UNIVERSITY OF COPENHAGEN

Introduction to causal discovery: More on (T)PC in practice

Anne Helby Petersen

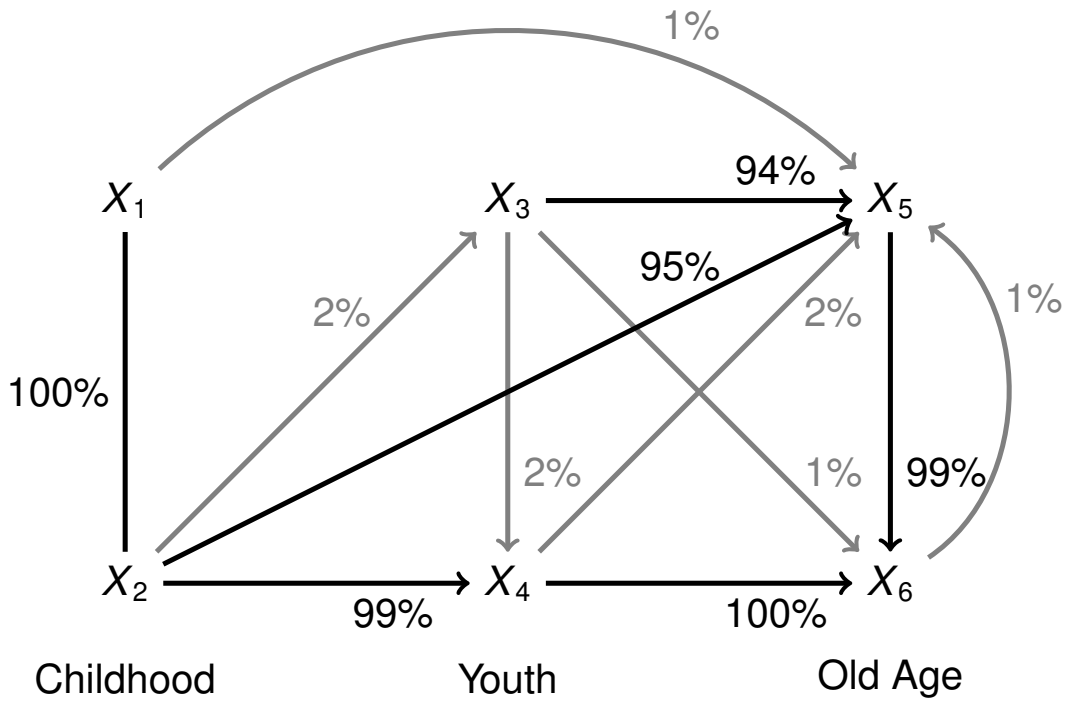
TPC in practice

We'll now dive into a few scattered topics regarding TPC in practice:

- **Edge retention:** What happens when α decreases?
- What type of **temporal information** is most useful?
- How does TPC **compare to traditional approaches** for constructing DAGs?
- What happens if there is **unobserved confounding**?

Edge retention: TPC applied to simulated data

$n = 200$ in each simulated dataset, $b = 100$ repetitions.

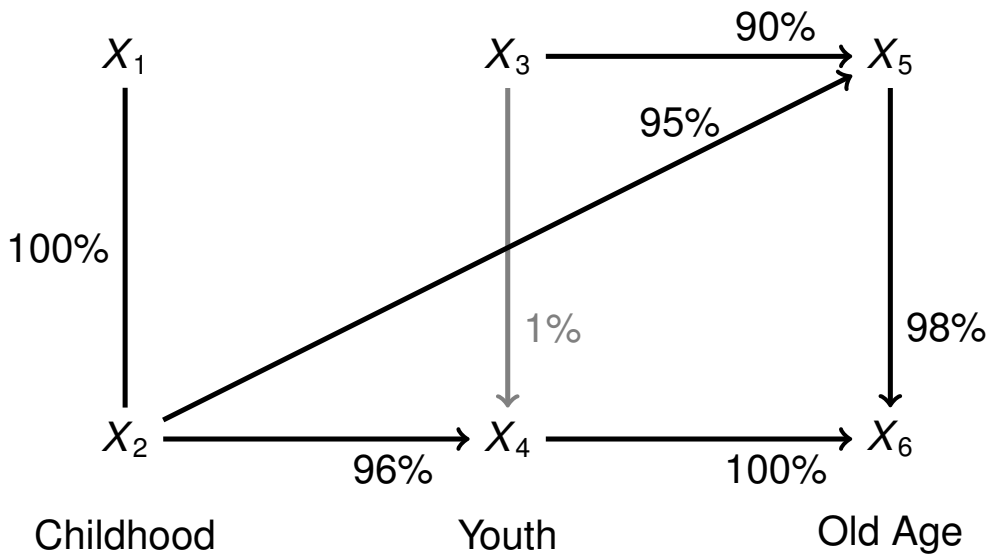


$\alpha = 0.1$

Black edge: True edge. **Gray edge:** Spurious edge. **Percentage:** Percentage of simulations that included this edge.

Edge retention: TPC applied to simulated data

$n = 200$ in each simulated dataset, $b = 100$ repetitions.

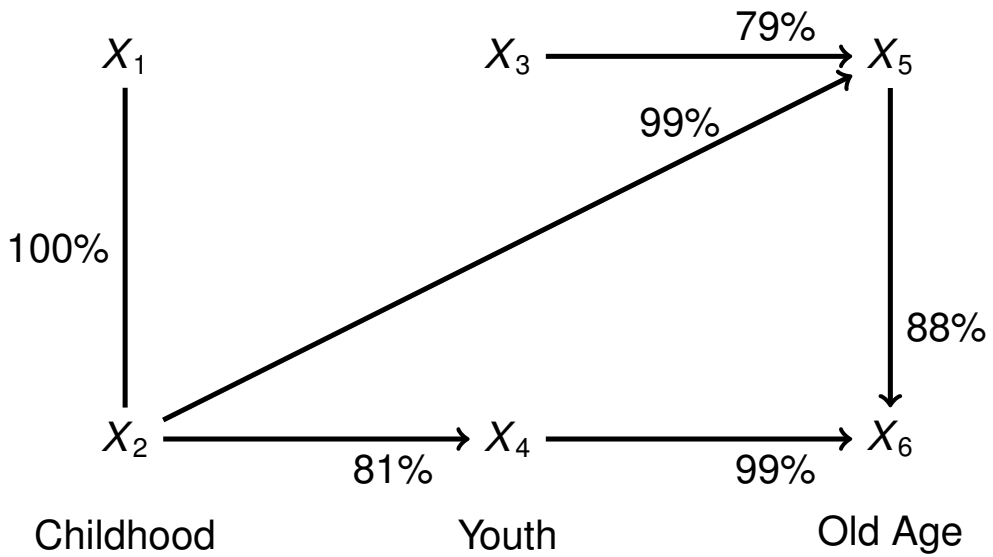


$\alpha = 0.01$

Black edge: True edge. **Gray edge:** Spurious edge. **Percentage:** Percentage of simulations that included this edge.

Edge retention: TPC applied to simulated data

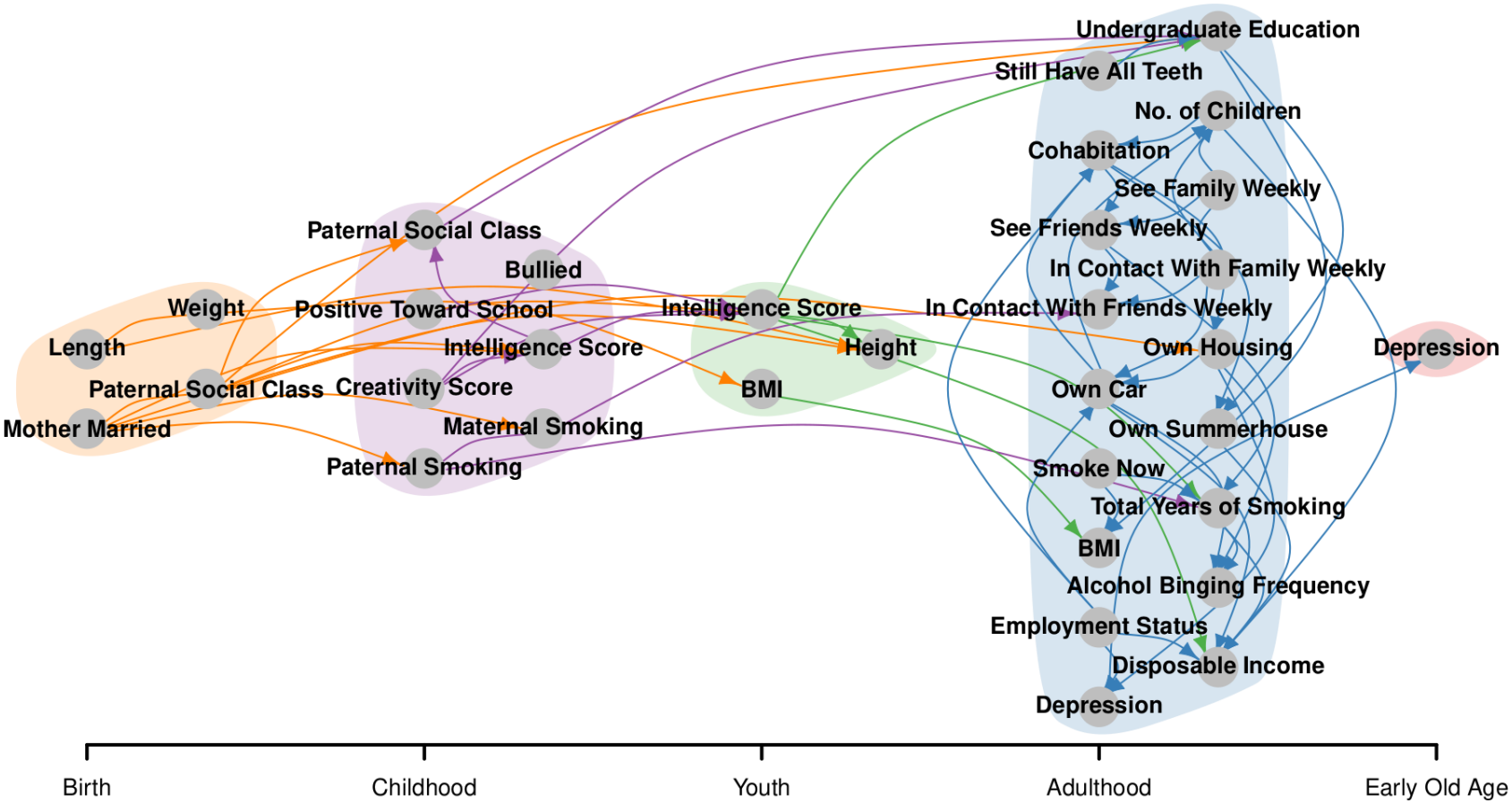
$n = 200$ in each simulated dataset, $b = 100$ repetitions.



$\alpha = 0.001$

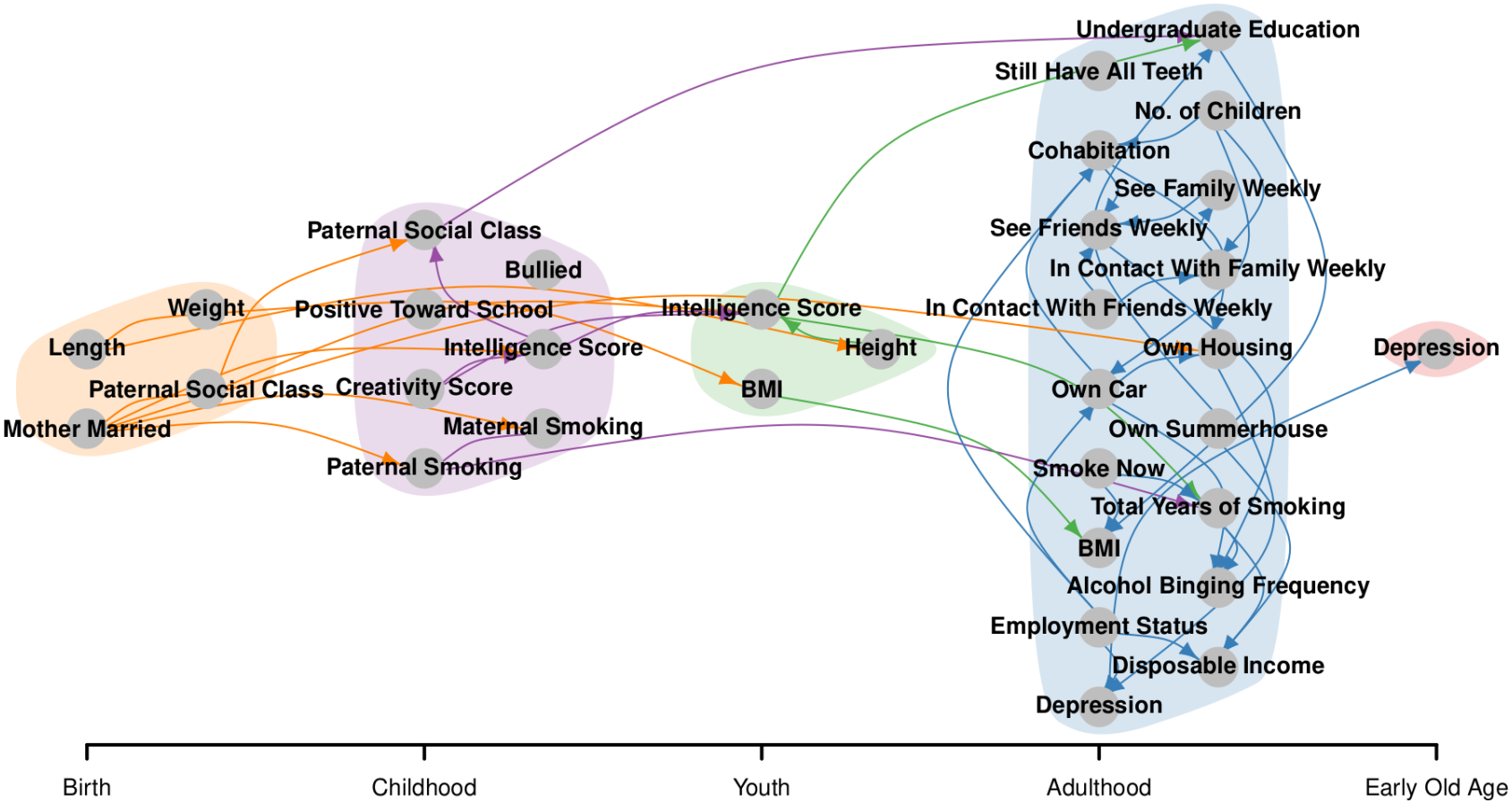
Black edge: True edge. **Gray edge:** Spurious edge. **Percentage:** Percentage of simulations that included this edge.

Edge retention in application (Petersen, Osler & Ekstrøm 2021)



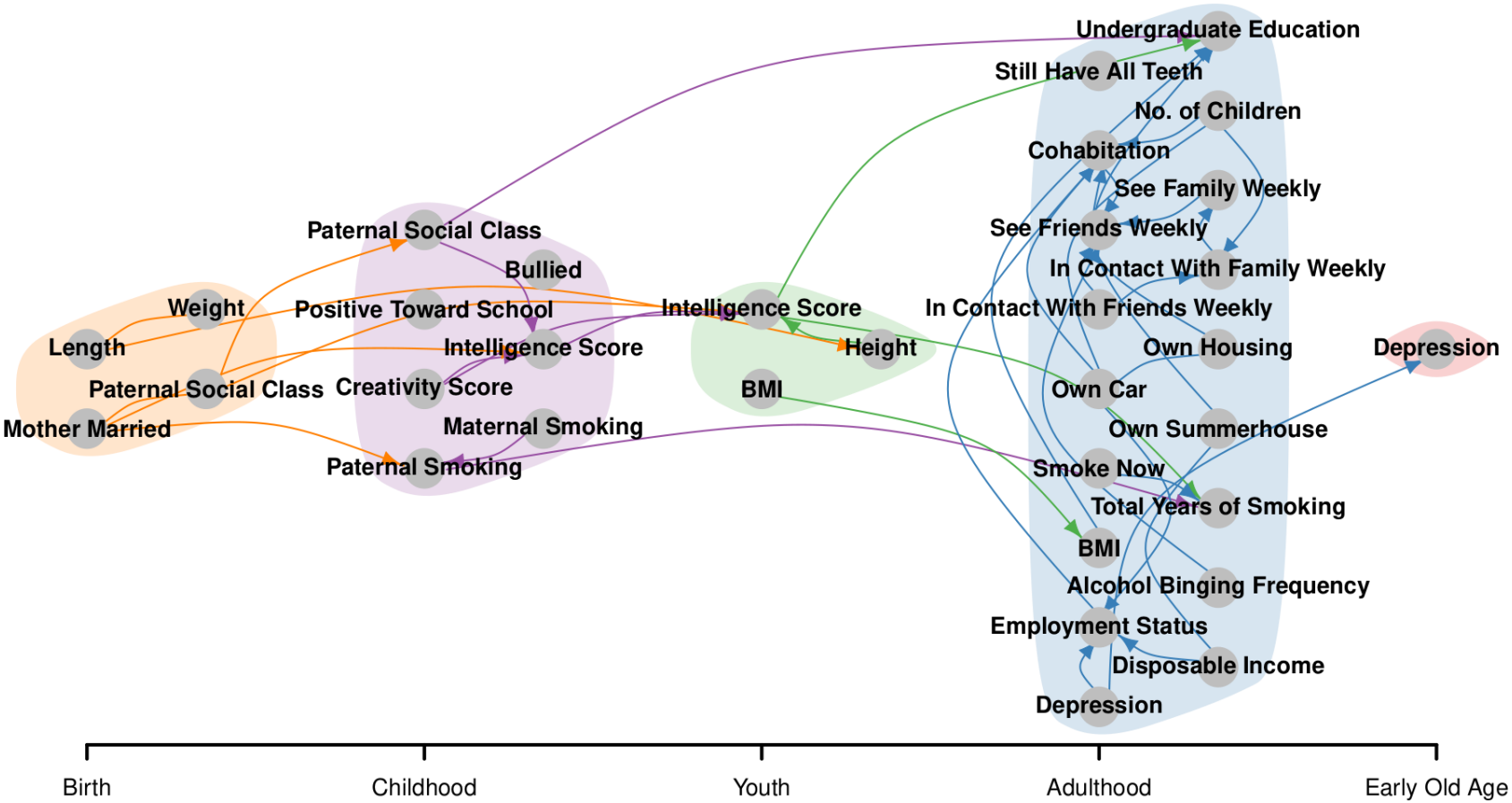
$$\alpha = 10^{-2}$$

Edge retention in application (Petersen, Osler & Ekstrøm 2021)



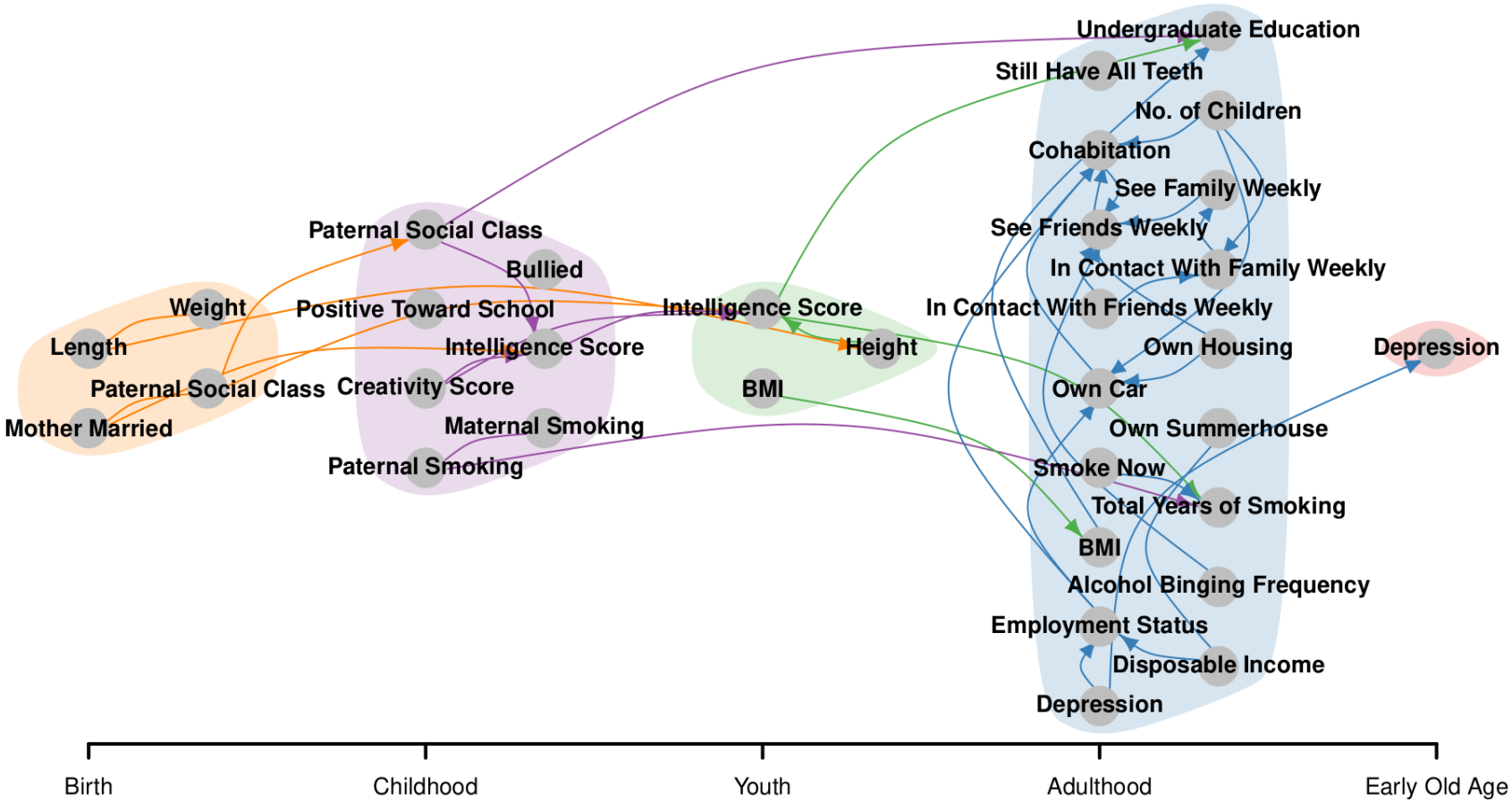
$$\alpha = 10^{-3}$$

Edge retention in application (Petersen, Osler & Ekstrøm 2021)



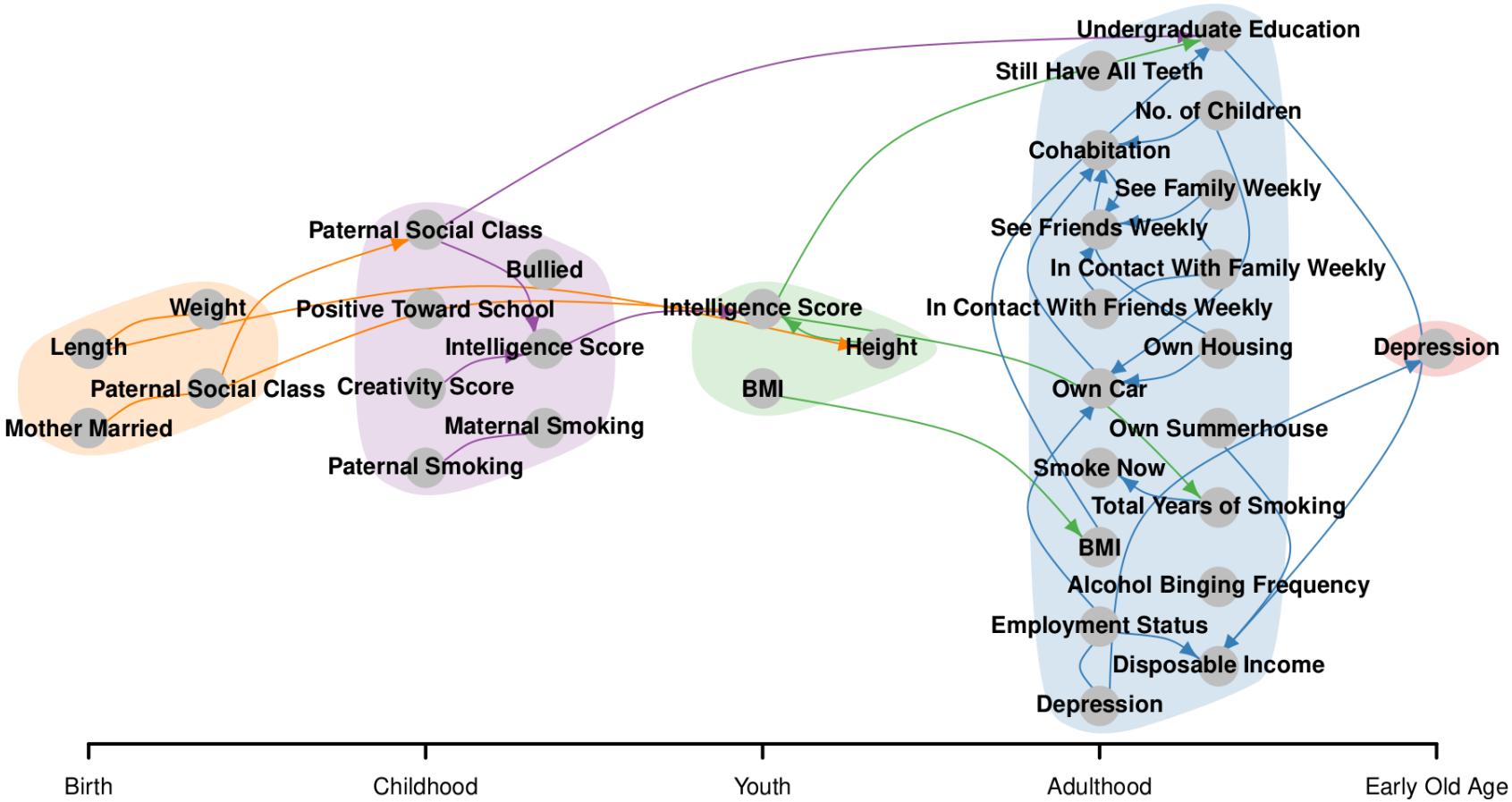
$$\alpha = 10^{-4}$$

Edge retention in application (Petersen, Osler & Ekstrøm 2021)



$$\alpha = 10^{-5}$$

Edge retention in application (Petersen, Osler & Ekstrøm 2021)



$$\alpha = 10^{-6}$$

Edge retention in application (Petersen, Osler & Ekstrøm 2021)

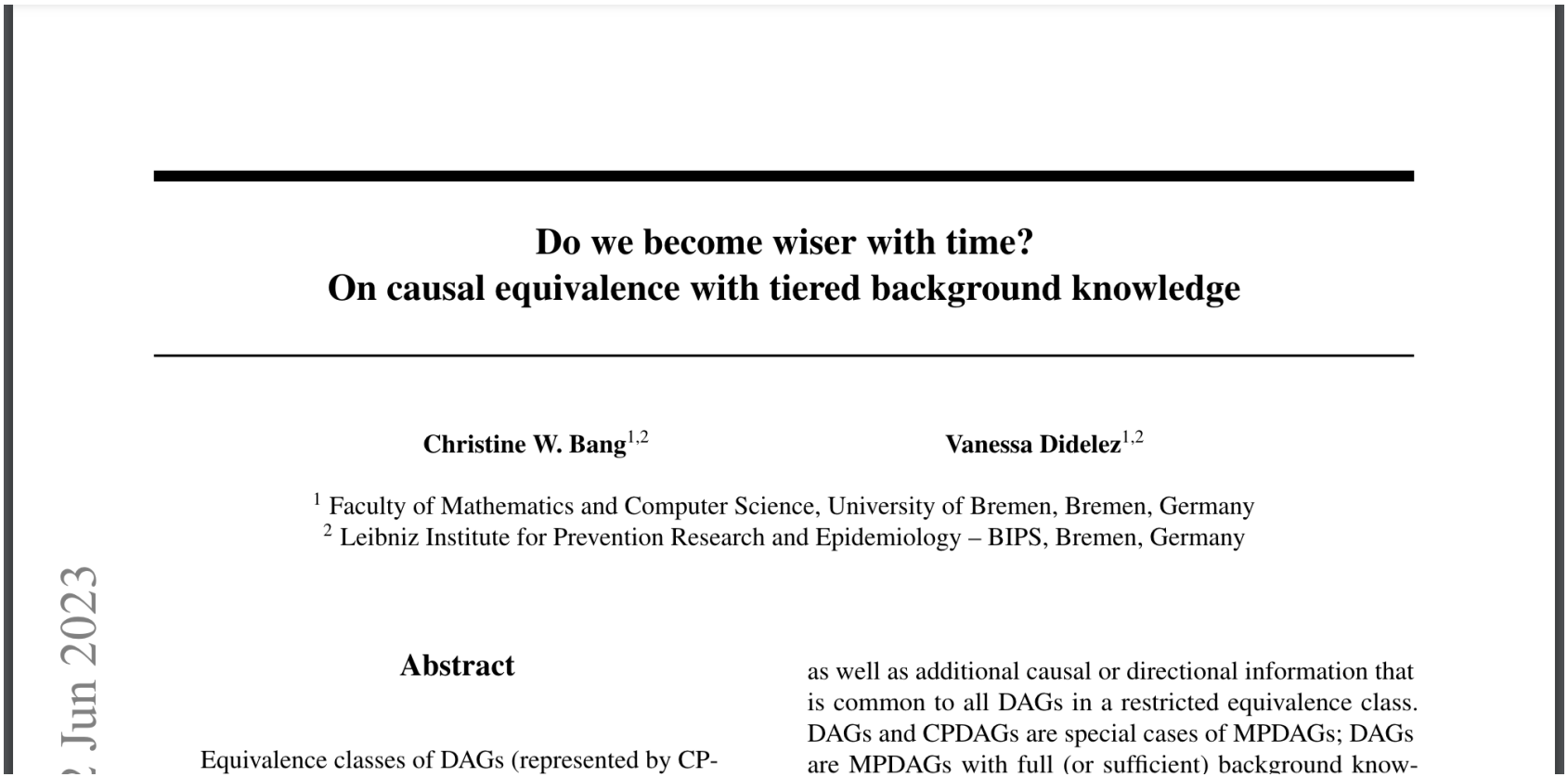
α	d_{total}	d_{new}	d_{removed}	Retention (%)
10^{-2}	61			
10^{-3}	47	0	14	100.00
10^{-4}	39	0	8	100.00
10^{-5}	37	0	2	100.00
10^{-6}	32	1	6	96.88
10^{-7}	27	0	5	100.00
10^{-8}	23	0	4	100.00
10^{-9}	22	0	1	100.00
10^{-10}	22	0	0	100.00

Edge retention in application (Petersen, Osler & Ekstrøm 2021)

α	d_{total}	d_{new}	d_{removed}	Retention (%)
10^{-2}	61			
10^{-3}	47	0	14	100.00
10^{-4}	39	0	8	100.00
10^{-5}	37	0	2	100.00
10^{-6}	32	1	6	96.88
10^{-7}	27	0	5	100.00
10^{-8}	23	0	4	100.00
10^{-9}	22	0	1	100.00
10^{-10}	22	0	0	100.00

Conclusion: As α decreases, more edges are pruned away (monotonically).

What type of temporal information is most useful?



Do we become wiser with time?
On causal equivalence with tiered background knowledge

Christine W. Bang[1,2] **Vanessa Didelez**[1,2]

[1] Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany
[2] Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany

Abstract

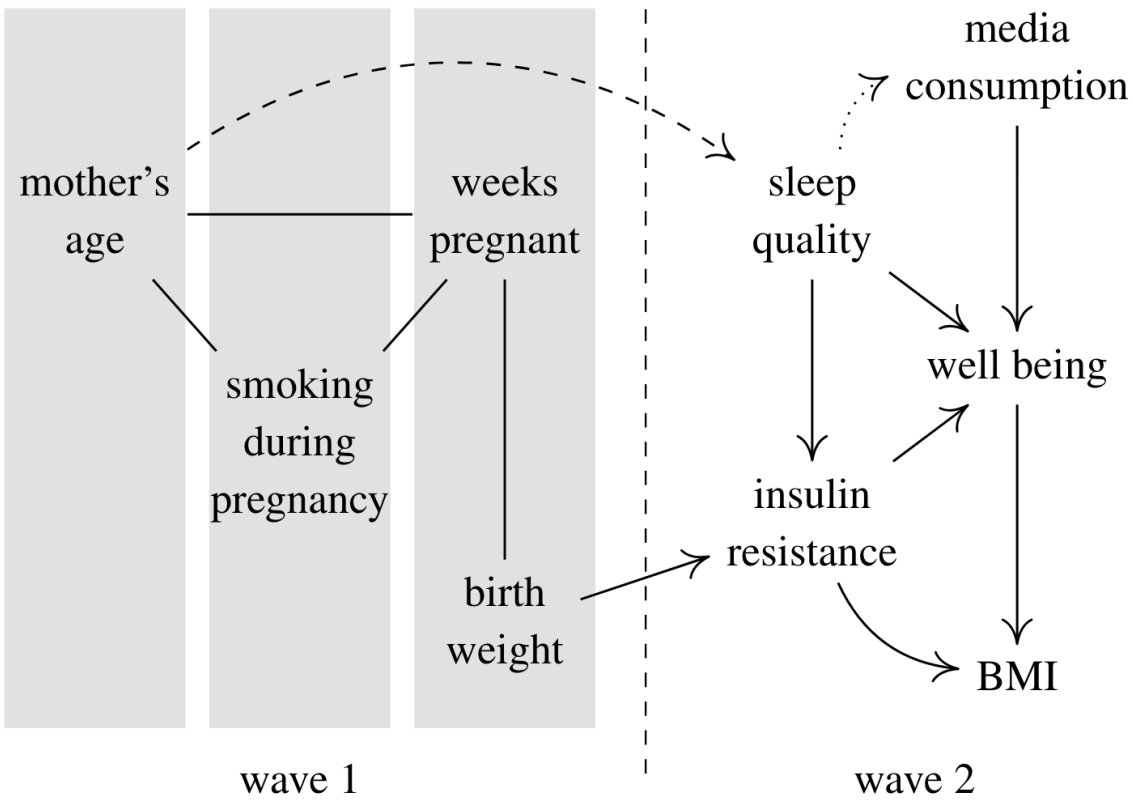
Equivalence classes of DAGs (represented by CP-

as well as additional causal or directional information that is common to all DAGs in a restricted equivalence class. DAGs and CPDAGs are special cases of MPDAGs; DAGs are MPDAGs with full (or sufficient) background know-

Bang & Didelez (2023)

What type of temporal information is most useful?

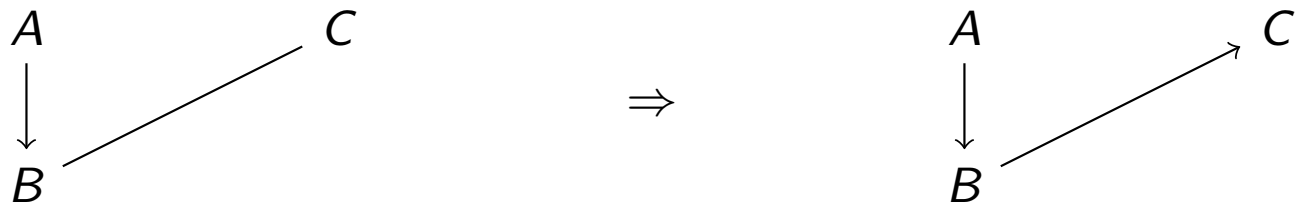
Bang & Didelez show mathematically (large sample limit): Early temporal information is the most useful.



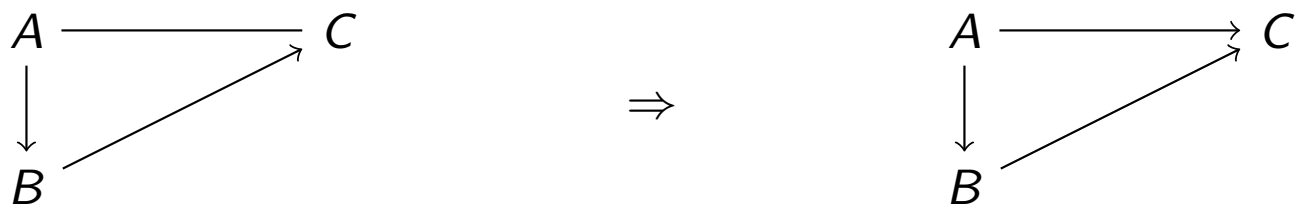
From Bang & Didelez 2023.

Recall: Orientation rules

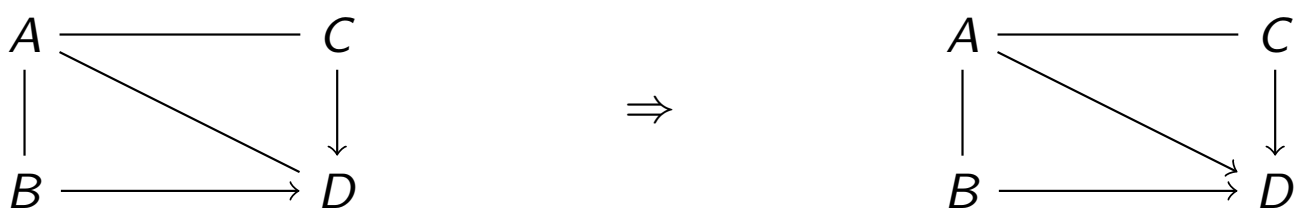
R1: Avoid introducing new v-structures (directly):



R2: Avoid introducing cycles.

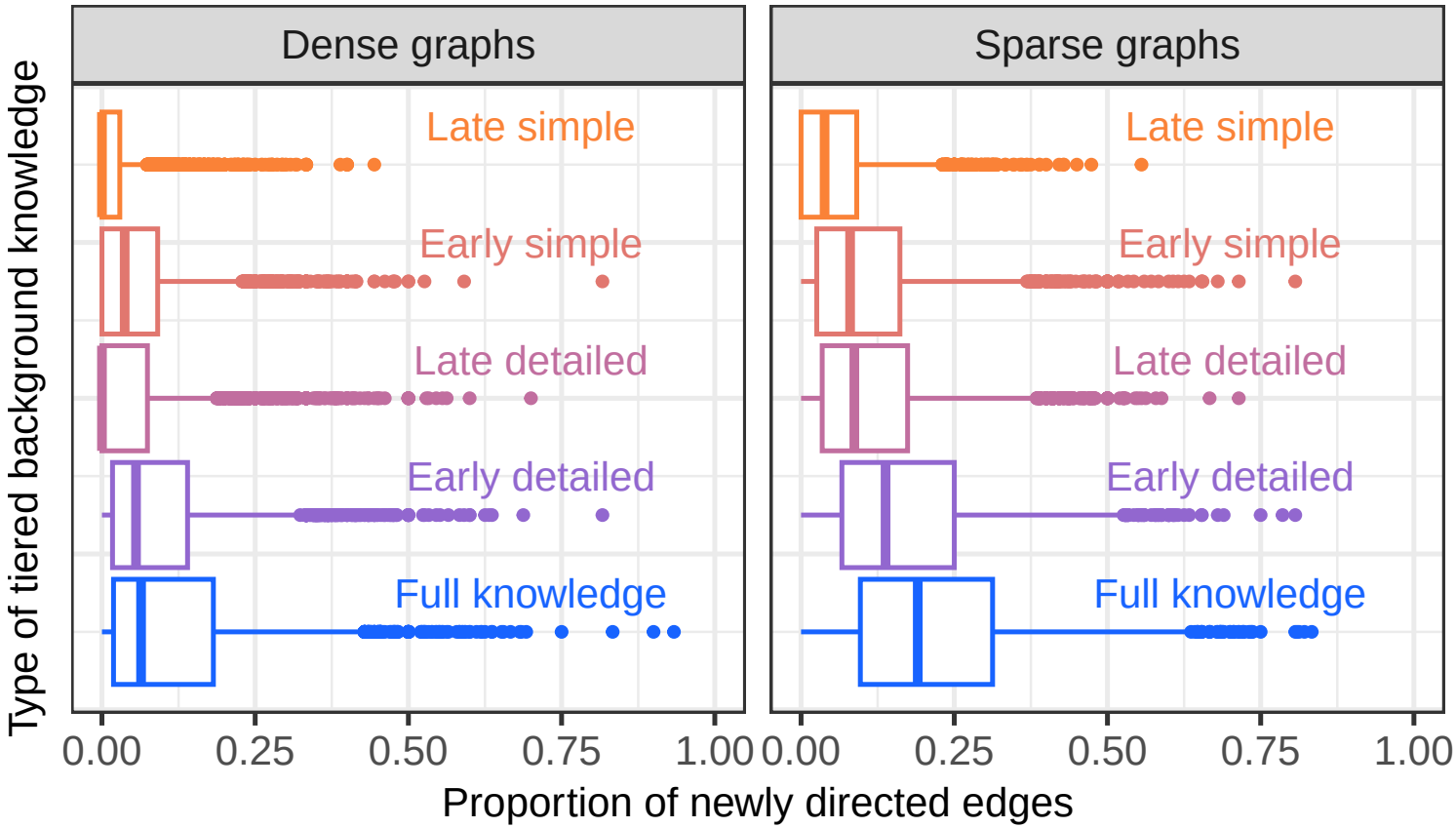


R3: Avoid introducing new v-structures (indirectly).



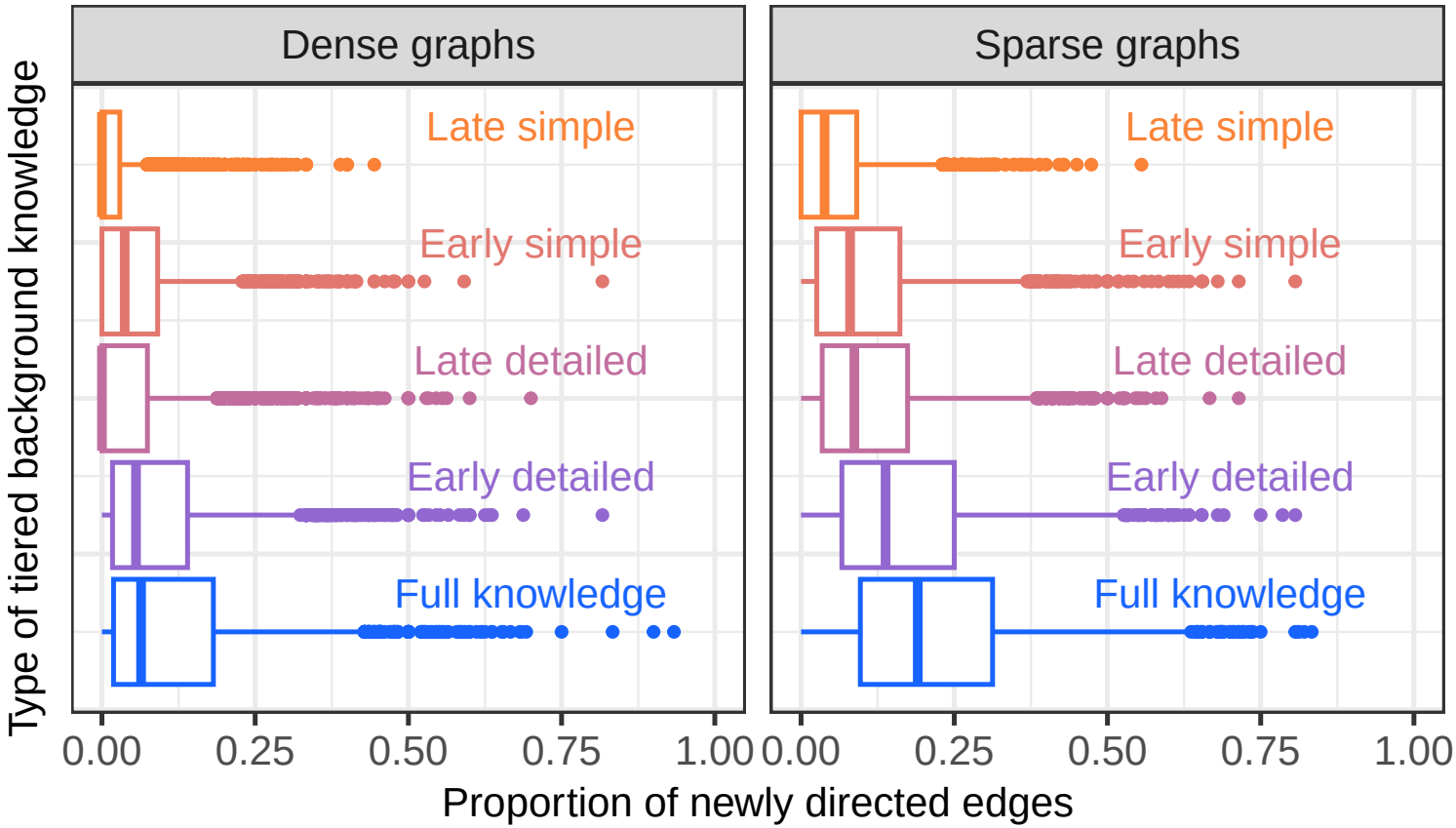
Note: Need "incoming" information to deduce further orientations.

What type of temporal information is most useful?



From Bang & Didelez 2023. Based on simulated graphs (but perfect knowledge about conditional independence).

What type of temporal information is most useful?

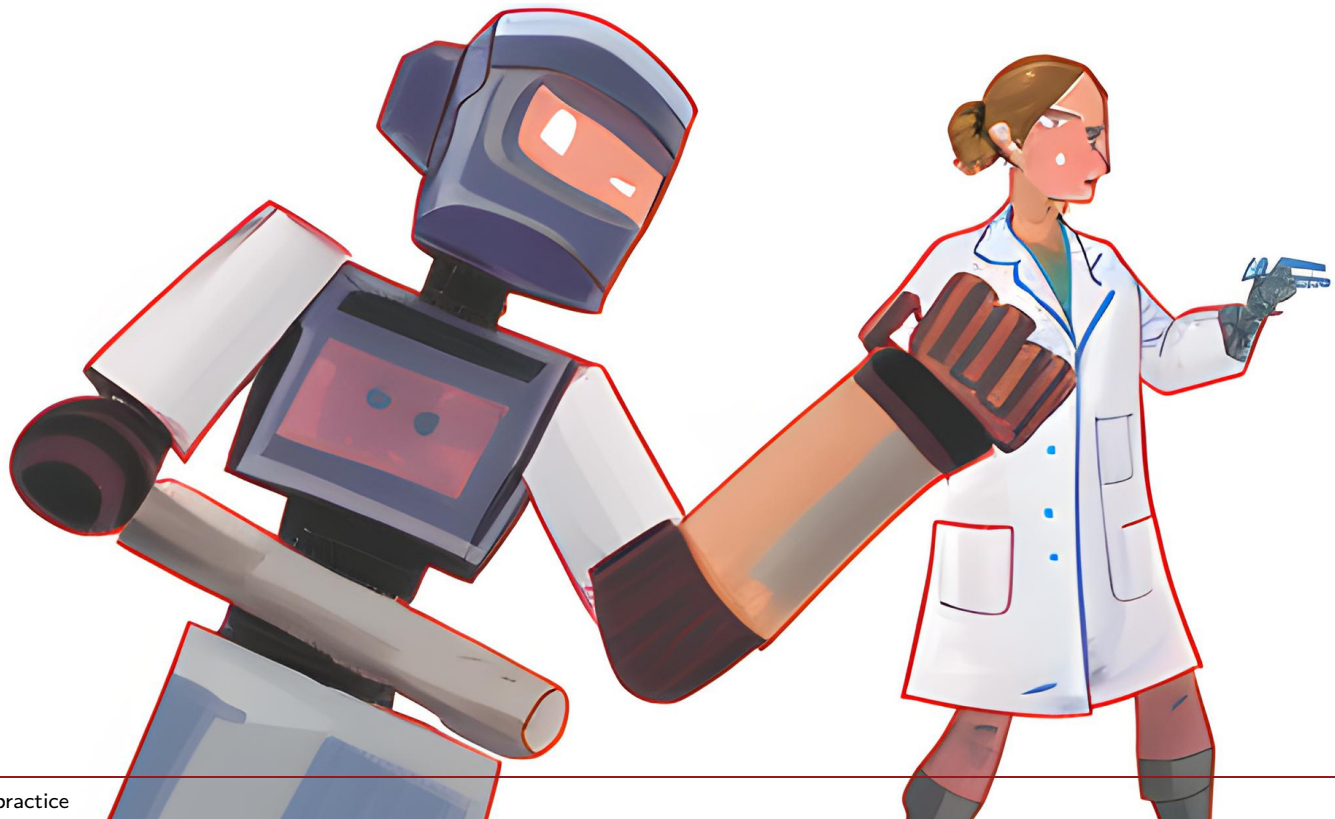


From Bang & Didelez 2023. Based on simulated graphs (but perfect knowledge about conditional independence).

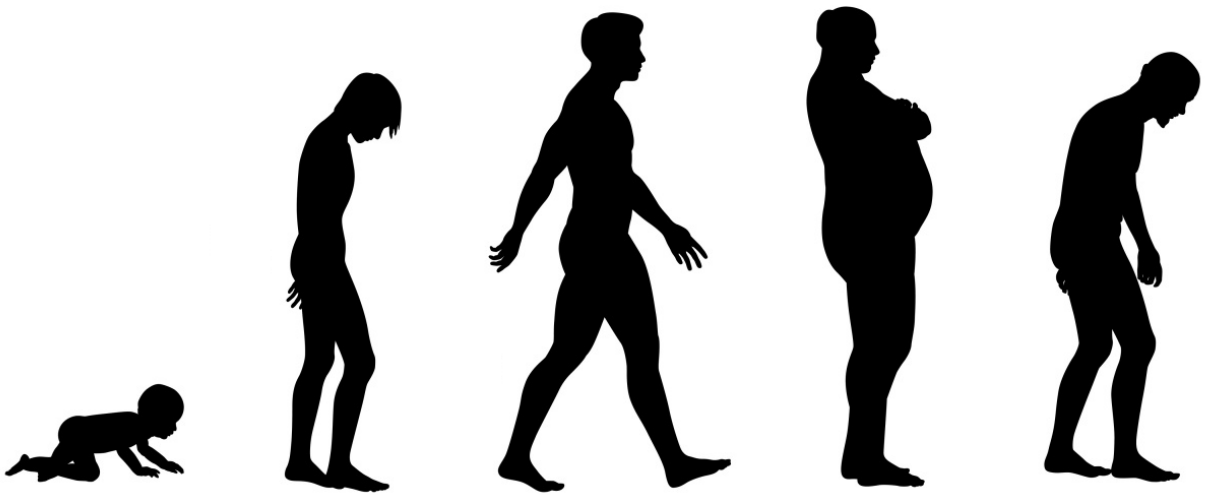
But unclear what happens on real data. . .

How does causal discovery compare with traditional approaches?

Petersen, Ekstrøm, Spirtes & Osler (2023). Constructing causal life course models: Comparative study of data-driven and theory-driven approaches. *American Journal of Epidemiology.*

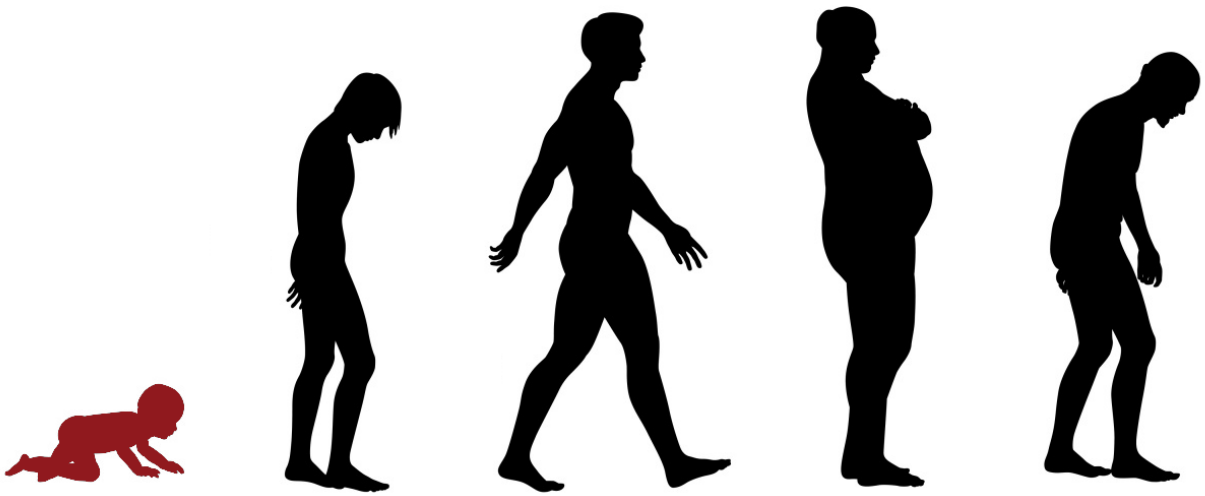


Depression etiology in the Metropolit cohort



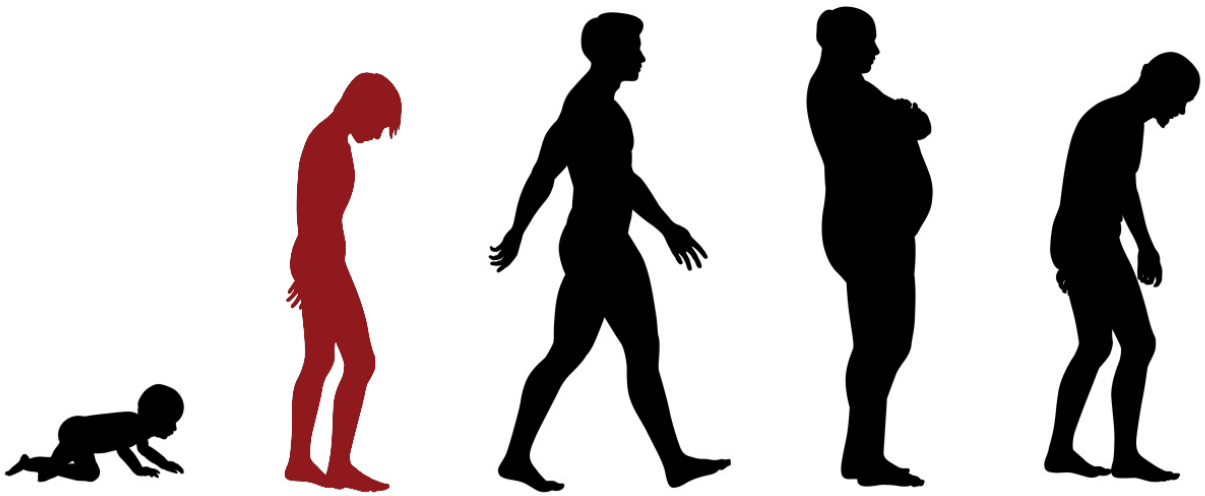
- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.

Depression etiology in the Metropolit cohort



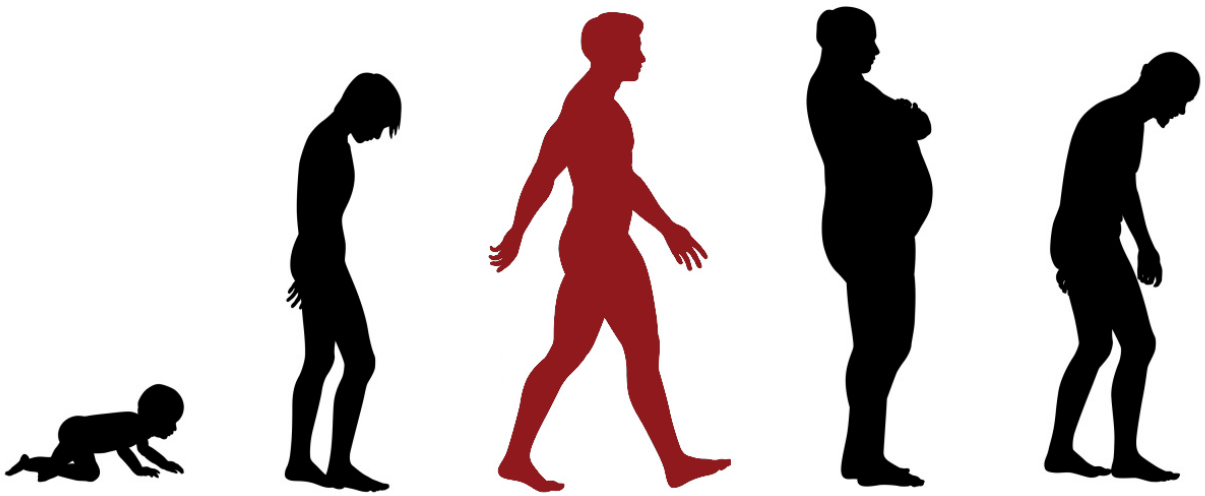
- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.

Depression etiology in the Metropolit cohort



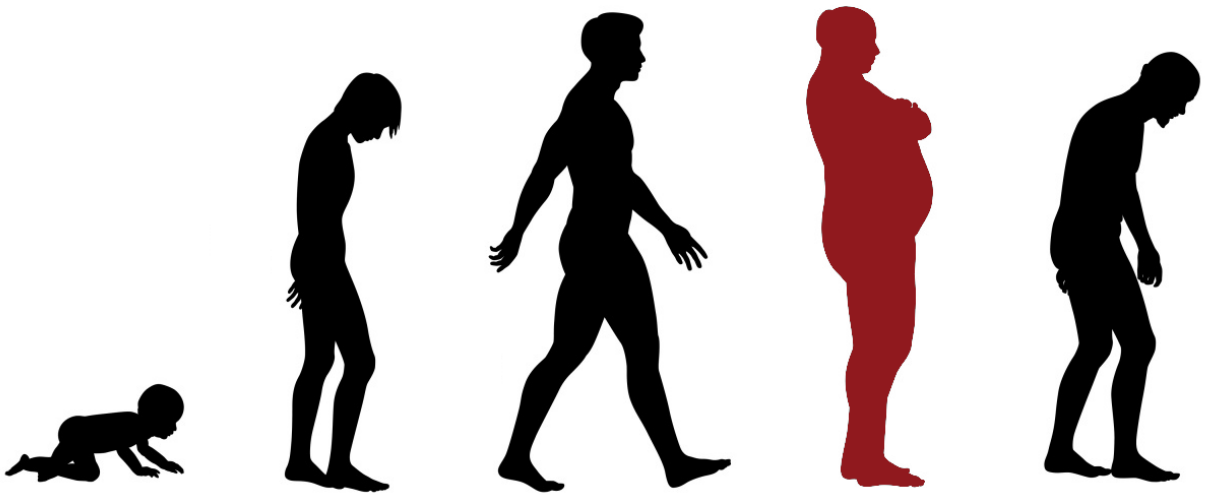
- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.

Depression etiology in the Metropolit cohort



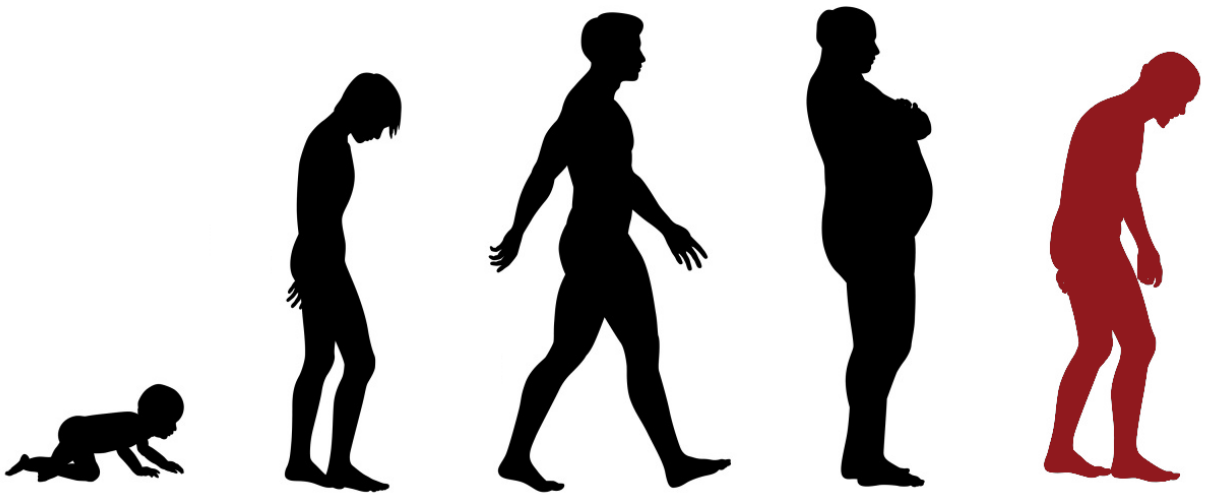
- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.

Depression etiology in the Metropolit cohort

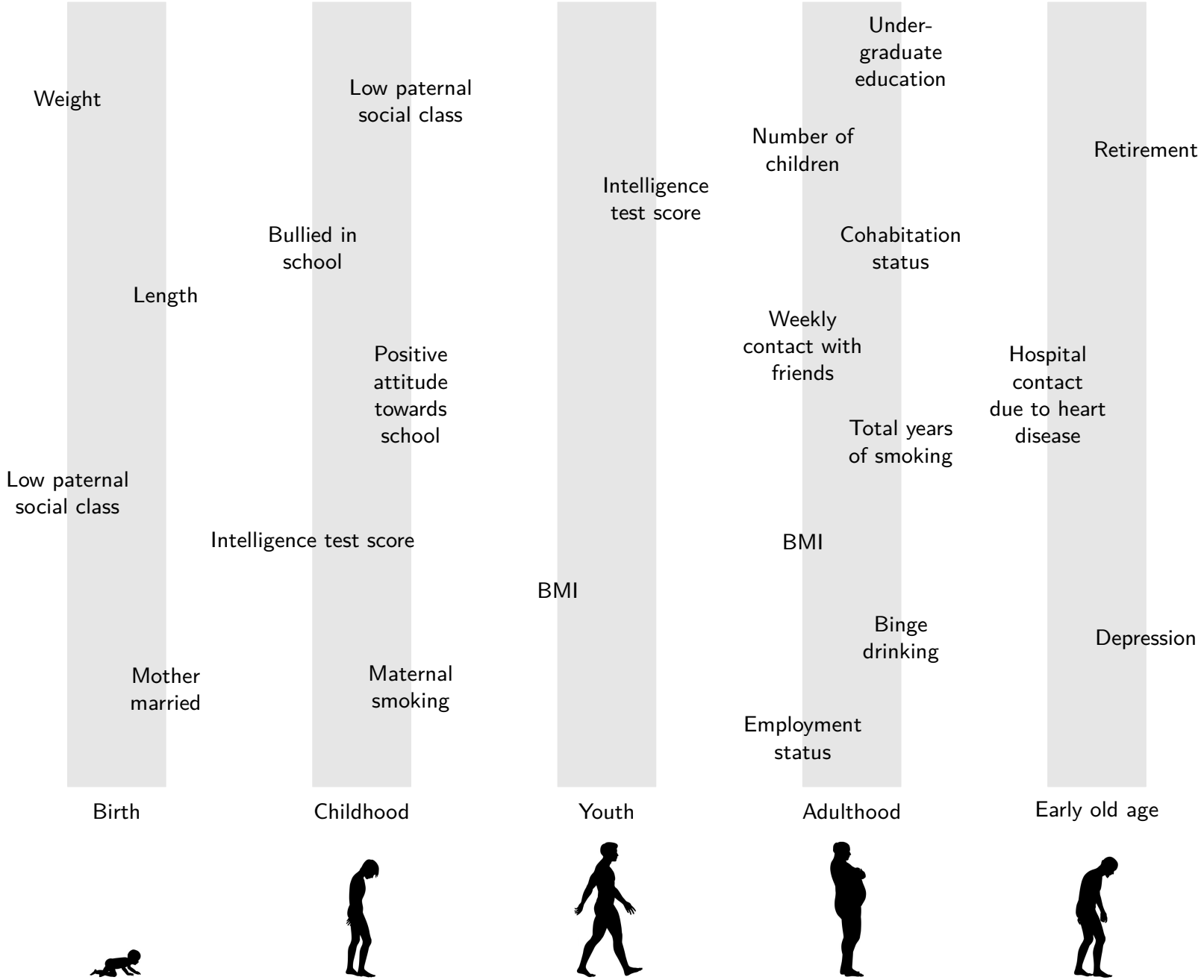


- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.

Depression etiology in the Metropolit cohort



- Cohort encompassing all boys born in the Copenhagen area in 1953 ($n = 12270$).
- Numerous data collections through time and linkage with health registers, social registers etc.
- Retrospective study design: Condition on being alive and residing in Denmark at end-of-followup (2018), and participation.
- We consider 22 variables and $n = 3145$ complete observations.


Weight
Length
Low paternal social class
Mother married
Low paternal social class
Bullied in school
Positive attitude towards school
Maternal smoking
Intelligence test score
Intelligence test score
BMI
BMI
Under-graduate education
Number of children
Cohabitation status
Weekly contact with friends
Total years of smoking
Binge drinking
Employment status
Retirement
Hospital contact due to heart disease
Depression
Birth
Childhood
Youth
Adulthood
Early old age

Study design

- **Focus on case:** Life course epidemiological study regarding etiology of depression and heart disease in early old age
- **Theory-based model construction:** DAGs constructed by epidemiologists (*experts*)
- **Data-driven model construction:** Apply temporal PC algorithm to dataset based on the Metropolit cohort (n = 3145)
- **Compare** these models
  - Assume that expert model is (mostly) correct, but possibly incomplete
  - Expect that data-driven model may or may not be correct, but perhaps more likely to be complete

Theory-based model construction: Expert DAGs

- Recruited two **experts** (health researchers with experience in epidemiology of heart disease and psychiatry)
- Experts were given:
  - List of 22 variables (no data) with temporal information
  - Information about the intended study population
  - Written instructions for DAG construction
- Each edge was annotated with **label of confidence**: Moderate/high
- One individual model from each expert + joint **consensus model**

Expert instructions

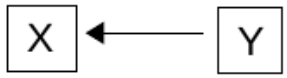
Guidelines for theory-driven model construction: DAG

Using the list of variables provided below, we ask you to construct a directed acyclic graph (DAG) for the data generating mechanism behind these variables. Constructing a DAG involves suggesting a number of causal relationships between the variables. In order to decide on which potential causal relationships exist between two variables, we ask you to consult general theory, relevant literature and previous empirical studies of the involved variables. However, *you are not allowed to refer to previous empirical studies conducted on the same dataset (the Metropolit cohort)*. For some of the variables, there may not exist specific theory or studies to help you in determining causal relationships. In these cases, we ask you to provide your best educated guess for what causal relationships may or may not exist. Please make sure you do not propose causal relationships that go against the direction of time (see temporal information on the variable list).

How to add arrows to the DAG

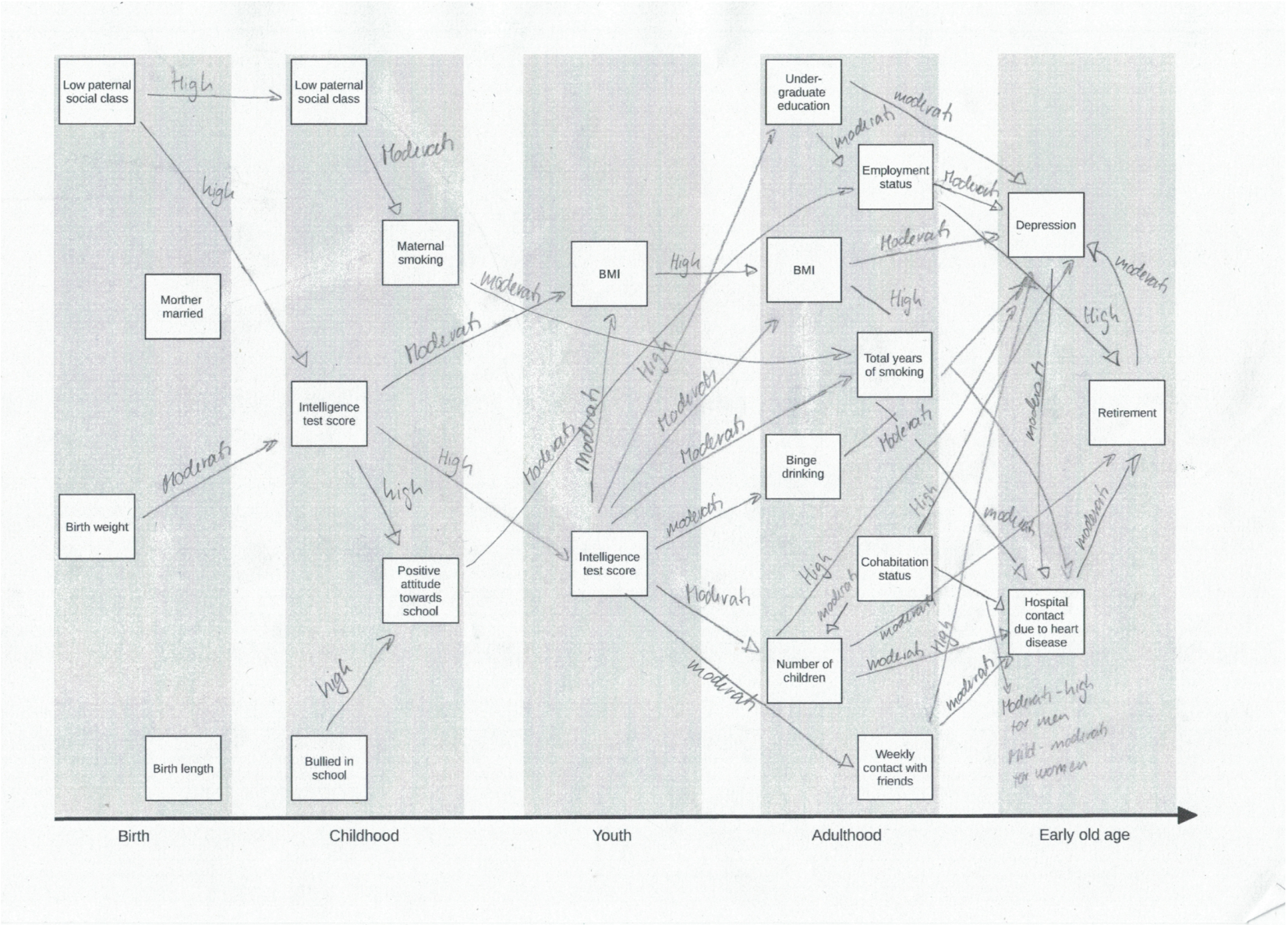
In order to construct the DAG, you need to add arrows between variables in the attached DAG template. For each pair of variables, we ask you to draw an arrow between them if you believe that one is a potential direct cause of the other. A direct causal effect is a causal effect that is not mediated via other variables. For example, for two variables X and Y, where X is a direct cause of Y, you should draw the following arrow:

If you draw the arrow in the opposite direction, it means that Y is a potential direct cause of X:



Finally, if you do not draw an arrow between X and Y it means that you do not believe that there is any direct causal relationships between the two variables: X is not a potential direct cause of Y, *and* Y is not a potential direct cause of X:

Example: An expert graph



Data-driven model construction: Temporal PC algorithm

We used TPC with GLM-based test of non-association.

We considered two strategies for choosing test significance level (α):

TPC-S: **Search** for α such that the number of edges equals the number of edges in the expert consensus graph.

TPC-P: **Pre-specified** value of $\alpha = 0.01$.

Data-driven model construction: Temporal PC algorithm

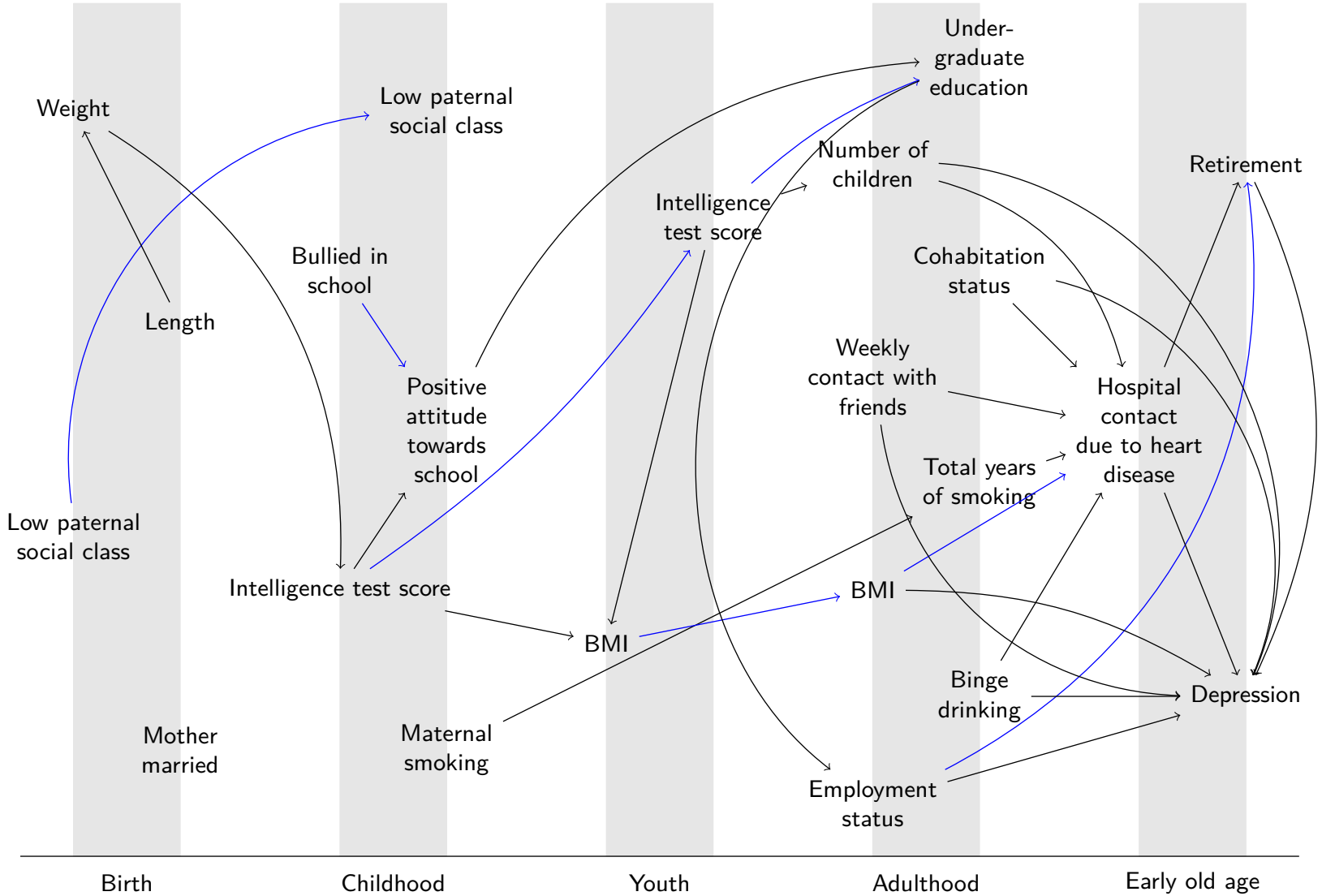
We used TPC with GLM-based test of non-association.

We considered two strategies for choosing test significance level (α):

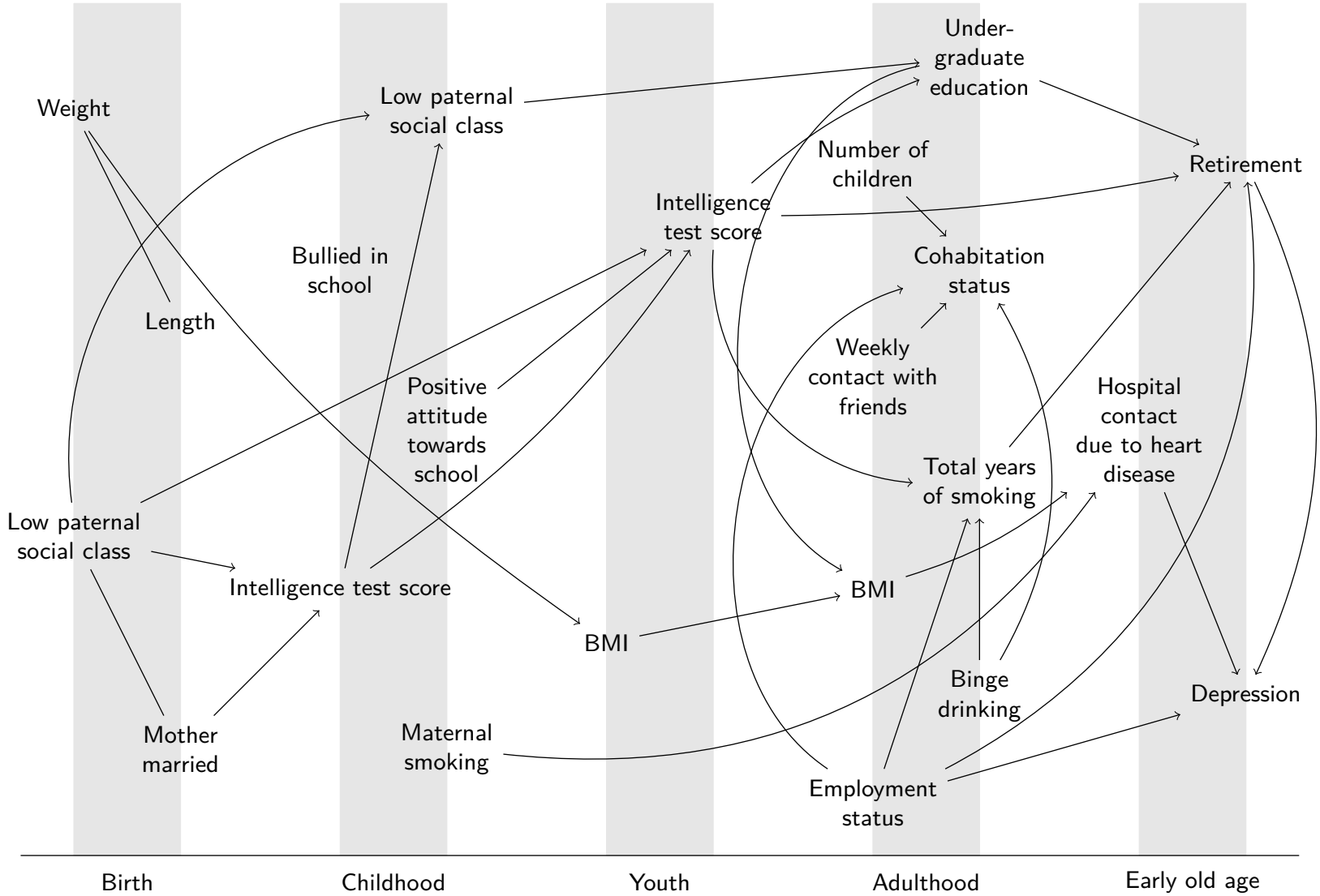
TPC-S: **Search** for α such that the number of edges equals the number of edges in the expert consensus graph.

TPC-P: **Pre-specified** value of $\alpha = 0.01$.

Results: Expert consensus model



Results: TPC-S (α-search) model



Comparison: Expert consensus and TPC-S

		Expert consensus	
		Adjacency	Non-adjacency
TPC-S	Adjacency	10	20
	Non-adjacency	20	181

- Among shared adjacencies, no disagreement on orientation (although 1 unoriented by TPC-S)
- Overall test of fit (Petersen 2025): $p = 0.002$ (comparing with random guessing), expected no. true adjacencies found under random guessing: 3.9, 95% CI: (1; 7).
- High confidence edges: 6 out of 7 found by TPC-S, all oriented in same direction as experts.

Comparison: Expert consensus and TPC-S

		Expert consensus	
		Adjacency	Non-adjacency
TPC-S	Adjacency	10	20
	Non-adjacency	20	181

- Among shared adjacencies, no disagreement on orientation (although 1 unoriented by TPC-S)
- Overall test of fit (Petersen 2025): $p = 0.002$ (comparing with random guessing), expected no. true adjacencies found under random guessing: 3.9, 95% CI: (1; 7).
- High confidence edges: 6 out of 7 found by TPC-S, all oriented in same direction as experts.

Plausibility of additional edges in TPC-S model

Post-hoc assessment of plausibility of additional edges in TPC-S model:

- All 20 additional edges classified into low/moderate/high plausibility by reference to epidemiological theory and literature.
- Results:

  Low plausibility: 3 edges.

  Moderate plausibility: 6 edges.

  High plausibility: 11 edges.

$\Rightarrow$ Additional suggestions from TPC-S seem mostly useful.

Stability of TPC-S results

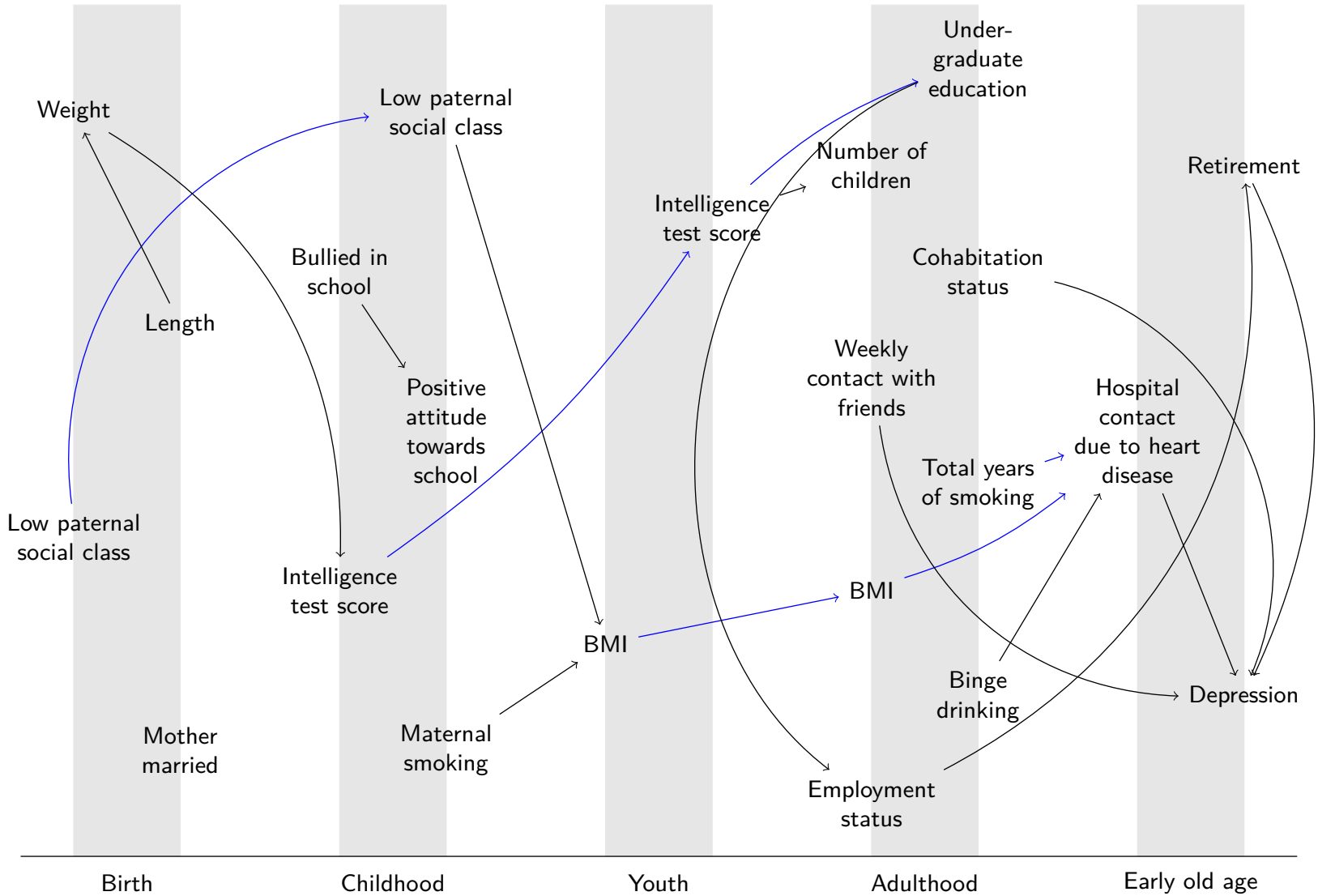
			Count	In full?
Low paternal social class (B)	→	Low paternal social class (C)	100	×
Intelligence test score (C)	→	Intelligence test score (Y)	100	×
BMI (Y)	→	BMI (A)	100	×
Intelligence test score (Y)	→	Total years of smoking (A)	100	×
Intelligence test score (Y)	→	Undergraduate education (A)	100	×
Employment status (A)	→	Retirement (E)	100	×
Length (B)	–	Weight (B)	100	×
Low paternal social class (B)	→	Intelligence test score (Y)	95	×
Mother married (B)	–	Low paternal social class (B)	94	×
Low paternal social class (C)	→	Undergraduate education (A)	93	×
Binge drinking (A)	→	Total years of smoking (A)	88	×
BMI (A)	→	Hospital contact due to heart disease (E)	88	×
Undergraduate education (A)	→	Retirement (E)	86	×
Employment status (A)	→	Total years of smoking (A)	84	×
Number of children (A)	→	Cohabitation status (A)	82	×
Mother married (B)	→	Intelligence test score (C)	79	×
Employment status (A)	→	Cohabitation status (A)	76	×
Undergraduate education (A)	→	BMI (A)	70	×
Total years of smoking (A)	→	Retirement (E)	69	×
Positive attitude towards school (C)	→	Intelligence test score (Y)	67	×
Low paternal social class (B)	→	Intelligence test score (C)	66	×
Weekly contact with friends (A)	→	Cohabitation status (A)	66	×
Intelligence test score (Y)	→	Retirement (E)	65	×
Employment status (A)	→	Depression (E)	64	×
Hospital contact due to heart disease (E)	→	Depression (E)	64	×
Weight (B)	→	BMI (Y)	61	×
Depression (E)	→	Retirement (E)	58	
Low paternal social class (C)	→	Intelligence test score (C)	52	
Maternal smoking (C)	→	Hospital contact due to heart disease (E)	43	×
Undergraduate education (A)	→	Total years of smoking (A)	42	
Retirement (E)	→	Depression (E)	42	×
Binge drinking (A)	→	Cohabitation status (A)	38	×

Inter expert agreement

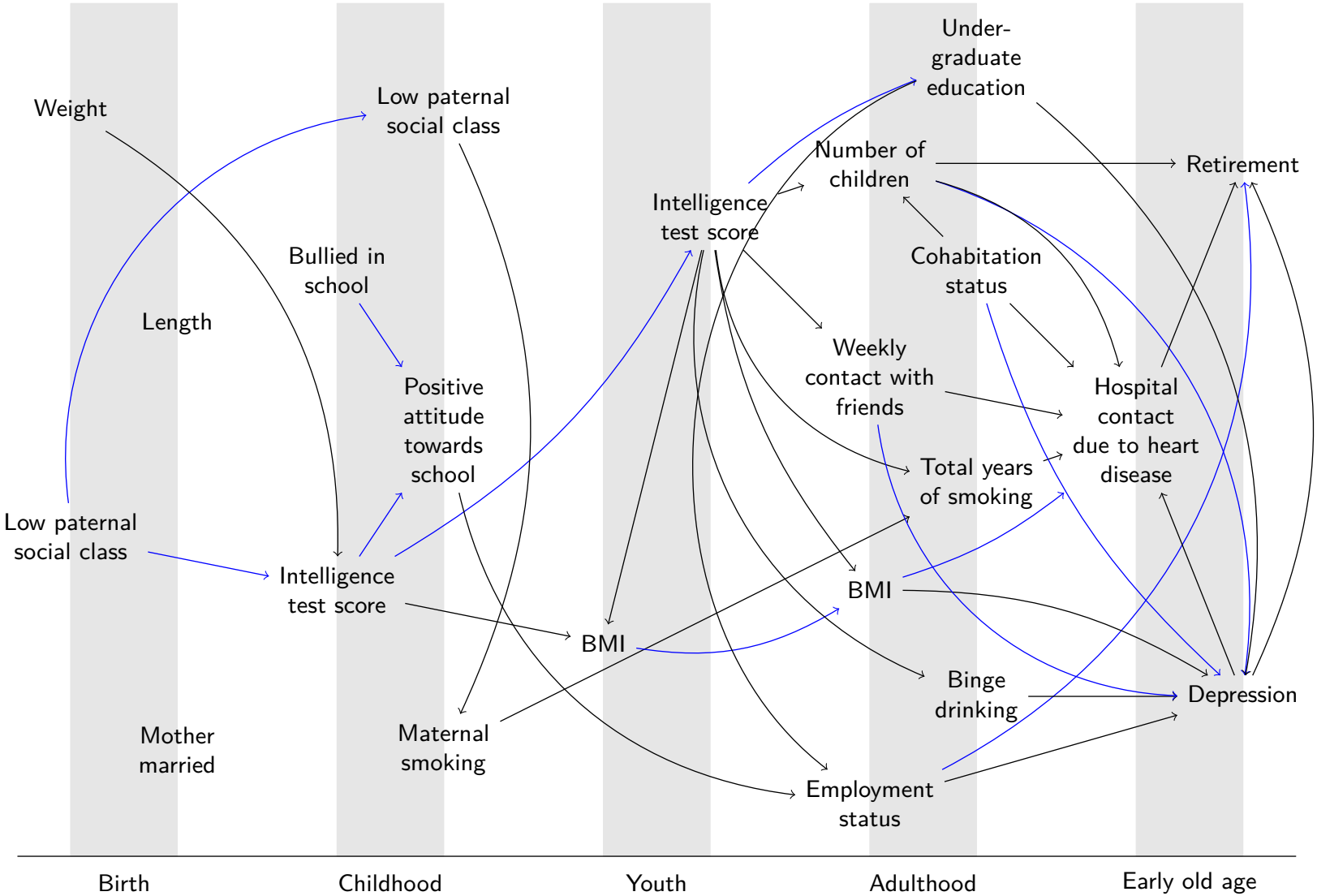
		Expert 1	
		Adjacency	Non-adjacency
Expert 2	Adjacency	15	22
	Non-adjacency	4	190

- Large disagreement about the number of edges (expert 1: 19, expert 2: 37).
- Agreement about orientation for 13 out of 15 shared edges.
- 5 edges marked with high confidence by both experts, agreement on orientation for all of these.

Expert 1 model



Expert 2 model



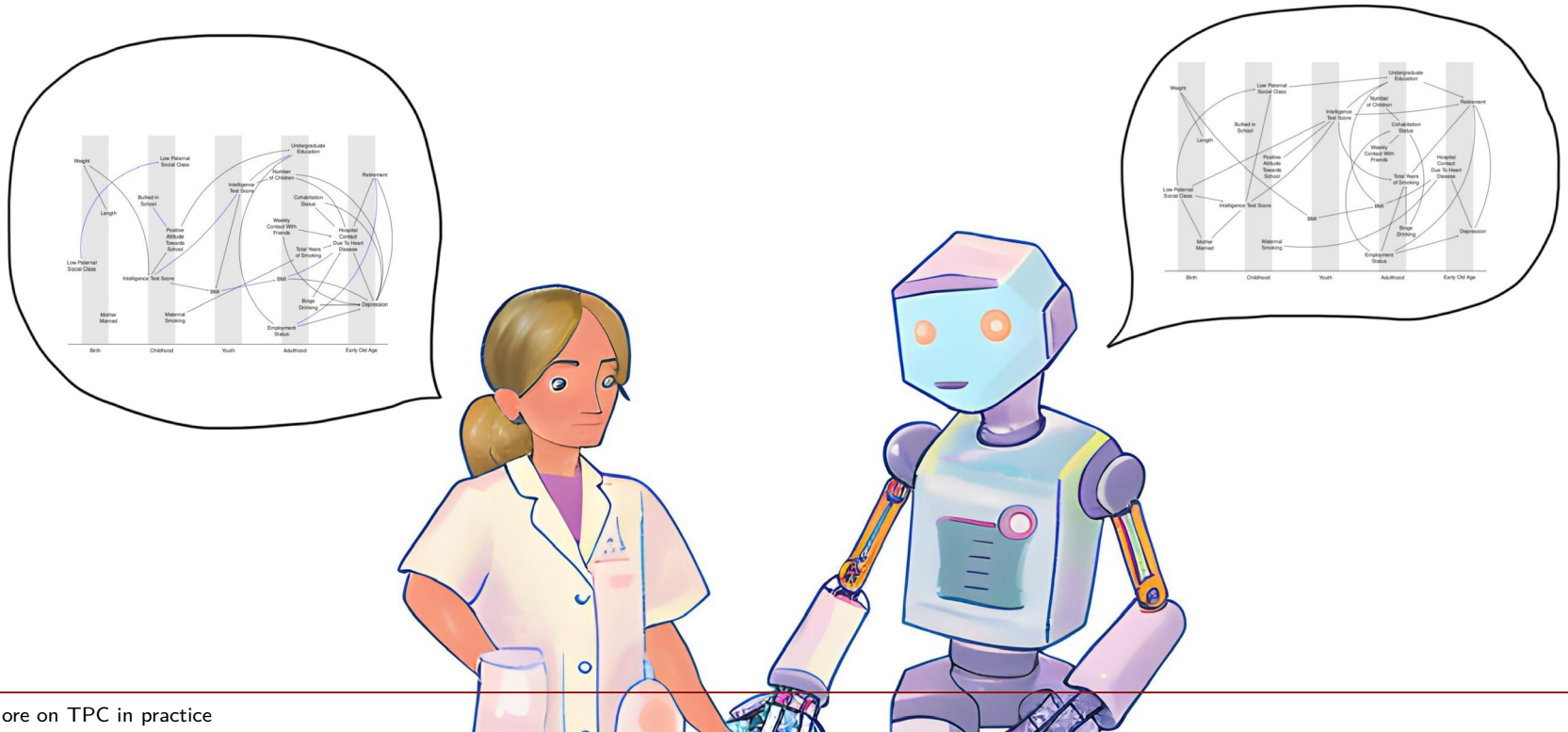
Conclusions

- TPC recovers parts of the causal model
  - Especially good at recovering "high confidence" causal links
- TPC gives rather stable results, especially for "high confidence" causal links
- Experts seem to overlook some plausible causal links at first
- Experts don't fully agree! Room for improvement over existing approach (often 1-2 experts)

Recommendation: Combine and conquer!

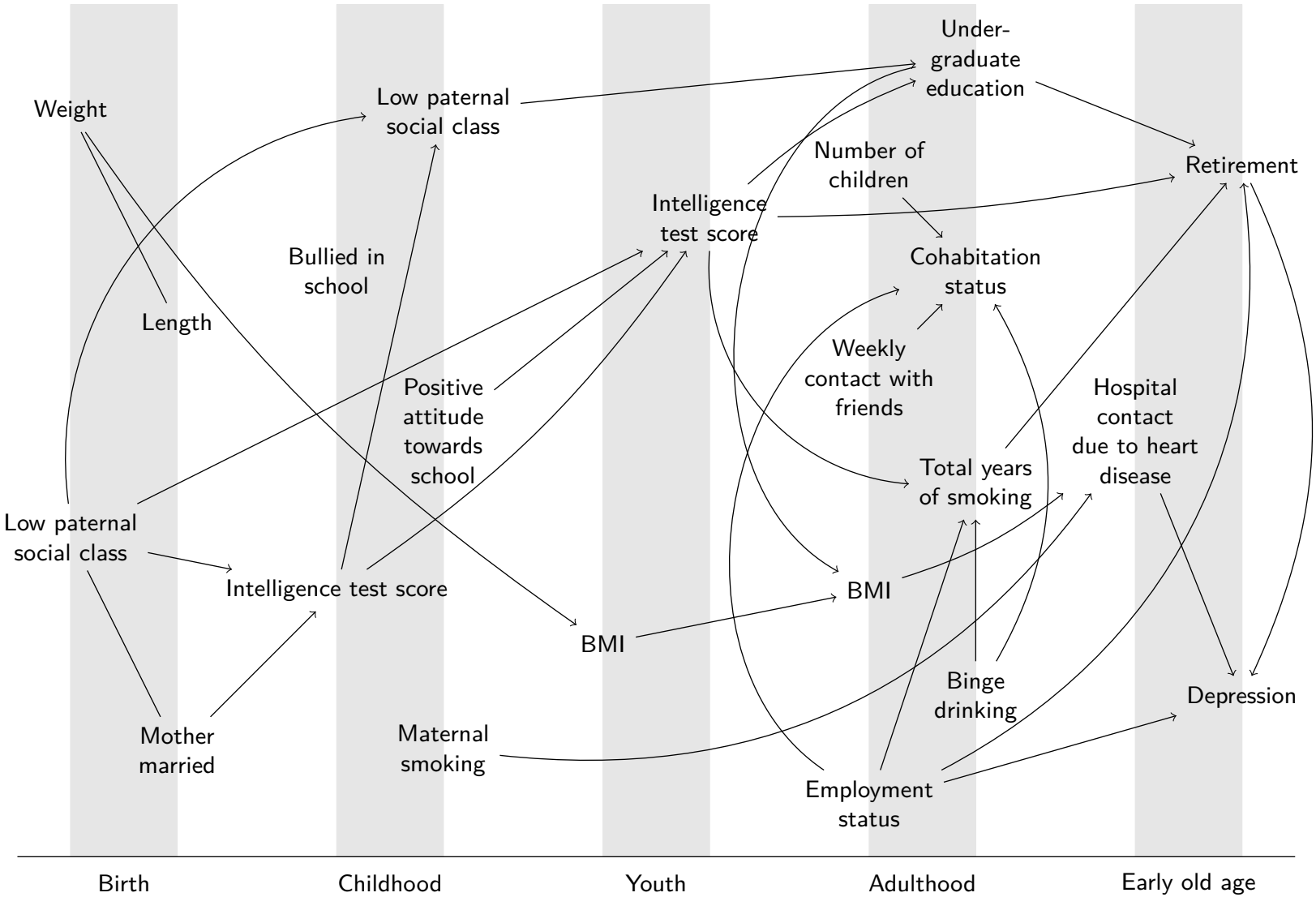
Idea for pipeline:

1. Construct expert (consensus) model
2. Use TPC-S with edge number from expert model
3. Assess TPC-S results critically, add plausible new suggestions to expert model draft $\Rightarrow$ Final combined model



But we assumed unobserved confounding...

Plausible assumption?



Unobserved confounding in PC

- If there **is** unobserved confounding, and we have infinite data, we know (mathematically) that the output from PC gets too many edges, not too few (Spirtes, Glymour & Scheines 2001).
- On finite data PC is generally biased towards sparse graphs, i.e. too few edges, due to the way statistical errors propagate (Petersen, Ramsey, Ekstrøm & Spirtes 2023).

Unobserved confounding in PC

- If there **is** unobserved confounding, and we have infinite data, we know (mathematically) that the output from PC gets too many edges, not too few (Spirtes, Glymour & Scheines 2001).
- On finite data PC is generally biased towards sparse graphs, i.e. too few edges, due to the way statistical errors propagate (Petersen, Ramsey, Ekstrøm & Spirtes 2023).
- We don't know how these two points interact on finite data.

Unobserved confounding in PC

- If there **is** unobserved confounding, and we have infinite data, we know (mathematically) that the output from PC gets too many edges, not too few (Spirtes, Glymour & Scheines 2001).
- On finite data PC is generally biased towards sparse graphs, i.e. too few edges, due to the way statistical errors propagate (Petersen, Ramsey, Ekstrøm & Spirtes 2023).
- We don't know how these two points interact on finite data.
- We don't know what happens to edge orientations, neither on "infinite" or finite data.

References

Bang & Didelez (2023). Do we become wiser with time? On causal equivalence with tiered background knowledge. In *Proceedings of Uncertainty in Artificial Intelligence*.

Petersen, Ekstrøm, Spirtes & Osler (2023). Constructing causal life course models: Comparative study of data-driven and theory-driven approaches. *American Journal of Epidemiology*.

Petersen, Ramsey, Ekstrøm & Spirtes (2023). Causal Discovery for Observational Sciences Using Supervised Machine Learning. *Journal of Data Science*.

Spirtes, Glymour & Scheines (2001). Causation, prediction, and search. *MIT press*.

Petersen (2025): Are you doing better than random guessing? A call for using negative controls when evaluating causal discovery algorithms. To appear in *Proceedings of Uncertainty in Artificial Intelligence*.