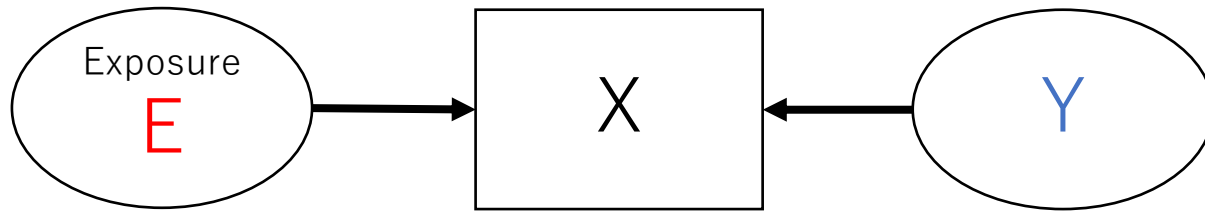
交絡因子を調整すれば、パスが閉じる



$$Y = \beta_0 + \beta_1 E + \beta_2 X + \epsilon \quad \rightarrow \beta_1 \text{ が } 0$$

他のDAGのパスの特徴：Colliderの調整2

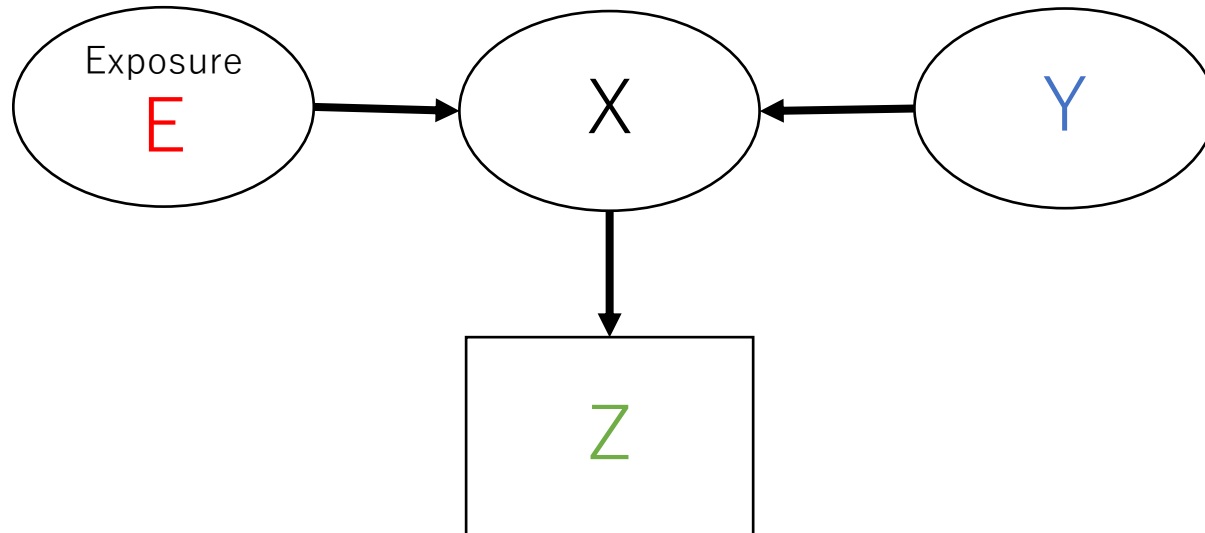
Colliderの調整はパスを開く



$$Y = \beta_0 + \beta_1 E + \beta_2 X + \epsilon$$

→ β_1 は0以外

Colliderから伸びているパスの調整はパスを開く

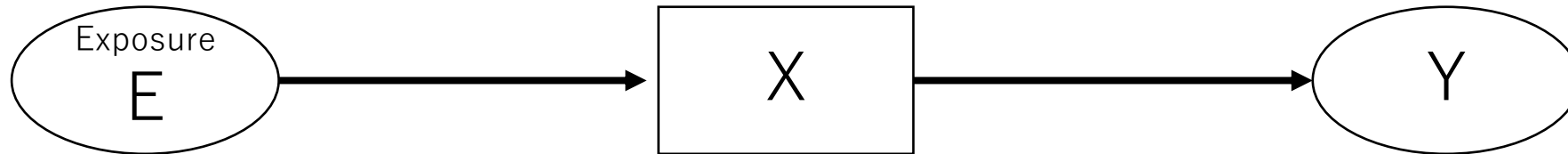


$$Y = \beta_0 + \beta_1 E + \beta_2 Z + \epsilon$$

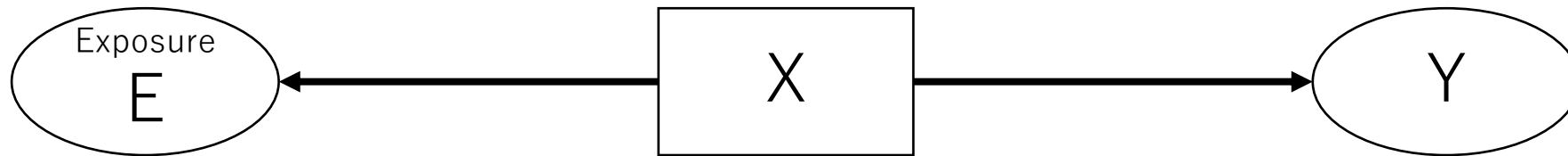
→ β_1 は0以外

課題：Rで次の事象が発生するかを確認してみてください

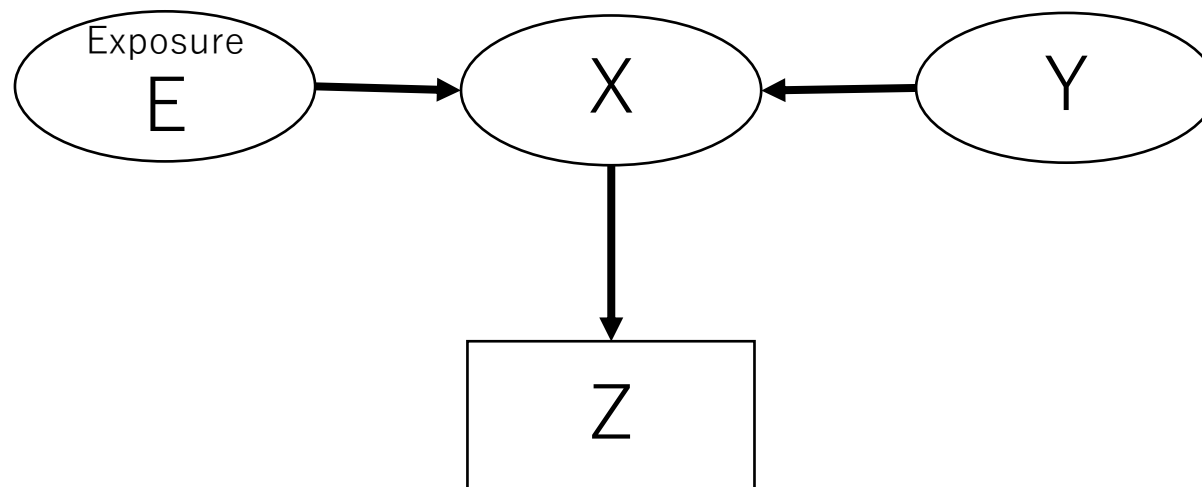
1 中間変数を調整すれば、パスが閉じる



2 交絡因子を調整すれば、パスが閉じる

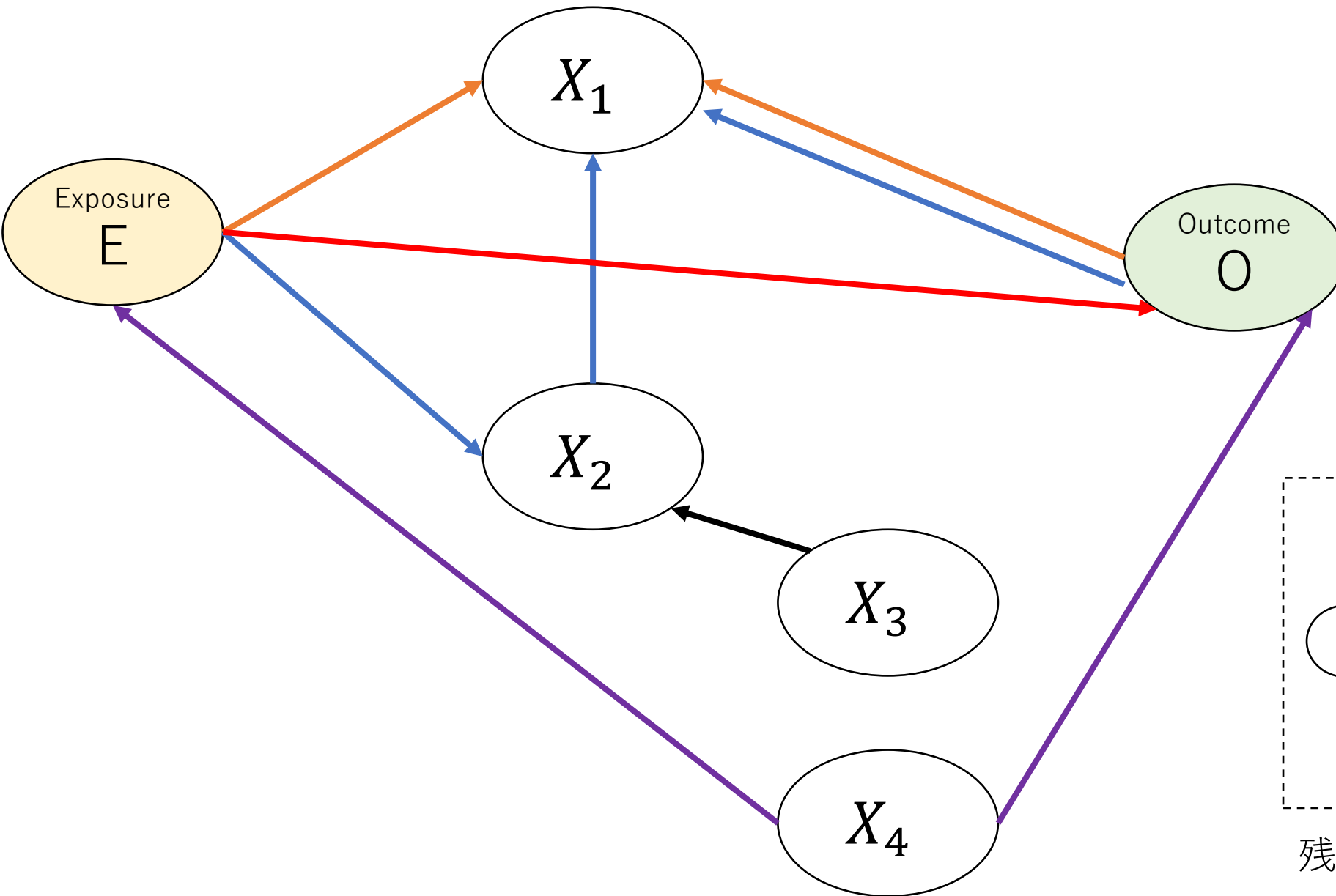


3 Colliderから伸びているパスの調整はパスを開く

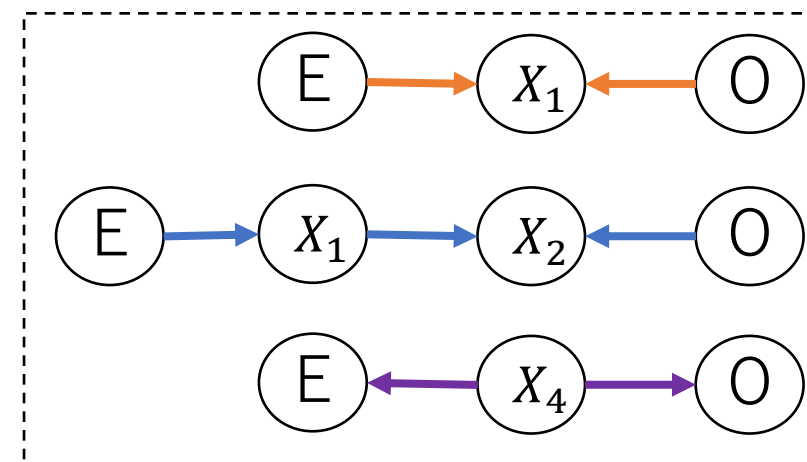


バックドアパス

DAGは「どの変数を調整するか」という議論で有用

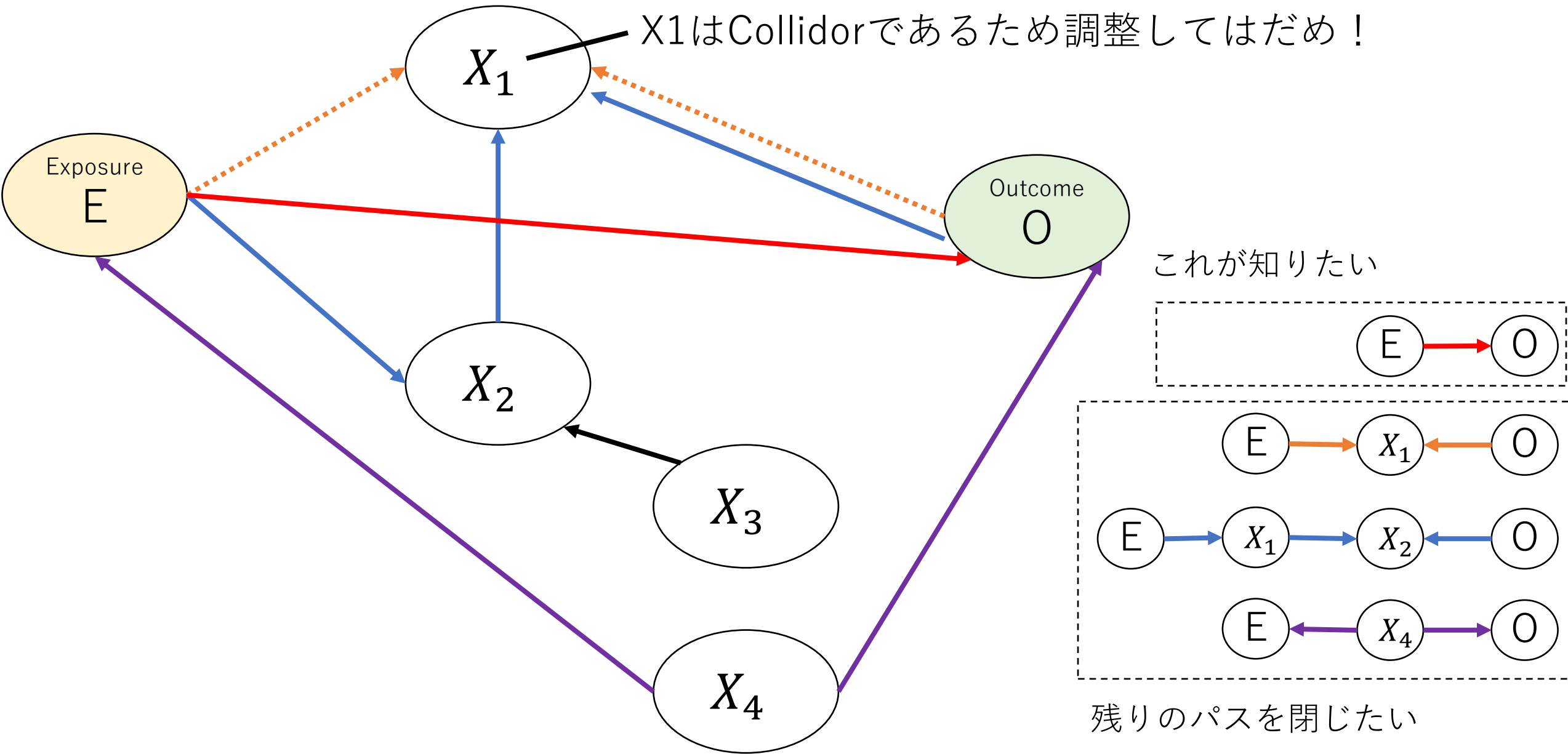


これが知りたい

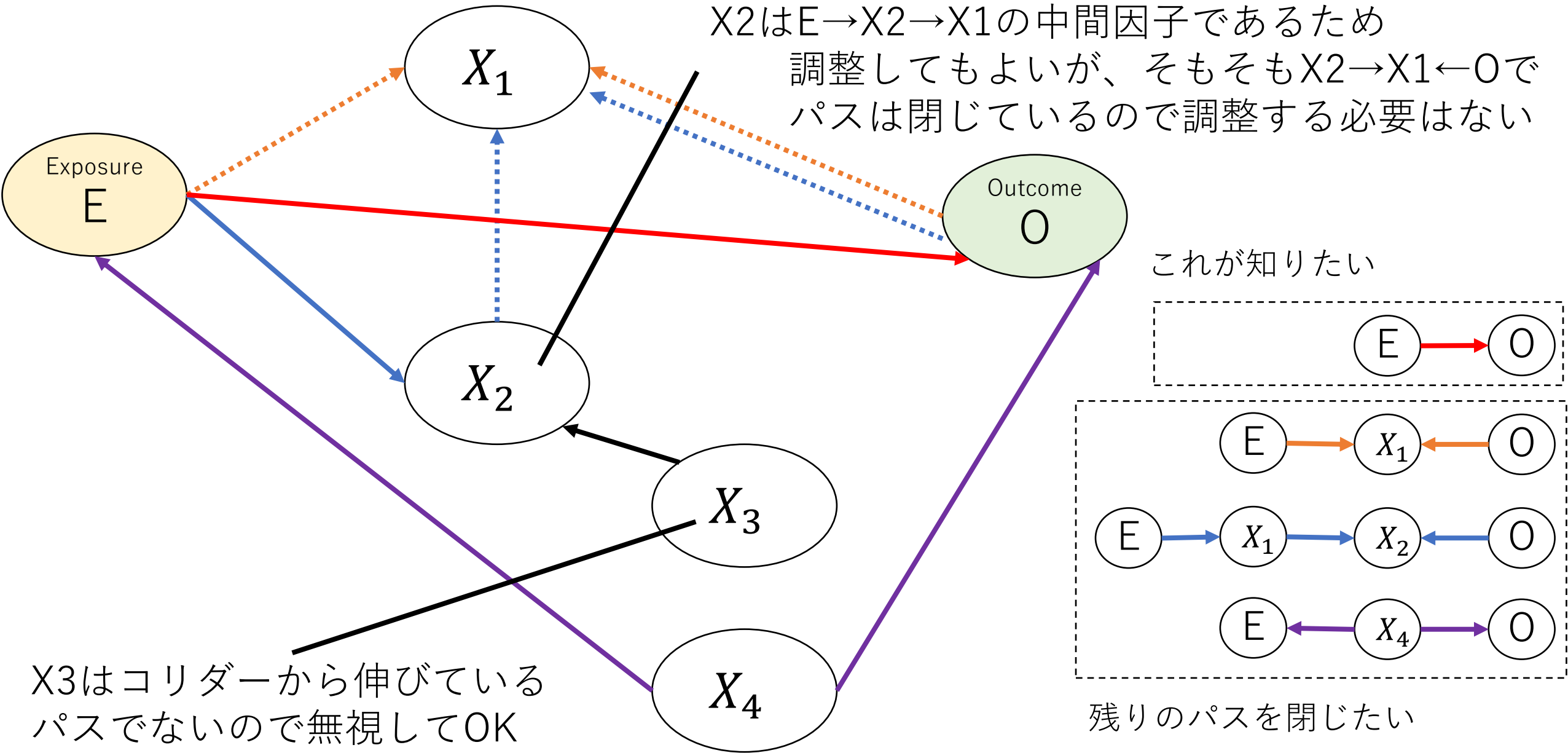


残りのパスを閉じたい

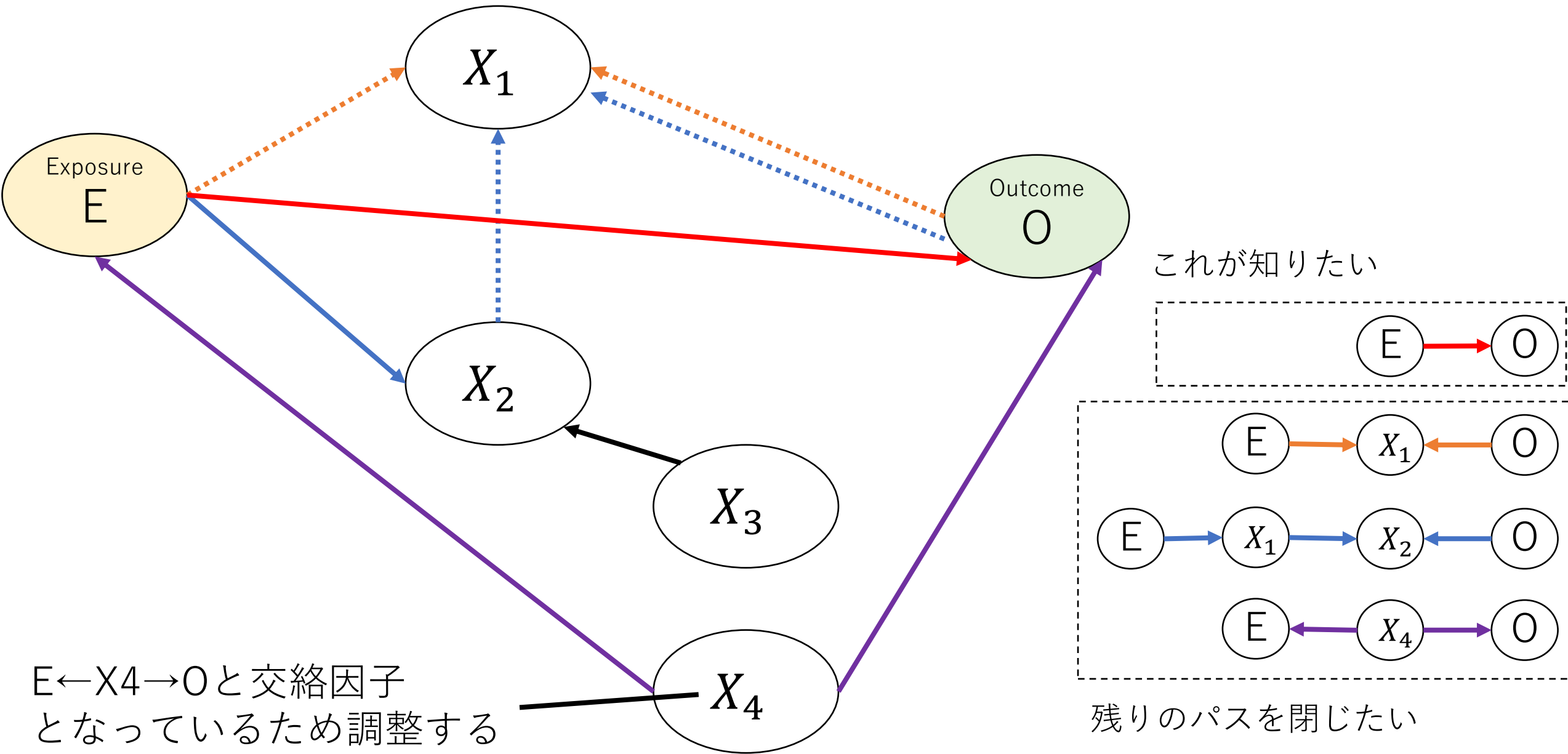
DAGは「どの変数を調整するか」という議論で有用



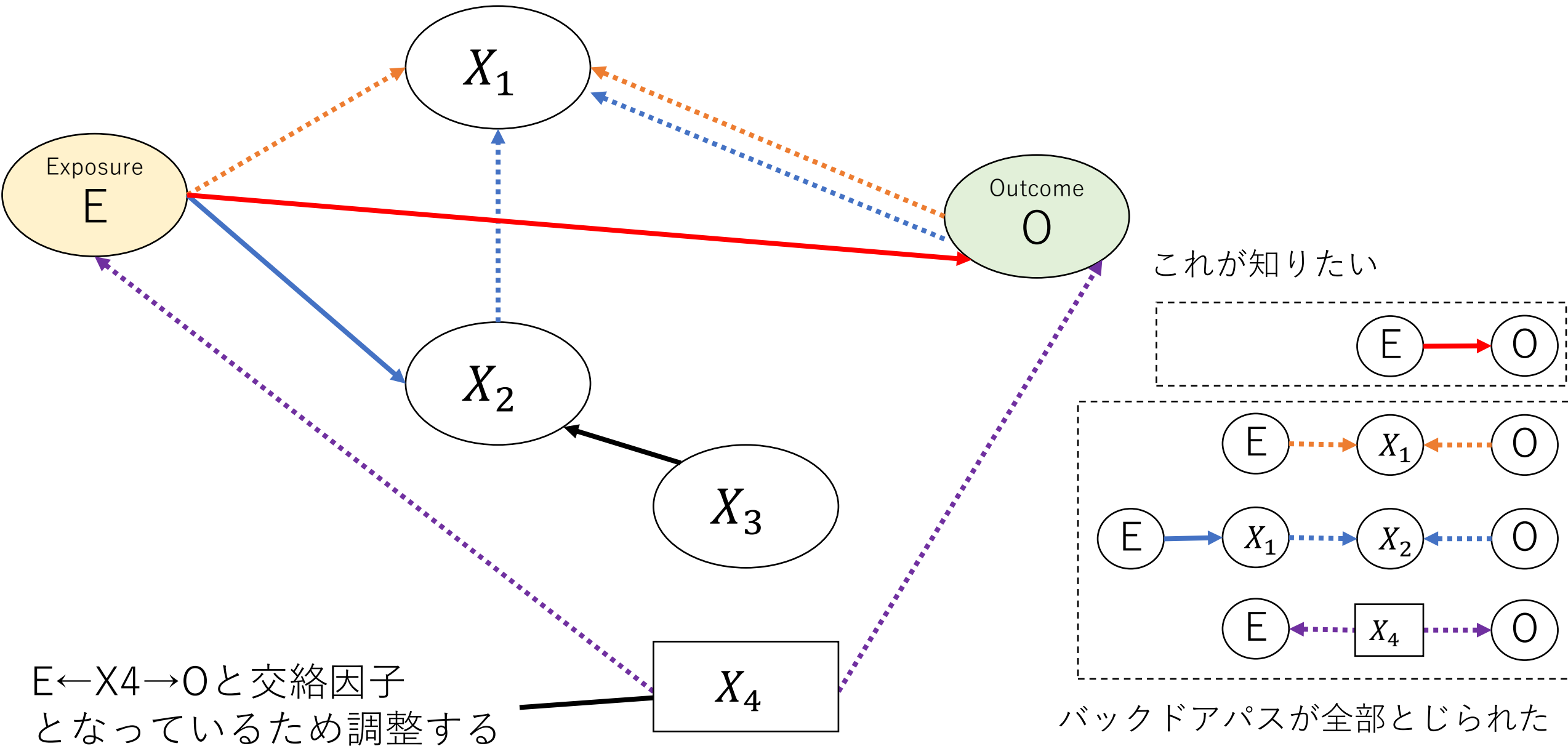
DAGは「どの変数を調整するか」という議論で有用



DAGは「どの変数を調整するか」という議論で有用



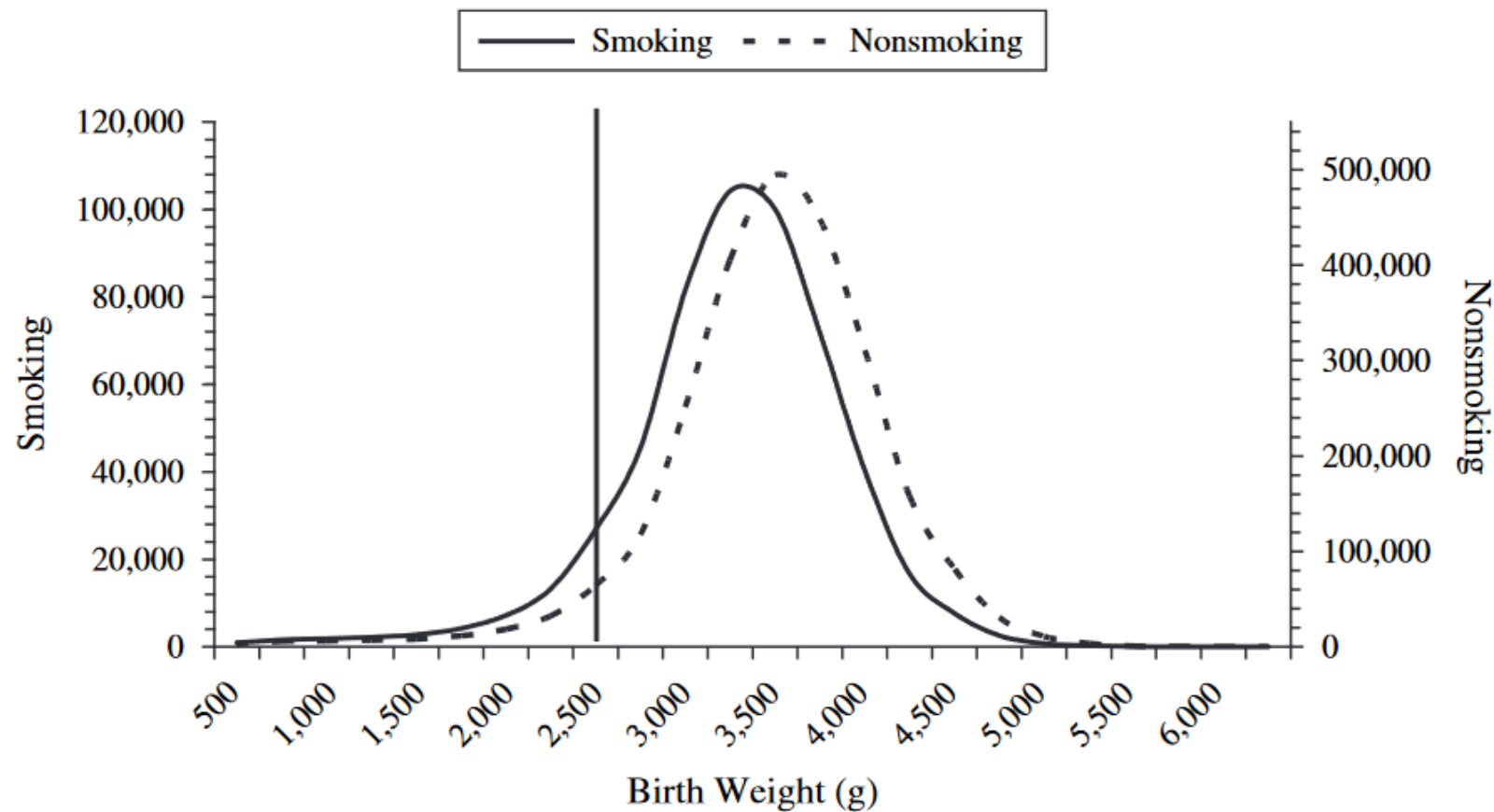
DAGは「どの変数を調整するか」という議論で有用



時間が余れば

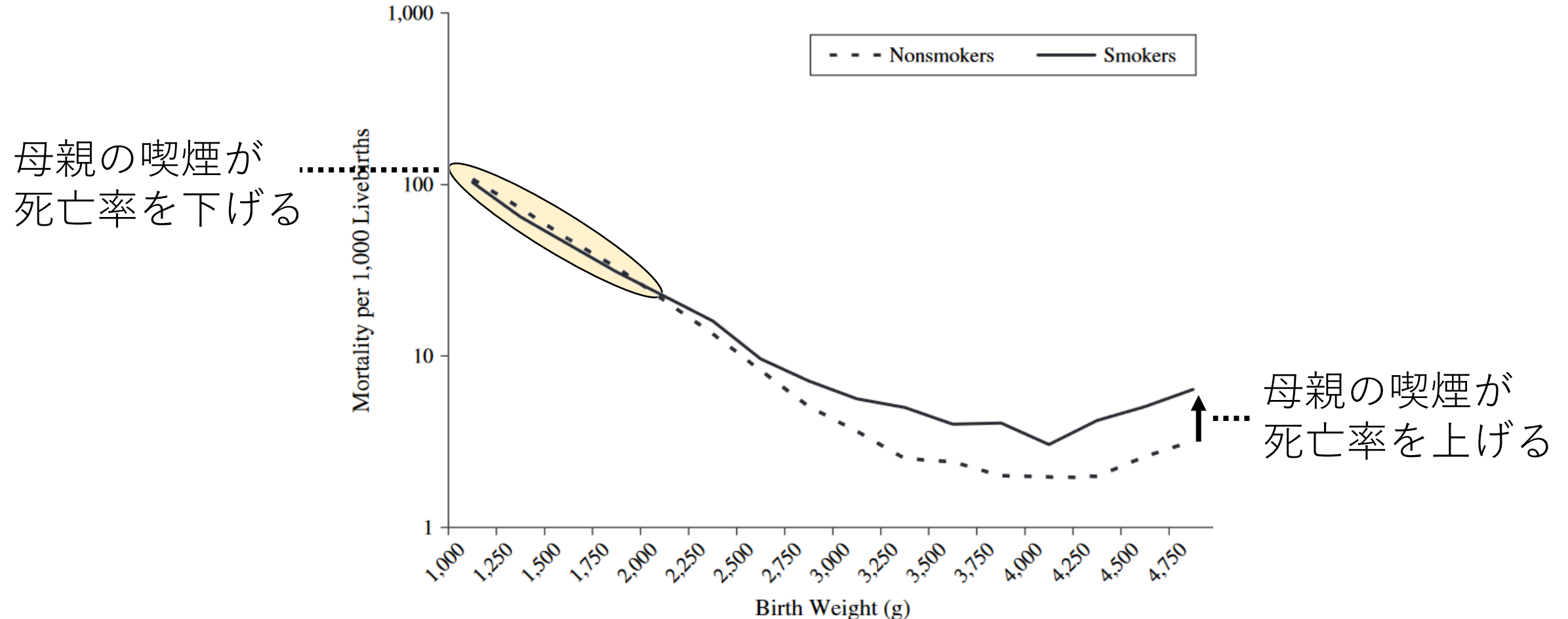
母親が喫煙者の場合、低体重児は死亡率低下に寄与する？

母親が喫煙する場合、低体重出生児が多い



母親が喫煙者の場合、低体重児は死亡率低下に寄与する？

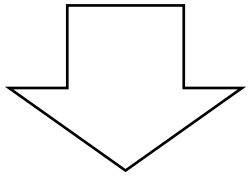
乳児の死亡率が<2000g前後で母親が喫煙する方(点線)が低い



母親が喫煙者の場合、低体重児は死亡率低下に寄与する？

母親が喫煙すると、低体重出生児が多い

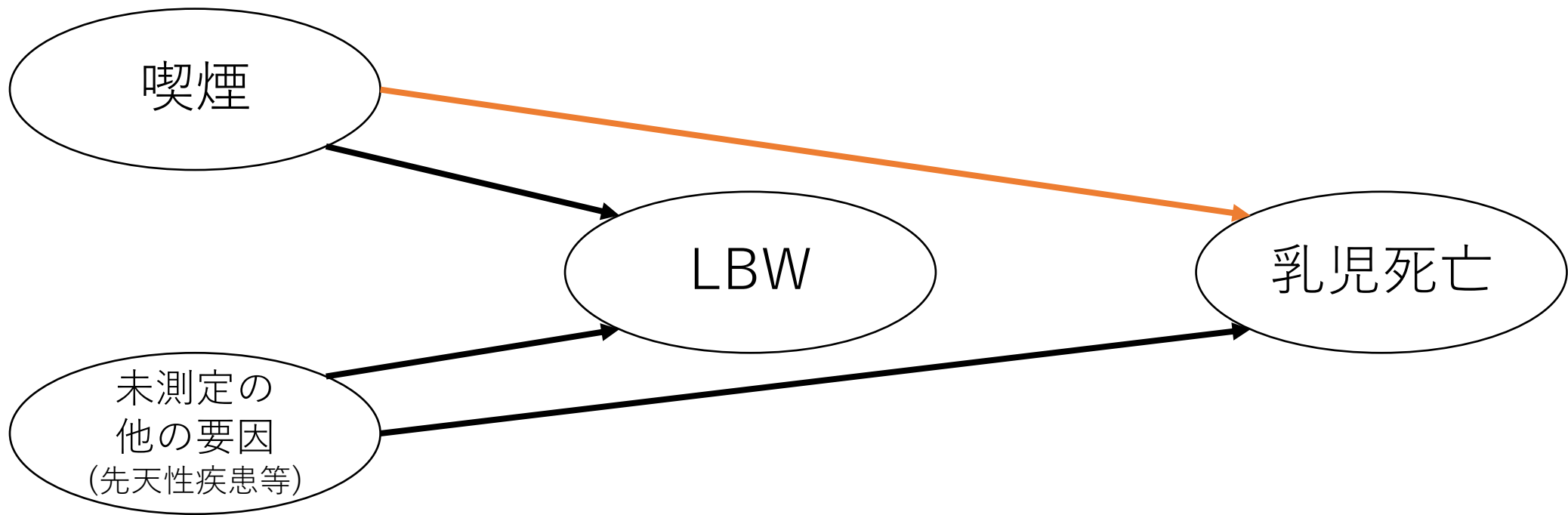
乳児の死亡率が<2000g前後で母親が喫煙する方が低い



母親の喫煙が実は低体重出生児(LBW)である場合の
乳児死亡リスクを下げる？

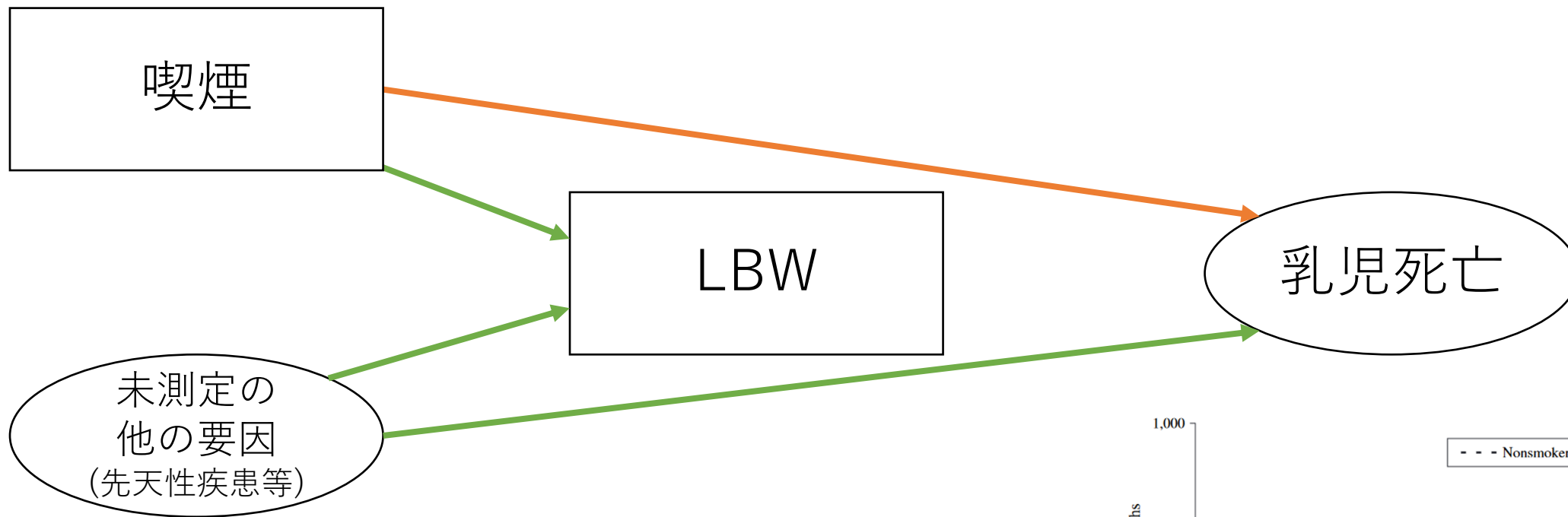
Birth Weight Paradox

パラドックスをDAGで説明してみる



低出生体重(LBW)と乳児死亡が直接の関係がなく、
未測定の他の要因と母親の喫煙がLBWと死亡との間に交絡として
存在する場合「喫煙→乳児死亡」が唯一のパス
(ここで、LBWがCollidorになっていることに注目)

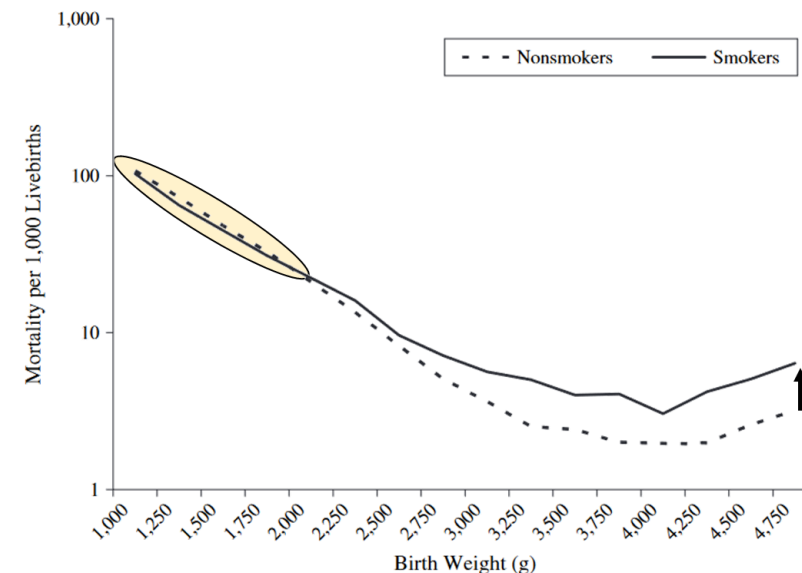
パラドックスをDAGで説明してみる



グラフのように層別化(喫煙とLBW)すると
「喫煙→LBW←他の要因」のパスが開く

喫煙→乳児死亡

喫煙→LBW←他の要因→乳児死亡



Rでこの構造をもつデータを再現できるか試してみましょう

参考文献 & 教科書 & その他

参考文献

▷ DAGの文献についてのReviewとRecommendation

Tennant, Peter W G, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, et al. “Use of Directed Acyclic Graphs (DAGs) to Identify Confounders in Applied Health Research: Review and Recommendations.” *International Journal of Epidemiology* 50, no. 2 (April 1, 2021): 620–32. <https://doi.org/10.1093/ije/dyaa213>.

Supplementary DataにDAGが掲載されている論文が多数（眺めていると色々な使い方がされています）

▷ 今日の「調整」についてわかりやすく書かれている和書

岩波データサイエンス刊行委員会編：岩波データサイエンスVol.3特集「因果推論—実世界のデータから因果を読む」，岩波書店(2016)

因果推論の導入の一冊としておすすめ

▷ 英語の教科書（現時点ではオンラインで無料で入手可能）

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

<https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>

（<https://zenn.dev/shuntarosato/articles/0316df77e19858>に有志の勉強会がまとめた日本語の解説があります。（作成には私は関わっていません））

参考文献

▷喫煙と低体重出生児のパラドックスについての文献

Hernández-Díaz, Sonia, Enrique F. Schisterman, and Miguel A. Hernán. “The Birth Weight ‘Paradox’ Uncovered?” *American Journal of Epidemiology* 164, no. 11 (December 2006): 1115–20. <https://doi.org/10.1093/aje/kwj275>.

スライドでこの論文の内容を説明しました

▷DAGを動画で学ぶなら

Causal Diagrams: Draw Your Assumptions Before Your Conclusions

What ifなどの著者であるHernan先生のオンラインコースです（無料）
オススメ

<https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your>

Rの学習資料

▷疫学のためのRハンドブック（ウェブサイト）

<https://epirhandbook.com/jp/>

無料でRの基礎からかなり発展的な内容まで学べます
（日本語版の翻訳に関わりました）

宣伝！

▷医師が教えるR言語での医療データ分析入門シリーズ

弊ブログへのリンク：<https://www.bunseki-data.com/coupon.html>

動画で学べるオンラインコースです。本日から11日間、クーポンコード
HEALTHDESIGNを入力いただければ無料で登録できるようにしてあります。

▷Rでらくらくデータ分析入門 ～効率的なデータ加工のための基礎知識～技術評論社

<https://gihyo.jp/book/2022/978-4-297-12514-1>

Rでデータを加工（前処理）することに特化した、これまでありそうでなかった本です

Rについてのご質問やコンサルティングは、ma06063@gmail.comまでご連絡ください