Mediation analyses with **imperfectly defined mediators** may be biased if there is significant unexplained heterogeneity, and their results should be interpreted with caution.

Mediation Analysis with Imperfectly Defined Mediators: a Microsimulation Experiment with Breast Cancer Survival, Socio-Economic Status, and Stage at Diagnosis

Alessandro Gasparini[1,2,*], Keith Humphreys[2], Sarah Booth[3], Mark J. Rutherford[3], Elisavet Syriopoulou[2]

Background

Mediation analysis in cancer epidemiology is often used for studying whether survival differences between socio-economic groups can be explained by tumour stage at diagnosis acting as a mediator. Commonly used stages at diagnosis (i.e., categories I-IV) may be too crude to capture the heterogeneity between patients, resulting in significant leftover within-stage variation. In this study, we explore the impact of such ill-defined mediators on the estimation of mediated effects.

Microsimulation Setup

Using microsimulation methods, we generated birth cohorts of women according to a breast cancer natural history model developed in Swedish settings. We independently simulated socio-economic status (SES) and imposed a screening programme mimicking that used in Sweden.

We design the following simulation scenarios:

I. No differences between socio-economic status groups in terms of rate of symptomatic detection and screening attendance;
II. Socio-economic status groups have different rates of symptomatic detection;
III. Socio-economic status groups have different probabilities of attending each screening visit;
IV. Socio-economic status groups have both different rates of symptomatic detection and probabilities of attending each screening visit. Thus, scenario IV is a combination of scenarios II-III.

Across all scenarios, we assume no direct effect of SES on breast cancer survival. A DAG illustrating the data-generating model is included in Figure 1.

Mediation Analyses

We perform five different mediation analyses, with different covariates adjustment in the survival model and definitions for breast cancer stage at diagnosis (Table 1).

The survival model for time to breast cancer is a flexible parametric model on the log-cumulative hazard scale with 5 degrees of freedom, with study time as the time scale (e.g., time since diagnosis). The models for stage (the mediator) are multinomial logistic regression models with non-linear age as a covariate, stratified by socio-economic status.

The estimand of interest is the natural direct effect (NDE) of socio-economic status on breast cancer survival, defined as the difference in marginal survival at time t between exposure levels while fixing the stage distribution. Identification of the NDE is possible under standard mediation analysis assumptions, and we estimate it using regression standardisation.

Figure 1: Illustration of the data-generating model.

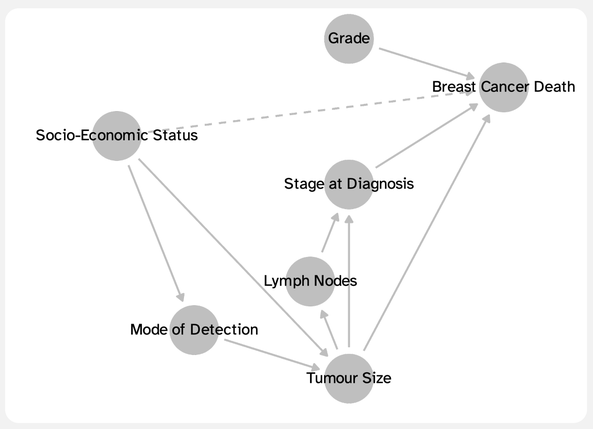
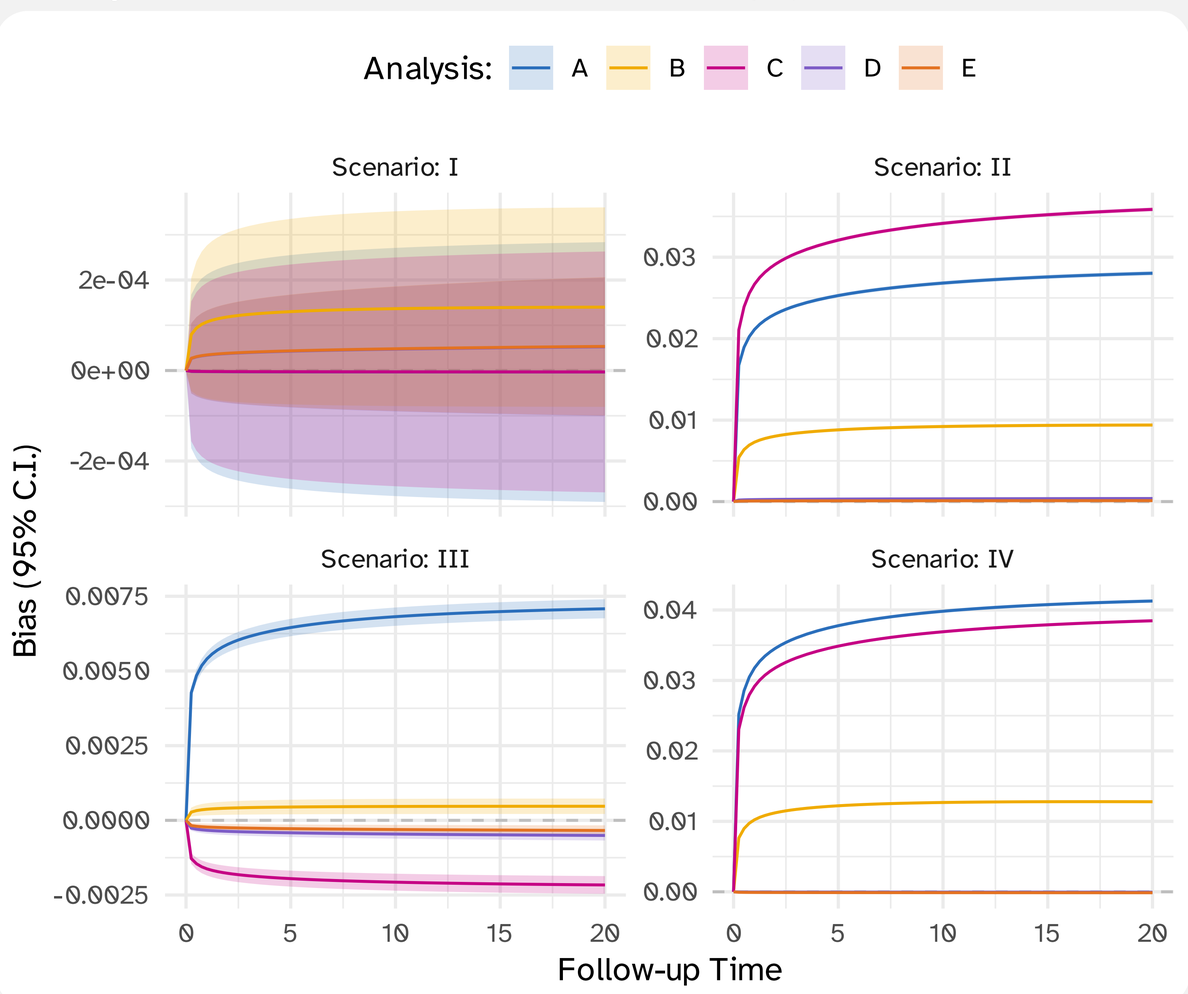


Table 1: Summary of the five mediation analyses.

Analysis	Survival Model Covariates	Stage Definition
A	Age, SES, Grade, Stage	Standard (I-III)
B	Age, SES, Grade, Stage	T-N-M Product
C	Age, SES, Grade, Stage, Detection Mode	Standard (I-III)
D	Age, SES, Grade, Stage, Tumour Diameter	Standard (I-III)
E	Age, SES, Grade, Stage, Detection Mode, Tumour Diameter	Standard (I-III)

Figure 2: Bias of estimated NDE across simulation scenarios and analyses, with 95% confidence intervals.



Results

Bias in NDE estimates, i.e., how much they deviate from their true value of zero, is depicted in Figure 2 by scenario and analysis, with 95% confidence intervals.

- We did not expect any bias in Scenario I and this is indeed what we observed.
- Significant biases were observed for scenarios II, III, and IV.
- Analysis A showed the largest bias across all time points considered in scenarios III and IV, with positive bias up to about 0.04 in Scenario IV.
- Analysis B (more detailed stage) improved on Analysis A across all scenarios.
- Analysis C performed worst in Scenario II, and between A and B otherwise.
- Analysis D (adjusting for tumour size at diagnosis) performed best overall.
- Analysis E (adjusting for tumour size at diagnosis and mode of detection) performed second best, with a performance close to that of Analysis D.

Notably, we observed some bias for all five analyses in scenarios II-IV, even though bias was negligible – in practice – for some settings.

Conclusions

The results suggest that mediation analyses with stage at diagnosis as the mediator may be significantly biased, as tumour stage is often too crude to fully capture the heterogeneity between cancer patients. Considering the definition of categorical mediators (or adjusting the analysis for additional characteristics) can help ensure that there is no significant leftover, unexplained heterogeneity.

Our simulation results rely on a specific microsimulation model. Nonetheless, we showed large estimated NDEs of socio-economic status on breast cancer survival when there was no direct effect to begin with: thus, we recommend caution when interpreting the results of mediation analyses with imperfectly defined mediators.

1. Red Door Analytics AB, Stockholm, Sweden
2. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
3. Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom
* E-mail: <alessandro.gasparini@reddooranalytics.se>

Karolinska Institutet

Red Door Analytics