Casual inference group assignment

Yufang Wang[1,✉], **Vinson Ciawandy**[2,✉], **and Jia Song**[3,✉]

[1]Statistics and Data Science, Leiden University

This study is to investigate the causal relationship between education level and long-term mortality by using the part of Framingham Heart Study's dataset. During this study, we employed two approaches (i.e., G-computation and Propensity score method with inverse probability weighting) to enable us to obtain a causal interpretation between EDUC and DEATH in the Framingham Heart Study's dataset. Both methods show that highly educated people have lower long-term mortality relative to lower educated people, suggesting that a good education can lead to lower long-term mortality. Our study can provide evidence for policymaking in terms of education.

Correspondence: *y.wang100@umail.leidenuniv.nl*, *vinson.ciawandy@umail.leidenuniv.nl*, *j.song.10@umail.leidenuniv.nl*

1. Introduction

Being highly educated is a good words always full of beautiful imagination. People in daily life are very willing to attribute a lot of things, including length of life span to a good education. However, whether a good education indeed can result in a longer life span remains to be addressed. In the current study, we did a causal inference analysis on the data from Framingham Heart Study [add reference]. This is a long-term cohort study on cardiovascular diseases among people in the community of Framingham, Massachusetts in 1948.

1.1. Research Question. Is a high education linked to low long-term mortality?

2. Methods

2.1 Information about the dataset.

This analysis used an enclosed dataset of the Framingham Heart Study, focusing on 4,434 participants who were selected from the 5209 volunteers since 1948. The data were collected through multiple ways such as laboratory, clinical, questionnaire, and event data on 4,434 participants during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. Each participant was followed up by tracking a 24-year death, which can be caused by any reason.

2.2. Variables in the dataset. The dataset includes 16 variables with 9 exposures or confounds and 6 outcomes. The exposures/confounds are participant biological sex (SEX), age measured in years (AGE), systolic blood pressure with mmHg as unit (SYSBP), diastolic blood pressure with mmHG as unit (DIABP), use of anti-hypertensive medication (BPMEDS), current cigarette smoking (CURSMOKE), education (EDUC), body mass index with kilograms/heights in meters squared as unit (BMI) and heart rate in beats/min (HEARTRTE). For the aim of this study, we only focus on the outcome of death(DEATH).

2.3. Protocol of this study.

Eligibility criteria:. Adults in Framingham, Massachusetts consented to this study started in 1948.

Exposure definition:. highly educated (equal or more than college-level, EDUC = 1) vs. low educated (less than college-level, EDUC = 0).

Assignment procedures. No randomization was described in this study.

Follow-up period:. 24 years

Outcome definition. 24-year mortality (death vs. still alive)

Causal contrast of interest. E(Y(X=1)) - E(Y(X=0))

Outcome Measures:. The 24-year mortality measured in the current report is death from any cause (DEATH), which takes values of either 1 (dead) or 0 (alive).

2.4. Data Analysis.

2.4.1. Expert-defined direct acyclic graph (DAG).

Instead of learning from the current data, we defined an expertise-based DAG, which is shown in *Fig. 1* to prevent the so-called double dipping issues, i.e., using the same dataset to define the DAG and conduct further tests based on the DAG. With this expertise-based DAG, we obtained the minimal adjustment set when trying to get the total causal relationship between EDUC and DEATH with backdoor criteria (or backdoor adjustment rules). As a result, we obtained that the minimal adjustment set is SEX, AGE.

2.4.2. missing values issues checking.

Table 1 shows that only the 'EDUC' variable has missing data, considered a predictor in the current study. We assume that some people with low education are less willing to fill in their education information. That is, the missing data mechanism here is missing not at random (MNAR) depending on the predictors. However, due to the research interest in the current study (i.e., to obtain causal interpretation from EDUC to DEATH), this missing data mechanism will not affect our estimation. Thus, we leave it be.

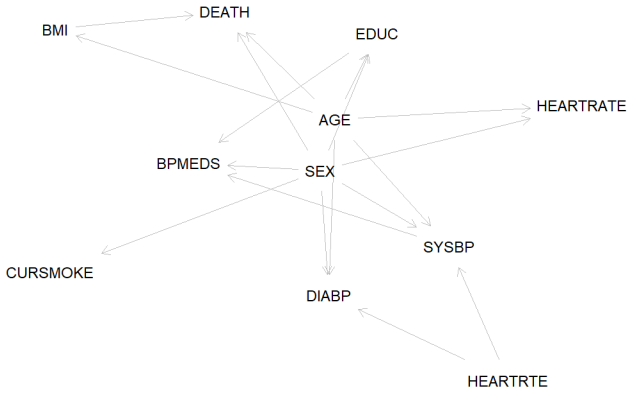

Fig. 1. DAG with all variables

Table 1. Missing data information

EDUC	SEX	AGE	DEATH
113	0	0	0

2.4.3. Assumptions of Identifiability Checking.

Note that the causal effect is based on the assumptions *consistency*, *positivity*, and *exchangablity*. Only if these three assumptions hold, the causal interpretation of the result would be valid and reliable. Therefore, it is crucial to check whether these three assumptions hold before conducting an analysis.

Consistency:. The exposure (EDUC) in this study is very well-defined. Meanwhile, as all participants are adults and thus we assumed that there's no interference in the population, that is, one person's 24-year death does not depend on another person's education level.

Positivity:. As shown in Table 2, each cell of the EDUC and DEATH has a positive value. Thus, the assumption of positivity holds.

Table 2. Posivity checking for EDUC and DEATH

	Alive	Death
HighSchoolorBelow	1943	1160
CollegeorMore	869	349

Exchangeability:. As shown in Table 3, the percentage of observations with Sex = female is much larger in the exposed group (highly educated). That is, there is no exchangeability between highly vs. lower-educated groups, which further necessitates the adjustment of the confounder SEX in the current study and therefore moving on to the conditional exchangeability.

3. Data Analysis.

Table 3. observations of confounder (SEX) between EDUC levels

	Male	Female
HighSchoolorBelow	1352(43.5%)	1751(56.4%)
CollegeorMore	535(43.9%)	683(56.1%)

To obtain the causal interpretation from observational data, we employed two analysis approaches (i.e., the G-computation and the propensity score method with inverse probability weighting)) to enable us to obtain the causal interpretation between EDUC and DEATH.

Approach 1: G-computation.

G-computation was invented from the traditional regression model when obtaining the causal relationship. More concretely, the old-fashioned way of using the regression model is to include all confounders in the regression model and then report the estimate of the exposure. By including all confounders as covariates in the regression model, it was argued that the effects of confounders on the outcomes have been accounted for already. This method has been widely used and is a good approach. However, it typically gives a conditional effect. However, in research, people care more about the marginal effect. Furthermore, the traditional regression method cannot take the interaction between exposure and confounders into account, which is particularly common in actual life.

The G-computation can exactly address the mentioned two flaws of the traditional regression model in obtaining a causal effect by assuming the regression model in G-computation is correctly defined. Therefore, we conducted the G-computation (see the result in Table 3) and obtained that being highly educated results in lower 24-year mortality (E(Y(1)) - E(Y(0)) = -0.05 with se(E(Y(1)) - E(Y(0))) = 0.0148, 95% Confidence interval = [-0.0832, -0.0252]). The assumption of the regression modelling was checked by visualizing the to be addressed (see Appendix 1).

Table 4. Output of the G-computation modelling

	Estimate	se	L 0.95CI	U 0.95CI
HighSchoolorBelow	0.000	0.000	0.000	0.000
CollegeorMore	-0.0542	0.0148	-0.0832	-0.0252

Approach 2: propensity score method with inverse probability weighting.

The propensity score method comes from the situation that certain values of confounds result in a higher chance of exposure. This method can give a summary of the confounds. The inverse probability weighting of the propensity score method is assumption-free. However, the propensity score method itself asks for the propensity score model to be correctly defined, i.e., the conditional positivity and conditional exchangeability hold after modelling. These assumptions were checked by visualizing the propensity

score and standardized mean difference (SMD)(See Appendix 2).

The output of our propensity score method with inverse probability weighting shows (see Table 5)the same results with approach 1, i.e., G-computation. That is, being highly educated can indeed result in lower 24-year mortality ($\hat{\beta}$ = -0.057, $se(\hat{\beta})$ = 0.016, t-value = -3.52, p < 0.001).

Table 5. Output of the propensity score method with inverse probability weighting

	Estimate	**se**	**t value**	**Pr(>|t|)**
Intercept	0.364	0.009	42.34	< 0.001
EDUC CollegeorMore	-0.057	0.016	-3.52	<0.001

Conclusions

This study used the sub data set of the Framingham Heart Study to analyse the relationship between different education level and mortality, while considering the other variables "Age" and "Sex". By excluding the mediators, "Age" and "Sex" become crucial predictors. When people's age increase, the risk ratio of dying also increase with age. Additionally, women have the lower probability of mortality compared to men. Furthermore, applying the G-computation and the propensity score method allow to analyse and reduce the useless information. According to the analysis results, these findings prove and explain the research question. And it clearly shows the important of education by explaining the association between education and the mortality. Therefore,

Reference/literature cited

1. World Health Organization. Regional Office for Europe. (2013). The European health report 2012: charting the way to well-being. World Health Organization. Regional Office for Europe. https://iris.who.int/handle/10665/326381

2. Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. Cureus. 2022 Oct 10;14(10):e30119. doi: 10.7759/cureus.30119. PMID: 36381818; PMCID: PMC9644238.

3. Ross, C. E., Wu, C. (1995). The Links Between Education and Health. American Sociological Review, 60(5), 719–745. https://doi.org/10.2307/2096319

Appendix

Result of Assumption checking of G-computation.

Result of Assumption checking of Propensity Score Method with Inverse Probability Weighting.

Code to conduct this analysis.