

DISCUSSION PAPER SERIES

DP14940

COLLIDER BIAS IN ECONOMIC HISTORY RESEARCH

Eric Schneider

ECONOMIC HISTORY



COLLIDER BIAS IN ECONOMIC HISTORY RESEARCH

Eric Schneider

Discussion Paper DP14940

Published 25 June 2020

Submitted 24 June 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Economic History

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Copyright: Eric Schneider

COLLIDER BIAS IN ECONOMIC HISTORY RESEARCH

Abstract

Economic historians have long been aware that sample-selection bias and other forms of bias could lead to spurious causal inferences. However, our approach to these biases has been muddled at times by dealing with each bias separately and by confusion about the sources of bias and how to mitigate them. This paper shows how the methodology of directed acyclical graphs (DAGs) formulated by Pearl (2009) and particularly the concept of collider bias can provide economic historians with a unified approach to managing a wide range of biases that can distort causal inference. I present ten examples of collider bias drawn from economic history research, focussing mainly on examples where the authors were able to overcome or mitigate the bias. Thus, the paper highlights how to diagnose collider bias and also strategies for managing it. The paper also shows that quasi-random experimental designs are rarely able to overcome collider bias. Although all of these biases were understood by economic historians before, conceptualising them as collider bias will improve economic historians' understanding of the limitations of particular sources and help us develop better research designs in the future.

JEL Classification: N01, N30

Keywords: sample-selection bias, collider bias, directed acyclical graphs

Eric Schneider - e.b.schneider@lse.ac.uk
LSE and CEPR

Acknowledgements

Thanks to Vellore Arthi, Jason Lennard, Mary Morgan, Felix Schaff and participants in the LSE Economic History Reading Group for helpful comments on an earlier draft. Any remaining errors are my own.

Collider Bias in Economic History Research*

Eric B. Schneider[†]

24 June 2020

Abstract

Economic historians have long been aware that sample-selection bias and other forms of bias could lead to spurious causal inferences. However, our approach to these biases has been muddled at times by dealing with each bias separately and by confusion about the sources of bias and how to mitigate them. This paper shows how the methodology of directed acyclical graphs (DAGs) formulated by Pearl (2009) and particularly the concept of collider bias can provide economic historians with a unified approach to managing a wide range of biases that can distort causal inference. I present ten examples of collider bias drawn from economic history research, focussing mainly on examples where the authors were able to overcome or mitigate the bias. Thus, the paper highlights how to diagnose collider bias and also strategies for managing it. The paper also shows that quasi-random experimental designs are rarely able to overcome collider bias. Although all of these biases were understood by economic historians before, conceptualising them as collider bias will improve economic historians' understanding of the limitations of particular sources and help us develop better research designs in the future.

Keywords: sample-selection bias, collider bias, directed acyclical graphs

JEL Codes: N01, N30

*Thanks to Vellore Arthi, Jason Lennard, Mary Morgan, Felix Schaff and participants in the LSE Economic History Reading Group for helpful comments on an earlier draft. Any remaining errors are my own.

[†]London School of Economics, e.b.schneider@lse.ac.uk, <https://www.ericbschneider.com/>

1 Introduction

Economic historians have long worried about sample-selection bias in historical sources because many historical sources are subject to sample selection either through their initial collection or survival. The recent debate sparked by Bodenhorn et al. (2017)'s critique of the anthropometric history literature has brought sample-selection bias back to the fore and prompted substantial new research on the topic including a special issue of *Social Science History* (Komlos and A'Hearn 2019; Schneider 2020; Steckel and Ziebarth 2016; Zimran 2019). Aside from typical sample-selection bias, economic history research is also subject to many other forms of bias that arise from the nature of our sources. Non-random attrition is common in longitudinal studies as is the use of truncated variables. Data may only survive for successful cases and we often rely on imperfect proxy measures that may be subject to bias. Economic historians have typically been astute at understanding these biases, but we can benefit and learn from interdisciplinary understandings of sample selection bias and other forms of bias that affect our work.

This paper introduces an approach to detecting and managing bias developed originally by the computer scientist Judea Pearl. Pearl has developed a graphical method for causal inference using directed acyclical graphs (DAGs) (Pearl 2009). This method has been adopted by epidemiologists and sociologists (Elwert 2013) but has seen slower take-up among economists who already had a very developed set of methodologies for determining causality, though c.f. Spiegler (2016). DAGs are very useful for teaching (especially undergraduates) because they explain some of the main threats to causal inference in a very intuitive and non-technical way (Cunningham 2018). However, DAGs can also be helpful to researchers in economic history through the concept of collider bias, which incorporates many common forms of bias that we face.

DAGs have several advantages relative to existing approaches. First, they are simpler, faster and more intuitive than many of the typical ways of modelling sample selection in economics. Second, modelling the bias in a DAG provides insight into how the researcher might mitigate its effect in different empirical settings. Economic historians obviously have less flexibility with research design than those working with more recent data or

experimental methods, which means that any insights on how to mitigate the effects of various forms of bias are helpful. Finally, DAGs make it clear that randomisation or other quasi-experimental research designs are not a panacea for collider bias. In fact, in most cases they are not able overcome collider bias because the collider bias is unrelated to the randomisation of the treatment (independent variable of interest). To be clear, the insights provided by DAGs about collider bias are not necessarily novel, i.e. traditional economics approaches would lead you to the same conclusions. However, their non-technical and intuitive structure may help to clarify these issues because many forms of bias can be summarised in a similar form.

This paper follows the structure of Elwert and Winship (2014), presenting ten examples of collider bias from economic history research. The examples are mainly drawn from historical economic demography since that is my own field of interest. Each example has its own DAG.¹ These DAGs are simplified versions that seek to highlight the potential for collider bias instead of illustrating all the potential causal paths between each exposure and outcome. Rather than seeking to undermine the papers discussed, I have almost exclusively chosen examples where the authors were able to manage the potential bias. Thus, the examples highlight both the forms collider bias can take in economic history research and techniques for overcoming (or seeking to mitigate) the bias. The forms of collider bias can be split into three broad categories: colliders that are direct descendants of the outcome variable, colliders that relate to intermediate variables and pre-treatment colliders. The paper gives a brief introduction to DAGs and collider bias before dealing with the three larger types of collider bias in turn.

2 Directed Acyclical Graphs and Collider Bias

There is not room in this paper for a full explication of DAGs and the do-calculus of causal inference developed by Judea Pearl in relation to them.² However, I will briefly

¹<http://www.dagitty.net/> is a useful web-based tool for drawing DAGs and getting a feel for how they work (Textor et al. 2016).

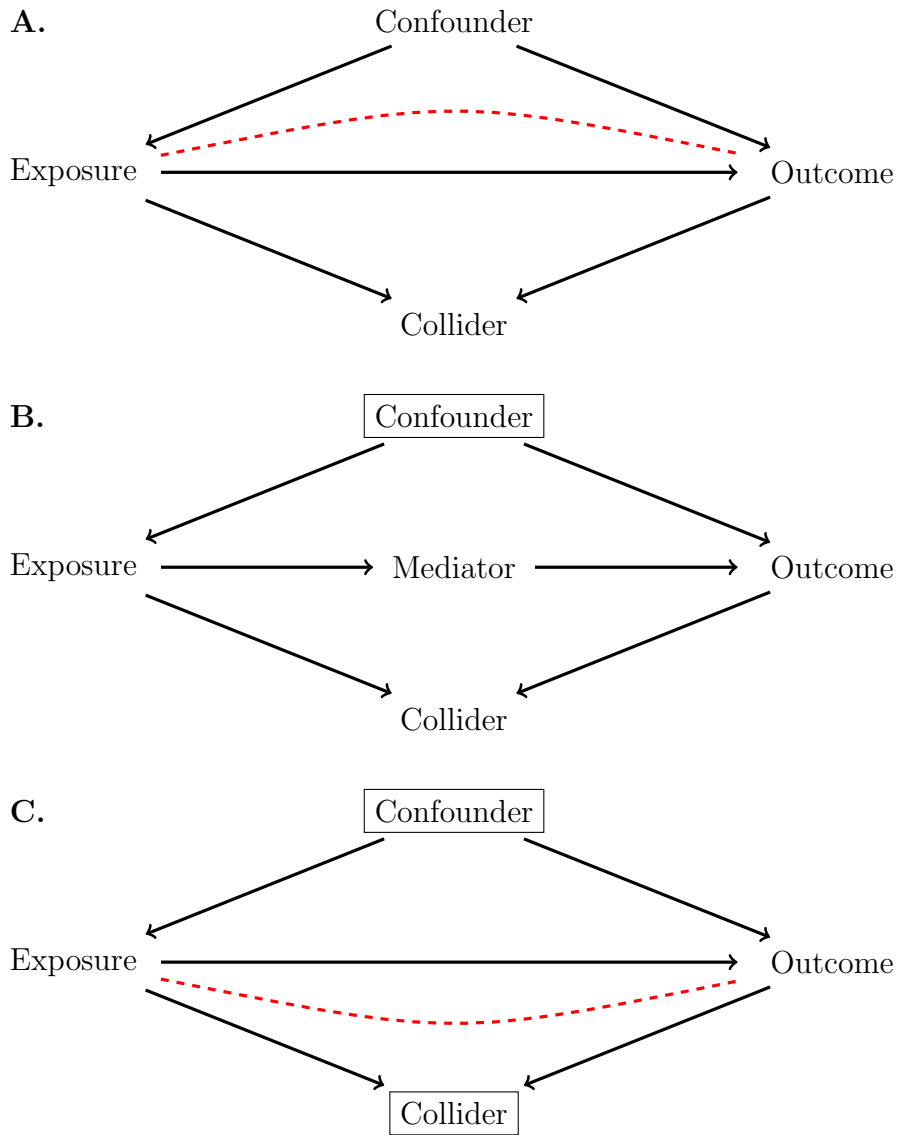
²For a brief introduction, see Cunningham (2018) chapter 4. For more advanced and technical discussions see Pearl (2009) or Elwert (2013).

present the language of DAGs to help the reader understand the points made in relation to collider bias below (see Figure 1). DAGs seek to illustrate the causal effect of an exposure variable, the main independent variable of interest, on an outcome variable, the dependent variable. Aside from the exposure and outcome, DAGs include three other types of variables: ancestor variables of the exposure or outcome, variables that cause the exposure or outcome; descendant variables of the exposure or outcome, variables that are caused by the exposure or outcome; and mediator variables between the exposure and outcome, variables caused by the exposure and that cause the outcome. Figure 1 presents these three types of variables.

Figure 1A presents a classic example of a confounding variable, which is simply an ancestor of both the exposure and outcome. In the language of DAGs, not controlling for a confounder leaves an open, backdoor (or non-causal) path between the exposure and outcome: $\text{Exposure} \leftarrow \text{Confounder} \rightarrow \text{Outcome}$. This creates bias equivalent to omitted variable bias in the relationship between exposure and outcome, represented in the DAG by the red dashed line. To get rid of this bias, one must close this non-causal path by controlling for the confounding variable, represented in the DAGs in this paper by putting a box around the variable. Thus, by controlling for the confounder in Figure 1B, we eliminate the bias in the effect between exposure and outcome because we have closed the non-causal path: $\text{Exposure} \leftarrow \boxed{\text{Confounder}} \rightarrow \text{Outcome}$.

Figures 1A and 1B also include a collider variable, which in this case is a descendent of both the exposure and outcome, i.e. a variable that is caused by both the exposure and outcome. Collider variables are different from confounders in that they naturally block non-causal paths from the exposure to the outcome. Thus, the path from exposure to outcome through the collider is naturally closed in Figures 1A and 1B: $\text{Exposure} \rightarrow \text{Collider} \leftarrow \text{Outcome}$. However, by controlling for or conditioning on the collider (or a descendent of the collider), one would open this non-causal path between exposure and outcome, inducing bias in the effect between exposure and outcome (see Figure 1C). This bias is known as collider bias or endogenous selection bias. Thus, in order to estimate a causal effect between exposure and outcome, the researcher must control for all confounding

Figure 1: Basic Structure of DAGs



Notes: Names in each DAG represent variables and the arrows show the direction of causation between the variables in the DAG. The exposure is the main independent variable (treatment variable) of interest, and the outcome is the dependent variable. Variables in a box have been controlled for or conditioned on in the DAG. This could mean that they are included as a control in a regression or that they are conditioned on by design since the data is only collected for a subset of the variable or sample selection is correlated with the variable. The red, dashed lines show the bias that is induced by conditioning on a collider (except in the case of Figure 1A, which shows bias induced by omitting a confounding variable).

variables and not condition on colliders.

In historical research, conditioning on a collider is often a product of the study design since historical sources may have only been collected or survive for selected samples and is especially likely to occur when sample selection is related to the outcome. However, collider bias can also arise in cases where sampling is not the underlying problem as in the cases described in Section 4 below where conditioning on an intermediate collider creates bias. Rather than discussing forms of bias at the theoretical level, the rest of the paper presents specific examples from economic history research of various types of collider bias to make the forms of bias clear.

3 Conditioning on an Outcome Collider

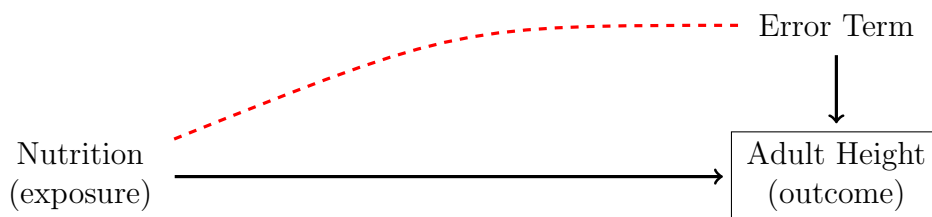
Before moving to complex forms of collider bias, it is best to start with the simplest form: colliders that are descendants of the outcome variable. I present four types of colliders that are direct descendants of the outcome variable.

3.1 Truncation Bias

Beginning with truncation bias, Figure 2 presents a very simple DAG following Floud et al. (1990) where the authors are (in the first instance) trying to understand the causal relationship between nutrition in childhood and adult height. The adult height measurements come from enlistment records of the British military, which had a minimum height requirement. This means that the outcome variable (adult height) is truncated, i.e. we are implicitly adjusting on the outcome variable. This creates collider bias because the outcome variable is always a collider on the path between the exposure and the error term. Thus, without accounting for the truncation via maximum likelihood estimators or other methods, we would estimate a biased effect of nutrition on adult height (A’Hearn 2004).

The fact that the outcome is always a collider on the path from the exposure to the error term yields another important insight. If one conditions on a variable that is a

Figure 2: Bias caused by Truncation of the Outcome



Sources: Empirical setting taken from Floud et al. (1990).

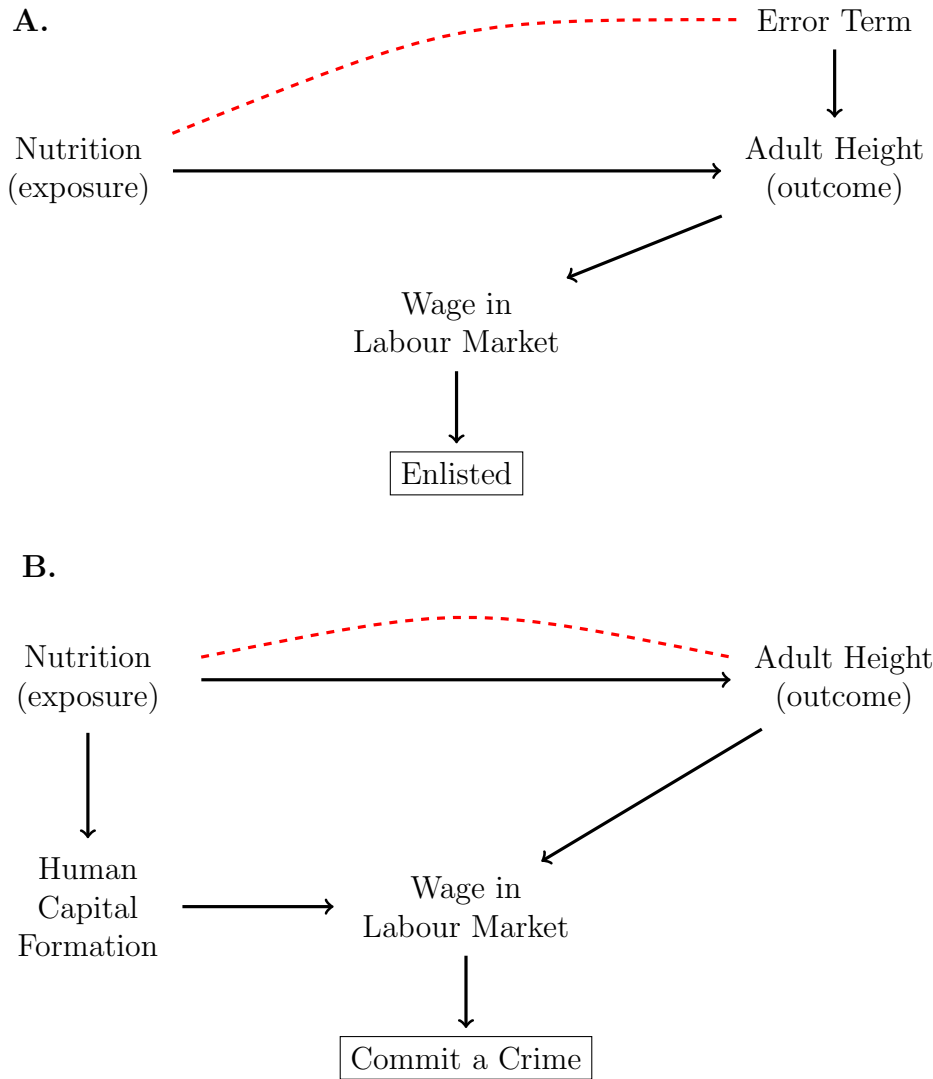
descendant of the outcome, this will always create collider bias whether the variable is related to the exposure or not. We will explore this in more detail in the next section.

3.2 Sample-Selection Bias

Collider bias can also be used to explicate the classic example of Heckman sample-selection bias. This applies to Bodenhorn et al. (2017)'s critique of the historical height literature. Bodenhorn et al. argue that since taller individuals enjoyed a wage premium in the labour market and the military was not an attractive employer in relative terms in the economy, changes in economic conditions would affect the types of men enlisting in the army. Specifically, they argue that as conditions improved during the Industrial Revolution, taller men had better employment opportunities in the civilian labour market, and therefore shorter men were more likely to enlist in the military. They argue that this bias, instead of changes in nutrition as had been argued before (Komlos 1998), produces the 'Industrial Growth Puzzle' or 'Antebellum Puzzle' where mean heights of adult men fell during the nineteenth century in Britain and the United States even as GDP per capita and real wages grew.

Seen in the context of a DAG (Figure 3A), the Bodenhorn et al. critique can be represented as conditioning on the descendant of a collider, the outcome. Since the researcher can only observe the heights of men who enlisted in the military, they are implicitly conditioning on a descendant of the outcome, introducing a non-causal association between the exposure and error term and biasing the estimated effect between nutrition and adult

Figure 3: Sample-Selection Bias as Collider Bias



Sources: Empirical setting for A adapted from Bodenhorn et al. (2017) and B from Bodenhorn et al. (2012).

height. Although there is debate in the literature about whether the Bodenhorn et al. critique is valid, thinking about selection on unobservables in a DAG context is helpful because it clarifies what solutions there are to the bias. Given that enlisted is not a confounder and is unrelated to the exposure in the DAG, the collider bias cannot be overcome by controlling for enlistment year or randomizing nutrition as the treatment (Hughes et al. 2019).³ Finding an instrumental variable (IV) for nutrition would not change the fact that enlistment could still be driven by height, especially since enlistment would occur after the exposure. Thus, one either has to model the selection explicitly as Zimran (2019) did, or seek to show that there is no causal link between the outcome and collider as Komlos and A’Hearn (2019) did. As a general note, one must always break the causal link between the outcome and the collider if the collider is a direct descendant of the outcome.

We can extend this example to nineteenth-century American prison records as in Bodenhorn et al. (2012). Because prison records are only collected for those convicted of a crime, they can create the same kinds of selection discussed above. Figure 3B presents a simplified DAG of this relationship. Again, the researcher is seeking to understand the relationship between nutrition in childhood and adult height. However, collider bias is present. The exposure variable, nutrition, affects human capital formation, which in turn influences an individual’s wage in the labour market in adulthood. On the other hand, height directly affects an individual’s wage in the labour market, and an individual’s wage affects their propensity to commit a crime since it affects the opportunity cost of crime. Thus, conditioning on committing a crime is conditioning on the descendant of a collider (wage in the labour market) and introduces bias in the causal path between nutrition in childhood and adult height.

The extent of the bias that this creates depends on the strength of the relationships between the exposure and outcome and the collider. For instance, it is possible that nutrition in childhood explains a very small share of human capital formation in the nineteenth-century economy because most human capital was gained through learning by

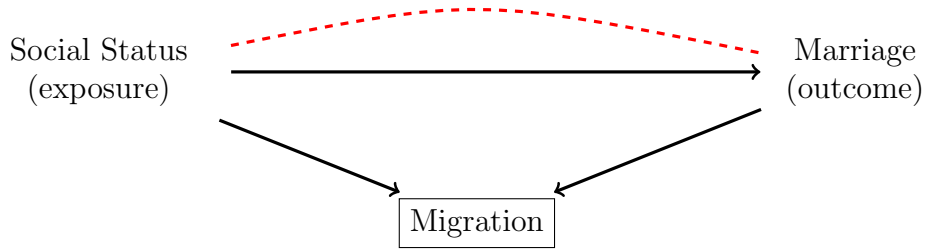
³Bodenhorn et al. (2017) propose a diagnostic test for selection on unobservables that includes enlistment year dummies in the regression, but this does not eliminate the potential bias.

doing and cognitive ability was less important. Thus, the correlation between nutrition in childhood and the wage in the labour market might be very small and therefore introduce only a small amount of bias. Likewise, one could argue that women's wages were not as strongly determined by their height as men's wages since men tended to work in the most physically demanding jobs where physical size might be most valued. This would mean that the bias introduced by conditioning on a collider might be smaller for women than for men, which is reassuring since prison records are one of the very few sources of women's heights that economic historians have. As Zimran (2019) shows, just because some bias is detectable in the data does not mean that the bias is large enough to overturn findings from the raw data that did not formally account for the collider. Finally, many researchers focus on local gaols like Houses of Correction rather than penitentiaries which housed criminals convicted of serious offences. By focusing on minor crimes such as drunkenness and disorderly conduct, these historians have argued that the link between an individual's wage in the labour market and their propensity to commit a crime was very small, reducing the potential bias (Horrell et al. 2009, p. 104). One could extend this approach by showing that height trends by birth cohort were similar among those committing crime strongly related and unrelated to labour market conditions. These solutions are not formally as convincing as modelling the selection directly, but they could help to assuage fears of substantial bias.

3.3 Nonresponse Bias

Another common form of collider bias in economic history research is nonresponse bias in longitudinal studies. One example of nonresponse bias arises in my recent paper with David de la Croix and Jacob Weisdorf (de la Croix et al. 2019). We were seeking to understand (in part) how social status affected women's propensity to marry versus to remain celibate across their lifecourse. We found that upper class women had much lower marriage rates than other social groups, i.e. upper-class women were more likely to remain celibate across their lifecourse. We analysed the Cambridge Group family reconstitution data, which uses records of baptisms (births), marriages and burials (deaths) in parish

Figure 4: Nonresponse Bias



Sources: Empirical setting taken from de la Croix et al. (2019).

registers to reconstruct the population of 26 parishes in early modern England (Wrigley et al. 1997). These vital events are recorded for people in a particular parish but not for surrounding parishes, so there are limits to what the data can tell us about the movement of people across parishes.

Figure 4 presents a simplified DAG for our empirical approach highlighting the potential collider bias. Social status is a categorical variable with seven occupational groups based on the wealth attributed to occupations in probate records. Marriage is a binary variable equal to one if we observed the woman’s birth, marriage and death in one of our parishes and zero if we observed a woman’s birth and death but no marriage or children, implying that the woman remained single across her lifecourse. We can only compute marriage rates for women who remained in the same parish their entire life, i.e. women who we observe both at birth and at death, because women born in our parish may have moved to a different parish to get married, leaving our dataset. These women are lost to follow up or non-respondents in our data, which could produce bias. If there were differential rates of marriage and celibacy between migrants and non-migrants, for instance because people migrated across parishes to get married and to start a new household, then this would mean that migration was a descendant of marriage, our outcome variable.

On the other hand, social status may also cause migration because women of higher social status might have migrated ‘for marriage more frequently since they had a smaller pool of eligible bachelors locally, had more money to search for potential partners and would have benefited more from their husband’s inheritance’ (de la Croix et al. 2019,

p. 246). Therefore, it is possible that we were conditioning on a collider, migration, which would have created a biased relationship between social status and marriage.

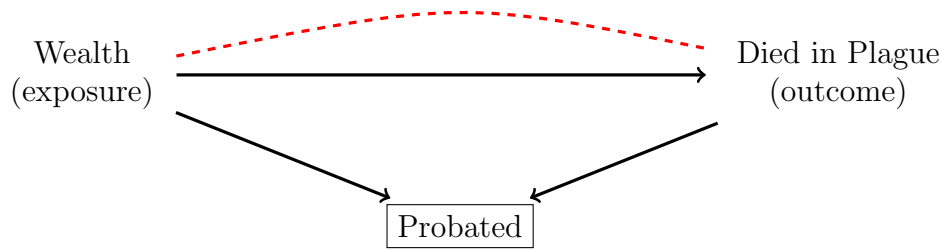
Unfortunately, there was no perfect solution for this problem because the data does not allow one to compare directly the marriage rates of migrants and non-migrants. To deal with this issue, we first tried to show that social status did not influence migration. There were no statistically significant or meaningful differences in outmigration rates of women by social class. Thus, upper-class women were not more likely to migrate than their middle- and lower-class counterparts. However, disproving this link is not sufficient because even if there were no causal link between social status and migration, migration would still be the descendent of a collider (the outcome) as in Figure 3A.

Second, we tried to disprove the causal connection between marriage and migration. Although we could not observe the marriages of out-migrants from our parishes, we could observe and study women who migrated into our parishes for marriage. We showed that the share of in-migrant brides was fairly constant across social groups and that upper class women were not more likely to migrate for marriage. Thus, it seems unlikely that there was a strong causal relationship between marriage and migration, and it certainly did not vary substantially by social group. This allowed us to argue that conditioning on non-migrants did not substantially bias our results. Although I have focused on this one, narrow example, nonresponse bias applies to many research designs in economic history including record linking between censuses or other administrative data where individuals who cannot be linked are nonrespondents.

3.4 Ascertainment Bias

Ascertainment bias is one of the most famous forms of collider bias. It is the example described by Cunningham (2018, pp. 78-79) looking at the correlation between acting talent and beauty. Cunningham (following Rossman) posits that acting talent and beauty are independent in the population. However, when one looks at the correlation between the two variables among Hollywood Stars, one would find a negative correlation. The reason for this is that becoming a Hollywood Star is caused by both being beautiful and

Figure 5: Ascertainment Bias



Sources: Hypothetical empirical setting.

talented. Some can overcome a lack of talent through exceptional beauty and others can overcome being less attractive through extraordinary talent. Thus, limiting the analysis to Hollywood Stars is conditioning on a collider by design and produces a spurious negative correlation between beauty and talent.

Economic historians sometimes face this issue as well. Consider a hypothetical example of a historian trying to understand whether plague mortality in early modern England was higher among the rich or the poor (Figure 5). To do this, the author might compare the wealth of individuals in probate records who died during the plague to those who died before or after the plague in ‘normal’ years. This empirical setup implicitly conditions on a collider because not all individuals were probated: the wealthy were more likely to be probated and people were probably less likely to leave wills and be formally probated during plague epidemics. Thus, conditioning on those who were probated could create a spurious finding that those dying during plague epidemics were wealthier than those dying in normal times, even if there was no difference in wealth between people dying during the plague and in normal years. Ascertainment bias cautions against looking at the causes of success (or failure) among the successful (or failures).

4 Conditioning on an Intermediate Collider

Having discussed several examples where the collider is a descendant of the outcome, we now turn to the more nuanced and trickier cases where the collider is an intermediate variable. By an intermediate variable, I mean a variable whose causal process occurs

between the exposure (treatment) and outcome and that is not a direct descendant of the outcome. I present four examples. Two are a product of selective attrition of individuals in longitudinal studies, another is the result of using an imperfect proxy and the fourth arises from conditioning on a mediator.

4.1 Informative/Dependent Censoring and Attrition Bias in Longitudinal Studies

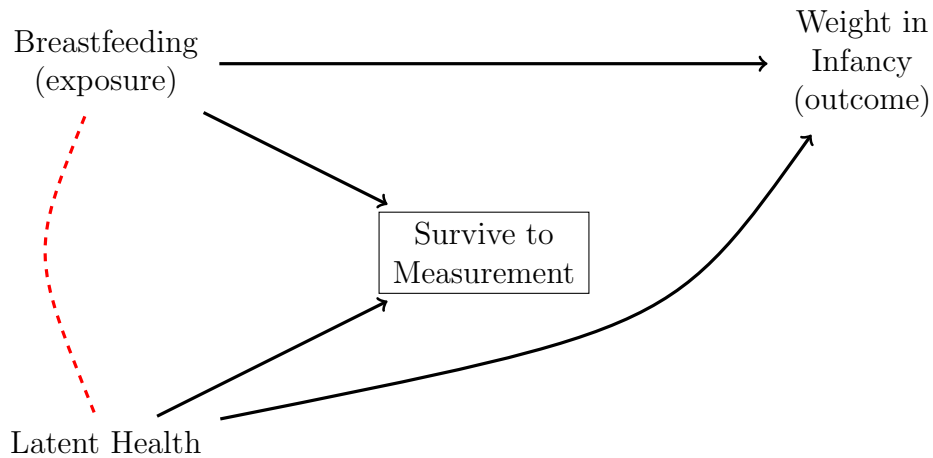
4.1.1 Survival Bias

Similar to nonresponse bias described above, collider bias also arises when the researcher can only observe outcomes for individuals who survived to a certain age and this attrition is caused by the exposure and an ancestor of the outcome. Survivor bias is very common in historical economic demography and is difficult to overcome. For example, in a recent paper with Vellore Arthi, we wanted to understand how infant feeding practices in early infancy (in particular breastfeeding) affected the weights of children later in infancy (Arthi and Schneider 2020). To test this relationship, we used a sample of children from the London Foundling Hospital. Figure 6 presents a simplified version of the DAG for this relationship. Since breastfeeding protects against mortality in infancy, there was a clear causal link between breastfeeding and survival. An individual's latent (initial) health status would clearly also affect their propensity to survive to measurement and would also influence their weight in infancy. We were forced to condition on a collider because we only observed weights in infancy and infant feeding information for children who survived to be measured in infancy (a number of months after birth). Conditioning on the collider, survive to measurement, opens a non-causal path from breastfeeding to weight in infancy and biases our estimate of the effect of exposure on outcome.

In this case, the collider bias would lead us to underestimate the benefits of exclusive breastfeeding on weight in infancy. The simplest way to see this is if we compare two categories of breastfeeding as a binary variable: exclusively breastfed = 1 and never breastfed = 0.⁴ Assume that the two groups had the same latent health at birth, and

⁴The paper takes a more nuanced perspective than this, but this simplifies the explanation.

Figure 6: Survivor Bias



Sources: Empirical setting taken from Arthi and Schneider (2020).

that breastfeeding did not respond to this latent health status.⁵ Mortality would be higher among those children who were never breastfed, and it is possible that this mortality would have selectively culled weaker individuals. At its most extreme, this would amount to removing the left tail of the initial health distribution, which would mean that the survivors later in infancy in the never breastfed group would have had higher mean initial health than the original group of never breastfed children. Thus, any difference in weight in infancy between the exclusively and never breastfed groups would be equal to the causal effect of breastfeeding minus the positive survival bias.

In order to combat this bias, we were able to find birth weights for a sub-sample of the Foundling Hospital children and compare their birth weights to a random sample of children born in London during the same period. Our results suggest that although mortality may have been higher among the never breastfed group, this mortality did not appear to be related to latent health, proxied by birth weight. Birth weights of breastfed and never breastfed children observed later in infancy were not statistically different from one another or from birth weights in the comparative sample of children. This casts doubt on the causal relationship between latent health and survival within

⁵We test this using a subsample and find no significant differences between the birth weight distributions of children who were breastfed and those who were not. This suggests that mothers did not seek to compensate or reinforce initial health endowments with their breastfeeding decisions.

breastfeeding categories at least insofar as birth weight captures latent health. We also controlled directly for birth weight as a proxy of latent health, which closes the non-causal path between exposure and outcome through the collider survival, eliminating the collider bias. This example highlights the importance of key controls in managing collider bias on intermediate variables.

A sceptical reader, however, would note that breastfeeding and latent health are not the only two variables that would affect the outcome, weight in infancy, and any common ancestor of the outcome and the collider (survival) would open its own biasing, non-causal path. For example, disease exposure and non-feeding health investments would both influence survival and weight in infancy and create biasing non-causal paths.⁶ Note that randomising breastfeeding at birth either experimentally or quasi-experimentally through an IV or differences-in-differences strategy would not necessarily overcome the collider bias. Randomisation of breastfeeding at birth would close non-causal, confounding paths between the exposure and outcome but would not remove the collider bias since the attrition from the sample occurs after the treatment. Thus, to remove the collider bias, randomisation would have to somehow break the causal link between breastfeeding and survival. It is difficult to think of any randomisation that could break that link.

To address the open non-causal paths described above, we added as many controls as possible. Fortunately, the Foundling Hospital Dataset we use is extraordinarily rich, allowing us to control for a wide variety of variables. For instance, we proxied disease exposure using birth year, birth district, and birth season fixed effects. In unreported specifications, we went even further controlling for individual-level exposure to some diseases and the infant mortality rate in the birth district in the year of birth. These additional controls did not substantially affect our results, suggesting that disease exposure (independent of infant feeding) did not produce significant bias. However, non-feeding health investments remained unobserved, so this biasing path and any relationship between the disease environment and the outcome not captured by our proxies remained open. To provide bounds for the importance of this bias, we checked for selection on unobservables using

⁶By non-feeding health investments, I mean washing the child, keeping a clean home, comforting the child, etc.

the test proposed by Oster (2019), which uses changes in the coefficient of the exposure variable and R-square when adding controls to give bounds on the potential importance of selection on unobservables. This test suggested that there was very little selection on unobservables biasing our main coefficient, alleviating concerns that there was substantial bias in our estimate causal effect of breastfeeding on weight in infancy. Our strategy for dealing with collider bias for this paper highlights the importance of key controls and the limitations of randomisation of treatment in mitigating this form of intermediate variable collider bias.

4.1.2 M-Bias

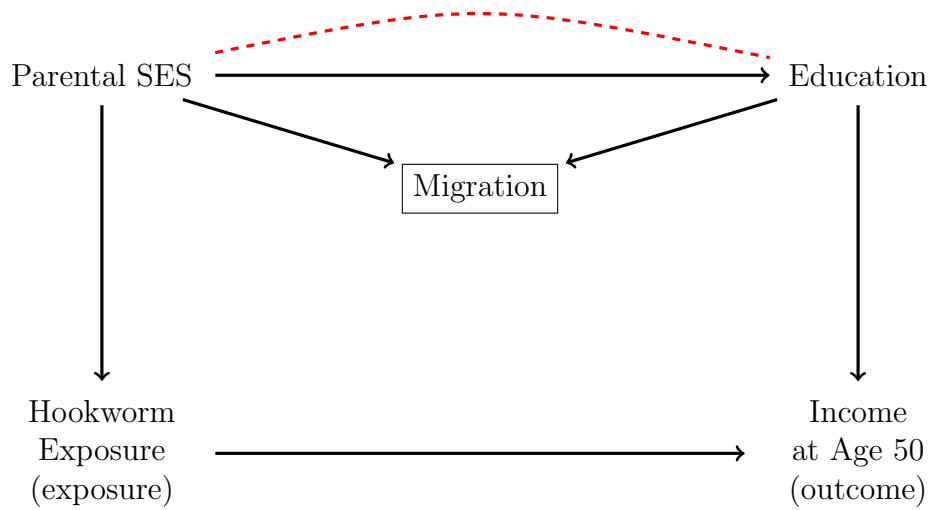
A particularly tricky form of collider bias is known as M-bias because the DAG structure tends to look like an M. In this case, the collider variable is not a direct descendant of either the exposure or the outcome variable but is instead a descendant of ancestors of the exposure and outcome. However, conditioning on an M-like collider still opens up a non-causal path from the exposure to the outcome biasing the estimated causal effect of interest (Liu et al. 2012).

For an example of M-bias, we can borrow the empirical setting of Bleakley (2007) in a slightly altered form. Bleakley wanted to understand how exposure to disease in childhood affected outcomes of children later in life (Figure 7). He focussed on hookworm in the US South in the early twentieth century. Hookworm is a parasitic disease that saps children's nutrients, making it difficult for them to concentrate in school.⁷ He showed that exposure to hookworm substantially decreased children's earnings in later life (say age 50).

This setting could potentially suffer from M-style collider bias if Bleakley were only able to observe hookworm exposure in childhood for non-migrants at age 50 in the South. This empirical setup would lead to the attrition of migrants by design, and although migration is not caused by the exposure variable (hookworm) or the outcome variable (income at 50) directly, it is caused by ancestors of the exposure and outcome. Migration

⁷Note that hookworm was very rarely fatal, so there should not be any survivor bias issues as described in Figure 6 above.

Figure 7: M-Bias due to Attrition related to Ancestors of Exposure and Outcome



Sources: Empirical setting adapted from Bleakley (2007).

is particularly important in the US South in this period as a large share of the Black population moved to the North during the Great Migration. These migrants tended to be positively selected on income and education and were more likely to come from a non-farm origin (Collins and Wanamaker 2015).⁸

Returning to our DAG, parental socio-economic status (SES) would clearly influence individual children’s exposure to hookworm since better off parents could buy their children shoes, maintain higher levels of hygiene and live in areas with better sanitary conditions. However, parental SES would also influence the propensity of individuals to migrate since migrants were more likely to come from non-farm employment. On the other hand, an individual’s education would influence their income at age 50 and would also influence their propensity to migrate since migrants were positively selected on education. The attrition of high SES migrants from the sample leads to a biased estimate of the influence of hookworm exposure on income at age 50.

Simulations conducted by epidemiologists have generally found that M-bias is small relative to typical confounding bias unless the correlation between the collider and its ancestors is very high (Liu et al. 2012), suggesting that M-bias may be less problematic

⁸The size of the positive selection was somewhat small, but the example is still valid.

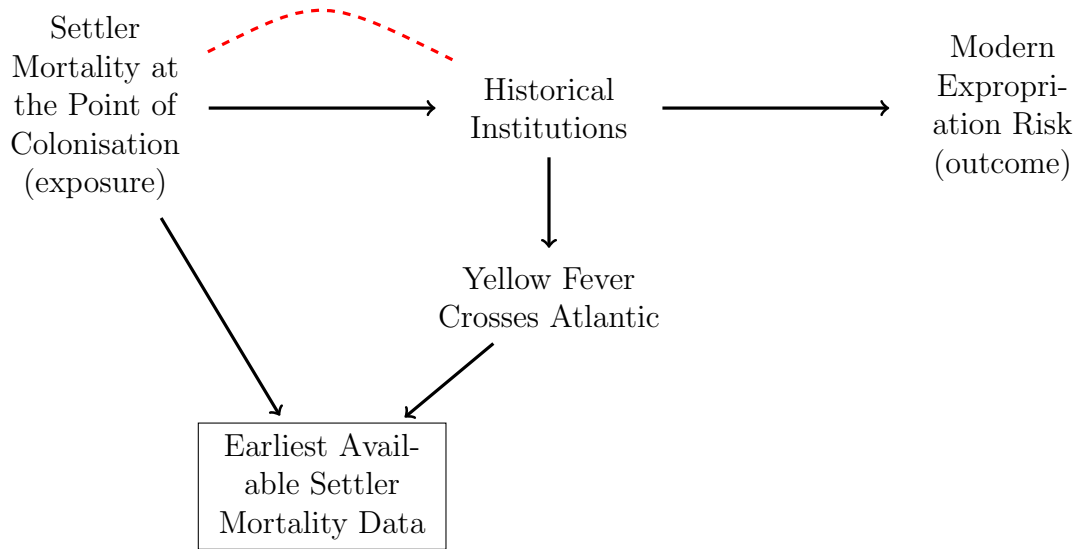
than other forms of collider bias. However, Bleakley (2007) avoids this collider bias in a number of ways. First, he uses a difference-in-difference methodology to break the causal link between parental SES and hookworm exposure. He uses the deworming campaigns carried out by the Rockefeller Sanitary Commission for the Eradication of Hookworm Disease after 1910 as an exogenous shock to hookworm exposure across cohorts. He also takes account of the different initial levels of hookworm exposure since the eradication of hookworm would have been more important in areas where hookworm was more prevalent. He compares cohorts born before and after hookworm eradication and in places where hookworm was very prevalent and not prevalent in order to produce quasi-randomised variation in hookworm exposure. Thus, randomisation of treatment is a way of eliminating M-style collider bias since the collider is not a direct descendant of the treatment or the outcome as in the examples above.

Bleakley also avoids this bias by including all individuals born in states where hookworm was prevalent. Thus, he does not exclude migrants from his analysis. This choice helps to avoid the collider bias above, but it also introduces other sources of endogeneity. The US census only records the state of birth of individuals, so this means that the hookworm exposure variable cannot be as precise as it would have been at the county level. This could introduce measurement error into the exposure variable and cause attenuation bias. Thus, there are often trade-offs with imperfect data about the types of bias one is willing to accept.

4.2 Imperfect Proxy Measures

Another form of intermediate collider bias arises from using imperfect proxies that may be descendants of other variables in the DAG. This is one of the forms of a ‘bad control’ described by Angrist and Pischke (2009, pp. 66-68). Figure 8 presents an example based on the first stage of Acemoglu et al. (2001)’s famous paper. Acemoglu et al. argue that settler mortality at the point of colonisation of a territory by Europeans determined whether the Europeans set up extractive or inclusive institutions. Where Europeans settlers faced high mortality rates, they set up extractive institutions such as the *encomienda* and slavery

Figure 8: Collider Bias due to Imperfect Proxy Variables

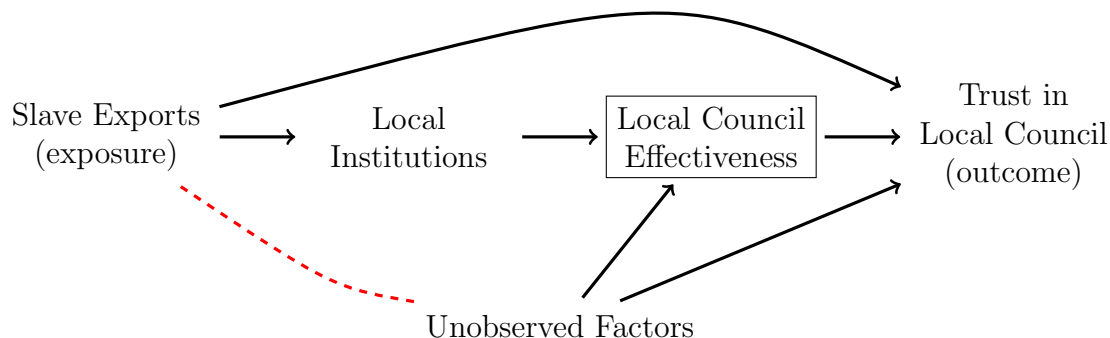


Sources: Empirical setting taken from Acemoglu et al. (2001).

in the Americas, but where Europeans faced low mortality rates, they settled importing their inclusive, participatory institutions. Albuoy (2012) has raised a number of issues about this paper, but one of his critiques is that the mortality data that Acemoglu et al. use mainly comes from non-combat deaths of European militaries in the late eighteenth and nineteenth centuries long after colonisation. This is especially true of mortality in Latin America and the Caribbean.

We can illustrate Acemoglu et al.’s use of a proxy for settler mortality in the DAG in Figure 8. The earliest available settler mortality data is a proxy for settler mortality at the point of colonisation. However, it is also a descendant of historical institutions. As McNeill (1999) has argued, yellow fever was an African disease that was transmitted across the Atlantic through the slave trade and gained a hold in the Americas in part because sugar cane production led to the use of small clay pots that often sat with stagnant water in them near population centres. Yellow fever epidemics were extraordinarily deadly for European troops at the end of the eighteenth century and therefore were an important cause of the earliest settler mortality data available to Acemoglu et al. for Latin America and the Caribbean. Thus, this imperfect proxy variable induces collider

Figure 9: Collider Bias due to Conditioning on a Mediator



Sources: Empirical setting taken from Nunn and Wantchekon (2011).

bias in the estimate of the causal effect of settler mortality at the point of colonisation on modern expropriation risk. If Acemoglu et al. wanted to overcome this bias, they could go back to the original sources and attempt to exclude yellow fever deaths from their mortality estimates in order to produce a more accurate proxy of settler mortality at the point of colonisation.

4.3 Mediation Analysis and Causal Mechanisms

Mediation analysis is less common in economic history than in other social science disciplines, but economic historians will at times condition on mediators when trying to understand the mechanisms behind a particular causal effect. This type of analysis is particularly common in persistence papers that analyse the long-run consequences of a particular historical institution. Collider bias arises in mediation analysis if the mediator shares a common cause with the outcome. If this is the case, then conditioning on the mediator is conditioning on a collider, which opens a non-causal path between exposure and outcome. This case is the second form of ‘bad control’ discussed by Angrist and Pischke (2009, pp. 64-66).

Nunn and Wantchekon (2011) provides an example of this kind of collider bias in economic history research (Figure 9). In the paper, Nunn and Wantchekon wanted to understand the direct effect of exposure to the historical slave trade on trust among

Africans today. To measure this effect, they regressed individual-level responses to a number of questions related to trust in the Afrobarometer Surveys on the exports of slaves in the historical slave trade. One of their results showed that the slave trade affected trust in local government councils, but they wanted to know whether this effect was a direct effect of the slave trade on trust or whether the effectiveness or quality of the local council mediated this effect on trust, i.e. did people trust the local council less because the slave trade directly affected the quality of local government. Thus, they controlled for several indicators of local council quality to see how this changed the coefficient on slave exports. They found that the mediated effect explained about half of the total effect of slave exports on trust in the local council.

However, by conditioning on the mediator, they also opened a non-causal path between exposure and outcome because unobserved factors may have affected both trust in the local council and the effectiveness of the local council. This induces a spurious correlation between slave exports and these unobserved factors, which biases the analysis. One such factor would have been local public goods. In actuality, Nunn and Wantchekon control for local public goods in a second specification, which closes the backdoor path through the collider, but as mentioned above, any variable that is a common ancestor of the mediator and outcome will lead to bias when conditioning on the mediator. This type of bias does not arise frequently in economic history research, but those trying to understand causal mechanisms by controlling for mediators should consider it carefully.

5 Pre-Treatment Colliders: Conditioning on Endogenous Location Choices

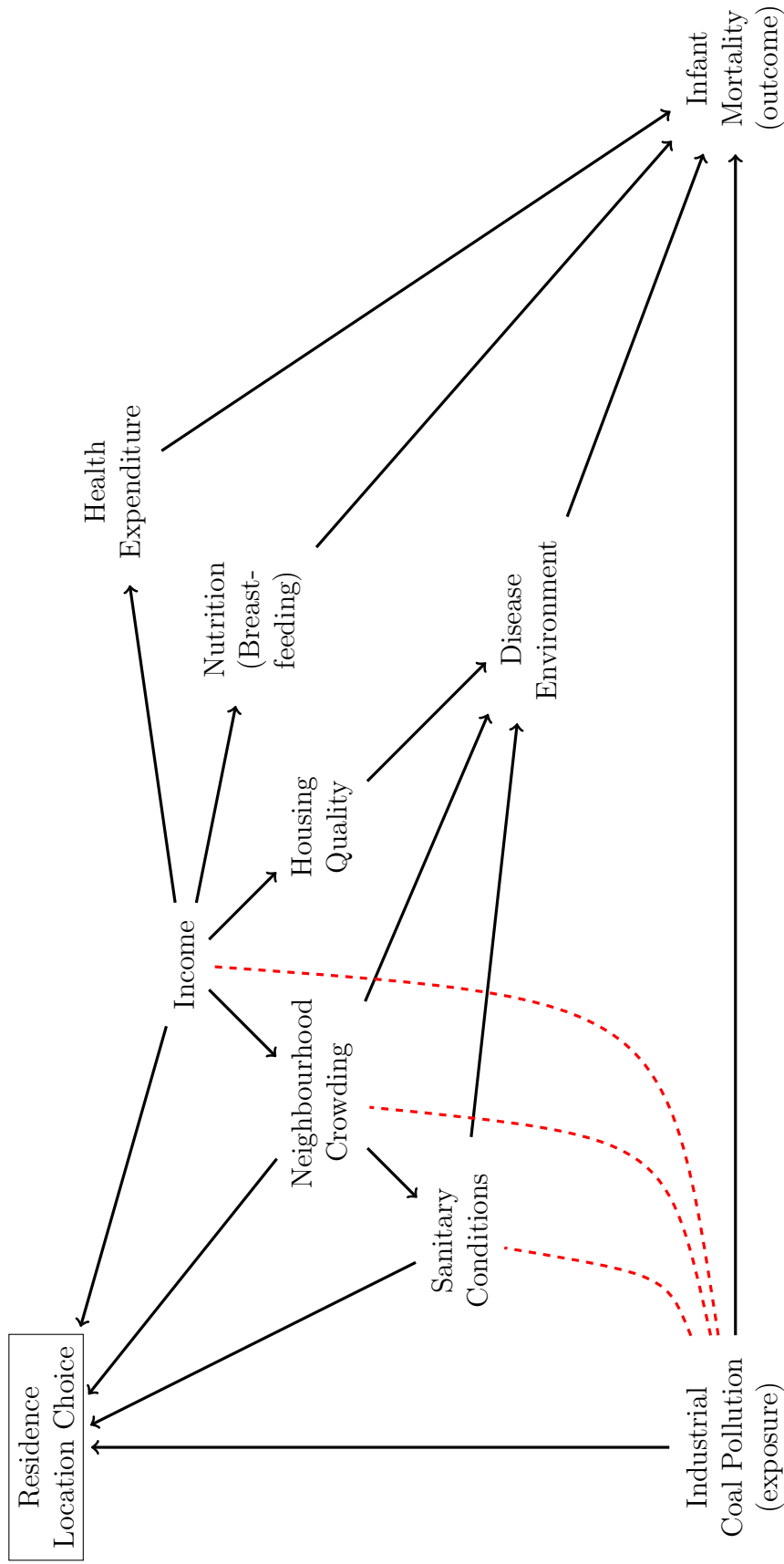
A final type of collider bias may occur when the researcher conditions on a pre-treatment collider. Elwert and Winship (2014) highlight homophily bias in social network analysis as an example, and this could certainly apply to economic history research (Esteves and Geisler Mesevage 2019). However, another example arises when the researcher only observes individuals (or firms) after they have made a choice about where to live (or

locate). An example of this is the recent paper by Beach and Hanlon (2018), which tests the effects of industrial coal smoke pollution on infant mortality using cross-sectional data from England and Wales, 1851-60. If we simplify for a moment and assume that the treatment effects of coal pollution would not be confounded by an individual's underlying health status, then Figure 10 presents a DAG that captures the causal pathway that Beach and Hanlon wanted to identify. As the DAG shows, there are many variables that influence mortality, but none of these would directly affect the location of industrial pollution in 1851-60. One might then feel comfortable simply regressing coal pollution in a registration district on the infant mortality rate in a registration district. However, this ignores the fact that individuals sort across locations based on observable amenities. In Figure 10, individuals choose locations based on their income, the crowding in a neighbourhood, the sanitary conditions and the observable pollution. Because Beach and Hanlon only observe the deaths of infants where their parents had chosen to live, they are implicitly conditioning on a collider. Their direct causal estimate is biased by all of the non-causal paths that run from pollution through the collider (residence location choice) to mortality, i.e. the raw association captures not only the effects of pollution on mortality but also that of disease exposure, neighbourhood crowding and income.

There are two strategies for mitigating the collider bias in this case. First, the researcher could attempt to control for one variable on each path from the collider to the outcome, i.e. control for (at minimum) income, neighbourhood crowding and sanitary conditions. This would close these non-causal paths and allow for the estimation of the causal effect of coal pollution on infant mortality. In actuality, it would be very difficult to find this kind of data for each registration district.

A second strategy would be to break the causal link between the exposure and collider. Note again that finding quasi-random variation in the exposure (industrial pollution) will not necessarily get rid of the bias. Only randomisation that breaks the causal link between the exposure and the collider would eliminate the collider bias. For instance, using industrial coal pollution in year $t - 50$ as an IV for industrial pollution in year t would provide quasi-randomised variation in the exposure, but because residence location

Figure 10: Conditioning on a Pre-Treatment Collider: Endogenous Location Choices



Notes: Beach and Hanlon (2018) estimate this DAG as a cross-section of registration districts across England and Wales. As mentioned in the text, the collider bias would be resolved if they could observe short-run changes in pollution that would be unlikely to affect longer-term residence location choices.

Sources: Empirical setting adapted from Beach and Hanlon (2018).

choices would occur after randomisation, part of the predicted variation in the endogenous variable would still be causally linked to the sorting that is the source of the collider bias. To break the causal link, one would need exogenous variation in coal pollution that came after residence location choices had been made: for instance a sudden shock to coal pollution such as the Lancashire cotton famine;⁹ changes in the weather that would temporarily intensify or ameliorate the existing pollution of an area but would not affect longer-run residential sorting (Hanlon 2019); or the introduction of ‘clean coal’ technology, such as taller chimneys, that would suddenly reduce the impact of coal production locally (c.f. Clay et al. (2019)). As Elwert and Winship (2014) highlight, when dealing with pre-treatment colliders, the researcher needs to develop a clear understanding of the data generating process in order to understand the bias and mitigate its effect.

Beach and Hanlon overcome this issue by comparing the effect of coal pollution upwind and downwind of a particular district. They argue that pollution upwind of a district will affect mortality but will be less noticeable in terms of residence location choices than the smoke pollution emitted in a district. Upwind and downwind districts were also likely to be similar to the target district, eliminating other unobservable confounders. This strategy breaks the causal link between the exposure and collider, eliminating the collider bias on the upwind part of the pollution in each district.

This form of collider bias is rather common in historical demography and economic geography. We know that people were not static across space (Schürer and Day 2019) and that migratory flows could respond quickly to economic conditions (Arthi et al. 2019). In addition, there is substantial evidence that internal migrants tended to be positively selected (Collins and Wanamaker 2015; Humphries and Leunig 2009) and that urban disamenities affected the growth of cities (Hanlon 2019). Thus, researchers need to consider the potential bias of endogenous location choices more carefully in the future.

⁹Though this would create other issues since the shock would affect income and migration directly Arthi et al. (2019).

6 Conclusion

Collider bias is common in research designs employing historical sources. Following Elwert and Winship (2014), this paper has presented ten examples of different forms of collider bias in economic history research. The forms of bias described in this paper are all well-known to economists and economic historians, but economic historians could benefit from recognising and analysing these forms of bias as collider bias for several reasons. First, because DAGs force the researcher to make their assumptions about the causal relationships between variables explicit, they can help the researcher both diagnose potential sources of collider bias and understand the type of evidence or empirical strategy necessary to mitigate their effects. Since economic historians have often have less flexibility in their research design, we have more to gain from using DAGs than other economists who can simply opt for an empirical strategy unaffected by collider bias.

Second, DAGs clarify when randomisation or quasi-randomisation of the treatment does and does not overcome various forms of collider bias. Randomisation cannot overcome collider bias from a direct descendent of the outcome since randomising the treatment will have no effect on the bias. Likewise, randomisation is only ensured to mitigate collider bias on an intermediate variable if the collider is a descendant of ancestors of the exposure and outcome (M-bias discussed in Section 4.1.2). Otherwise, randomisation only eliminates intermediate variable collider bias if it breaks the causal link between the exposure and the collider. Whether this occurs often depends on the timing of randomisation of the treatment relative to the collider generating causal path. If the causal path generating the collider bias must occur after the treatment, as in the case of survival bias presented in Section 4.1.1, it is almost impossible for randomisation to overcome this bias. However, if randomisation of the treatment can occur after the causal path generating collider bias, then randomisation may be able to break this causal link.

Third, the examples have highlighted a number of strategies (other than randomisation) for mitigating collider bias. Since the causal relationships represented in the DAGs are often drawn from theoretical understandings, one can attempt to show that there is not a strong causal link between the collider and the other variables in a particular

historical setting. For sample selection dependent on the outcome, one can model sample selection explicitly to formally account for the bias. For collider bias on an intermediate variable, one can control for key variables that break the non-causal path from the exposure through the collider to the outcome. For the most part, though, avoiding collider bias may simply require being more careful about the types of causal relationships that can be studied using a particular dataset, an extension of understanding the limits of data at which economic historians excel.

As a final note, I would like to highlight how useful DAGs can be in teaching causal inference, especially to undergraduates. Obviously, this method will never replace standard econometrics. However, DAGs demonstrate forms of endogeneity like omitted variable bias and reverse causality in a very clear, intuitive and non-technical way and can also be used to explain the most common identification strategies used by economic historians. As illustrated in this paper through collider bias, they can also incorporate sample-selection bias and other forms of bias that students often struggle to understand. Thus, DAGs could be another powerful tool for teaching the fundamentals of causal inference (Cunningham 2018).

References

- Acemoglu, D., Johnson, S., and Robinson, J. (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review*, 91(5):1369–1401.
- A’Hearn, B. (2004). A restricted maximum likelihood estimator for truncated height samples. *Economics and Human Biology*, 2(1):5–19.
- Albouy, D. Y. (2012). The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *The American Economic Review*, 102(6):3059–3076.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, Princeton.
- Arthi, V., Beach, B., and Hanlon, W. W. (2019). Recessions, Mortality, and Migration Bias: Evidence from the Lancashire Cotton Famine. *Working Paper*, pages 1–73.
- Arthi, V. and Schneider, E. B. (2020). Infant Feeding and Post-Weaning Health: Evidence from Turn-of-the-Century London. *Working Paper*, pages 1–78.
- Beach, B. and Hanlon, W. W. (2018). Coal Smoke and Mortality in an Early Industrial Economy. *The Economic Journal*, 128(615):2652–2675.
- Bleakley, H. (2007). Disease and Development: Evidence from Hookworm Eradication in the American South. *The Quarterly Journal of Economics*, 122(1):73–117.
- Bodenhorn, H., Guinnane, T. W., and Mroz, T. A. (2017). Sample-Selection Biases and the Industrialization Puzzle. *Journal of Economic History*, 77(1):171–207.
- Bodenhorn, H., Moehling, C., and Price, G. N. (2012). Short Criminals: Stature and Crime in Early America. *Journal of Law and Economics*, 55(2):393–419.
- Clay, K., Lewis, J., and Severnini, E. (2019). Canary in a Coal Mine: Infant Mortality and Tradeoffs Associated with Mid-20th Century Air Pollution. *Working Paper*, pages 1–64.

- Collins, W. J. and Wanamaker, M. H. (2015). The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *Journal of Economic History*, 75(04):947–992.
- Cunningham, S. (2018). *CAUSAL INFERENCE: THE MIXTAPE (V. 1.7)*.
- de la Croix, D., Schneider, E. B., and Weisdorf, J. (2019). Childlessness, celibacy and net fertility in pre-industrial England: the middle-class evolutionary advantage. *Journal of Economic Growth*, 24(3):223–256.
- Elwert, F. (2013). Graphical Causal Models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*, pages 1–29. Springer, Dordrecht.
- Elwert, F. and Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40(1):31–53.
- Esteves, R. and Geisler Mesevage, G. (2019). Social Networks in Economic History: Opportunities and Challenges. *Explorations in Economic History*, 74:101299.
- Floud, R., Wachter, K. W., and Gregory, A. (1990). *Height, health and history : nutritional status in the United Kingdom, 1750-1980*. Cambridge University Press, Cambridge.
- Hanlon, W. W. (2019). Coal Smoke, City Growth, and the Costs of the Industrial Revolution. *The Economic Journal*, pages 1–27.
- Horrell, S., Meredith, D., and Oxley, D. (2009). Measuring misery: Body mass, ageing and gender inequality in Victorian London. *Explorations in Economic History*, 46(1):93–119.
- Hughes, R. A., Davies, N. M., Davey Smith, G., and Tilling, K. (2019). Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. *Epidemiology*, 30(3):350–357.
- Humphries, J. and Leunig, T. (2009). Was Dick Whittington taller than those he left behind? Anthropometric measures, migration and the quality of life in early nineteenth century London? *Explorations in Economic History*, 46(1):120–131.

- Komlos, J. (1998). Shrinking in a Growing Economy? The Mystery of Physical Stature during the Industrial Revolution. *The Journal of Economic History*, 58(3):779–802.
- Komlos, J. and A’Hearn, B. (2019). Clarifications of a Puzzle: The Decline in Nutritional Status at the Onset of Modern Economic Growth in the United States. *The Journal of Economic History*, 79(4):1129–1153.
- Liu, W., Brookhart, M. A., Schneeweiss, S., Mi, X., and Setoguchi, S. (2012). Implications of M Bias in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 176(10):938–948.
- McNeill, J. (1999). Ecology, Epidemics and Empires: Environmental Change and the Geopolitics of Tropical America, 1600-1825. *Environment and History*, 5(2):175–184.
- Nunn, N. and Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American Economic Review*, 101(7):3221–3252.
- Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business and Economic Statistics*, 37(2):187–204.
- Pearl, J. (2009). *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, second edition. edition.
- Schneider, E. B. (2020). Sample Selection Biases and the Historical Growth Pattern of Children. *Social Science History*, pages 1–31.
- Schürer, K. and Day, J. (2019). Migration to London and the development of the north–south divide, 1851–1911. *Social History*, 44(1):26–56.
- Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations *. *The Quarterly Journal of Economics*, 131(3):1243–1290.
- Steckel, R. H. and Ziebarth, N. (2016). Trader Selectivity and Measured Catch-Up Growth of American Slaves. *Journal of Economic History*, 76(01):109–138.

- Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. H. (2016). Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *International Journal of Epidemiology*, 45(6):1887–1894.
- Wrigley, E. A., Davies, Oeppen, J., and Schofield, R. (1997). *English Population History from Family Reconstitution*. Cambridge University Press, Cambridge.
- Zimran, A. (2019). Sample-Selection Bias and Height Trends in the Nineteenth-Century United States. *Journal of Economic History*, 79(1):99–138.