Health Services and Outcomes Research Methodology (2023) 23:354–375
https://doi.org/10.1007/s10742-022-00291-x

Bias? Clarifying the language barrier between epidemiologists and economists

AN Varga[1] · AE Guevara Morel[1] · JM van Dongen[1] · J Lokkerbol[2] · M Lindeboom[3,4] · MW van Tulder[5,6] · LM Bouter[7,8] · JE Bosmans[1]

Received: 14 January 2022 / Revised: 5 October 2022 / Accepted: 22 October 2022 /
Published online: 3 November 2022
© The Author(s) 2022

Abstract
In health intervention research, epidemiologists and economists are more and more interested in estimating causal effects based on observational data. However, collaboration and interaction between both disciplines are regularly clouded by differences in the terminology used. Amongst others, this is reflected in differences in labeling, handling, and interpreting the sources of bias in parameter estimates. For example, both epidemiologists and economists use the term selection bias. However, what economists label as selection bias is labeled as confounding by epidemiologists. This paper aims to shed light on this and other subtle differences between both fields and illustrate them with hypothetical examples. We expect that clarification of these differences will improve the multidisciplinary collaboration between epidemiologists and economists. Furthermore, we hope to empower researchers to select the most suitable analytical technique from either field for the research problem at hand.

Keywords Selection bias · Confounding · Omitted variable bias · Endogeneity · Epidemiology · Health economics

1 Introduction

Causal inference has always been an important topic in health research (Bach 2019; Thaul et al. 1994). In essence, causal inference aims to explain the effect of the occurrence(s) of an event (often a health intervention) on an outcome of interest (Grobbee and Hoes 2014; Rothman and Greenland 2005). Experimental designs have traditionally been considered the "gold standard" for establishing causality (Heckman 2008). Consequently, health intervention research traditionally relied on randomized controlled trials (RCTs) as the most optimal design for studying the causal effects of an intervention (van Leeuwen et al. 2016; Bouter et al. 2017). Although RCTs can be considered the gold standard for establishing

AN Varga and AE Guevara Morel have contributed equally to this work.

✉ AN Varga
a.n.varga@vu.nl

Extended author information available on the last page of the article

causality, they also have some limitations. Therefore, RCTs may not always be the most optimal study design for answering research questions in the fields of epidemiology or economics. First, the possibility of conducting a high-quality, well-powered RCT is bound by financial constraints, because of their relatively high costs. If an RCT is conducted, its sample size is relatively small compared to observational studies[1]. Second, the generalizability of RCT results may be limited due to strict inclusion and exclusion criteria resulting in a group of individuals that is not (fully) representative of the population of interest (van Leeuwen et al. 2016). Amongst others, some groups (e.g., elderly, lower educated, female, or individuals with many comorbidities or co-medications) are less likely to participate in scientific research than others and are therefore often underrepresented in RCTs (Franklin et al. 2017; van Leeuwen et al. 2016). Third, RCTs may not follow up participants for sufficiently long periods of time (e.g., longer than 12 months). Because many outcomes do not occur directly after receiving an intervention but later in time, longer follow-up periods are essential for being able to observe these outcomes (Sanson-Fisher et al. 2007). In some situations, randomizing individuals to an intervention or control condition may even be impossible or unethical. For instance, randomization may not be possible when one is interested in the effect of adverse conditions very early in life on later-life health and mortality, because it is unethical to intentionally expose individuals to adverse conditions (Lindeboom et al. 2010). Also, randomization may not be possible when the effect of a widely implemented cancer treatment is evaluated, because withholding treatment is considered unethical. In these situations, epidemiologists and economists have to rely on observational data. Observational data suffers from these limitations to a lesser extent, although the risk of bias is increased as compared to an RCT.

There are also situations in which both epidemiologists and economists prefer the use of routinely collected data over RCTs. For example, when evaluating the effectiveness and/or cost-effectiveness of a policy measure (e.g., a region-wide or nationwide public health measure) to improve health outcomes, the use of routinely collected data may be preferred. The increasing availability of routinely collected electronic data (e.g., electronic health records) has raised the interest of both epidemiologists and economists in using this type of data for causal inference. Although routinely collected data are often not of good quality or complete, and can also be selective as the result of a choice for a specific data collecting source, they also have important advantages. Amongst others, routinely collected data are typically readily available, often have larger sample sizes, have long follow-up periods, and can be obtained at relatively low-cost (Ho et al. 2020; Cave et al. 2019; Tugwell et al. 2013).

One of the biggest challenges in assessing causal effects when using routinely collected data is confounding or selection (Grimes and Schulz 2002). An example of the latter is that a patient's choice to participate in a health intervention under study may not be independent of their health status. If, for instance, healthier (or sicker) individuals are more inclined to participate in the intervention than in the control group, the true effect of the intervention will be over- or underestimated. In these situations, researchers try to mimic an experimental design as closely as possible and use statistical techniques to assess the causal relationship.

Epidemiologists and economists also share a strong interest in methods to reduce bias when making causal inferences in routinely collected data. Still, both fields have different

[1] In epidemiology referred to as non-experimental studies.

habits and sometimes encounter different circumstances in which causal inference research is conducted. This has led to the development and use of (seemingly) different approaches within the two fields. Although some practices overlap, many of the practices employed by the two fields are complementary to each other. Hence, both fields can benefit by learning from each other. However, to facilitate an exchange of methods between both fields, it is essential that epidemiologists and economists understand each other's language, particularly their language concerning bias.

Epidemiologists define bias as an error in the study design (e.g., in the process of data collection, analysis, interpretation, reporting, publication) that leads to systematically different results or conclusions, and remains even in an infinitely large sample (Rothman 2012; Porta 2014). Similarly, economists refer to bias when the value of the parameter being estimated (a property of the population) and the expectation of the estimator[2] differ ($E[\hat{\beta}] \neq \beta$) (Wooldridge 2009). Although epidemiologists and economists often talk about the same types of biases with the same underlying mechanisms, the terms they use differ. As a result, the discussion of bias between the two fields is clouded, which might in turn lead to confusion.

Hernán (2017) and Haneuse (2016) previously reported that there are differences between epidemiology and economics in the way bias is defined. For example, Haneuse (2016) noted that bias occurring due to failure to adjust for the impact of an explanatory variable (i.e., independent variable) associated with both the treatment and the outcome (i.e., dependent variable) is labeled as confounding by epidemiologists, but in specific situations as selection bias by economists. The use of such field-specific terminology without providing explicit definitions may lead to confusion in interdisciplinary research groups (e.g., including both epidemiologists and economists). This, may also result in sub-optimal use of statistical methods from the two disciplines for causal inference in studies using observational data. Finally, this language barrier may limit comprehension and adequate use of research papers from each other's discipline.

So far, papers addressing the confusion in terminology between epidemiologists and economists either alert researchers to the differences between the fields, such as Hernán (2017), Haneuse (2016) and Vigneri et al. (2018), or provide a glossary of terms, such as Gunasekara et al. (2008) and the Catalogue of bias collaboration, Aronson JK, Bankhead C, Nunan D. (2018). However, papers that aim to bridge the terminology gap between epidemiology and economics are missing.

With this paper, we aim to address this gap by providing an overview of differences and similarities in how epidemiologists and economists describe and define different sources of bias. Our paper is structured as follows. Section 2 describes the concepts of bias from the perspective of epidemiologists and economists and identifies a common ground for the concepts of bias between both fields. Although an in-depth discussion of the methodology for resolving different biases is beyond the scope of the paper, commonly used methods are briefly mentioned in this section for illustrative purposes. Section 3 illustrates the similarities and differences in terminology between the two fields using a hypothetical example. Finally, Sect. 4 discusses implications for researchers and proposes a classification scheme of bias that can benefit multidisciplinary research groups as well as epidemiologists and economists.

[2] An estimator is a rule for calculating a quantity of interest. For example, a natural estimator of a population with mean μ would be the average of a random sample drawn from the population distribution (Wooldridge 2009).

Table 1 Summary table of the field-specific definitions in epidemiology

Epidemiology	
Bias (systematic error)	Bias is an error in the study design (e.g., in the process of data collection, analysis, interpretation, reporting, publication) that leads to systematically different results or conclusions, and remains even in an infinitely large sample (Rothman 2012; Porta 2014).
Confounding	Confounding occurs when the effect of an intervention on the outcome is mixed with the effect of another variable (Grimes and Schulz 2002; Rothman 2012; Catalogue of bias collaboration, Aronson JK, Bankhead C, Nunan D. 2018). The association between an intervention and outcome is distorted, because the third variable is independently associated with both the intervention and the outcome (Catalogue of bias collaboration, Aronson JK, Bankhead C, Nunan D. 2018). This is often referred to as **measured confounding**.
Residual confounding	Confounding caused by unobserved explanatory variable(s) that consequently cannot be successfully removed by restriction or adjusted for in the data-analysis, thereby leading to residual (uncontrolled) confounding (Rothman 2012; Porta 2014). This is often referred to as **unmeasured confounding**.
Confounding by indication	Special form of confounding that occurs when the outcome is caused by the indication for an intervention (i.e., exposure to a risk factor of a treatment) (Catalogue of bias collaboration, Aronson JK, Bankhead C, Mahtani KR, Nunan D. 2018; Miettinen and Cook 1981). Thus, individuals who are (self-) selected to get the intervention are typically different from those who do not (Rothman 2012; Ahlbom 2021).
Selection bias	Selection bias occurs when the association between intervention and outcome differs between participants and non-participants of the study as a consequence of using non-randomly selected samples (Rothman 2012; Ertefaie et al. 2015).

2 The terminology conundrum

In this section, we will define the different terms that are used in the fields of **epidemiology**[3] and *economics*[4] to describe the different types of bias. The definitions of different field-specific types of bias distinguished in epidemiology and economics are summarized in Tables 1 and 2, respectively.

2.1 Internal and external validity

Generally, assessing causal relations by studying all individuals in a population is not possible. Hence, researchers select subsets of the total population. Such subsets are often referred to as the **study population** by epidemiologists, whereas economists use the term *sample*. The *sample*/**study population** is defined by a set of in- and exclusion criteria that are applied to a **source population** (Grobbee and Hoes 2014; Bouter et al. 2017) (see Fig. 1), which is referred to as *population of interest* by economists. It is assumed that the *sample*/**study population** consists of a representative (i.e., random) selection of eligible individuals from the *population of interest*/**source population**, and therefore that the

[3] Epidemiology terms are written in **bold**.

[4] Economics terms are written in *italics*.

Table 2 Summary table of the field-specific definitions in economics

Economics	
Bias	Bias occurs when the value of the parameter being estimated (a property of the population) and the expectation of the estimator differ ($E[\hat{\beta}] \neq \beta$) (Wooldridge 2009).
Endogeneity	Endogeneity arises when the explanatory variables are correlated with the error term (Cameron and Trivedi 2005; Zohoori and Savitz 1997). Sources of endogeneity can be omitted confounding variables, simultaneity between a predictor and the outcome, and errors in regression covariates (Terza et al. 2008).
Measurement error	Measurement error occurs when an imprecise measure of a variable is included in the regression model (Wooldridge 2009).
Omitted variable bias	Omission of a variable whose coefficient is non-zero leads to a biased estimator. The asymptotic bias from omitting covariates is shown to be zero if the regression of the response variable on treatment and covariates is linear or exponential unless that relevant variable is correlated with both the treatment and outcome of interest (Hill et al. 2011; Gail et al. 1984).
Selection bias	Selection bias arises when a process other than random sampling is used to sample the underlying population. This non-random sampling can occur due to distorted sampling rules of the researcher or altered behavior of the patients, such as self-selection to the treatment, non-response to surveys, or attrition from treatment (Heckman 2010; Wooldridge 2001).
i. Sample selection bias	The sample selection bias arises due to selection on the outcome either directly or through the error term (Crown et al. 2011; Heckman 1979; Wooldridge 2009). This can also be referred to as endogenous sample selection.
ii. Treatment selection bias	Treatment selection bias is a result of systematic differences between treated and non-treated populations, which occurs because treatment selection is influenced by patient and provider characteristics (Hill et al. 2011; Austin 2011b, 2009; Austin and Stuart 2015).
iii. Self-selection bias	The treatment uptake is not random, individuals who opt for the treatment are different from the non-treated populations (Hill et al. 2011).

associations derived from the *sample*/**study population** can be generalized to the *population of interest*/**source population** (Grobbee and Hoes 2014; Bouter et al. 2017).

Before estimates based on observational data can be causally interpreted, there are general identifying assumptions that need to be met to ensure internal validity of causal effect estimates. Often, assumptions for causal inference are defined according to the Neyman-Rubin counterfactual framework (Sekhon 2008). In this framework, a causal effect is defined as the difference between the two potential outcomes that an individual can have in the two conditional states of receiving and not receiving the treatment, while keeping everything else equal (Sekhon 2008). However, in reality only one of the two potential outcomes is observed (Sekhon 2008). A recapitulation of the assumptions in this framework is listed in Appendix 1. When these assumptions hold, internally valid effects can be estimated and thus it is possible to obtain an asymptotically unbiased estimator of the treatment effect in the *sample*/**study population**.

On the other hand, external validity cannot be guaranteed unless a sample is randomly drawn from the *population of interest*/**source population** (Lesko et al. 2017). When

external validity is compromised, this generally[5] implies that the findings are not generalizable to the *population of interest*/**source population**. A recapitulation of the assumptions formulated by Lesko et al. required for external validity of the estimates is listed in Appendix 1. Knowledge of these assumptions is required for understanding the source of biases.

2.2 The general concept of bias: biased versus unbiased estimator

In order to define the terms used to describe the different types of bias, we will assume that the following linear regression model describes the underlying mechanism of the association of interest within the entire *population of interest*/**source population**. In economics, this equation is known as the *population equation*, whereas in epidemiology there is no equivalent term. For expositional purposes we use a linear model that describes the underlying mechanism of the *population of interest*/**source population**,

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni} + \varepsilon_i \tag{1}$$

where y_i represents the outcome for the individual i, $x_1..x_n$ represent the explanatory variables, and ε stands for the unobservable random error term. $\beta_1..\beta_k$ are the coefficients for the explanatory variables $x_1..x_n$.

When the *sample*/**study population** is representative of the *population of interest*/**source population**, estimation of the linear regression model described in Eq. (1) in the *sample*/**study population** results in parameter estimates that on average coincide with the *population equation*.

In epidemiology, **bias** can result from the way in which individuals are selected (**selection bias**), the way in which the variables are measured (**information bias**/*measurement error*), or a failure to control for the impact of explanatory variables (**confounding**) (Bouter et al. 2017; Grimes and Schulz 2002; Rothman 2012). Similarly, economists refer to bias when the value of the parameter being estimated (a property of the population) and the expectation of the estimator differ ($E[\hat{\beta}] \neq \beta$) (Wooldridge 2009). For an estimator $\hat{\beta}$ to be unbiased, it is required that $E[\hat{\beta}] = \beta$, meaning that the expected value of estimator is equal to the parameter value β. In economic terms, the latter holds if the following conditional expectation of ε_i given $x_{1i}...x_{ni}$:

$$E(\varepsilon_i \mid x_{1i}...x_{ni}) = E(\varepsilon_i \mid X_i) = 0. \tag{2}$$

Economists also refer to this as the *zero conditional mean assumption* (Wooldridge 2010). Simply put, this means that if the *sample*/**study population** is randomly selected from the *population of interest*/**source population**, the error term has a mean of zero and is uncorrelated with each of the explanatory variables in the model. If this assumption holds, then each explanatory variable is necessarily *exogenous* (Wooldridge 2010). That is, the variable is not influenced by other variables in the association.

The *zero conditional mean assumption* is violated when an included regressor is *endogenous*, meaning that it is dependent on the error term (ε_i). This phenomenon is referred to as *endogeneity* by economists (see also Table 2). *Endogeneity* often occurs as a result of self-selection of individuals (Wooldridge 2010). To complicate matters even further, in economics the term *endogeneity* is often also used as an umbrella term for various different

[5] External validity can englobe both generalizability and transportability problems (Lesko et al 2017).

problems that cause a violation of the *zero conditional mean assumption*: *omitted variable bias*, *simultaneity* or *measurement error* (Wooldridge 2009).

2.3 Confounding versus selection bias

In epidemiology, **confounding** (Table 1) implies that the effect of x_i on y_i is mixed with the effect of a third factor, also known as the **confounder** (Ahlbom 2021). When this **confounder** is not included in the regression model, this leads to bias (Grimes and Schulz 2002; Rothman 2012). Thus in the terminology used in economics—the *zero conditional mean assumption* is violated due to an *omitted variable*. It is important to keep in mind, that there can be more than one **confounder** that can be either observed or unobserved. In order for a **confounder** to distort the effect of x_i on y_i, it should be associated with both y_i (as a cause or proxy of the cause) and x_i. However, the **confounder** should be correlated with x_i, but should not be an effect of x_i (Rothman 2012). When **confounding** is present, the validity[6] of a study is compromised, because the estimated association does not reflect the true relationship between x_i and y_i (Bouter et al. 2017; Grimes and Schulz 2002; Rothman 2012; Miettinen and Cook 1981). Within the concept of **confounding**, a distinction is often made between **measured confounding** or **unmeasured confounding** (Ahlbom 2021). **Measured confounding** is defined as confounding resulting from variables that are observed and measured. However, when **confounding** is not successfully removed or corrected for, or results from a variable that is not observed, this leads to uncontrolled or **residual confounding**. When **residual confounding** is caused by a failure to observe the confounder, we will refer to the resulting bias as **unmeasured confounding**. Economics does not have a term to describe the overall concept of **confounding**, but it does have equivalent terms for **measured confounding** and **unmeasured confounding**, which will be explained in further detail in the next sections.

The epidemiological concept of **measured confounding** is present when an unbiased estimator of the treatment effect cannot be obtained by directly comparing outcomes between treatment groups, due to the presence of an observed **confounder**, and a correct specification of the outcome model can resolve this problem (Austin 2011a). Economists would then say that the model coincides with the *population equation*, and hence the estimator is unbiased. However, researchers can never be sure that the specified model is equal to the *data generating process* of the population, because the population equation is unknown. Let us assume that the data generating process of the population is described by the following *population equation*:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_1 x_{1i} * \beta_2 x_{2i} + \beta_2 x_{2i}^2 + \beta_3 x_{3i} + \varepsilon_i. \tag{3}$$

If one wrongly models the association with an ordinary least squares (OLS) according to Eq. (1) instead of the *true population equation* (3), the ***interaction term*** $(\beta_1 x_{1i} * \beta_2 x_{2i})$[7] and the quadratic term $(\beta_2 x_{2i}^2)$ become part of the error term and thus induce *endogeneity*. On the other hand, if the omitted variable is independent from a specific x of interest, the estimator of the effect of that specific x on y remains unbiased and only the standard errors will be compromised. In other words, not taking the interaction and the quadratic

[6] Referring specifically to internal validity.

[7] Epidemiologists may also refer to this as effect modification, which implies that $x_1 i$ has not one effect on y_i but two: one when $x_2 i$ is present and one when it is absent.

terms into account leads to bias. This is true because omitting a relevant term from the regression model results in a correlation between the error term (ε_i) and the explanatory variables $\beta_1 x_{1i}$ and $\beta_2 x_{2i}$. As a result, the assumption of conditional independence of the error term no longer holds and the estimator is ***biased***. This refers to the assumption that distinguishes **measured** from **unmeasured confounding**. The assumption of no **unmeasured confounding** is sometimes also referred to as *selection on observables*, *exogeneity*, *conditional independence assumption* or *ignorability* in economics. When the assumption of *selection on observables* holds, correct inferences for causal parameters can be obtained by using methods such as regression-adjustment, matching, re-weighting, and the doubly-robust estimator (Cerulli 2015).

As indicated above, the epidemiological term **unmeasured confounding** refers to **confounding** that is the result of a **confounder** that was not or poorly measured, and therefore, not taken into account in the data analysis. In some situations, this type of **confounding** can be considered to be equivalent to the economics term *selection bias*, which arises due to incomplete observation of the population. Economists often use instrumental variable analysis (Angrist and Pischke 2009) and epidemiologists use its analogue, Mendelian randomization in genetic research, to adjust for this type of bias (Streeter et al. 2017).

In economics, *selection bias* can take up various forms. For instance, *self-selection bias* might occur because participation is not randomly determined, thus the selection occurs based on an explanatory variable (x). The term *self-selection bias* is generally used when an indicator of participation might be systematically related to unobserved factors (Wooldridge 2009). In the following example, we represent the situation where the treatment indicator is related to the wage variable (i.e., a person's monthly salary). Let us assume that we observe only individuals with a wage below 2500 ($x_{1i} < 2500$). Wage x_{1i} is an explanatory variable of the outcome y_i. This implies that the error term becomes the following:

$$\eta_i = \varepsilon_i + \beta_1(x_{1i} \geq 2500), \tag{4}$$

which consists of a random error as well as the unobserved part of the population (i.e., individuals with a wage above 2500 do not occur in our sample). The conditional expectation of interest is the following:

$$\begin{aligned} &E(y_i \mid x_{1i} < 2500, ..., x_{ki}, ..., x_{ni}) \\ &\quad = \beta_0 + \beta_1(x_{1i} < 2500) + ...\beta_n x_{ki} + ... + \beta_k x_{ni} \\ &\qquad + E(\eta_i \mid x_{1i} < 2500, ..., x_{ki}, ..., x_{ni}). \end{aligned} \tag{5}$$

If

$$E(\eta_i \mid x_{1i} < 2500, ..., x_{ki}, ..., x_{ni}) \neq 0, \tag{6}$$

we are facing *selection bias*. If the non-observed characteristics are correlated with any other observed term, then

$$corr(x_{1i} \geq 2500, x_{1i} < 2500, ..., x_{ki}, ..., x_{ni}) \neq 0.$$

Ergo, if selection is not random, that is, either influenced by the individual or by the sampling researcher, conditional independence does not hold. If the entire variable is unobserved, *selection bias* is said to be equal to *omitted variable bias*. Hence, in economics the distinction between *self-selection bias* and *omitted variable bias* is based on the degree of observation of the confounder variable. However, in epidemiology this type of bias is referred to as **confounding** (Ahlbom 2021). Instrumental variable analysis can be used to

adjust for both *self-selection bias* and *omitted variable bias*, when a valid instrument is available (Angrist and Pischke 2009). Since **unmeasured confounding** cannot be easily resolved with any available data, economists frequently use quasi-experimental designs, such as difference-in-difference (Wing et al. 2018) or regression discontinuity analyses (Robin et al. 2012), to deal with this.

To increase confusion even more, epidemiologists also use the term **selection bias**, but their definition is not necessarily equivalent to the aforementioned economic definition of *selection bias*. In epidemiology, **selection bias** occurs due to the procedures or processes used to select the *sample*/**study population** in observational studies (Table 1) (Bouter et al. 2017; Rothman 2012; Hernan and Robins 2019). **Selection bias** is present when the association between x_1 and y_i differs between the *sample*/**study population** and *population of interest*/**source population** (Grimes and Schulz 2002; Rothman 2012). In these cases, the magnitude and direction of the bias are difficult to determine (Ertefaie et al. 2015). As a result, the study's external validity is compromised because the identified association cannot be generalized to the *population of interest*/**source population**. This is in line with another form of *selection bias* in economics, which can arise due to *endogenous sample selection*, and is referred to by some researchers as *sample selection bias*. The latter implies that the non-random sample selection from the population occurs based on the outcome variable y. For instance, if we intend to estimate the relationship between frailty (y) and several other factors in the population of adults:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i. \tag{7}$$

Sample selection bias occurs in this case if there is selective attrition from the *panels survey*/**cohort study**, meaning that those who remain in the *sample*/**study population** have on average better (or worse) outcomes. In our case, those who are frail may not continue to participate in the study, and therefore, the resulting *sample*/**study population** is not random anymore but rather a selective subset of the population. This will as well result in biased and inconsistent estimator, due to the fact that the *population equation* is not in line with the expected value conditional on the outcome being less than a cutoff value (e.g., given frailty scale 0-5 $E(y_i > 3 \mid x_{1i}, x_{2i}) \neq E(y_i \mid x_{1i}, x_{2i})$).

Thus, in economics the term *selection bias* incorporates different forms of *selection bias* (e.g., *sample selection bias*, *treatment selection bias*, and *self-selection bias*), but these specific terms are not used as often as their overarching term *selection bias*. This implies that the definition of *selection bias* in economics is broader than in epidemiology. The epidemiological concept **selection bias** is equivalent to the economic concept *sample-selection* bias, which occurs due to *endogenous sample selection*. *Treatment selection bias* and *self-selection bias* on the other hand occur due to *endogenous* treatment allocation. The economic terms *treatment selection bias* and *self-selection bias*, which refer to non-random treatment uptake and individual self-selection to treatment, respectively, encompass the epidemiological term **confounding by indication**. **Confounding by indication** does not have one equivalent term in economics. While **selection bias** defined by an epidemiologist will be understood by an economist, the reverse might not be true and can lead to confusion.

The epidemiological concept **confounding by indication** (Table 1) is a special form of **confounding** that occurs when y_i is causally related to the indication for x_1 (Catalogue of bias collaboration, Aronson JK, Bankhead C, Mahtani KR, Nunan D. 2018; Miettinen and Cook 1981). In other words, individuals in the intervention group are different from those in the control group, based on an underlying factor(s) that influenced their choice

for the intervention (Rothman 2012). Randomization is the best way to ensure the prevention of **confounding by indication** (Bhide et al. 2018). However, when using non-experimental data, the decision to allocate or start an intervention (i.e., x_1), may be influenced by a wide variety of underlying factors (e.g., therapist or patient preferences, the severity of the disease, prognosis, availability) (Grobbee and Hoes 2014). If these underlying factors are positively or negatively associated with the outcome, **confounding by indication** is present. As a consequence, the validity of the study is compromised. It is important to note that **confounding by indication** is in fact a form of **selection bias** that cannot be fully adjusted for, because factors that drive the choice for an intervention are often not completely known or difficult to measure. This implies that there will be a substantial amount of **residual confounding**.

In epidemiology, **time-varying confounding** is said to occur when confounders have values that change over time. Examples of time-varying confounders can be labor market status, body mass index, and depression severity. **Time-varying confounding** can also occur with changes in a time-varying intervention (i.e., an intervention that is not fixed in time), like, for example, a treatment dose (Platt et al. 2009). This type of **confounding** can be resolved by using marginal structural models, g-computation, targeted maximum likelihood estimation, or G-estimation of structural nested models (Clare et al. 2019). In economics, this phenomenon is another example of the violation of the zero conditional mean assumption in a longitudinal context and it is categorized as an *endogeneity* problem. Most frequently, it is dealt with by applying instrumental variable techniques or fixed effects.

Finally, the term **collider bias** can create confusion, especially when compared to the term **confounding**. In **collider bias**, similar to **confounding**, the effect of x_i on y_i is distorted. The difference lies in the fact that in the case of **collider bias** both x_i and y_i independently cause a third factor, also known as the **collider** (i.e., a collider is a variable that is caused by two other variables: one that is (or is associated with) the treatment, and another one that is (or is associated with) the outcome) (Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D. 2019; Griffith et al. 2020; Elwert and Winship 2014). When **confounding** is present, the **confounder** variable is associated with both x_i and y_i, but x_i and y_i do not independently cause the confounder (like in the case of a **collider**). When the **collider** is not adjusted for (either in the study design or in the statistical analysis), it may influence the likelihood of being selected into a study, leading to bias (Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D. 2019; Griffith et al. 2020). Thus, in some cases **selection bias**, can be considered a type of **collider bias** (Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D. 2019). This is because just like selection bias, **collider bias** stems from conditioning (e.g., controlling, stratifying, or selecting) on the collider variable. In economics, *endogenous selection bias* is equivalent to **collider bias** (Elwert and Winship 2014).

2.4 Terms in a nutshell

In order to have a clear panorama, we will summarize the equivalent concepts of bias between epidemiology and economics. In Table 3, we summarize the field-specific terms with their proposed equivalents. Furthermore, Fig. 2, maps the economic terms for bias on epidemiological terms.

The economic term *endogeneity* implies a correlation between the error term and the explanatory variables, and essentially indicates the violation of the zero conditional mean assumption. The term has no exact equivalent in epidemiology. According to

Table 3 *Field-specific terms with proposed equivalents*. The equivalent terms that have exact equivalents are displayed in parallel lines. Terms that have no exact equivalents are not part of the table

Epidemiology	Economics
Bias	Bias
–	Endogeneity
Unmeasured confounding	Self- or treatment- selection bias (if the variable partially unobserved)/ Omitted variable bias (if variable fully unobserved)
Measured confounding	Selection on observables
Information bias	Measurement error
Selection bias	Sample selection bias or Endogenous sample selection
Confounding by indication	Treatment selection bias
	Self-selection bias

epidemiologists, **confounding** occurs at the intervention uptake level, and can occur when intervention allocation is not random. Economists use the term *selection on observables* when referring to the assumption of no **unmeasured confounding**, which is the key assumption for inference in presence of **measured confounding** in epidemiology. **Unmeasured confounding** and *omitted variable bias* represent the same phenomenon[8] which occurs when intervention uptake is associated with unobserved characteristics. *Omitted variable bias* (if the variable is fully unobserved) and *self-selection bias* (if the variable is partially unobserved) **/ unmeasured confounding** can be considered as a subcategory of *endogeneity*. **Unmeasured confounding** and *omitted variable bias* both arise in the phase when the ***sample/* study population** is selected by the researcher or when the data collection is carried out (Fig. 2). Sampling from the population using a pre-defined set of inclusion and exclusion criteria will result in bias if the sample selection is related to characteristics that are associated with the outcome and the intervention allocation, i.e., *sample selection bias* or *endogenous sample selection* according to economists or **selection bias** according to epidemiologists (Fig. 1). Thus, in this case the terms are equivalent between fields. Economists, however, in general do not distinguish whether the bias occurs at sample selection or intervention uptake level, and use the overarching term of *selection bias* for all scenarios where an explanatory variable is related to both intervention or outcome (Fig. 2). The economics definition of *selection bias* can, thus, be equated with the term **confounding** used in epidemiology, when it refers to *treatment-selection bias* or *self-selection bias*. However, the terms *treatment selection bias* and *self-selection bias* are less often used than the overarching term *selection bias*.

3 Examples

In this section, we will illustrate how the different terms that were discussed in Sect. 2 are used in the fields of **epidemiology** and *economics*. This will be done using a simplified, hypothetical example. We refer to Tables 1 and 2 for an overview of the definitions

[8] If the regression of the response variable on treatment and covariates is linear or exponential.

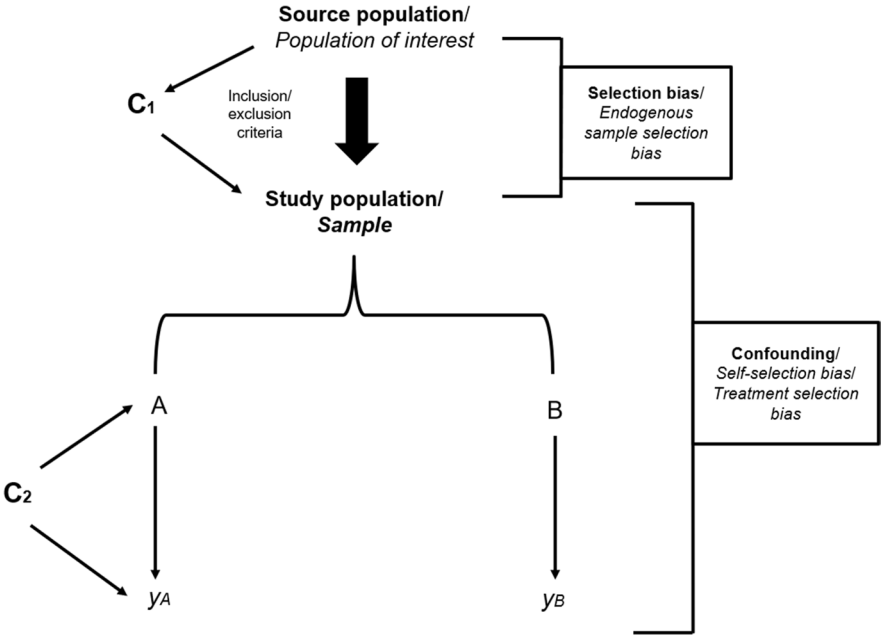

Fig. 1 Graphical representation of confounding and selection bias. **Source population/***Population of interest*: group that we would like to make inference about. **Study population/*Sample***: group we selected with inclusion/exclusion criteria from the **source population/***population of interest*. A: participants with intervention A. B: participants with intervention B. y_A: outcome of participants with A. y_B: outcome of participants with B. C1: pre-intervention factors that lead to **selection bias/** *endogenous sample selection bias*. C2: confounding factor. **Bold** terms refer to **epidemiology** terms and *italic* terms refer to *economics* terms

of different forms of bias as used by epidemiologists and economists respectively. Figure 1 gives a visual representation of the different forms of bias.

In this hypothetical example, the Bias National Health Institute was employed to evaluate the effect of a smoking cessation program (x_i) offered by general practices (GPs) on tobacco smoking cessation (y_i). To assess the effect of x_i on y_i, the research team retrospectively constructed a *sample*/**study population** from electronic health records by arbitrarily selecting a number of GPs that offered the smoking cessation program (intervention practices) and a number of GPs that did not offer the smoking cessation program (control practices). The *sample*/**study population** included individuals who smoked tobacco, were eligible for the smoking cessation program, and were registered at a participating general practice. An individual was assumed to be treated if they participated for at least one month in the smoking cessation program. Smoking cessation was a self-reported binary outcome, which was operationalized as non-smoking for at least three consequent months. Information on demographics (e.g., date of birth, gender, socioeconomic status), self-reported tobacco smoking cessation (yes/no), and additional treatments related to tobacco smoking cessation was extracted from electronic health records.

In a perfect scenario, the estimator would be unbiased if the following requirements hold (also see Appendix 1):

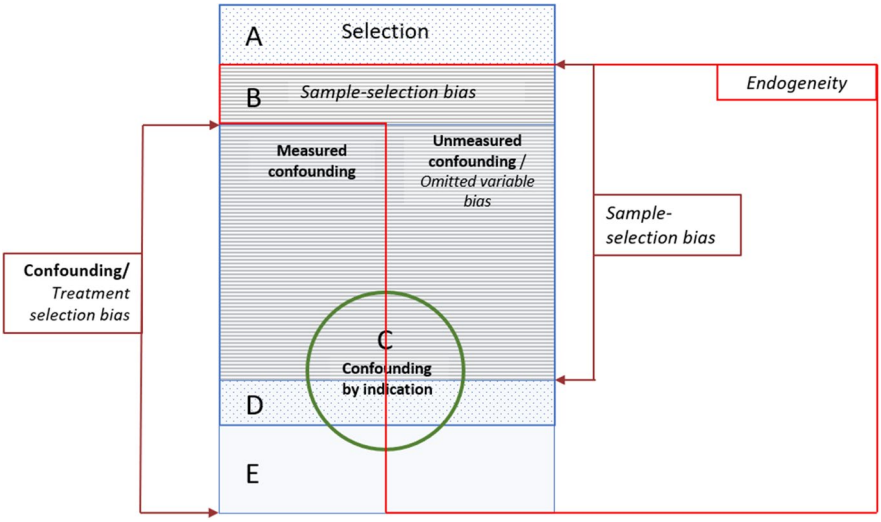

Fig. 2 Proposed mapping of the economics terms of bias on those of epidemiology. A: Selection only, no bias due to selection; B: Selection with *sample-selection bias*; C: Sample-selection with **confounding** and **confounding by indication**; D: Selection with **confounding** without *sample-selection bias*; E: **Confounding** without *selection* or *sample-selection bias*. **Bold** terms refer to **epidemiology** terms and *italic* terms refer to *economics* terms

- The *sample*/ **study population** is representative of the *population of interest*/ **source population**;
- The actual intervention and control groups are representative for the *population of interest*/ **source population** and the distributions of the explanatory variables are comparable in the two groups;
- Data is available on the outcome of interest as well as for every relevant explanatory variable;
- The data is modelled in accordance with the data-generating process, which implies that the analysis model is specified correctly;
- Intervention assignment is conditionally independent of the individual characteristics.

Below, we will discuss different scenarios in which at least one of these requirements fails, leading to imprecise and/or biased estimator. Appendix 2 illustrates the different described scenarios by means of simplified simulation examples.

Scenario 1: The smoking cessation program was offered during working hours on weekdays only. This will introduce ***selection bias***/*endogenous sample selection*, if individuals who attended the smoking cessation program were more likely to be individuals who did not have a nine to five job, for example because they were retired. This scenario would be labeled as ***selection bias*** by both epidemiologists and economists (referring to the term *endogenous sample selection* or *sample selection bias*). It is important to note that estimator on the sample level will not be biased as long as not having a nine to five job is not related to the outcome. However, the *sample*/**study population** is not representative of the *population of interest*/**source population**, hence the external validity is compromised and the results are not generalizable.

Scenario 2: It is possible that some individuals who would otherwise participate in the smoking cessation program are affected by negative personal circumstances (e.g., divorce, death of a close friend or family member) and are therefore less likely to enroll. These negative events are likely to also have an impact on the likelihood of success in smoking cessation. Comparison of outcomes of individuals participating in the smoking cessation program only will lead to biased estimator, because we are excluding a large group of individuals based on reasons that the researchers are not aware of. In this case, *self-selection bias* will occur and the success of the smoking cessation program will be overestimated (upward biased). In most cases, self-selection is driven by unobserved characteristics of the individual or healthcare provider, such as motivation or beliefs, which are characteristics that are not measured or cannot be measured well. Therefore, this kind of bias is referred to as **unmeasured confounding** and *omitted variable bias* in epidemiology and economics, respectively.

Scenario 3: In some intervention practices, nurses routinely prescribe the use of nicotine patches as a support to the smoking cessation program. Nicotine patches are prescribed to decrease the smoking cessation withdrawal symptoms, and consequently support individuals to continue the smoking cessation program while also increasing the probability of successfully stopping smoking (the primary outcome of the study). The use of nicotine patches was not a mandatory part of the program. However, since the use of nicotine patches affects both the participation in the smoking cessation program (the intervention) and smoking cessation (the outcome), it is necessary to control for nicotine patch use in order to obtain an unbiased effect of the smoking cessation program. Epidemiologists refer to this problem as **measured confounding** and in economics this phenomenon is referred to as *selection on observables*, when the confounder is observed and can be adjusted for.

4 Conclusions

Epidemiologists have a long tradition of using observational data. However, their interest in using routinely collected data for causality research when RCTs cannot or have not been done has increased over the past years (Bartlett et al. 2019). Economists have traditionally relied more on observational data and therefore have a vast experience in inferences based on routinely collected data. However, data collected within routine health care are typically not collected for research purposes, and therefore, carry a high risk of bias if researchers fail to use appropriate analytical methods to deal with bias (Nørgaard et al. 2017). Methods that are used to obtain unbiased estimator when using routinely collected data differ between epidemiologists and economists. Nevertheless, some practices overlap or complement each other. Hence, epidemiology and economics can benefit from exchange of methods to adjust for potential biases.

The way in which bias is classified does not necessarily have implications for the data-analysis (Hernan and Robins 2019), which is the case for confounding. However, in some situations, due to lack of a clear distinction between the different definitions of selection bias, it is more difficult to make an informed choice about which study design and analytical method to use, as the term can refer to both internal and external validity issues. This is because epidemiologists define **selection bias** as a generalizability problem, whereas in economics *selection bias* is a broader term that can indicate either an internal or external validity problem. This suggests that it would be beneficial for researchers from epidemiology and economics to understand the terminology used in each other's fields. For instance,

in this paper we show that **selection bias/***sample selection bias* and *treatment selection bias/endogenous sample selection* are two separate concepts that should be treated differently from each other, because they differ in the mechanism that creates the bias. In the simplified example, **selection bias/***treatment selection bias/endogenous sample selection* could have been prevented by offering the smoking cessation program at any hour and not during working hours only. Therefore, researchers are recommended to investigate which type of bias is most likely to be present in their data, so they can subsequently choose an analytical method to validly account for the bias.

Economists tend to prioritize methods for unmeasured confounding (e.g., instrumental variable methods) or a quasi-experimental setup (e.g., difference in differences, regression discontinuity). On the other hand, epidemiologists focus more on methods to deal with measured confounding (e.g., propensity score methods), while assuming that there is no unmeasured confounding. At the same time, there is also overlap between methods from the two fields. To be able to profit optimally from the range of methods in the fields of epidemiology and economics, it is important to have a correct understanding of the terminology used. In all situations, it is important to realize that without randomization of interventions, the interpretation of the "treatment effect" as being causal is up to the decision of the researcher and comes with the burden of proof. Strictly speaking, when the intervention assignment is non-random, a causal relation cannot be definitely established. However, it is possible to provide support credibility of a causal relationship between the intervention and the outcomes by adjusting for bias as much as possible (DeMaris 2014).

In conclusion, this paper provided an overview of differences and similarities in how epidemiologists and economists define bias to improve understanding of each other's definitions. This information will hopefully improve collaboration and support researchers in identifying the most suitable study designs and analytical methods from either field for the research question that is being dealt with.

Appendix 1

An estimator is internally valid when the estimator of sample average treatment effects is unbiased. The following assumptions are required for causal identification of treatment effects within the sample:

Recap of assumptions of Rubin's potential outcomes framework for causal inference (Neyman-Rubin Counterfactual Framework)

- **Positivity**: Each individual has a non-zero probability of being assigned to either treatment or control.
- **Strongly Ignorable Treatment Assumption (SITA)/exchangeability/no unmeasured confounders**: All relevant confounders are measured and adjusted for, which implies that the treatment estimate is independent of the error term. This assumption is also known as the combination of uncounfoundedness with overlap (Brady et al. 2008).
- **Stable Unit Treatment Value Assumption (SUTVA)**: Potential outcomes of an individual are not affected by the treatments assigned to any other individual (e.g., no interference), and there is no variation in the treatment (e.g., only a single version of each treatment level exists) (Constantine Gatsonis 2017) (Brady et al. 2008).
- **Consistency**: This assumption implies that an individual's potential outcome given a hypothetical exposure is equivalent to the outcome that would actually be observed

if this individual would have received the treatment. Rubin's definition of SUTVA includes the no-multiple-versions-of-treatment assumption and, thus, the consistency assumption (VanderWeele and Hernán 2013).

- **Correct model specification**: the covariate-outcome model is correctly specified and the covariates are correctly measured.

An estimator is externally valid when the estimator of average treatment effects of the population is unbiased.

Assumptions required for external validity based on Lesko et al. (2017)

- **Exchangeability**: Participants in the *sample*/**study population** are exchangeable with members of the *population of interest*/**source population**.
- **Positivity**: All individuals of the *population of interest*/**source population** have a non-zero probability of being part of the *sample*/**study population**.
- **Same distribution**: The distribution of treatment in the *sample*/**study population** equals to that of the *population of interest*/**source population**.
- **No interference**: Individuals are not affected by the any other individual's inclusion in the sample.
- **Correct model specification**: the covariate-outcome model is correctly specified.

Appendix 2

Randomized controlled trials (RCTs) have traditionally been considered the gold-standard for establishing causality. To be able to fully understand the different sources of bias in non-randomized data, it is important to first understand how bias may affect estimates when using randomized data, such as in an RCT setting. For the sake of clarity and completeness we will begin by presenting a simplistic simulation of a baseline RCT with the aim of obtaining an unbiased treatment effect. From there on, we will build up and illustrate the different sources of bias described in the example by means of a simplified simulation code that can be run in R Studio.

For expositional purposes, we define the following terms: x_1 is a binary treatment variable representing the participation in the smoking cessation program or not, x_2 is a binary variable indicating whether the individual has a nine to five job, x_3 is a binary variable indicating nicotine patch use and x_4 is a normally distributed variable ($x_4 \sim \mathcal{N}(55, 10)$) representing an unobservable mental health score. Y is the binary outcome of interest, which indicates smoking cessation.

Baseline scenario

In the following baseline scenario, the treatment (x_1) is randomly assigned to study participants allocating them to the treatment or control group, the outcome (Y) is dependent on treatment and there is no model misspecification present. In this scenario, the mean difference between treatment and control is an unbiased estimator of the treatment effect.

```
#Random treatment assignment, correct model specification
library(dplyr)
N=10000
X1=(runif(N)<=.5)+0 #smoking cessation program
X2=(runif(N)<=.2)+0 #having a 9 to 5 job
X3=(runif(N)<=.4)+0 #use of nicotine patches
X4=rnorm(N,55,10) #mental health score
Y=1+1.5*X1 #smoking cessation outcome
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X1==1,]$Y)-mean(dataset[X1==0,]$Y) # no bias
```

In the following extended baseline scenario, the treatment uptake (x_1) is still random. However, the outcome model is not only associated with the treatment status, but also with other variables (i.e., x_2 and x_3). This implies that comparing mean differences would lead to biased treatment effects due to misspecification of the outcome model. However, after adjusting for the observed covariate that is associated with the outcome, the treatment effect estimator becomes unbiased.

```
# Random treatment assignment with model misspecification
X1=(runif(N)<=.5)+0 #smoking cessation program
Y=1+1.2*X2+1.3*X3+1.5*X1 #outcome dependent on X1, X2, and X3
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X1==1,]$Y)-mean(dataset[X1==0,]$Y) #bias
lm(Y~X1+X2+X3) # adjusted model:no bias
```

Scenario 1

In this scenario, the treatment allocation (x_1) is not random anymore but dependent on the value of x_2 (having a nine to five job). This implies that a conditional selection on observables occurs, thus the treated group and the non-treated group differ not only in treatment status but also by a set of observable characteristics (x_2), which can lead to ***selection bias***. However, in this scenario, the outcome model is solely dependent on the treatment status (x_1) and not on having a nine to five job (x_2). Therefore, estimation of a mean difference would lead to an unbiased treatment effect estimator, because the outcome model is correctly specified. In this scenario, there is no **confounding** present, because there is no association between x_2 and Y.

```
psc_logit<-log(0.3)+(0.5*X2) #X1 uptake dependent on X2
p <- 1/(1 + exp(-psc_logit))
X1 <- rbinom(N, size = 1, prob = p)
Y=1+1.5*X1 #outcome is not dependent on X2
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X1==1,]$Y)-mean(dataset[X1==0,]$Y) #no bias
```

Scenario 2

In this scenario, the treatment allocation (x_1) is not random and is dependent on the value of the mental health score (x_4), which is also associated with the outcome. The outcome model is dependent on treatment uptake (x_1), but also on mental health score (x_4). This implies that x_4 becomes a **confounder**, because of its association with both treatment (x_1) and outcome (Y). However, x_4 represents an unobservable score (because it was either not measured or cannot be measured well), thus exclusion of this confounder leads to model misspecification and there is bias present due to **unmeasured confounding**/*endogeneity*.

```
psc_logit<-log(0.3)+(0.01*X4) #treatment uptake dependent on X4
p <- 1/(1 + exp(-psc_logit))
X1 <- rbinom(N, size = 1, prob = p)
Y=1+1.5*X1 #outcome is not dependent on X2
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X1==1,]$Y)-mean(dataset[X1==0,]$Y) #no bias

Y=1+1.2*X2+1.3*X3+1.5*X1 #outcome dependent on X2 and X3
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
#dataset <- dataset[ which(dataset$X2>50), ]
ATEY=mean(dataset[X1==1,]$Y)-mean(dataset[X1==0,]$Y)
#bias larger than in RCT
lm(Y~X1+X2+X3) #no bias
```

Scenario 3

In this scenario, the treatment allocation (x_1) is not random and continuation of the treatment is dependent on the value of x_3 (nicotine patch use). The outcome model is dependent on the treatment status as well as x_3. This implies that x_3 becomes a **confounder**, because of its association with both treatment (x_1) and outcome (Y). Not adjusting for x_3 in the final model, can lead to bias due to **measured confounding**/*selection on observables*.

```
psc_logit<-log(0.3)+(0.5*X3) #treatment uptake dependent on X3
p <- 1/(1 + exp(-psc_logit))
X1 <- rbinom(N, size = 1, prob = p)
Y=1+1.5*X1 #outcome is not dependent on X3
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X3==1,]$Y)-mean(dataset[X3==0,]$Y) #no bias

#dataset <- dataset[ which(dataset$X2>50), ]
dataset=as.data.frame(cbind(X1,X2,X3,X4,Y),nrow=N,ncol=5)
ATEY=mean(dataset[X3==1,]$Y,na.rm=TRUE)-mean(dataset[X3==0,]
$Y,na.rm=TRUE) # bias
lm(Y~X1+X2+X3) #adjusted model:no bias
```

Funding This funding was provided by ZonMw(Grant No. 91717368), Prof Dr Judith E Bosmans.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahlbom, A.: Modern epidemiology, tl lash, tj vanderweele, s haneuse, kj rothman wolters kluwer, 2021. Eur. J. Epidemiol. **36**(8), 767–768 (2021)

Angrist, J., Pischke, J.S.: Mostly harmless econometrics: an empiricist's companion, 1st edn. Princeton University Press (2009)

Austin, P.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res **46**, 399–424 (2011). https://doi.org/10.1080/00273171.2011.568786

Austin, P.: Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm. Stat. **10**, 150–61 (2011). https://doi.org/10.1002/pst.433

Austin, P.C.: The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med. Decis. Mak. **29**(6), 661–677 (2009). https://doi.org/10.1177/0272989X09341755

Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. **34**(28), 3661–3679 (2015). https://doi.org/10.1002/sim.6607

Bach, J.F.: Causality in medicine. Comptes Rendus Biol. **342**(3–4), 55–57 (2019)

Bartlett, V.L., Dhruva, S.S., Shah, N.D., et al.: Feasibility of using real-world data to replicate clinical trial evidence. JAMA Netw. Open **2**(10), e1912,869-e1912,869 (2019)

Bhide, A., Shah, P.S., Acharya, G.: A simplified guide to randomized controlled trials. Acta Obstet. Gynecol. Scand. **97**(4), 380–387 (2018). https://doi.org/10.1111/aogs.13309

Bouter, L., Zielhuis, G., & Zeegers, M.: Textbook of epidemiology. Bohn Stafleu van Loghum, https://books.google.nl/books?id=-J0EtAEACAAJ (2017)

Brady, H., Collier, D., Sekhon, J.: The neyman-rubin model of causal inference and estimation via matching methods. Oxf. Handb. Polit. Methodol. (2008). https://doi.org/10.1093/oxfordhb/9780199286546.003.0011

Cameron, A., Trivedi, P.: Microeconometrics: methods and applications. Cambridge University Press (2005)

Catalogue of bias collaboration, Aronson, J.K., Bankhead, C., Mahtani, K.R., Nunan, D.: Catalogue of bias-confounding by indication (2018)

Catalogue of bias collaboration, Aronson, J.K., Bankhead, C., Nunan, D.: Catalogue of bias-confounding (2018)

Catalogue of bias collaboration, Lee, H., Aronson, J.K., Nunan, D.: Catalogue of bias-collider bias (2019)

Cave, A., Kurz, X., Arlett, P.: Real-world data for regulatory decision making: challenges and possible solutions for europe. Clin. Pharmacol. Ther. **106**(1), 36–39 (2019)

Cerulli, G.: Methods based on selection on observables, pp. 49–159. Springer, Heidelberg (2015)

Clare, P.J., Dobbins, T.A., Mattick, R.P.: Causal models adjusting for time-varying confounding-a systematic review of the literature. Int J Epidemiol **48**(1), 254–265 (2019)

Constantine Gatsonis SCM: methods in comparative effectiveness research. CRC Press (2017)

Crown, W.H., Henk, H.J., Vanness, D.J.: Some cautions on the use of instrumental variables estimators in outcomes research: how bias in instrumental variables estimators is affected by instrument strength, instrument contamination, and sample size. Value Health **14**(8), 1078–1084 (2011). https://doi.org/10.1016/j.jval.2011.06.009

DeMaris, A.: Combating unmeasured confounding in cross-sectional studies: evaluating instrumental-variable and heckman selection models. Psychol. Methods (2014). https://doi.org/10.1037/a0037416

Elwert, F., Winship, C.: Endogenous selection bias: the problem of conditioning on a collider variable. Annu. Rev. Sociol. **40**, 31–53 (2014)

Ertefaie, A., Small, D., Flory, J., et al.: Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. Int. J. Biostat. (2015). https://doi.org/10.1515/ijb-2015-0006

Franklin, J., Eddings, W., Austin, P., et al.: Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Stat. Med. (2017). https://doi.org/10.1002/sim.7250

Gail, M.H., Wieand, S., Piantadosi, S.: Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika **71**(3), 431–444 (1984)

Griffith, G.J., Morris, T.T., Tudball, M.J., et al.: Collider bias undermines our understanding of Covid-19 disease risk and severity. Nat. Commun. **11**(1), 1–12 (2020)

Grimes, D.A., Schulz, K.F.: Bias and causal associations in observational research. Lancet **359**(9302), 248–252 (2002)

Grobbee, D., Hoes, A.: Clinical epidemiology. Jones & Bartlett Learning, https://books.google.nl/books?id=ZvbpAgAAQBAJ (2014)

Gunasekara, F.I., Carter, K., Blakely, T.: Glossary for econometrics and epidemiology. J. Epidemiol. Community Health **62**(10), 858–861 (2008)

Haneuse, S.: Distinguishing selection bias and confounding bias in comparative effectiveness research. Med. Care **54**(4), e23–e29 (2016). https://doi.org/10.1097/MLR.0000000000000011

Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)

Heckman, J.J.: Econometric causality. Int. Stat. Rev. **76**(1), 1–27 (2008). https://doi.org/10.1111/j.1751-5823.2007.00024.x

Heckman, J.J.: Selection bias and self-selection, pp. 242–266. Palgrave Macmillan, London (2010)

Hernan, M., Robins, J.: Causal inference. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, https://books.google.nl/books?id=_KnHIAAACAAJ (2019)

Hernán, M.: Invited commentary: selection bias without colliders. Am. J. Epidemiol. **185**, 1–3 (2017). https://doi.org/10.1093/aje/kwx077

Hill, R., Griffiths, W., Lim, G.: Principles of econometrics, 4th edn. Wiley (2011)

Ho, Y.F., Hu, F.C., Lee, P.I.: The advantages and challenges of using real-world data for patient care. Clin. Transl. Sci. **13**(1), 4 (2020)

Lesko, C.R., Buchanan, A.L., Westreich, D., et al.: Generalizing study results: a potential outcomes perspective. Epidemiology **28**(4), 553–561 (2017)

Lindeboom, M., Portrait, F., Van den Berg, G.J.: Long-run effects on longevity of a nutritional shock early in life: the dutch potato famine of 1846–1847. J. Health Econ. **29**(5), 617–629 (2010)

Miettinen, O.S., Cook, E.F.: Confounding: essence and detection. Am. J. Epidemiol. **114**(4), 593–603 (1981)

Nørgaard, M., Ehrenstein, V., Vandenbroucke, J.P.: Confounding in observational studies based on large health care databases: problems and potential solutions—a primer for the clinician. Clin. Epidemiol. **9**, 185–193 (2017)

Platt, R.W., Schisterman, E.F., Cole, S.R.: Time-modified confounding. Am. J. Epidemiol. **170**(6), 687–694 (2009)
Porta, M.: A dictionary of epidemiology. Oxford University Press (2014)
Robin, J., Pei, Z., Marie-Andrée, S., et al.: A practical guide to regression discontinuity. MDRC pp 113–132 (2012)
Rothman, K.: Epidemiology: an introduction. OUP USA, https://books.google.nl/books?id=tKs7adtH-_IC (2012)
Rothman, K.J., Greenland, S.: Causation and causal inference in epidemiology. Am. J. Public Health **95**(S1), S144–S150 (2005)
Sanson-Fisher, R., Bonevski, B., Green, L., et al.: Limitations of the randomized controlled trial in evaluating population-based health interventions. Am. J. Prev. Med. **33**(2), 155–161 (2007). https://doi.org/10.1016/j.amepre.2007.04.007
Sekhon, J.S.: The Neyman-Rubin model of causal inference and estimation via matching methods. Oxf. Handb. Polit. Methodol. **2**, 1–32 (2008)
Streeter, A.J., Lin, N.X., Crathorne, L., et al.: Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. J. Clin. Epidemiol. **87**, 23–34 (2017)
Terza, J.V., Basu, A., Rathouz, P.J.: Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J. Health Econ. **27**(3), 531–543 (2008). https://doi.org/10.1016/j.jhealeco.2007.09.009
Thaul, S., Lohr, K.N., Tranquada, R.E.: Health services research: opportunities for an expanding field of inquiry—an interim statement. National Academies (1994)
Tugwell, P., Knottnerus, J.A., Idzerda, L.: Has the time arrived for clinical epidemiologists to routinely use 'routinely collected data'? J. Clin. Epidemiol. **66**(7), 699–701 (2013)
van Leeuwen, N., Lingsma, H., Craen, T., et al.: Regression discontinuity design simulation and application in two cardiovascular trials with continuous outcomes. Epidemiology (2016). https://doi.org/10.1097/EDE.0000000000000486
VanderWeele, T., Hernán, M.: Causal inference under multiple versions of treatment. J. Causal Inference **1**, 1–20 (2013). https://doi.org/10.1515/jci-2012-0002
Vigneri, M., Masset, E., Clarke, M., et al.: Epidemiology and econometrics: Two sides of the same coin or different currencies? Centre of Excellence for Development Impact and Learning Inception Paper 10 : London (2018)
Wing, C., Simon, K., Bello-Gomez, R.A.: Designing difference in difference studies: best practices for public health policy research. Annu. Rev. Public Health **39**, 453–469 (2018)
Wooldridge, J.M.: Econometric analysis of cross section and panel data. no. 0262232197 in MIT Press Books, The MIT Press (2001)
Wooldridge, J.M. Introductory econometrics: a modern approach. ISE-International Student Edition, South-Western, http://books.google.ch/books?id=64vt5TDBNLwC (2009)
Wooldridge, J.M.: Econometric analysis of cross section and panel data. The MIT Press, http://www.jstor.org/stable/j.ctt5hhcfr (2010)
Zohoori, N., Savitz, D.A.: Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. Ann. Epidemiol. **7**(4), 251–257 (1997)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

AN Varga[1] · AE Guevara Morel[1] · JM van Dongen[1] · J Lokkerbol[2] · M Lindeboom[3,4] · MW van Tulder[5,6] · LM Bouter[7,8] · JE Bosmans[1]

AE Guevara Morel
a.e.guevaramorel@vu.nl

[1] Department of Health Sciences, Faculty of Science, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] Centre of Economic Evaluation & Machine Learning, Netherlands Institute of Mental Health and Addiction (Trimbos Institute), Utrecht, The Netherlands

[3] Department of Economics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[4] Centre for Health Economics, Monash University, Melbourne, Australia

[5] Faculty Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[6] Aarhus University Hospital, Aarhus, Denmark

[7] Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

[8] Department of Philosophy, Faculty of Humanities, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands