



博士課程進学者数に関する 統計的因果探索と交絡因子の取り扱い

高山 正行^{A, B}, 小松 尚登^B, ファム テトン^{B, C}, 前田 高志 ニコラス^{A, B, C, D, F},
三内 顕義^{A, B, E, G, H}, 小柴 等^{A, B}, 清水 昌平^{A, B, C, E}

A: 科学技術・学術政策研究所 (NISTEP)

B: 滋賀大学

C: 理化学研究所 革新知能統合研究 (AIP) センター

D: 東京電機大学

E: 京都大学

F: 学習院大学

G: 東京大学

H: 国立情報学研究所



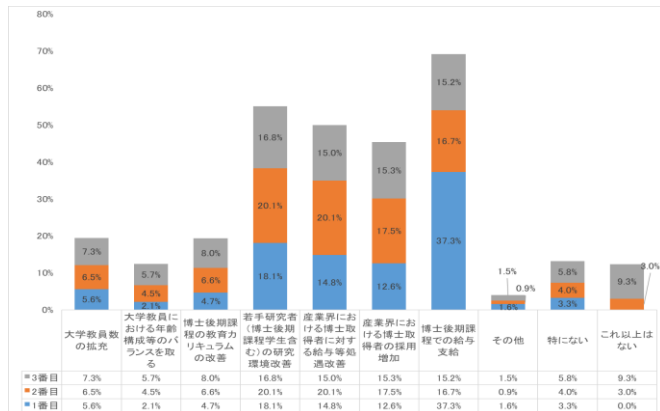
CREST
Core Research for Evolutional Science and Technology

Japan Science and Technology Agency

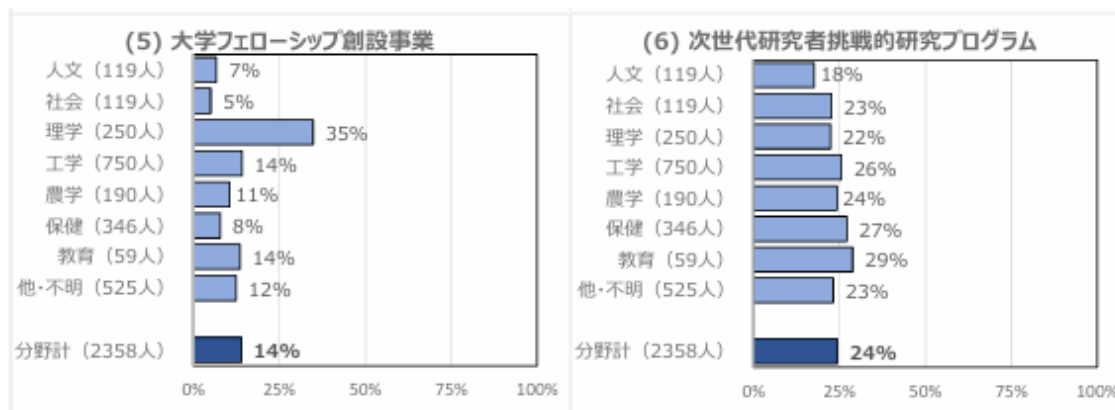


信頼される AI システム
Trusted quality AI systems

**NISTEP「修士課程（6年制学科を含む）
在籍者を起点とした追跡調査」（2021）**



NISTEP「博士(後期)課程 1年次における進路意識と経済的支援状況に関する調査 -令和4年度(2022年12月～2023年1月)実施調査-」(2023)



- 研究力強化の文脈の中でも重要視される、（イノベーションの担い手としての）**博士課程進学者数の増加**
- **経済的支援・研究環境改善・アカデミア&産業界のポスト拡充、雇用条件改善等の重要性は、**博士課程進学者数増の要因に関するアンケート結果で上位
- 最近、文部科学省等による経済的支援施策も充実してきた

- 今後、継続的に予見性をもって、研究力強化のための博士人材を安定供給できるようにするかが重要
 - しかし、他の資源状況の兼ね合いの中で、経済的支援等の政策効果がどの程度のものか、全体像が分かっているわけではない
- ⇒ 定量的な議論が可能な変数に基づいて、マクロスコピックな統計データから、**博士課程進学率に関する各種要因の因果関係の全体像を、未観測変数もあたりをつけながら、把握できないか？**

【先行研究】

- 国全体での博士課程進学要因の関係について、データから統計的に因果関係を見つけ出す統計的因果探索アルゴリズムのうち、I/Oの定量分析にもつなげやすいLiNGAMを用い試行的に分析

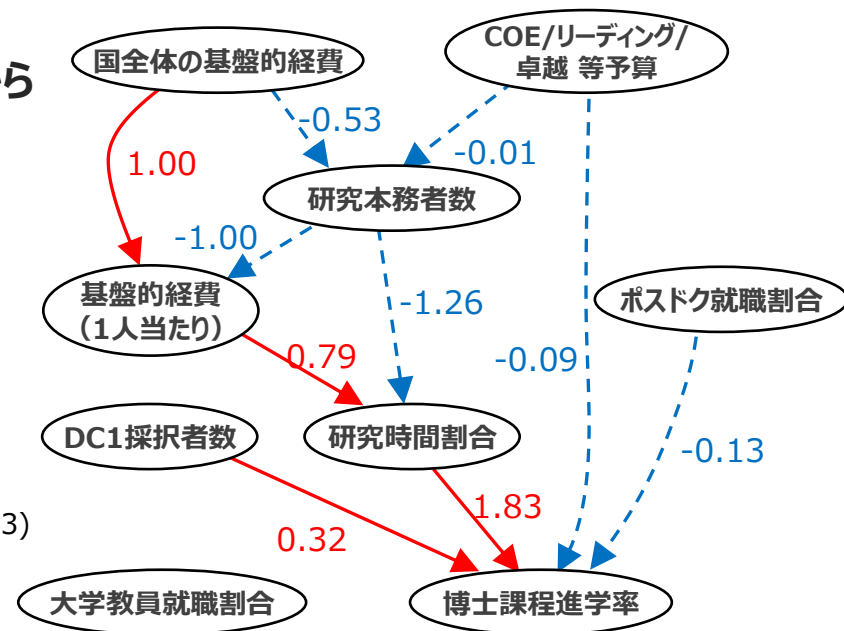
高山・小柴・前田・三内・清水・星野、Jxiv.1(2022)など

- 国立大学別のデータセットを構築し、LiNGAMで因果分析することで、大学別の特徴を捉えながら、博士課程進学等のメカニズムに関する示唆

高山ほか、研究・イノベーション学会 年次大会, 2D20(2023)

【示唆された課題】

- ✓ 未観測変数は複数想像される一方、統計的にどこにその影響が出うるか不明で、疑似的な因果関係が出力される恐れもある。



高山・小柴・前田・三内・清水・星野、Jxiv.1(2022)

❗ 本研究はこの克服を試みる

本研究の概要

博士課程進学に関する政策研究において、よりの確な因果探索を可能とするため…

- ✓ (公開された統計値を基に) **大学別のデータセットを先行研究よりも拡大**
- ✓ **未観測共通原因の存在可能性に配慮した、因果探索アルゴリズムの適用**

を行い、博士課程進学に関する統計的因果推論について、より精緻な議論のための道を切り開く。

データセットの基本構造

✓ (国立86大学) × (2012-2022年度の11か年度) に関する、博士課程進学に関連する変数のデータセットを構築 (**最大946点**)

変数名	内容	出典
x0(t)	修士課程等修了者数	大学改革支援・学位授与機構 (NIAD) のHPに掲載されている大学基本情報を基に加工
x1(t)	博士課程進学者数	
x2(t)	博士課程修了者数	
x3(t)	博士課程修了直後のポストドク就職者数	
x4(t)	博士課程修了直後の大学教員就職者数	
x5(t)	運営費交付金収益額	各国立大学法人の財務諸表等をもとに加工
x6(t)	教員一人当たり学生数	NIADのHPに掲載されている国立大学の財務指標等を基に加工
x7(t)	DC1採択者数	日本学術振興会から過去に公開されたDC1採択者一覧のデータを基に加工

本研究での基本的な使用条件

- 文部科学省における**運営費交付金の重点支援の3類型に基づきグループ化**し、グループごとに分析
- 時系列は、いったん1年先の遅延効果までを取り入れる
- 「積の構造的因果モデル」に基づいて分析
- ➡ データは全て対数変換した上でアルゴリズムに投入

【重点支援3類型の分類】

重点支援 類型	定義・特徴	大学数 (データ点数)	大学例
類型Ⅰ	大雑把には下記以外	55 (605)	下記以外
類型Ⅱ	専門分野の特性に応じた強み・特色のある分野で教育研究を推進	15 (165)	医科歯科大、政研大…
類型Ⅲ	全学的に卓越した教育研究、社会実装を推進	16 (176)	東北大、京大…

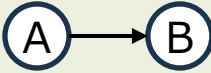
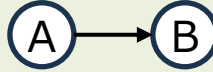
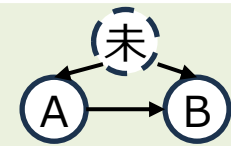
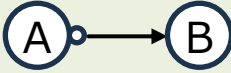
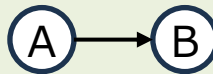
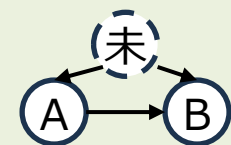
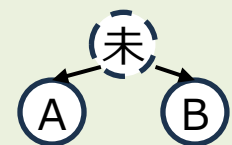
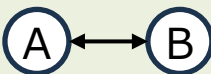
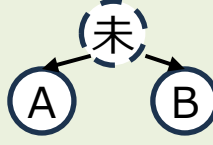
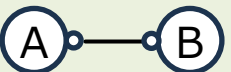
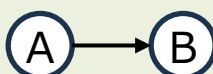
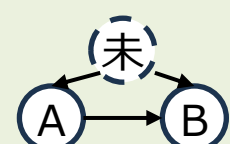
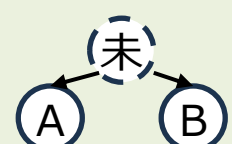
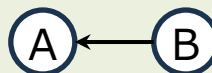
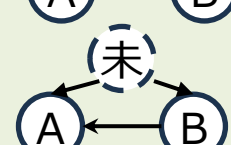
分析の条件・手法② 因果探索のアルゴリズム

- ✓ 未観測共通原因の存在を許した因果探索として、主にFCIによる分析を試行
- ✓ レファレンスとして、PC、LiNGAMの結果とも比較し…
 - ・PCと比較することで、未観測共通原因がないと仮定した場合との**共通構造の確認・特定**
←今回の発表では略（未観測共通原因以外の部分はほとんどPCとFCIで一致）
 - ・LiNGAMを比較することで、**構造方程式に基づいた因果効果の解釈** ←**本日は主にここを発表**
 …を実施
- ➡ 特にどこに未観測共通要因があり得るか、統計的観点から議論することを可能にし、
 他の変数間の詳細な因果効果の議論も議論可能になる

アルゴリズムの種類	探索の考え方	仮定① 巡回性	仮定② 未観測共通要因	仮定③ 関数系の仮定	変数の種類	アウトプット
FCI (Fast Causal Inference) P. Spirtes, PMLR, 2001	PCを未観測共通原因がある場合にも拡張（m-分離の構造特定）	非巡回	○ （あり得るものとして因果グラフを探索）	×	離散/連続	因果グラフ （辺の種類はPCよりも多く、未観測共通要因の存在可能性について細かくパターン分けされている）
PC (Peter-Clark) P. Spirtes, C. N. Glymour, <i>et al.</i> , MIT press, 2001.	変数間の条件付き独立性に基づくd-分離の構造特定 + ルールベースでの向き付け	非巡回	×	×	離散/連続	因果グラフ
LiNGAM S. Shimizu, JMLR, 2006 S. Shimizu, <i>et al.</i> , JMLR, 2011.	構造方程式の誤差変数の依存性の最小化	非巡回	×	線形構造方程式	連続	因果グラフ + 構造方程式の因果係数の推定値

未観測共通要因がある場合の特殊な因果グラフの表記

FCIは、PCベースでありつつも、未観測共通要因がある場合も含め、より慎重に様々なパターンを考慮
 ➡ 因果グラフの表記が、（未観測共通要因を仮定しない場合の）通常の因果グラフと比べて複雑

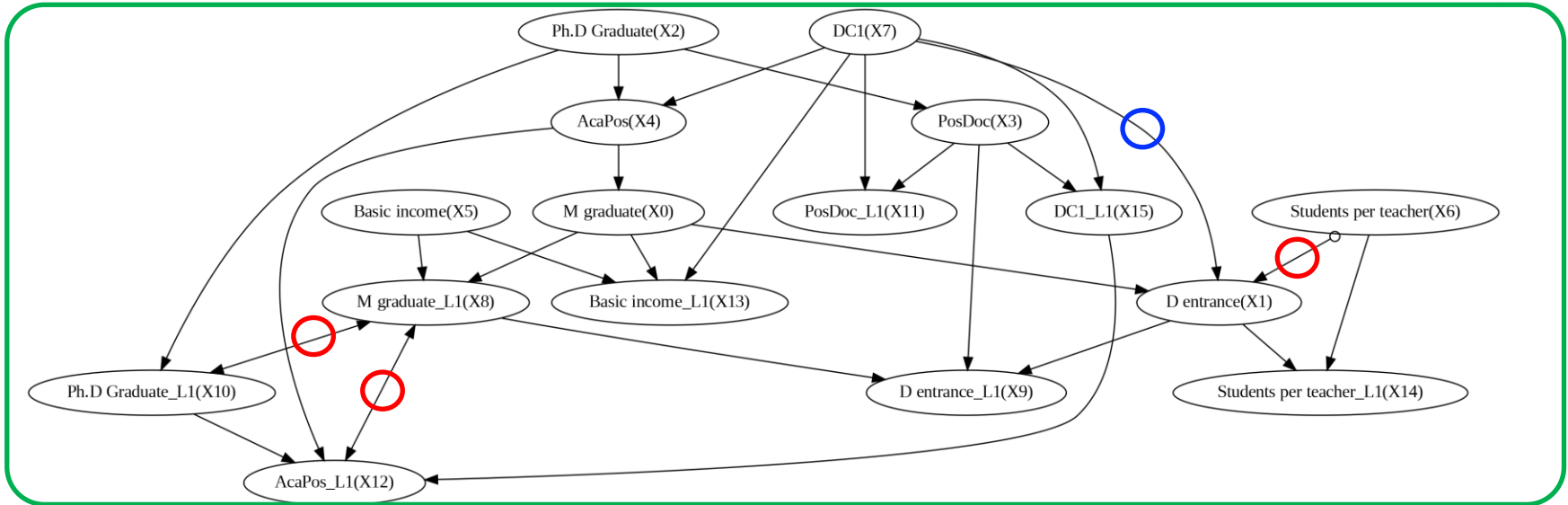
FCIでの辺のパターン	辺の意味	通常の未観測共通要因なしの因果グラフで書かれる候補
パターン1 	AがBの原因である。ただし、別途両者に未観測共通原因が存在する可能性あり	 
パターン2 	BがAの原因になることはない。また、「AがBの原因となる」もしくは「AとBの間に未観測共通原因が存在する」の少なくとも一方が成り立つ	  
パターン3 	AとBは互いに相手の原因にならないが、未観測共通原因がある	
パターン4 	「AがBの原因」「BがAの原因」「AとBの間に未観測共通要因が存在する」のうち、少なくとも1つが成り立つ	    

- ✓ 特に、線の端に○がついている、パターン2/4は、幅広い解釈可能性があり得ることに注意。
- ✓ (体系的な理解があるわけではないが) パターン2～4は、実際に操作してみると、データ点数が少ないと出にくい傾向
 ➡ 本研究でも、類型I(55大学)ではパターン2～4も出るが、類型II(15大学)・III(16大学)ではパターン1のみになりがち

分析結果①: FCIでの因果探索の結果

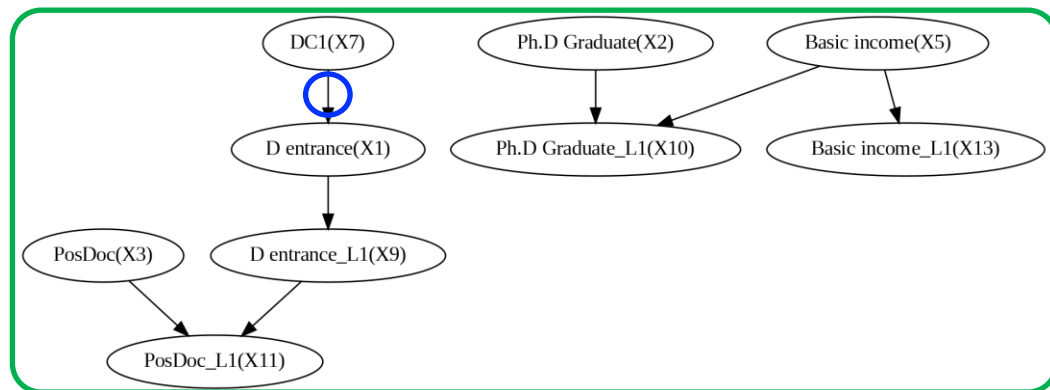
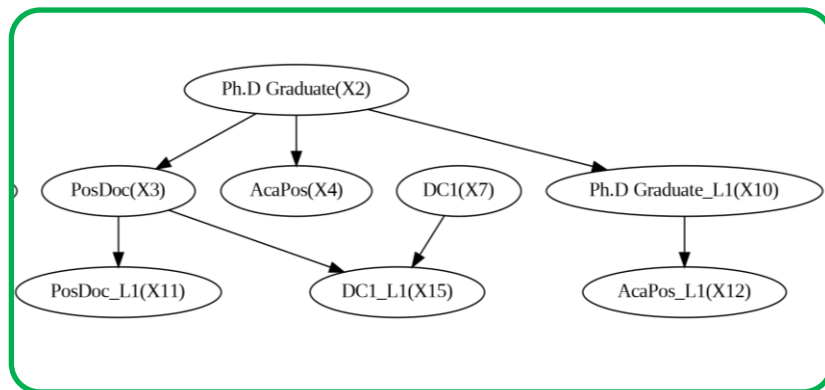
【類型Ⅰ（55大学）の因果探索結果】

※ 他変数との辺が一切出ない、もしくは自己回帰のみの場合は省略して掲載。



【類型Ⅱ（15大学）の因果探索結果】

【類型Ⅲ（16大学）の因果探索結果】

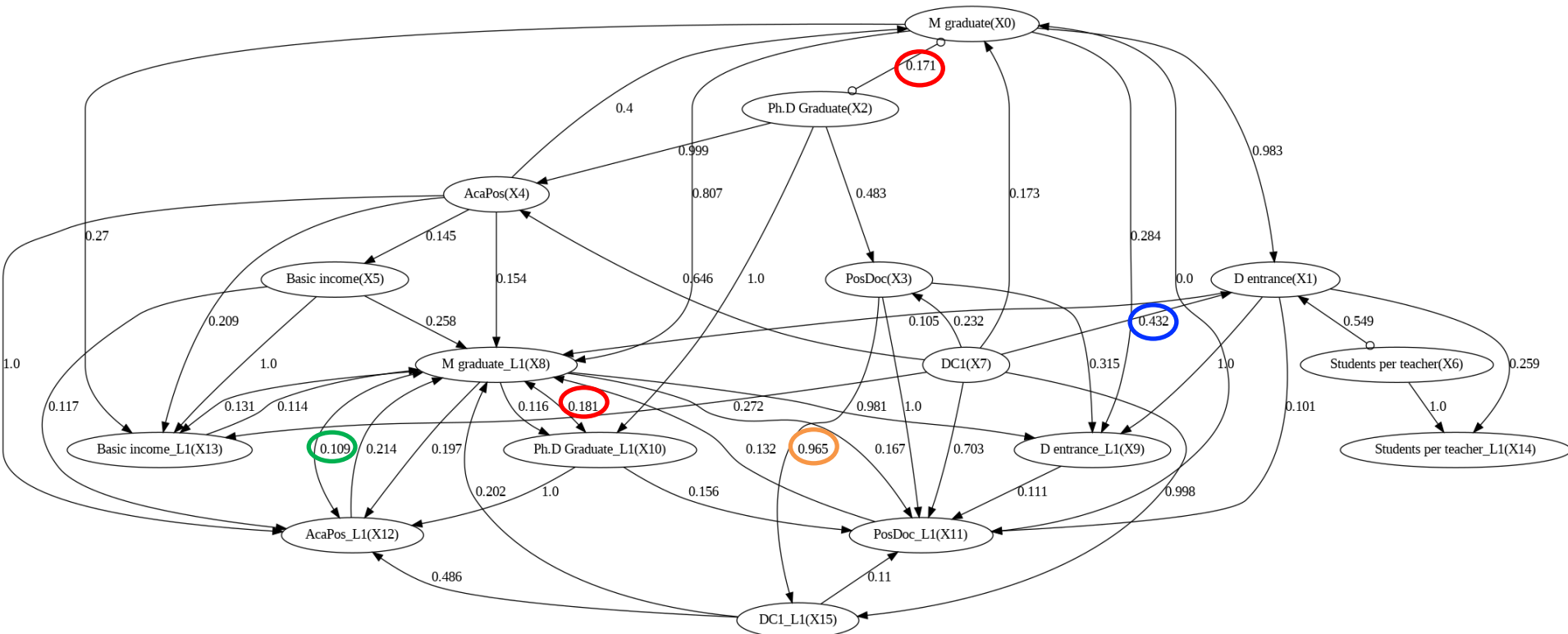


- ✓ 類型Ⅰ/Ⅲは、DC1→博士課程進学者数 という自然なパスが見られる
- ✓ 類型Ⅰは、未観測共通原因が存在する可能性を示すパスも現れている
- ✓ 類型Ⅱは、特徴的なパスが観測されなかった ➡ ここからは分析対象から除外

分析結果②: FCIでのブートストラップの結果-類型I

重点支援類型 I でのFCIのブートストラップの結果

※ データの無作為復元抽出の回数は1000回。また、各辺の数字はブートストラップ確率。

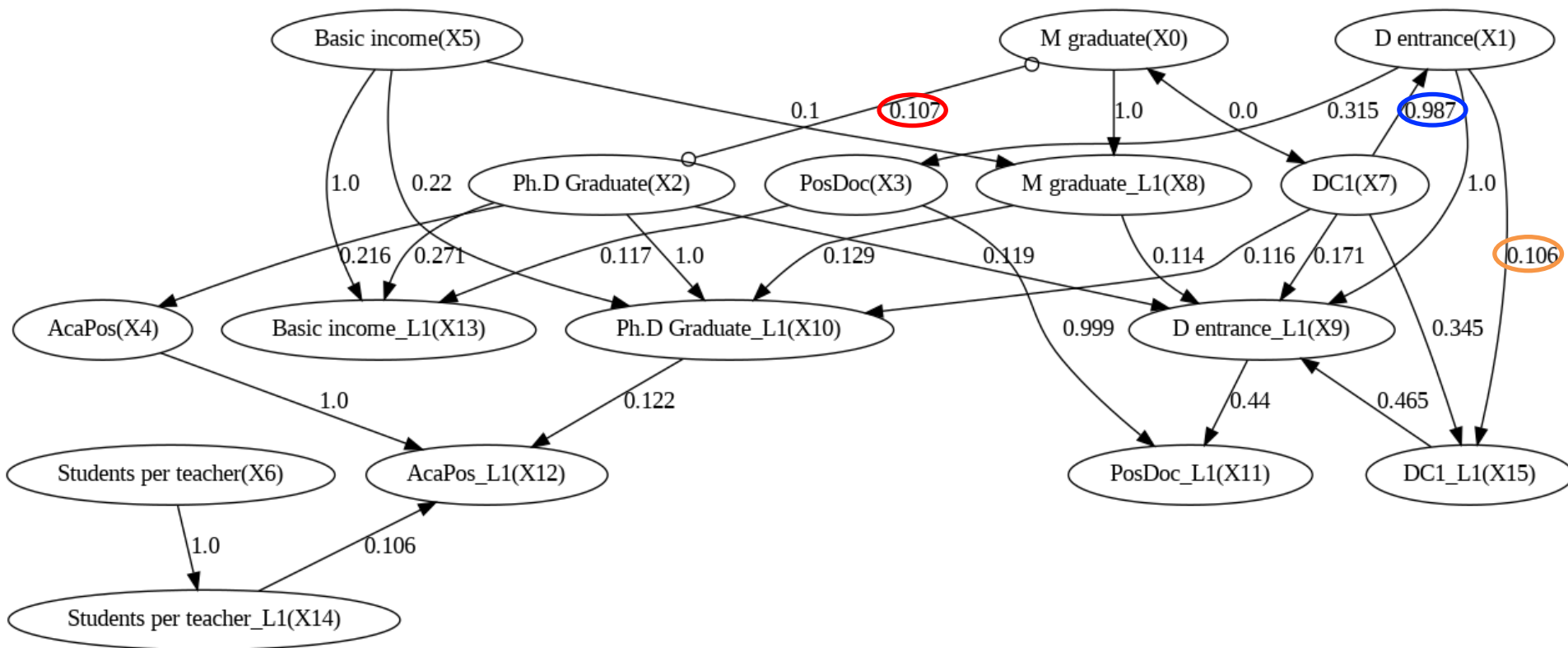


- ✓ DC1採択者数→博士課程進学者数 の確率は0.432とやや少なめ
- ✓ 博士課程修了後のポスドク就職者数→1年後のDC1採択者数 の確率は0.965と高い
- ✓ 修士課程修了者-博士課程修了者の関係は、0年目は $\bullet \longrightarrow \bullet$ で確率0.171、1年後は \longleftrightarrow で確率0.181出ている
 ➡ 未観測共通原因が存在する可能性
- ✓ 1年後については、修士課程修了者-博士修了直後の大学教員修了者数の間も、確率0.109で未観測共通要因が示唆される

分析結果③: FCIでのブートストラップの結果-類型Ⅲ

重点支援類型ⅢでのFCIのブートストラップの結果

※ データの無作為復元抽出の回数は1000回。
また、各辺の数字はブートストラップ確率。

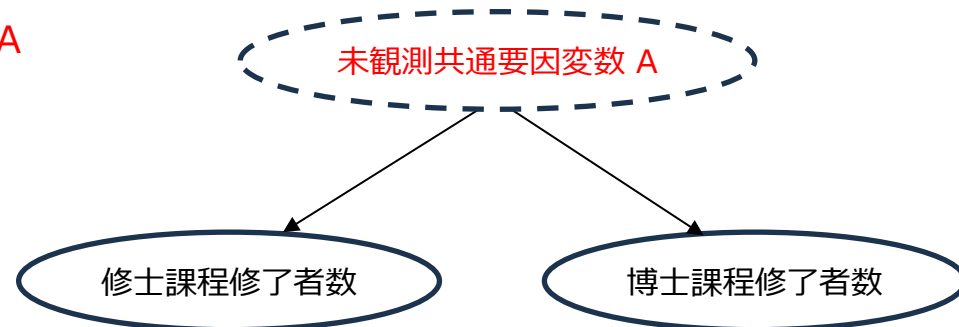


- ✓ DC1採択者数→博士課程進学者数 の確率は0.987と極めて高い
- ✓ 博士課程進学者数→1年後のDC1採択者数 の確率が0.106だけある
- ✓ 修士課程修了者-博士課程修了者の関係は、0年目は $\bullet \longrightarrow \bullet$ で確率0.107と、様々な可能性あり
➡ 仮に定性的には類型Ⅰと同様のメカニズムと仮定すると、未観測共通原因の可能性

考察① FCIの結果から示唆される未観測共通原因

博士課程修了者と修士課程修了者の未観測共通原因の可能性

✓ 右図の因果グラフを満たすような、**未観測共通原因変数A**が存在する場合、それは一体何か？



直観的な発想に基づく候補のピックアップ

修士も博士も、（修了の難易度は違うが）どちらも成果をまとめて学位論文を出すことが前提であることに着目すると、

- ① 大型研究費の獲得による研究PJの進展
- ② 大学院生への指導体制の強化

といった、学位取得に足る成果の創出に係る何かである可能性

➡ ①なら、例えば国立大学別の競争的資金の受け入れ金額等をまとめて、変数に追加することは可能であり、この位置に来うる変数が、因果探索で調べるという方法もありうる

✓ ただし一般に、未観測共通原因があることがわかって、それだけは**完全な特定に至るのは非常に難しい**

理由1: 仮に因果関係を単純な正/負の影響で考えても、未観測共通要因の変動がどうなっているのかまではわからないわかって、その挙動に対応すると考えられる変数が領域知識から見つかるか不明

理由2: 未観測共通原因が、1つとは限らないので、1つ見つけて変数を加えられても、より良いモデルになるとは限らない

 **探索的因子分析や、非構造化テキストデータ等に基づく因果表現学習の活用がカギ？**

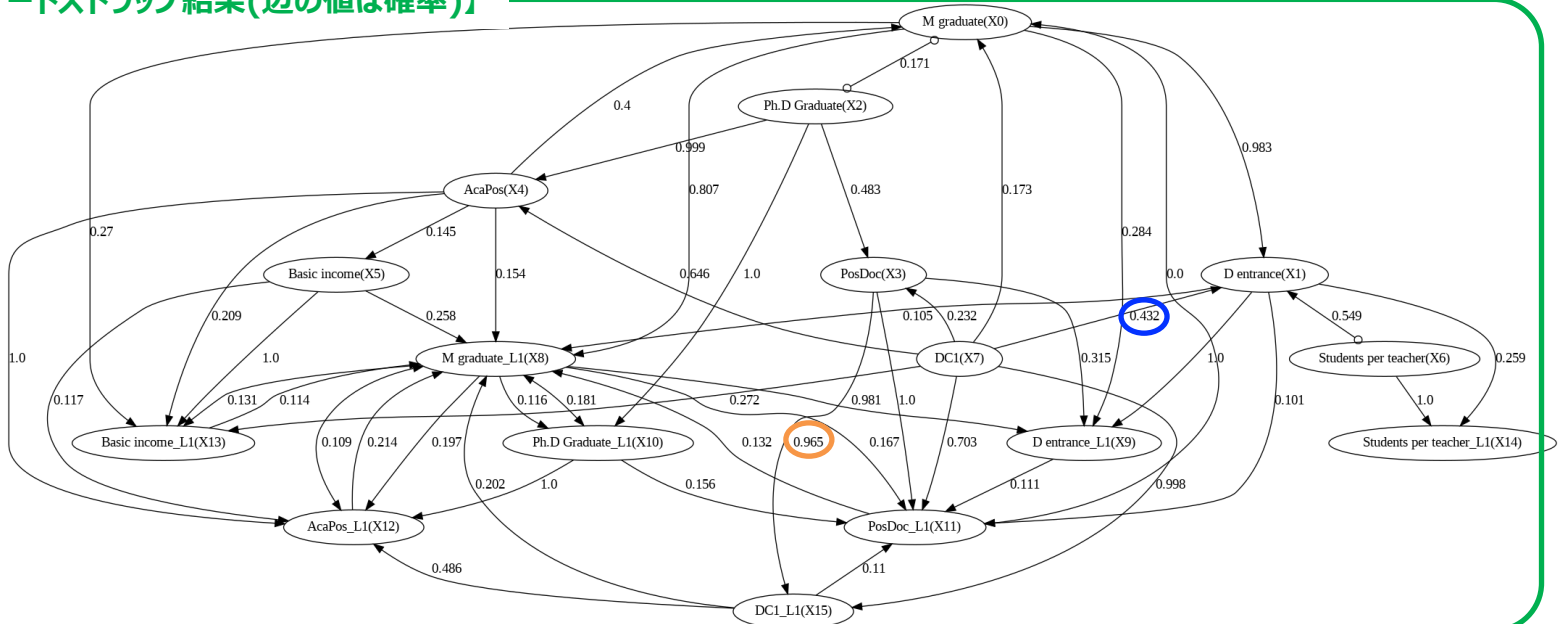
修士課程修了者数 への影響	博士課程修了者数 への影響
正	正
負	負
負	正
正	負

両パターンあり得ても、どっちかなのか、両方効く場合にどちらがどれほど、というのも難しい

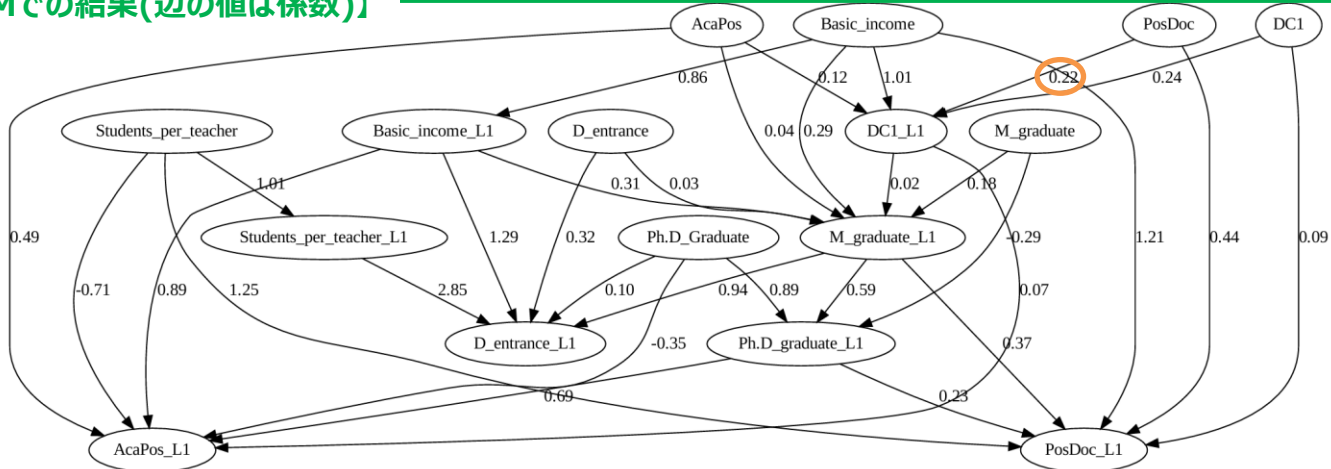
逆の振る舞いに繋がる変数は思いつきにくい

分析結果④: DirectLiNGAMとの比較-類型I

【FCIでのブートストラップ結果(辺の値は確率)】



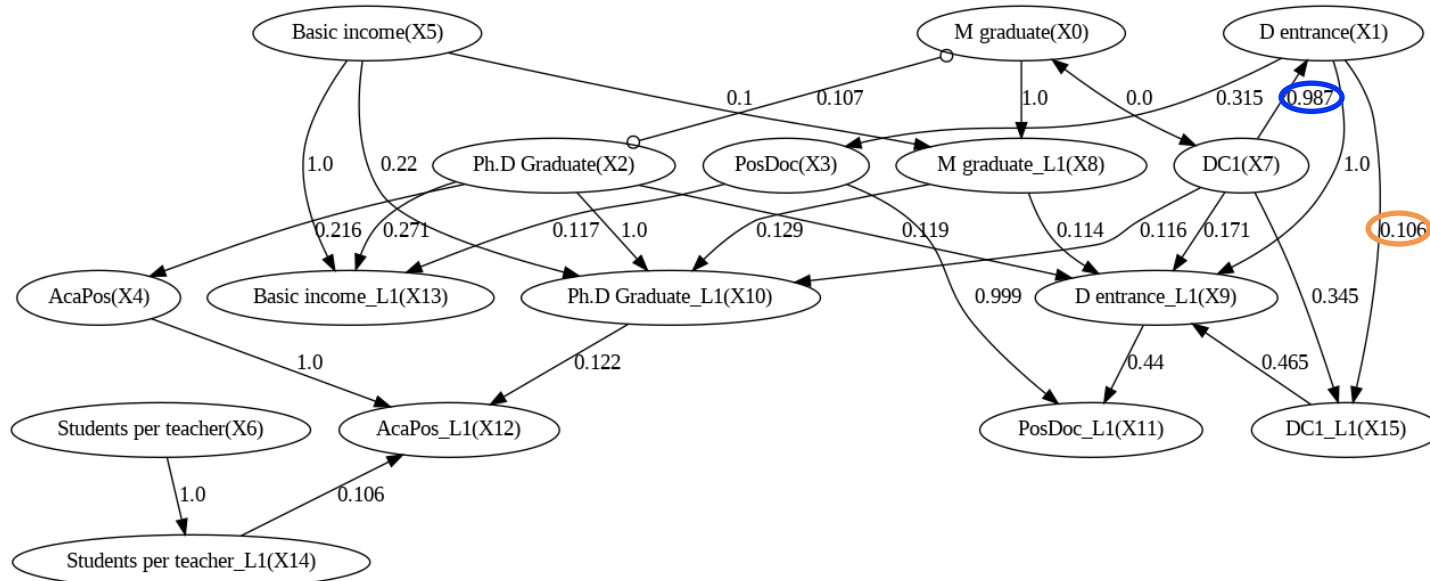
【LiNGAMでの結果(辺の値は係数)】



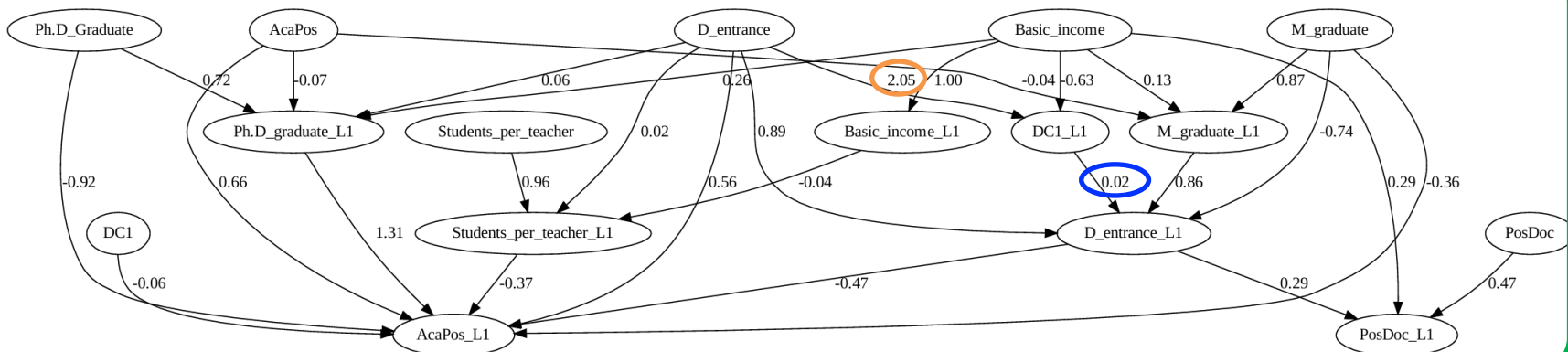
- ✓ DC1採択者数→博士課程進学者数は、データセット全体でのLiNGAMでは現れない（非常に弱い関係？）
- ✓ 博士課程修了後のポスドク就職者数→1年後のDC1採択者数は、LiNGAMでも係数は0.22で出ている

分析結果⑤: DirectLiNGAMとの比較-類型Ⅲ

【FCIでのブートストラップ結果(辺の値は確率)】



【LiNGAMでの結果(辺の値は係数)】



- ✓ DC1採択者数→博士課程進学者数 は、データセット全体でのLiNGAMでも、0.02と小さいが正の値として出ている
- ✓ 博士課程進学者数→1年後のDC1採択者数 は、LiNGAMでも係数は2.05と大きな値で出ている

考察② 博士課程進学者数・DC1採択者数に関する 類型Ⅰ/Ⅲでの差異

DC1採択者数に影響する変数の違い

○○ →1年後のDC1採択者数	類型Ⅰ (55大学)	類型Ⅲ (16大学)
○○に入る変数	博士課程修了後の ポスドク就職者数	博士課程進学者数
FCIでのブートストラップ確率	0.965	0.106
LINGAMでの係数	+0.22	+2.05

- ✓ DC1採択者数が増える（ような申請書が書けるようになる）メカニズム・環境が、類型Ⅰと類型Ⅲで異なる可能性
- ✓ 本学会で当グループが昨年報告した内容（https://dSPACE.jaist.ac.jp/dSPACE/bitstream/10119/19271/1/kouen38_190.pdf）に整合

DC1採択者数→博士課程進学者数の表れ方の比較

DC1採択者数 →博士課程進学者数	類型Ⅰ (55大学)	類型Ⅲ (16大学)
FCIでのブートストラップ確率	0.432	0.987
LINGAMでの係数	0 (辺自体が出現せず)	+0.02

- ✓ DC1採択者数→博士課程進学者数 が特に類型Ⅲで**正の影響**として現れていることは、
 - ・博士課程進学の要因として、経済的支援（DC1は年240万円）が重要視されていること
 - ・JSPS特別研究員のDC1が、採択者の博士課程進学を前提にしたものであること
 を考えると極めて自然
- ✓ 類型Ⅲのような研究大学では、博士課程進学者数へのDC1採択者数の影響がかなり確からしく見えるが、
類型Ⅰではそこまででもなく、むしろ他の要因（ポスドク就職者数）の方が主として効く可能性

本研究の成果

- 全86の国立大学11か年分の博士課程進学に関連する各変数で構築した、公開情報ベースでのデータセットで、**未観測共通原因の存在にも配慮しながら、FCI等の複数のアルゴリズムで統計的因果探索を実施した。**
- 「博士課程進学者数」と「修士課程進学者数」の間に未観測共通原因が存在する可能性が、統計的に示唆された（詳細は不明 & 今後要検討だが、学位論文に含め得る成果創出につながるものとして、例えば競争的研究費による研究PJの活性化等が考えられる）
- DC1採択者数→ 博士課程進学者数 というパスが、類型 I ではやや確率が低め、類型 III ではかなりはっきりと正の影響として示唆されている
- DC1採択者数に正の影響を与うる変数が、類型 I（前年の博士課程進学者数）と類型 III（前年の博士課程修了直後のポスドク終章者数）で異なり、メカニズム・環境が差異が表出している可能性。

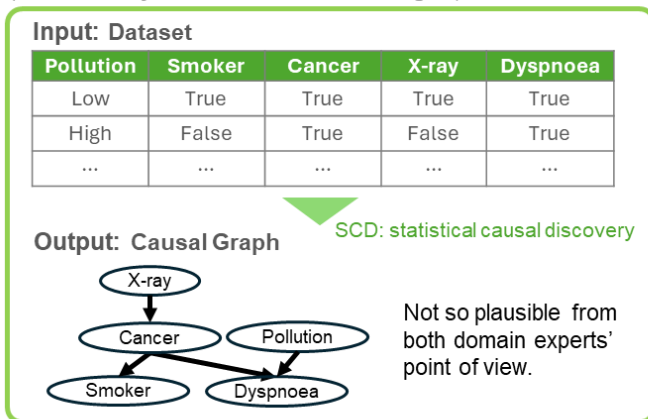
留意点・今後の展望

- ✓ 高等教育論等の政策研究の領域知識に基づいた、因果グラフの各辺の詳細な解釈
👉 **次ページp.15参照**
- ✓ 研究分野別など、別のドメインで分けた際の議論
👉 **次々ページp.16参照**
- ✓ **未観測共通原因の具体的な特定**（方法論の確立含む）
- ✓ 因果モデルを用いた**政策的なシミュレーション**（政策意志決定支援のためのマルチエージェントシミュレーション等）

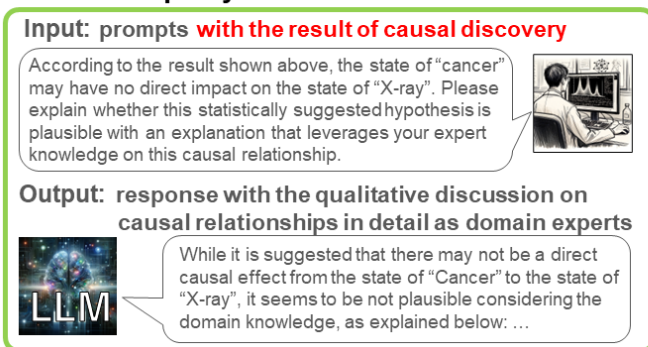
番外編①領域知識に基づく因果グラフ構築のための 大規模言語モデル（LLM）の活用

✓ 当グループでは別途、LLM（特にGPT-4）を活用し、領域知識から見ても統計的に見ても、妥当な因果グラフを構築する一般的な方法を最近提案したところ

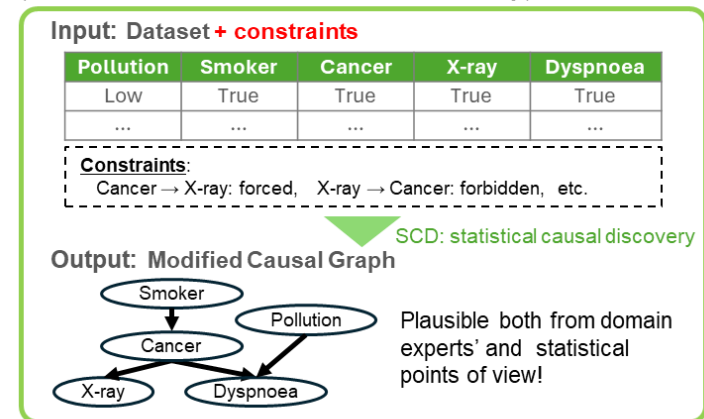
1st step: Data-Driven Causal Discovery (without any constraints on the edges)



2nd step: Knowledge Generation on Causal Relationships by the LLM with ZSCoT

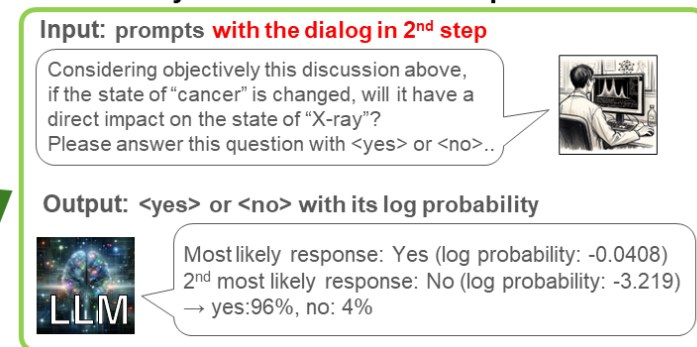


4th step: Retrying Causal Discovery (with the constraints determined in 3rd step)



Transforming the probability matrix generated in 3rd step into background knowledge for the causal discovery

3rd step: Knowledge Integration and Evaluation of the Probability of Causal Relationships with the LLM



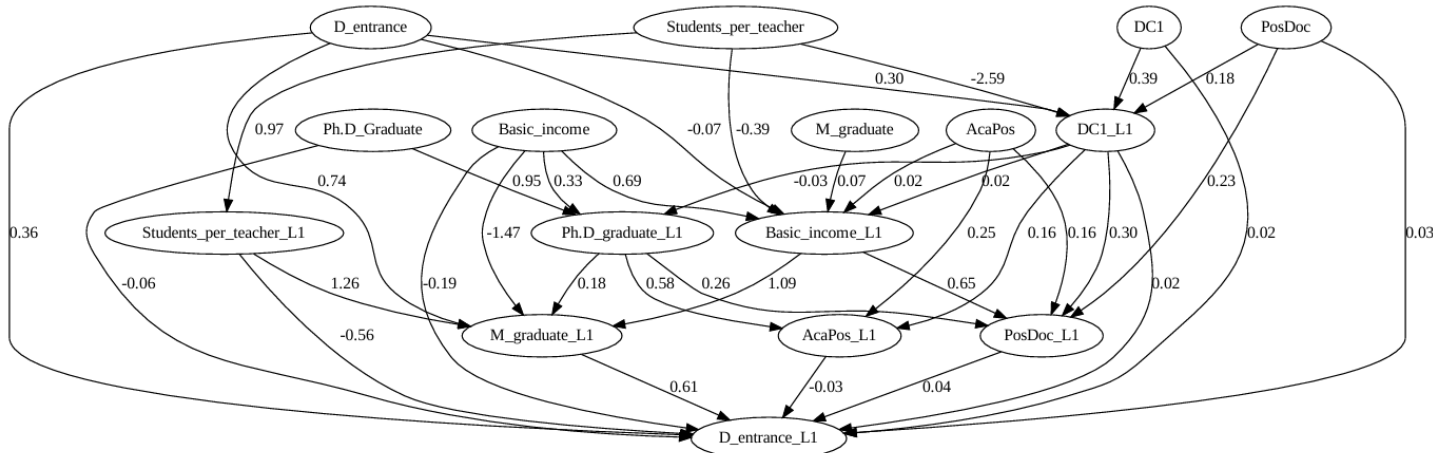
M. Takayama *et al.*, Large Language Models in Causal Discovery:
A Statistical Causal Approach, arXiv(2024)

<https://arxiv.org/abs/2402.01454>

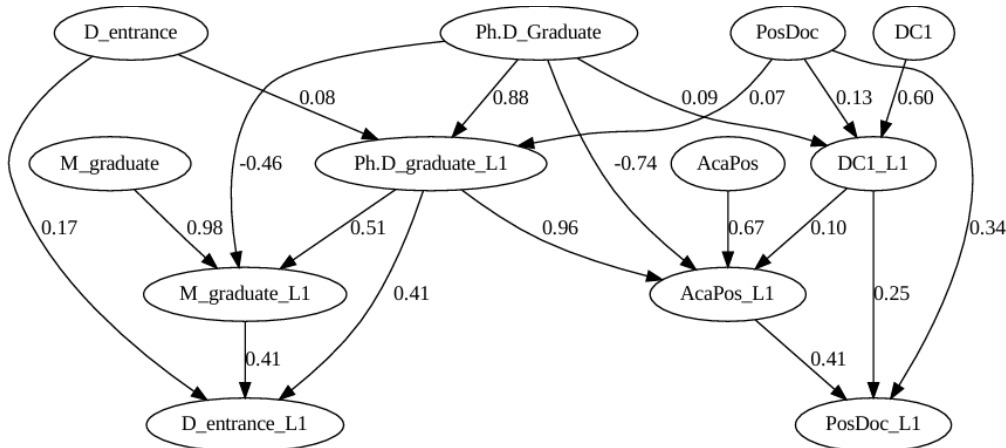
番外編②研究分野別での統計的因果探索（速報）

2018年度-2023年度の一部の国立大学(30程度)を対象に、分野別のデータセットを構築し、LiNGAMで試行的分析。
（運営費交付金の収益額が、2022年度以降の財務諸表から全国立大学で学部等のセグメント別に書かれるようになったため、ごく最近、分析が可能に！）

理学・工学 合算でのLiNGAMの結果（各辺の値は因果係数）



人文科学・社会科学・教育学 合算でのLiNGAMの結果（各辺の値は因果係数）



✓ 人社系だと、博士課程進学者数にDC1採択者数は効かず、博士課程修了者数が最も効く様子（出られるかどうかが大事？）

✓ 理工系だと、DC1採択者数に、教員一人当たり学生数から負の影響が現れたり、ミクロに解析すると興味深い現象が示唆されそう

一緒に本PJで、特に因果分析のためのデータ準備・加工、分析作業を実施し、研究を深化させてくれる方、絶賛大募集です！

- ・「もっと、こういうデータのこういう変数を加えてみたらいいんじゃないか」的なアイデアをお持ちの方
- ・博士課程進学以外にも、研究力強化の文脈で論文数とかも因果分析してみたいという方

ご関心がありましたら、

mas.takayama.babygrand@gmail.com

までご連絡ください！ 笑

ご清聴ありがとうございました！！