Directed Acyclic Graphs (DAGs) in epidemiology – why and how

Henrik Støvring

Senior researcher, MSc, PhD, DMSc

Steno Diabetes Center Aarhus - Denmark

hersto@rm.dk

October 26, 2023 , SDCA

Overview

- What is a DAG
- Two classic uses:
  - Identifying confounders (to be controlled for)
  - Identifying colliders (NOT to be controlled for)
- Application 1:
  RCTs and the concept of instrumental variables
- Application 2:
  Negative control outcome

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 142 (2022) 264–267

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Tutorial on directed acyclic graphs

Jean C. Digitale, Jeffrey N. Martin, Medellena Maria Glymour*

Department of Epidemiology and Biostatistics, University of California, 550 16th St, 2nd Floor, San Francisco, CA 94158

Received 8 May 2021; Received in revised form 28 July 2021; Accepted 2 August 2021; Available online 8 August 2021

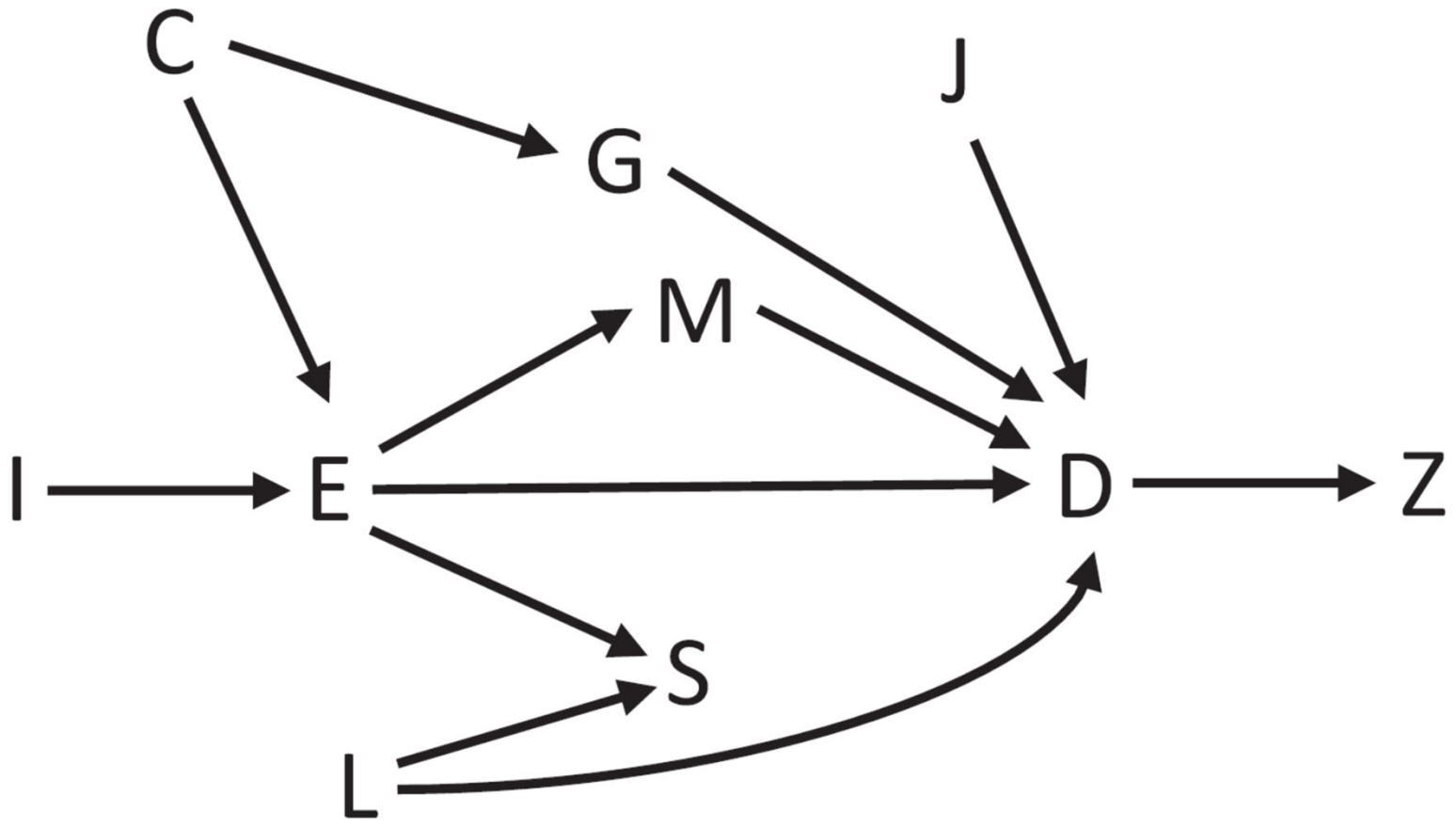
Abstract

Directed acyclic graphs (DAGs) are an intuitive yet rigorous tool to communicate about causal questions in clinical and epidemiologic research and inform study design and statistical analysis. DAGs are constructed to depict prior knowledge about biological and behavioral systems related to specific causal research questions. DAG components portray who receives treatment or experiences exposures; mechanisms by which treatments and exposures operate; and other factors that influence the outcome of interest or which persons are included in an analysis. Once assembled, DAGs — via a few simple rules — guide the researcher in identifying whether the causal effect of interest can be identified without bias and, if so, what must be done either in study design or data analysis to achieve this. Specifically, DAGs can identify variables that, if controlled for in the design or analysis phase, are sufficient to eliminate confounding and some forms of selection bias. DAGs also help recognize variables that, if controlled for, bias the analysis (e.g., mediators or factors influenced by both exposure and outcome). Finally, DAGs help researchers recognize insidious sources of bias introduced by selection of individuals into studies or failure to completely observe all individuals until study outcomes are reached. DAGs, however, are not infallible, largely owing to limitations in prior knowledge about the system in question. In such instances, several alternative DAGs are plausible, and researchers should assess whether results differ meaningfully across analyses guided by different DAGs and be forthright about uncertainty. DAGs are powerful tools to guide the conduct of clinical research.

What is a DAG

- Directed -> one variable is a cause of other variables
  - means that there is a causality and hence a time aspect
- Acyclic -> no closed loops
  - a cause can never be an effect of itself
- Graph -> in the mathematical sense of points and vertices
  - we can talk about groups, about ancestors and descendants, separation sets, moralizing a graph, etc.
  See https://en.wikipedia.org/wiki/Graph_theory

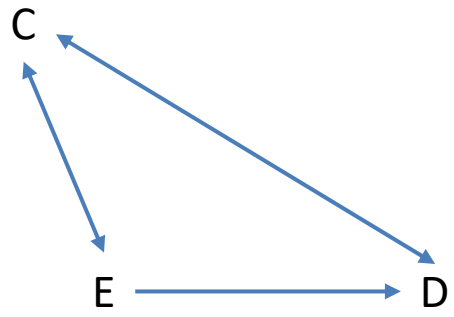
Example of DAG



To identify the causal effect of E on D, we must block all non-causal paths and none of the causal paths between the two variables.

Definition of confounding?

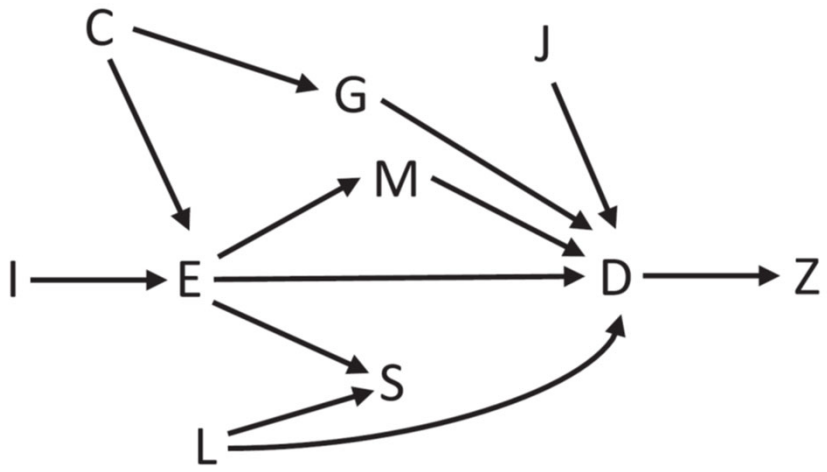
- Classic definitions



- Modern definition:
  Confounding is absence of exchangeability
- Exchangeability is defined via counter-factuals:
  Exposed group would have had same outcome as control group *had they not been exposed*

Example of DAG – and exercises

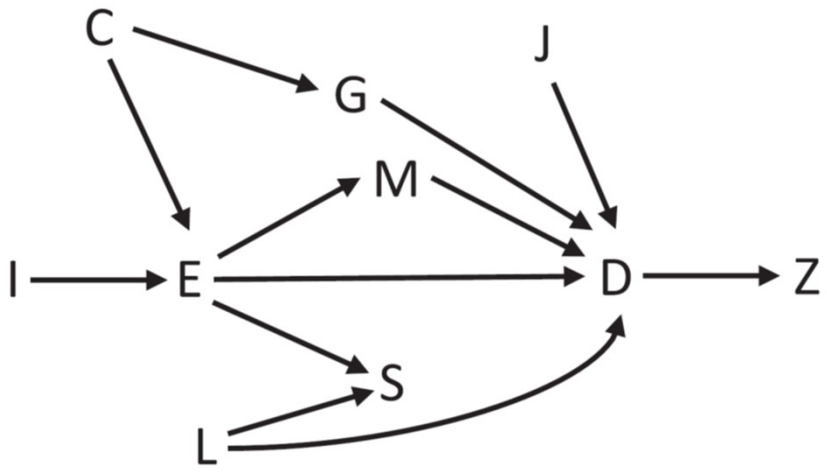
- Identify all variables which are confounders for the E→D effect
- Identify all variables which are *not* confounders for the E→D effect
- Identify minimum sets to control for confounding



To identify the causal effect of E on D, we must block all non-causal paths and none of the causal paths between the two variables.

Example of DAG

- Identify all variables which are confounders for the E→D effect
- Identify all variables which are *not* confounders for the E→D effect
- Identify all minimum sets to control for confounding
- Why is L not a confounder?

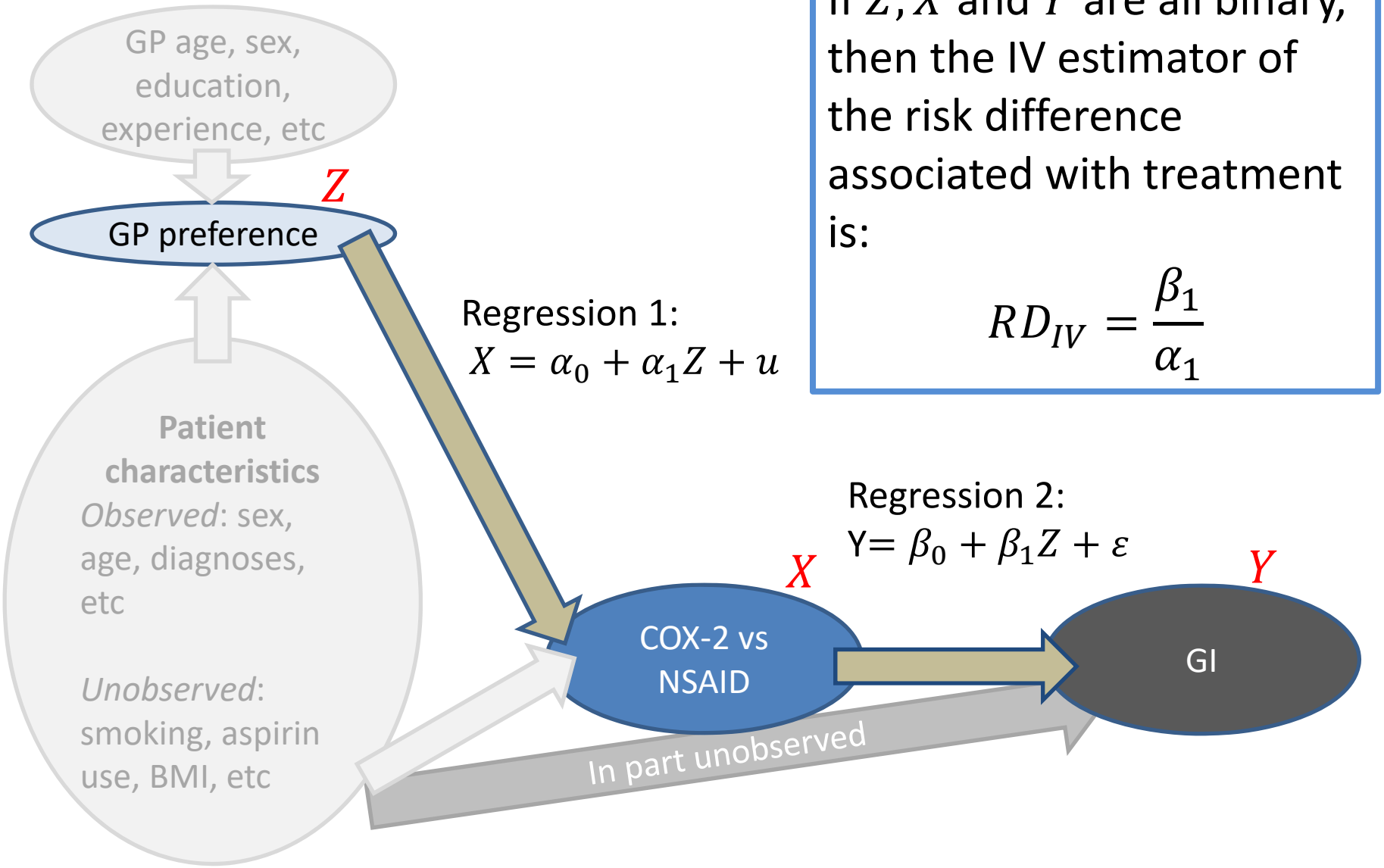


To identify the causal effect of E on D, we must block all non-causal paths and none of the causal paths between the two variables.

Two useful criteria

- The backdoor criteria:
  close all open backdoor paths to control for confounding
- Avoid collider bias:
  do *not* control for a collider or variables in their path (from exposure to outcome)

IV analysis – an example



If Z, X and Y are all binary, then the IV estimator of the risk difference associated with treatment is:

$$RD_{IV} = \frac{\beta_1}{\alpha_1}$$

Small example of IV estimation

- Table 3, p272 – Brookhart (2006)

	Treatment Received					Instrumental Variable—Type of Most Recent New NSAID Prescription Started by Physician				
	COX-2		Nonslective NSAID			COX-2		Nonselective NSAID		
	No. of Patients	No. of Events	No. of Patients	No. of Events	Risk Difference*	No. of Patients	No. of Events	No. of Patients	No. of Events	Risk Difference†
Event within 60 d										
All patients	32,273	211	17,646	110	0.03	25,363	148	12,479	99	−0.21
All patients of PCPs	24,336	154	11,748	61	0.11	20,416	112	9396	66	−0.15
Patients with OA or RA	16,298	112	6125	36	0.10	11,948	68	5349	49	−0.35

As required, the instrument was also related to treatment. Across the entire population, if the last prescription written by a physician was for a COX-2 inhibitor, then the probability that the next prescription would be for a COX-2 inhibitor was 77%. On the other hand, if the last prescription written by a physician was for a nonselective NSAID, then the probability that the next prescription would be for a COX-2 was only 55%. Among patients of primary care physicians, these probabilities were nearly identical (77% and 57%).

- $RD_{IV} = \frac{\beta_1}{\alpha_1}$ where
  - $\beta_1 = -0.21$
  - $\alpha_1 = 0.77 - 0.55 = 0.22$
- Thus: $RD_{IV} = -\frac{0.21}{0.22} = -0.95$
- Reported estimate in Table 4: $-0.92\ (-1.74;\ -0.10)$
- NB: $RD_{IV} = \frac{\beta_1}{\alpha_1}$ is the same as formula (1) in paper: $\hat{\delta} = \frac{\hat{E}[Y \mid Z=1] - \hat{E}[Y \mid Z=0]}{\hat{E}[X \mid Z=1] - \hat{E}[X \mid Z=0]}$,

Example 2: understanding negative control exposures

International Journal of Epidemiology, 2020, 1032–1042
doi: 10.1093/ije/dyaa029
Advance Access Publication Date: 25 March 2020
Original article

Misc

Ne
cor
ma
par

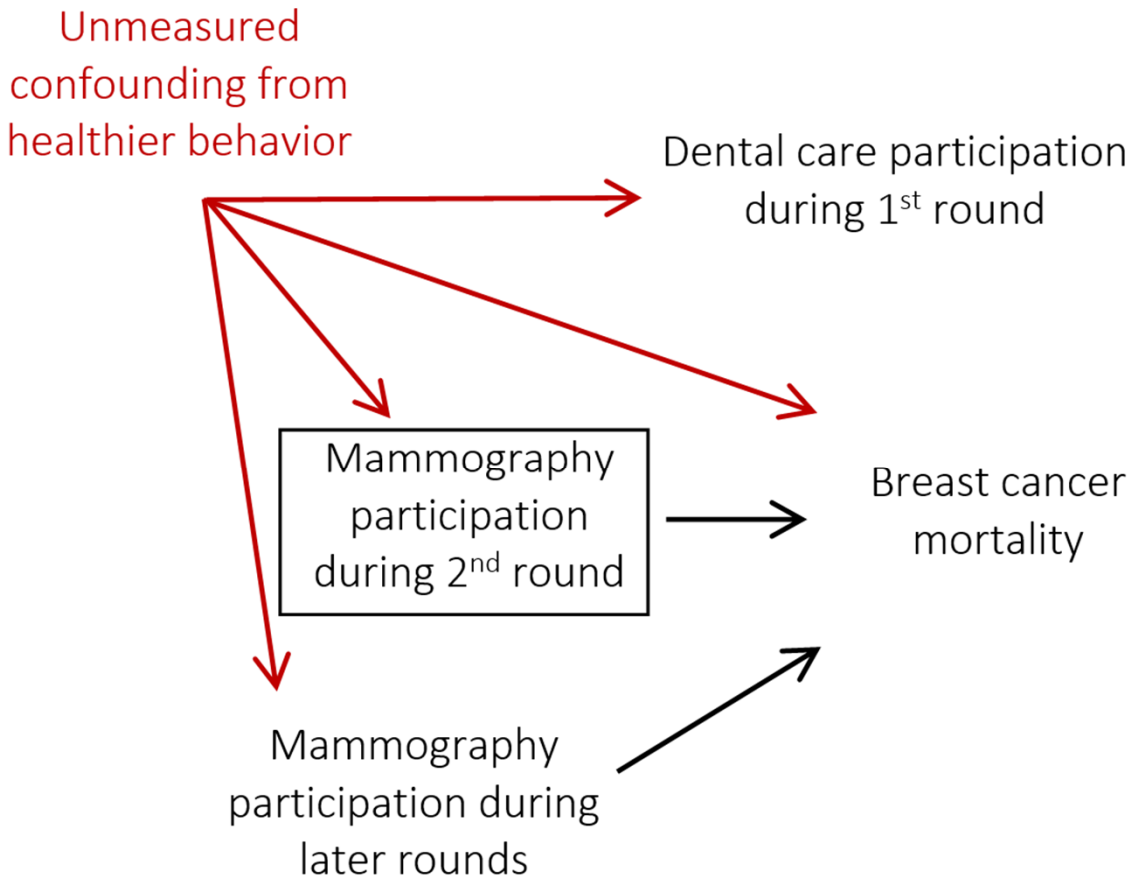
Mett
M Al
Henr

first. We estimated hazard ratios (HRs) of death from breast cancer, causes other than breast cancer and external causes. We added dental-care participation as an exposure to test for an independent association with breast-cancer mortality. We adjusted for civil status, parity, age at first birth, educational attainment, income and hormone use.

Results: Screening participants had a lower hazard of breast-cancer death [HR 0.47, 95% confidence interval (CI) 0.32, 0.69] compared with non-participants. Participants also had a lower hazard of death from other causes (HR 0.43, 95% CI 0.39, 0.46) and external causes (HR 0.35, 95% CI 0.23, 0.54). Reductions persisted after covariate adjustment. Dental-care participants had a lower hazard of breast-cancer death (HR 0.75, 95% CI 0.56, 1.01), irrespective of screening participation.

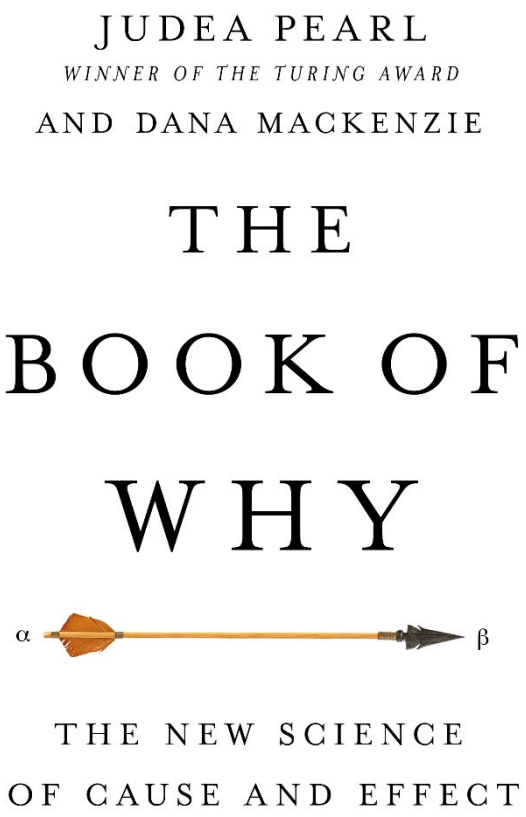
Conclusions: Negative-control associations indicated residual uncontrolled confounding

Supplemental Figure S1. Directed acyclic graph outlining the causal structures of the negative control exposure analysis. We only adjust for mammography participation during the 2nd round. Therefore, the negative control exposure of dental care participation may show an association with breast cancer mortality partly due to the unobserved associations with healthier behavior affecting breast cancer mortality and partly due to a true protective effect mediated via mammography participation during later rounds.



FYI

- http://bayes.cs.ucla.edu/WHY/



Time for discussion!

(Djursland, July 2015 – H Støvring)