Expert-driven versus Data-driven Causal Graph Construction in Epidemiology

Vanessa Didelez (jointly with Janine Witte, Claudia Börnhorst, Ronja Foraita)

Leibniz Institute for Prevention Research and Epidemiology – BIPS,
& Faculty of Mathematics/Comp.Sci, University of Bremen, Germany

23 April 2024

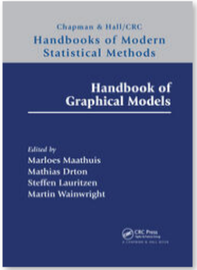
Applied Causal Graphs Workshop, Berlin

Overview

- Part 1: Introduction — what is a (causal) DAG?
- Part 2: Optimal adjustment sets
- Part 3: Expert-driven causal DAG construction
  - → Example: IDEFICS/I.Family children's cohort study — estimating effect(s) of hypothetical sustained interventions on health-behaviours on BMI/obesity
- Part 4: Data-driven (causal) DAG construction
  - → Example: IDEFICS/I.Family — discovering direct and indirect causal paths

Part 1

What is a (Causal) DAG?



Chapter

Causal Concepts and Graphical Models

By Vanessa Didelez

Book [Handbook of Graphical Models]

Edition	1st Edition
First Published	2018
Imprint	CRC Press

Aim of Causal Inference

Analyse data for inference on probabilistic behaviour under (hypothetical) interventions on a ‘system’

Here: aim is to **inform decision making**

– in medicine, for public health authorities, for doctors, and individuals
– assess risks and benefits of life-style, drugs, preventive measures etc.
– wanted: actionable / policy-relevant analyses

Challenge: for many important research questions / decision problems there are no, and never will be, RCTs

$\Rightarrow$ must use observational studies or otherwise available data

e.g. cohort data or routinely collected data (such as health claims)

What is a (Probabilistic) DAG?
aka Bayesian Network, Probabilistic Expert Sytem etc.

- $\mathcal{G} = (\mathbf{V}, \mathbf{E})$
  vertices/nodes (= variables), directed edges $\longrightarrow$, no directed cycles

- Imposes factorisation into factors $p(V \mid \mathrm{pa}(V))$, $V \in \mathbf{V}$, of joint distribution
  $\Rightarrow$ implies conditional independencies for every non-edge
  ... which can be read off by d-separation

What is a (Causal) DAG?

aka Causal Graph, Causal Diagram, Influence Diagram etc.

- the above $+$ various versions of 'causal' semantics / augmentation
- "Causal" if distribution under interventions accurately represented by truncated factorisation

  $\Rightarrow$ edge represents a 'controlled direct' causal effect *relative* to V

  $\Rightarrow$ directed paths = causal paths; other open paths are non-causal

- Little known fact: can define and work with *locally* causal DAGs

  – often more plausible as many nodes/variables not intervenable

What is a (Probabilistic / Causal) DAG?

Important:

- Model restrictions are imposed by
  - **absence of edges** (non-edges) and
  - **absence of further nodes** with ≥ 2 children (non-nodes)
- Edge = possible (direct) causal relation that is not restricted to be null

DAG Example

Informally:

- nodes represent variables
- non-edges represent conditional independencies in the underlying joint distribution
- d-separation to read off all (cond.) independencies

$$
\begin{tikzcd}
 & X_1 \arrow[rr] \arrow[dl] & & X_2 \arrow[dr] & \\
A \arrow[rr] & & X_3 \arrow[rr] & & Y
\end{tikzcd}
$$

Example:

$Y \perp\!\!\!\perp (A, X_1) \mid (X_2, X_3);$ but $X_2 \not\!\perp\!\!\!\perp X_3 \mid Y$ — aka ‘collider effect’

Causal DAG Example (ctd)

Informally:

- nodes represent variables
- edges represent direct causal effects
- directed paths represent some causal effects
- backdoor path from A to Y induces association blocked by $\{X_1\}$ or $\{X_2\}$ or $\{X_1, X_2\}$

$$
\begin{tikzcd}
 & X_1 \arrow[rr] \arrow[dl] & & X_2 \arrow[dr] & \\
A \arrow[rr] & & X_3 \arrow[rr] & & Y
\end{tikzcd}
$$

Causal DAG Example (ctd)

More formally:

- observational distribution factorises as:

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \mathrm{pa}(V))$$

- interventional distributions factorise as:

for $V_i \in \mathbf{V}$:

$$p(\mathbf{V} \mid \mathrm{do}(V_i = v_i)) = \prod_{V \in \mathbf{V} \setminus V_i} p(V \mid \mathrm{pa}(V)) \mathbb{1}(V_i = v_i)$$

$$\begin{tikzcd}
{} & X_1 \arrow[rr] \arrow[dl] & & X_2 \arrow[dr] & {} \\
A \arrow[rr] & & X_3 \arrow[rr] & & Y
\end{tikzcd}$$

Example:

$p(X_1, X_2, X_3, A, Y) = p(X_1)p(X_2 \mid X_1)p(A \mid X_1)p(X_3 \mid A)p(Y \mid X_2, X_3)$

Use of Causal DAGs

Systematic & transparent way to **represent assumed causal structure**

- Illustrate or examine possible sources of bias
  – e.g., due to bad design or analysis choices
  – Typically: expert-driven construction of (partial) DAG

- **Identification of causal parameters** via graphical characterization
  – e.g. explicit justification for choice of adjustment sets
  – Popular: backdoor criterion
  – Also: e.g. frontdoor criterion *(Piccininni et al. 2023 Epidemiology)*

Or: **DAG itself** is object of interest: "causal discovery"

$\Rightarrow$ data-driven construction of DAG(s)

Part 2

Optimal Adjustment Sets

Journal of Machine Learning Research 21 (2020) 1-45 Submitted 2/20; Revised 11/20; Published 12/20

On Efficient Adjustment in Causal Graphs

Janine Witte WITTE@LEIBNIZ-BIPS.DE
*Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany
and Faculty of Mathematics and Computer Science, University of Bremen, Germany*

Leonard Henckel HENCKEL@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zurich, Switzerland

Marloes H. Maathuis MAATHUIS@STAT.MATH.ETHZ.CH
Seminar for Statistics, ETH Zurich, Switzerland

Vanessa Didelez DIDELEZ@LEIBNIZ-BIPS.DE
*Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany
and Faculty of Mathematics and Computer Science, University of Bremen, Germany*

Adjustment Sets

Wanted: causal effect of A on Y
Using identifying functional

$$\mathrm{E}(Y \mid \text{do}(A = \tilde{a})) = \int \mathrm{E}(Y \mid A = \tilde{a}, \mathbf{C} = \mathbf{c}) p(\mathbf{c}) \, \mathrm{d}\mathbf{c}$$

- Backdoor criterion: $\mathbf{C}$ must be set of (measured) covariates s.t.
  – not descendants of A and
  – block all backdoor paths
- Note: $\mathbf{C}$ is not unique
  – often: focus on 'minimal' $\mathbf{C}$ (`Dagitty`)
  – but: more (and less) efficient choices possible

Adjustment Formula

Average causal effect: $\mathrm{E}(Y \mid \mathrm{do}(A=1)) - \mathrm{E}(Y \mid \mathrm{do}(A=0))$

in the example: $$p(Y \mid \mathrm{do}(A=a)) = \sum_{X_1} p(Y \mid X_1, A=a)p(X_1)$$
$$= \sum_{X_2} p(Y \mid X_2, A=a)p(X_2)$$
$$= \sum_{X_1}\sum_{X_2} p(Y \mid X_1, X_2, A=a)p(X_1, X_2)$$

A valid adjustment set is any set $\mathbf{C}$ that satisfies the adjustment formula:

$$p(Y \mid \mathrm{do}(A=a)) = \sum_{\mathbf{C}} p(Y \mid \mathbf{C}, A=a)p(\mathbf{C})$$

All valid adjustment sets can be read off from the causal DAG

Which Adjustment Set is Best?

If there is more than one valid adjustment set, which one should we choose?

Assumption: The causal DAG represents a linear system with normal errors

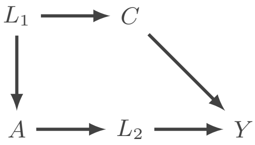
Criterion: Smallest asymptotic variance of the OLS estimator

can be considerably relaxed to cover large class
of np-regular estimators: Rotnitzky & Smucler (2020)

The Forbidden Projection: Motivation

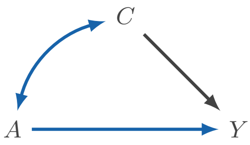
Latent projection *Verma & Pearl (1991); Shpitser et al. (2014)*

motivation: 'hide' latent nodes



$V_1 \leftarrow \mathbf{L} \rightarrow V_2$ becomes $V_1 \leftrightarrow V_2$

$V_1 \rightarrow \mathbf{L} \rightarrow V_2$ becomes $V_1 \rightarrow V_2$

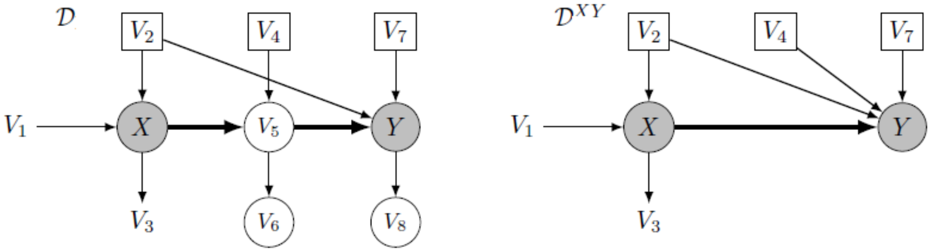


Forbidden projection

motivation: 'hide' forbidden nodes (mediators and descendants of mediators)

The Forbidden Projection

Forbidden projection
latent projection over mediators between X and Y and over descendants of such mediators



The Forbidden Projection

Properties of the forbidden projection: *(Witte, Henckel, Maathuis, Didelez, 2020 JMLR)*

- forbidden projection is a causal DAG
- forbidden projection represents a linear system with normal errors
- a set $\mathbf{C}$ is a valid adjustment set in the forbidden projection if and only if it is a valid adjustment set in the original graph

The Optimal Adjustment Set

With $\mathcal{G}$ original causal DAG, and $\mathcal{G}^{AY}$ forbidden projection wrt. (A, Y).

Let $\hat{\beta}_{ya.\mathbf{x}}$ be estimated coefficient of A in linear regression of Y on A and $\mathbf{X}$
Denote its asymptotic variance as $avar(\hat{\beta}_{ya.\mathbf{x}})$.

Define $\mathbf{O}(A, Y, \mathcal{G}) = \mathrm{pa}(Y, \mathcal{G}^{AY})$.

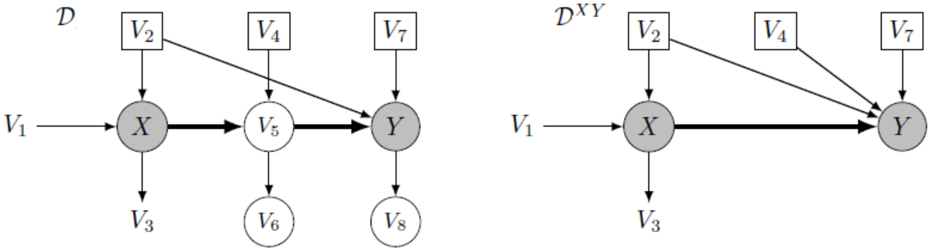
Key results:
1) $\mathbf{O}(A, Y, \mathcal{G})$ is a valid adjustment set.
2) $\mathbf{O}(A, Y, \mathcal{G})$ is optimal in the following sense: For any valid adjustment set $\mathbf{Z}$,

$$avar(\hat{\beta}_{ya.\mathbf{o}}) \leq avar(\hat{\beta}_{ya.\mathbf{z}})$$

(Henckel, Perković, Maathuis, 2019; Witte, Henckel, Maathuis, Didelez, 2020)

Example: Optimal Adjustment

Sufficient for adjustment: $\{V_2\}$; but optimal adjustment set is $\{V_2, V_4, V_7\}$



Idea: reduce residual variance, increase unexplained exposure variance

Remarks on Optimal Adjustment

- **Duality:**
  Set $\mathrm{pa}(A)$ is least efficient ('local' adjustment)
  while $\mathrm{pa}(Y)$ (in forbidden projection) is most efficient adjustment set

- In practice:
  efficiency usually not the first concern, avoiding bias more important
- But: including strong (pre-exposure) predictors of Y should be considered

- Criterion can be used with expert-constructed DAG or after data-driven DAG(s) selection — more: later!

Read More on Selecting Adjustment Sets:

Received: 1 December 2017 | Revised: 6 July 2018 | Accepted: 25 July 2018

DOI: 10.1002/bimj.201700294

RESEARCH PAPER

Biometrical Journal

Covariate selection strategies for causal inference: Classification and comparison

Janine Witte[1,2] | Vanessa Didelez[1,2]

Expert-Driven DAG Construction

American Journal of
EPIDEMIOLOGY

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Invited Commentary: Where Do the Causal DAGs Come From?

Vanessa Didelez

American Journal of Epidemiology, kwae028, https://doi.org/10.1093/aje/kwae028

Published: 03 April 2024 **Article history** ▾

Eliciting a Causal DAG using Expert Knowledge?

- No generally agreed procedure for eliciting / constructing / justifying causal DAG based on expert knowledge
- Typically: small number of experts *(Petersen et al, 2023 AJE)*
  - who screen (more or less systematically) the literature
  - somehow agree (or not) on one DAG
  - Danger: **confirmation bias**

Issues with Expert-Driven Construction

- Experts uncomfortable with specifying the *full* DAG

  ... full DAG not actually required

- Tendency: every ‘known’ association is represented by a directed edge
  i.e. direct/indirect & marginal/conditional dependencies not well-represented

- Focus on *measured* nodes and edges instead of *latent* nodes and *non-edges*

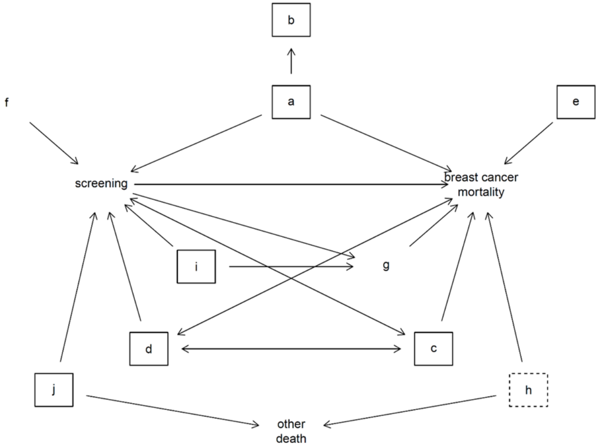
- Time-dependent variables are ‘lumped’ together,
  e.g. binary indicator for “smoking”

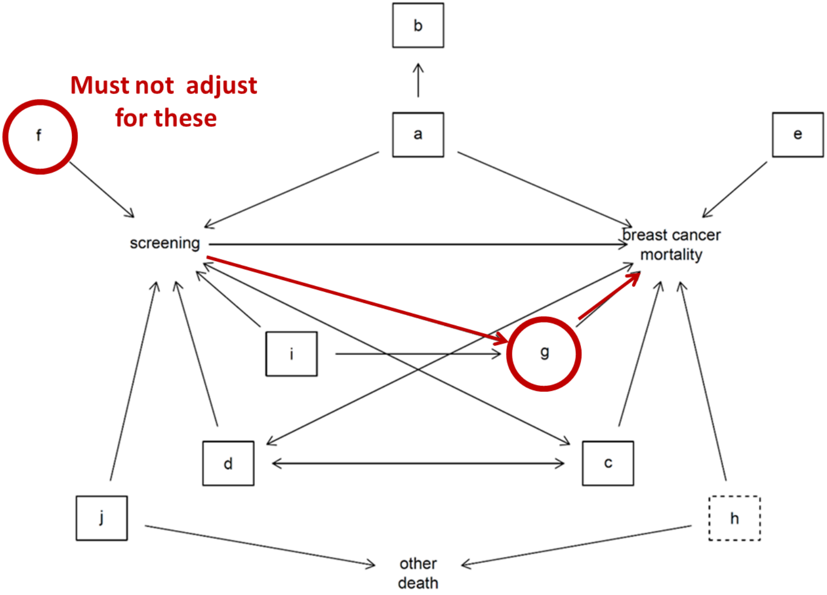
  $\Rightarrow$ should at leat represent ‘up-take’ and ‘continuation’ as separate nodes

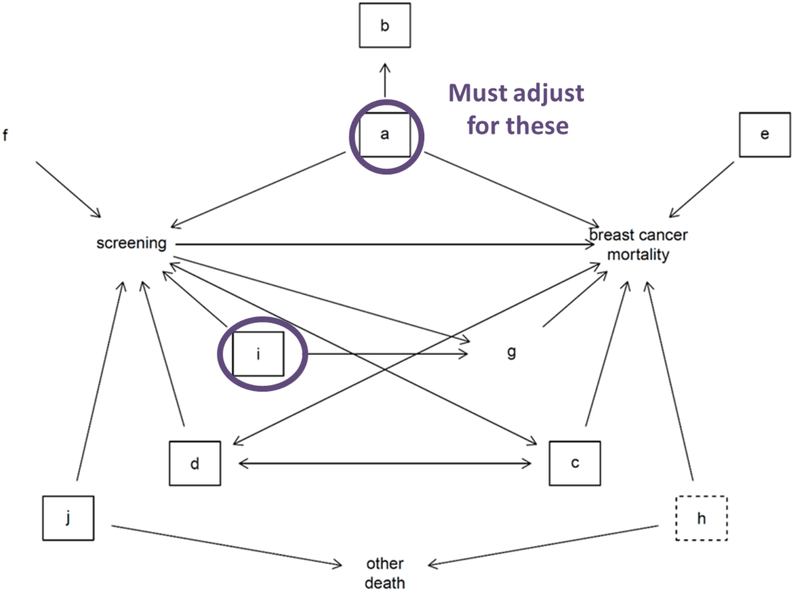
Tool to Prompt Relevant Variables

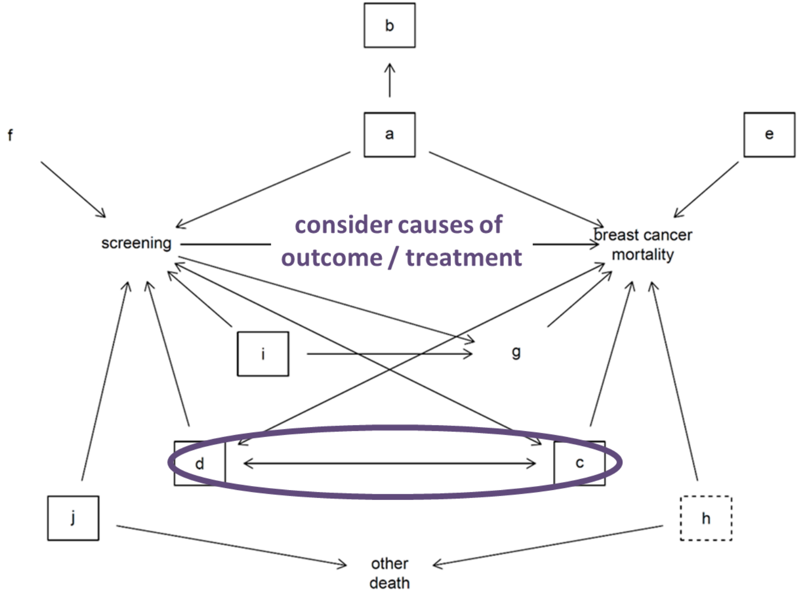
For study protocol to evaluate effectiveness of screening mammography wrt breast cancer mortality in German screening programme

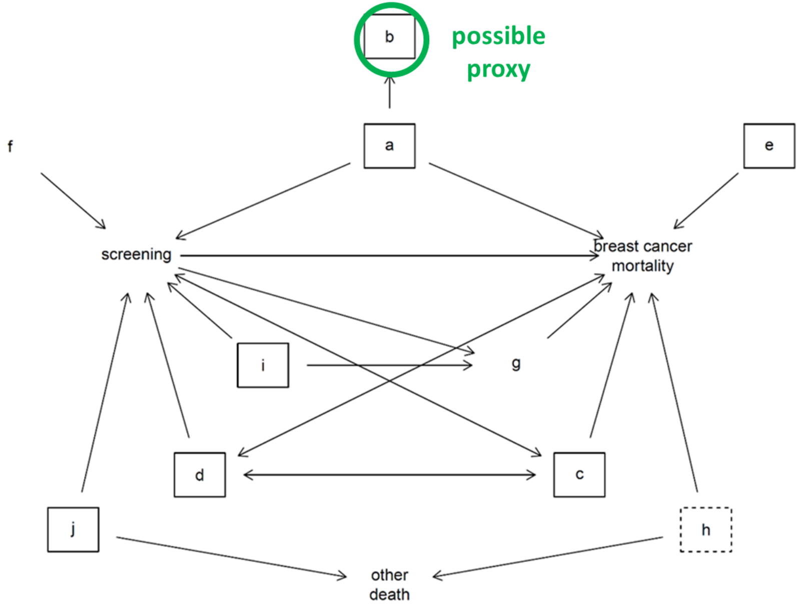
(Braitmaier& Didelez et al 2022:ClinEpi)

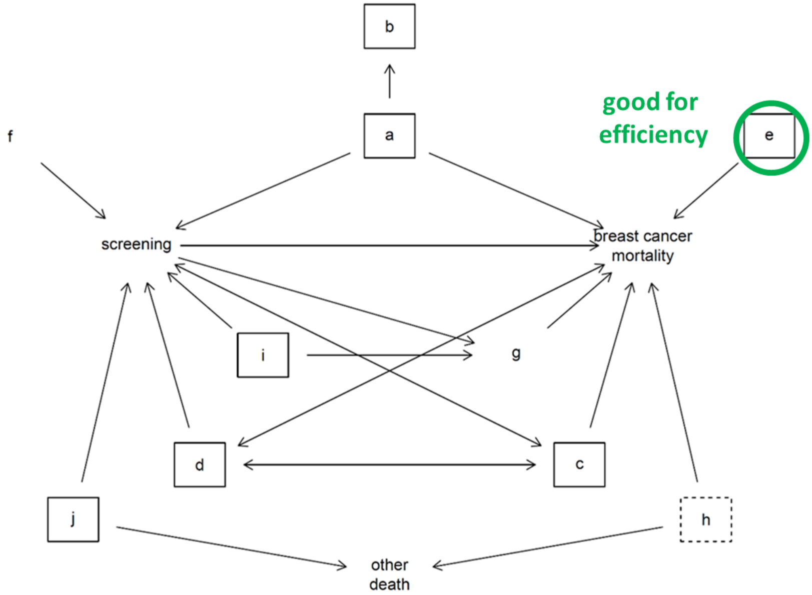


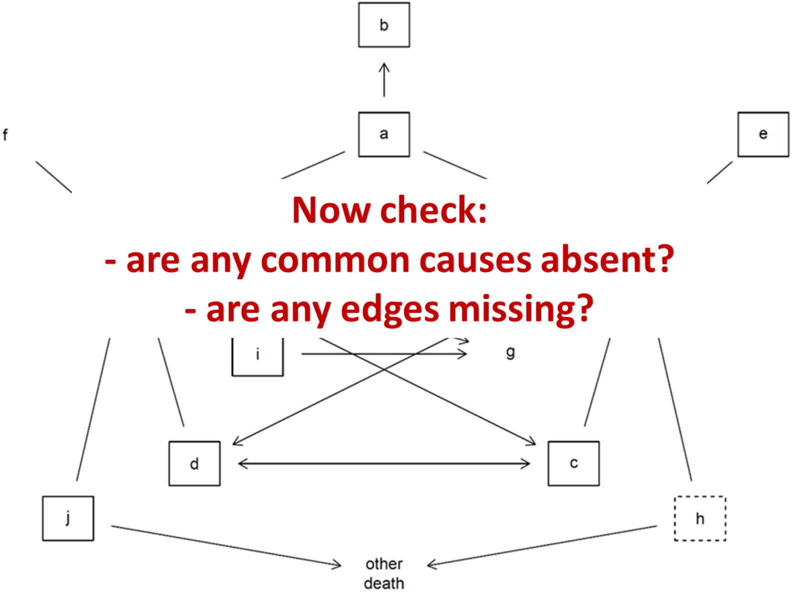

Must not adjust for these
b
a
e
f
screening
breast cancer mortality
i
g
d
c
j
h
other death


b
Must adjust
for these
a
f
e
screening
breast cancer
mortality
i
g
d
c
j
h
other
death


b
a
f
e
screening
consider causes of
outcome / treatment
breast cancer
mortality
i
g
d
c
j
h
other
death


b
possible
proxy
a
f
e
screening
breast cancer
mortality
i
g
d
c
j
h
other
death

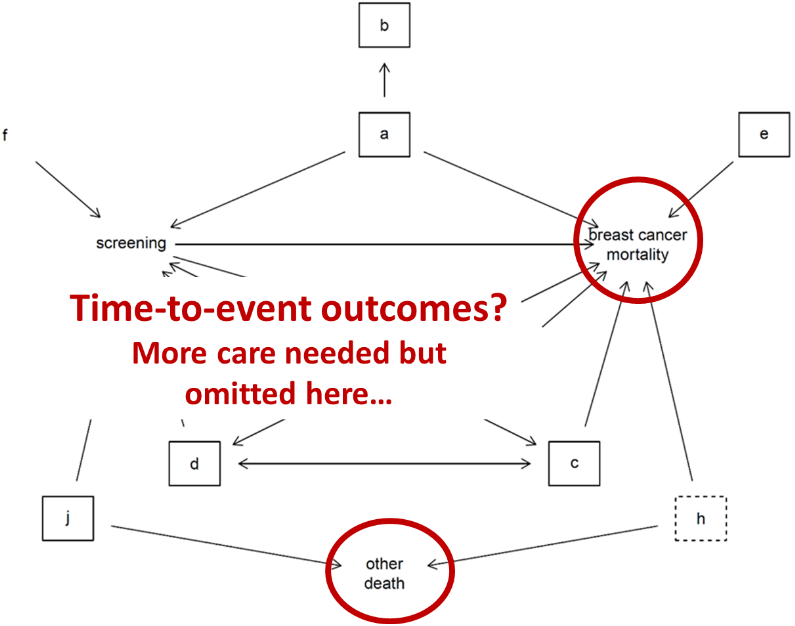

b
a
f
good for
efficiency
e
screening
breast cancer
mortality
i
g
d
c
j
h
other
death



Now check:

- are any common causes absent?

- are any edges missing?



Time-to-event outcomes?

More care needed but omitted here…

Sensitivity Analyses

Little known / used:

- When ambiguous / uncertain about individual edges or directions:
  $\Rightarrow$ carry out sensitivity analyses
- Compare results for each choice
  – either not much difference
  – or raises awareness of sensitivity of results to such assumptions

Example: Cohort Study with Sensitivity Analyses

Börnhorst *et al. Int J Behav Nutr Phys Act* (2023) 20:100
https://doi.org/10.1186/s12966-023-01501-6

International Journal of Behavioral Nutrition and Physical Activity

RESEARCH **Open Access**

The effects of hypothetical behavioral interventions on the 13-year incidence of overweight/obesity in children and adolescents

C. Börnhorst[1*], I. Pigeot[1,2], S. De Henauw[3], A. Formisano[4], L. Lissner[5], D. Molnár[6], L. A. Moreno[7,8], M. Tornaritis[9], T. Veidebaum[10], T. Vrijkotte[11], V. Didelez[1,2†], M. Wolters[1†] and on behalf of the GrowH! consortium

IDEFICS/I.Family Cohort Study

- eight European countries, $\approx$ 16000 children aged 2-9 at start;
- with three waves, 2007 – 2017; $n = 5112$ in all waves
- information collected on: health behaviours (diet and physical activity), socioeconomic factors, genetics, medication, peer networks, media consumption, cardiovascular / metabolic health, subjective well-being
  - repeated measures e.g. of BMI, PA etc.
  - at single times e.g. taste preferences, puberty stage, smoking etc.

(Ahrens et al., on behalf of the I.Family Consortium, 2017)

Obesity in Children: Causal Question?

Context: health behaviour and obesity in children *(Börnhorst et al, 2023)*

Eligible: children (ca. 5 years old), non-obese; using cohort-data (volunteers)

Health behaviours: sleep duration, screen time, sugar drinks, sports club / physical activity, active transport $\longrightarrow$ guidelines (GL) identified from literature

Outcome: 13-year risk of obesity

Estimand (a): population interventional effect

'Treatment' arm: ensure behaviours always meet GLs (sustained)

'Control' arm: natural behaviour (i.e. no intervention)

Shift Interventions

Context: health behaviour and obesity in children *(Börnhorst et al, 2023)*

Eligible: children (ca. 5 years old), non-obese; using cohort-data (volunteers)

Health behaviours: sleep duration, screen time, sugar drinks, sports club / physical activity, active transport $\longrightarrow$ guidelines (GL) identified from literature

Outcome: 13-year risk of obesity

Estimand (b): shift intervention

'Treatment' arm: shift behaviours by specific amount towards guidelines
whenever they don't already meet the GL

'Control' arm: natural behaviour (i.e. no intervention)

Interpretation

Example: health behaviour and obesity in children *(Börnhorst et al, 2023)*

Estimand (a): population interventional effect

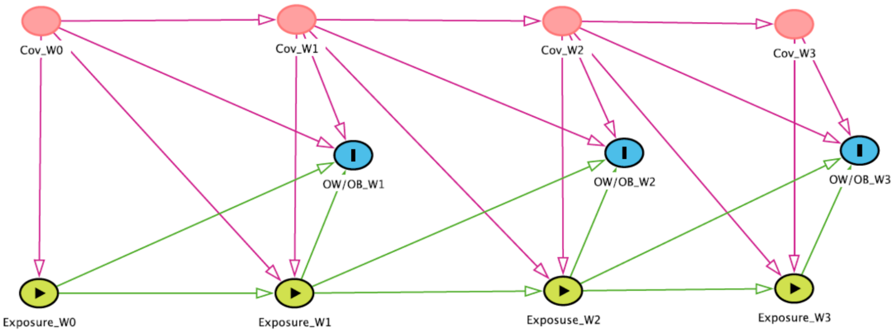
Estimand (b): shift intervention

Note: control arm reflects current behaviour of population;
shift-interventions considered less 'invasive' and thus more realistic.

Note also: analysis with parametric **g-formula** & numerous sensitivity checks
With only three waves, some 'heroic' assumptions involved!

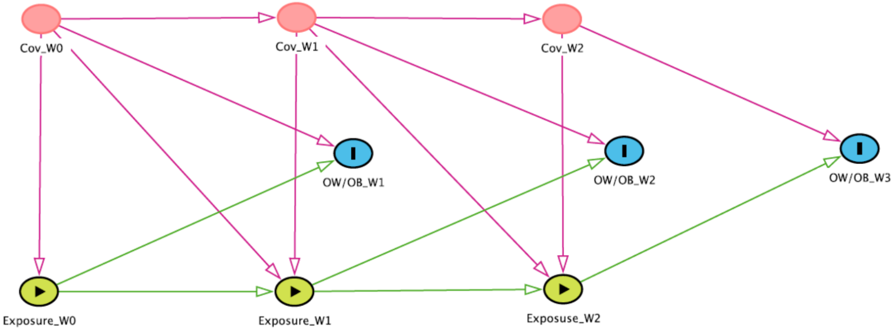
DAG with Cohort Structure

- Main model allowing contemporaneous effects



Alternative DAG with Cohort Structure

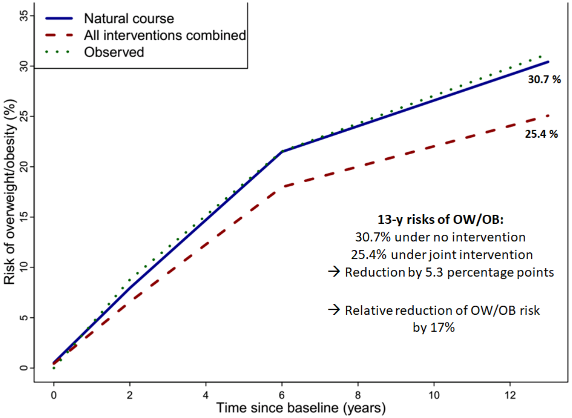
Sensitivity analysis: only time-delayed effects



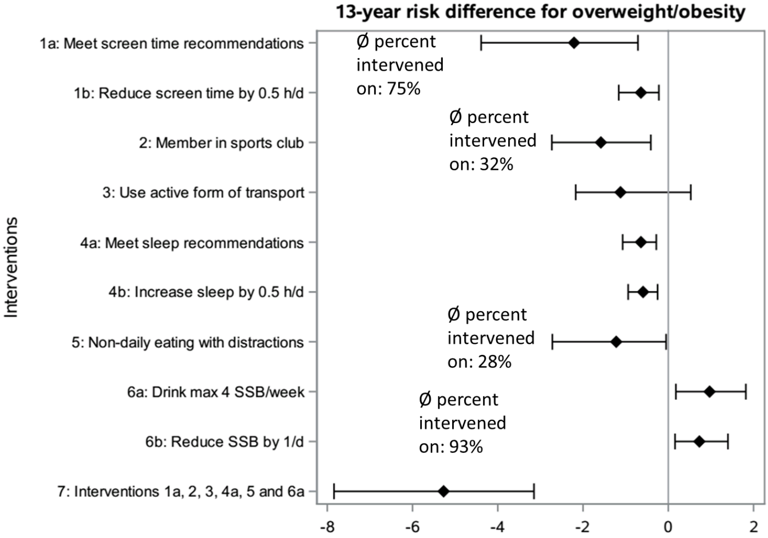
Results (Main Analysis)

Results

– jointly *all* hyp. intervention
– in line with (scarce) existing evidence
– reduction by 5.3 %-points or relative: 17%

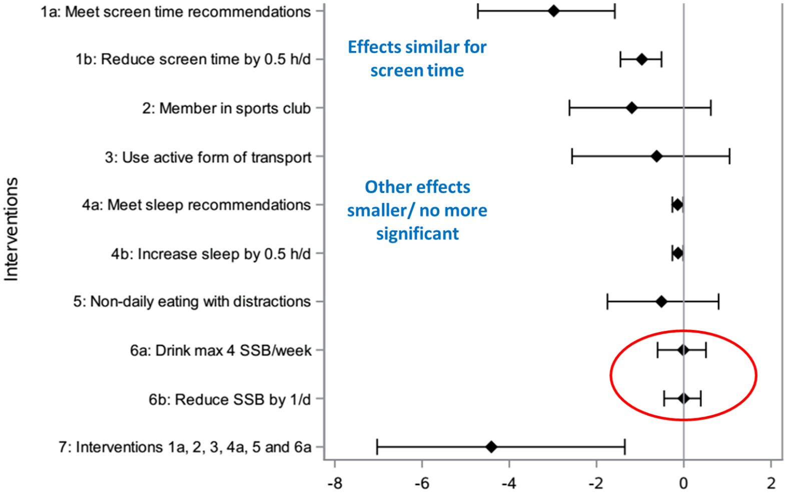


Detailed Results (Main)



Detailed Results (Sensitivity)

13-year risk difference for overweight/obesity - allowing only time-delayed effects



Remarks on Data Example

- Single intervention effects only modest, joint intervention slightly more effective
- Agrees with previous *actual* intervention studies, e.g. for diet and/or PA
- Implausible effect of SSB may, e.g., be due to reverse causation
- Potential sources of bias:
  – waves are few and far apart in time
  – measurement error (self-reporting)
  – large drop-out (g-formula: forces 'no drop-out' analytically)
  – social desirability
  – heroic assumption of 'no unmeasured confounding'
- But with g-formula: adequately accounting for time-dependent confounding, clear interpretation with immediate public-health interpretation

Part 4

Data-Driven Construction of (Causal) DAG(s)

J. R. Statist. Soc. A (2020)
183, *Part* 4, *pp.* 1747–1775

Causal discovery of gene regulation with incomplete data

Ronja Foraita,
Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany

Received: 9 September 2021 | Revised: 12 June 2022 | Accepted: 11 July 2022
DOI: 10.1002/sim.9535

RESEARCH ARTICLE

Statistics in Medicine

Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data

Janine Witte[1,2] | Ronja Foraita[1] | Vanessa Didelez[1,2]

A practical guide to causal discovery with cohort data

Ryan M. Andrews[1] Ronja Foraita[2] Vanessa Didelez[2,3] Janine Witte[2,

[1] Boston University School of Public Health
[2] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen
[3] University of Bremen

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

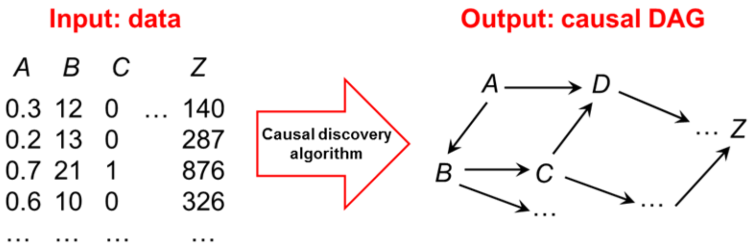
Christine W. Bang[1,2] Vanessa Didelez[1,2]

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

Causal Discovery

aka: causal search, (causal) structure learning, (causal) graph estimation, network inference ...

Causal Discovery



Actually:

→ need special assumptions (faithfulness, causal sufficiency, likelihood, additivity, ...)

→ output not a unique DAG, instead: equivalence class and: sampling uncertainty??

Causal Discovery
Caveats

DAGs for 10 variables $> 4 \times 10^{18}$

Number of DAGs superexponential in number of nodes

$\Rightarrow$ cannot evaluate all possible DAGs

There is no free lunch — all methods rely on *strong* assumptions

More modest:
interpret output as probabilistic DAG; generate some causal hypotheses;
absence of edge still absence of (direct) causation *(but for power)*

$\Rightarrow$ Consider causal discovery as **exploratory** data analysis

Types of Algorithms
there are very many

(1) Constraint-based

- principle: (conditional) **in**dependence $\Rightarrow$ no (direct) causation
- find (conditional) independencies (= constraints) in data
- construct graph to satisfy these constraints, e.g. PC algorithm

(2) Score-based

- define a score for fit between data and causal graph
  (often: penalised likelihood-based)
- optimise the score over space of graphs
- includes Bayesian approaches

Types of Algorithms
there are very many

(3) Exploiting structural asymmetries

- various 'modelling' assumption render $X \longrightarrow Y$ observationally different from $X \longleftarrow Y$, e.g. additive errors
- relies on some information-theoretic justification

(4) Reformulation as continuous optimisation problems ('NOTEARS', 'SAM')

- with smooth acyclicity constraints
- combine with black-box machine learning approaches
- *I would say: still work in progress...*

R Packages `micd` & `tpc`

Our work

- Combine constraint-based algorithms with multiple imputation or test-wise deletion for data with missing values
- Modified PC-algorithm for robust and efficient use of temporal (tiered) back-ground knowledge
- Additional tools for bootstrapping, mixed data types etc.

Remarks on Data-Driven Approaches

- Can (in principle) be combined with expert knowledge
  - ‘Temporal’ PC-algorithm (tPC): robust and efficient use of temporal background knowledge e.g. cohort studies *(Bang & Didelez, 2024 forthcoming)*
  - black/white-listing of edges (often ad-hoc)
  - weights in SAT-solver approaches
- Good solutions to represent sampling uncertainty of DAG(s) still needed
  - Can use resampling (bootstrap)
  - ... but often only edge-wise uncertainty reported
- Appropriate algorithm? More algorithms than real-data applications...
  - Validation on real data requires experimental data – rarely available
  - Validation on synthetic data: need *neutral* comparisons

Data-Driven Selection, then Estimation?

First find DAG(s), then estimate causal effect(s)?

- IDA (Intervention when the DAG is Absent) algorithm *(Maathuis et al., 2010)*
- Note: non-uniqueness of DAGs
  $\Rightarrow$ non-uniqueness of adjustment sets
  $\Rightarrow$ non-uniqueness of estimates

Caveat: same data for both steps $\Rightarrow$ post-selection inference problem

Application with IDEFICS/I.Family Data

scientific reports

www.nature.com/scientificreports

OPEN

A longitudinal causal graph analysis investigating modifiable risk factors and obesity in a European cohort of children and adolescents

Ronja Foraita[1], Janine Witte[1,2], Claudia Börnhorst[1], Wencke Gwozdz[3,4], Valeria Pala[5], Lauren Lissner[6], Fabio Lauria[7], Lucia A. Reisch[1,8], Dénes Molnár[9], Stefaan De Henauw[10], Luis Moreno[11], Toomas Veidebaum[12], Michael Tornaritis[13], Iris Pigeot[1,2] & Vanessa Didelez[1,2]

IDEFICS/I.Family Cohort Study

- eight European countries, $\approx$ 16000 children aged 2-9 at start;
- with three waves, 2007 – 2017; $n = 5112$ in all waves
- information collected on: health behaviours (diet and physical activity), socioeconomic factors, genetics, medication, peer networks, media consumption, cardiovascular / metabolic health, subjective well-being
  - repeated measures e.g. of BMI, PA etc.
  - at single times e.g. taste preferences, puberty stage, smoking etc.

(Ahrens et al., on behalf of the I.Family Consortium, 2017)

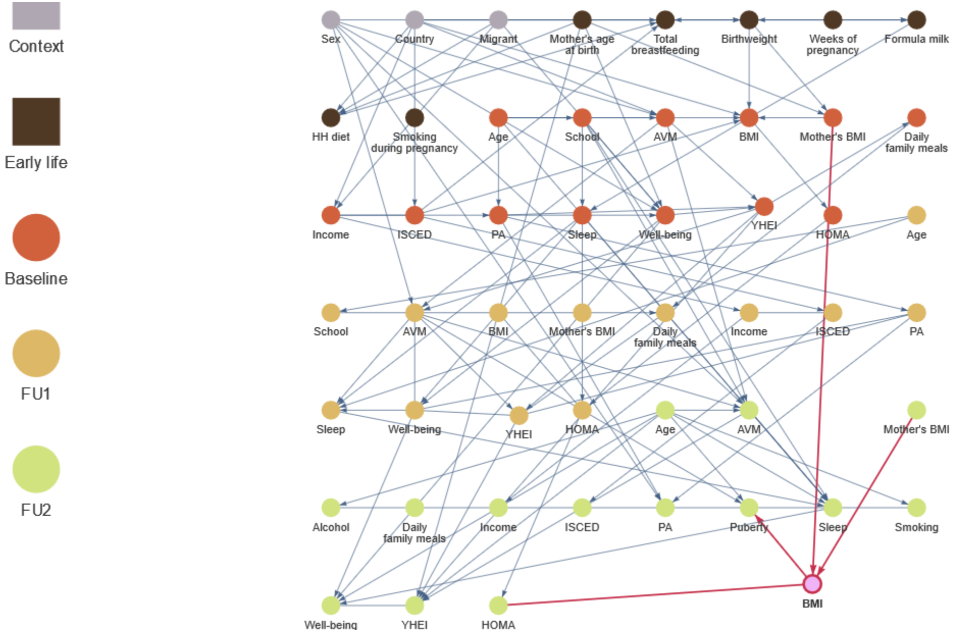
Cohort Causal Graph — Analysis

- Methods: PC-algorithm with MI (random forest imputation models), various sensitivity analyses

  PC-alg assumes causal sufficiency!
- Efficient use of temporal structure with tPC algorithm
- Apply local and optimal generalised IDA to determine adjustment sets for interesting exposure and outcome pairs *(Witte et al 2020 JMLR)*
- Nonparametric estimation ('double machine learning') of effects as rough guide (post-selection-inference issues here) *(Kennedy et al., 2017)*

Cohort Causal Graph — Results

https://bips-hb.github.io/ccg-childhood-obesity/



Context
Early life
Baseline
FU1
FU2
Sex
Country
Migrant
Mother's age at birth
Total breastfeeding
Birthweight
Weeks of pregnancy
Formula milk
HH diet
Smoking during pregnancy
Age
School
AVM
BMI
Mother's BMI
Daily family meals
Income
ISCED
PA
Sleep
Well-being
YHEI
HOMA
Age
School
AVM
BMI
Mother's BMI
Daily family meals
Income
ISCED
PA
Sleep
Well-being
YHEI
HOMA
Age
AVM
Mother's BMI
Alcohol
Daily family meals
Income
ISCED
PA
Puberty
Sleep
Smoking
Well-being
YHEI
HOMA
BMI

Cohort Causal Graph — Stability

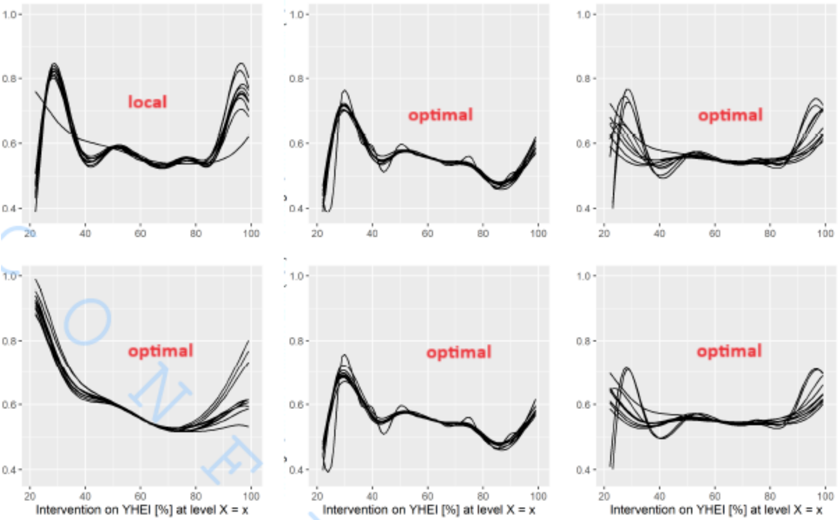
Based on 100 bootstrap graphs: consider stability of individual (non)edges but also of specific interesting graphical structures like causal paths

- Of 104 edges, 36 were stable ($> 80\%$) while 50 were instable ($\leq 50\%$)
- **All** graphs had multiple possibly causal paths from early modifiable behaviours to later BMI
  – youth-healty eating index (YHEI)
  – audio-visual media consumption
  – sleep-duration
  – physical activity
- **No** graph had a direct edge from early modifiable behaviours to later BMI

Cohort Causal Graph — Estimating Effects

- Example here:
  estimate causal effect of early YHEI (point exposure) on later BMI (2nd FU)
- *Non-parametric* causal response curves for continuously measured YHEI
- Local adjustment set (least efficient)
- Optimal adjustment sets – *non-unique* in equivalence class

Exposure: healthy-eating-index (baseline); outcome: BMI at 2nd FU
NP-estimates of average outcome under hypothetical intervention in exposure for different adjustment sets and 10 multiply imputed datasets



Conclusions

- Causal questions at the heart of much research, e.g. in epidemiology
  $\Rightarrow$ should use transparent formalism and suitable methods
- Causal inference & discovery relies on specific (mostly untestable) assumptions
  $\Rightarrow$ should make those explicit
- Expert-driven construction is transparent with DAGs, but can be unreliable
- Data-driven construction in practice rather unstable
  $\Rightarrow$ should incorporate back-ground or (most reliable) expert knowledge
- Validation of expert- / data-driven approaches is usually not possible

Thanks for your attention!

www.leibniz-bips.de/en

Contact
Vanessa Didelez
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
D-28359 Bremen
didelez@leibniz-bips.de

GeTTCausal !

BIPS Initiative to

- ... use the **Ge**PaRD database (German health insurance claims data)
- ... with **T**arget **T**rial emulation
- ... for **Causal** inference
- ... to support & improve health-related decision making

⇒ Joint work with many collaborators