Canadian Journal of Cardiology ■ (2018) 1–4

Editorial

Selection Bias in Cardiology Research: Another Thing to Worry About (and How to Correct for It)

Christopher Labos, MD, CM, MSc, and George Thanassoulis, MD, MSc

Preventive and Genomic Cardiology, Division of Cardiology, McGill University Health Centre and Research Institute, Montréal, Québec, Canada

See article by Banack et al., pages xxx-xxx of this issue.

Randomized trials are rightly considered the gold standard in medical research; however, observational data have provided some of our most important insights into risk factors for disease. Observational data from Doll and Hill first suggested that smoking caused lung cancer.[1] And observational data from the Framingham cohort demonstrated that risk factors—such as diabetes, hypertension, and hypercholesterolemia—could lead to heart disease.[2] Thus, it is not surprising that we are seeing an increase in observational research, especially given the high costs of randomized trials. Scientists today have access to larger and more diverse patient cohorts and population databases than ever before.

However, observational research is prone to problems not seen in randomized trials. Observational data once suggested that vitamin C might reduce cardiovascular events,[3] but subsequent trials showed that it did not.[4] There are many reasons why observational data might show a noncausal association. In most cases, confounding is often to blame. However, although most people are generally aware of confounding as a problem in observational research, far less time is spent discussing other biases, including selection bias.[5]

Confounding occurs when a third variable, related to both the exposure and the outcome, affects the association between the 2. For example, smoking is correlated with consumption of alcohol, and smoking also increases the risk of lung cancer. Therefore, alcohol consumption would appear to be causally linked to lung cancer, but this association would be confounded by smoking. Selection bias occurs when the sample of patients chosen for analysis is not representative of the population at large but is selected for having a specific characteristic, such as the presence of a disease (or for not having a disease; that is, healthy people). Although confounding is inherent to the sample under study and the specific association of interest, selection bias usually stems from the actual study design or analysis and can have pervasive effects on the associations identified.

One form of selection bias, commonly seen in cardiovascular studies, is recurrent event bias,[6] which occurs when the population under study is not representative of the general population but is selected to include patients who already have the disease. Consider the well-known example of the **T**hrombolysis **I**n **M**yocardial **I**nfarction (TIMI) risk score, which lists 1 of its 7 predictors of adverse events as having taken aspirin in the past week.[7] In the general population, aspirin prevents cardiovascular events, and yet in patients presenting with acute coronary syndrome (ACS), it seems to augur adverse outcomes. Why might this be? The TIMI score was developed to estimate the risk of recurrent cardiovascular events among patients who presented with initial ACS. A patient who presents with ACS despite taking aspirin is very likely a higher-risk person (either because of a much higher risk factor burden or a greater burden of coronary disease, or perhaps a genetic predisposition) and is therefore also more likely to have a recurrent event. Aspirin at the time of the ACS is simply a marker of the higher baseline risk. Aspirin does not cause subsequent adverse events, although it appears to associate with adverse events when the analysis is limited to patients presenting to hospitals with ACS.[8]

There are numerous other examples of selection bias. For example, it has been frequently observed that strong risk factors for a first event appear paradoxically protective for second and subsequent events. A patent foramen ovale (PFO) is a known risk factor for a first stroke but is often "protective" for a second stroke.[6] Several other biases have been described including the "smoker's paradox" and the "obesity paradox," which share many features. These paradoxical phenomena can often be explained by understanding the impact that conditioning on an event (ie, stroke or ACS) has on the risk factors of the participants of the study.

To illustrate this phenomenon, we will use the example proposed by Glymour[9] and consider the hypothetical situation of National Basketball Association (NBA) basketball players. It is obvious that being tall increases your chances of playing the in the NBA. It also true that being fast increases your chance of playing in the NBA. And, therefore, it is also true that basketball players in the NBA are both taller and

Received for publication February 9, 2018. Accepted March 19, 2018.

Corresponding author: Dr George Thanassoulis, McGill University Health Center, 1001 Decarie Boulevard, D5-5120, Montréal, Québec H4A 3J1, Canada. Tel.: +1-514-934-1934 x 35465.

E-mail: george.thanassoulis@mcgill.ca

See page 4 for disclosure information.

https://doi.org/10.1016/j.cjca.2018.03.010

faster, on average, than a random collection of people from the general population. But, if you were to look solely at NBA players, you would find a negative association between these 2 predictors of basketball prowess. An NBA player who is not that fast is likely very tall. Conversely, a player who is not that tall, if he made it into the NBA, must be very fast. Indeed, by "selecting" only for individuals who have experienced a particular event (in this case, being drafted into the NBA), tallness and speed seem to be inversely correlated, even though both predict entry into the NBA and are uncorrelated in the general population. Therefore, the act of selecting certain individuals from the general population distorts the expected distribution of risk factors and induces dependencies among them. It is these unusual risk factor distributions and dependencies that lead to the attenuated and, at times, paradoxical associations detected under these circumstances.

With PFO, we have a similar situation. A PFO is known to be a significant risk factor for a first stroke. Likewise, having a heavy burden of atherosclerotic risk factors is also a risk factor for stroke. Therefore, a person might have a stroke if he or she has a PFO or many traditional risk factors. If we select for people who have already had stroke (which is akin to having been drafted into the NBA), there will be a tendency for people with PFOs to have fewer traditional risk factors than those who do not (ie, the shorter NBA players are very fast). Thus, among patients with stroke, if we simply compare people who have PFOs with those who do not *for their risk of having a subsequent stroke*, we are effectively comparing people with few atherosclerotic risk factors for stroke with those who have many such risk factors, making it appear that a PFO is protective for recurrent stroke and leading to the "paradox." Although, theoretically, these differences can be controlled in the analysis, practically, this is not possible for several reasons. First, statistical adjustment can only adjust for *known* and *measured* factors that could be leading to bias. Second, because the risk factor distributions are distorted compared with the general population (due to the selection bias), this further clouds which variables should be included in any statistical adjustment. Therefore, standard approaches (ie, regression) will frequently fail to account fully for the differences in risk factors.

In the field of cardiovascular research, we are particularly prone to selection bias, especially when we study patients presenting to hospital with ACS. This occurs because most patient cohorts are based on patients who not only made it to hospitals alive but also survived to discharge. Therefore, by definition, such patient cohorts ignore pre- and in-hospital mortality. Thus, it is easy to see how patients who survive to discharge are different from those who do not because of a different distribution of risk factors between survivors and nonsurvivors.

Consider the recent study by Canto and coworkers that examined the risk factors (smoking, diabetes, hypertension, dyslipidemia, or family history) in patients presenting to hospitals with ACS. The authors found that patients with fewer risk factors had increased in-hospital mortality. However, patients suffering fatal out-of-hospital events likely had higher risk-factor burdens, greater severity of coronary disease, or lower ejection fractions than those who made it to hospital alive. If a higher risk-factor burden increased the probability of having a fatal ACS and not making it to hospital alive, the exclusion of these patients from any subsequent analysis of cardiac events would tend to attenuate the impact of risk factors and could, in extreme situations, make a higher risk-factor burden seem protective.[10,11]

Table 1. Association between risk factor burden and MACE in the general population

	MACE	No MACE	Total	Risk of MACE
High risk-factor burden	120	880	1000	120/1000 x 100% = 12%
Low risk-factor burden	100	900	1000	100/1000 x 100% = 10%
Total	220	1780	2000	

Risk ratio (RR) = 120/1000 ÷ 100/1000 = 1.2.
MACE, major adverse cardiac event.

Consider a simplified situation in which we examine patients with high and low risk- factor burdens and measured adverse cardiac events that occurred both before and after hospitalization. In this hypothetical situation, having more risk factors increases the risk of adverse cardiac events (Table 1).

Now let's consider what would happen if we excluded patients who died early. Let's imagine that, in our hypothetical example, 75 patients had fatal out-of-hospital events and were therefore excluded from the analysis. Let's further imagine that, of those 75 patients, 50 patients had high risk-factor burdens, and 25 patients had low burdens of risk factors (Table 2).

In our new 2 x 2 table, the risk of major adverse cardiac events (MACE) has fallen for both groups because the patients at high risk for early mortality have been excluded. But the risk for those with a high risk-factor burden has fallen by a greater degree because they were over represented in the early-mortality group. Thus, the risk ratio has fallen from 1.2 to 0.96 and makes a higher risk-factor burden appear less dangerous and even mildly protective. If even greater numbers of patients were excluded from the analysis, and if those patients preferentially have a high risk-factor burden, the association between risk factors and MACE will be further driven in the reverse direction, leading to a "paradoxical protective" association (Fig. 1).

Clearly, conditioning on an index event, such as ACS or survival to discharge, poses problems in observational research that are not easily overcome. Patients who die out of hospitals will not be captured in most observational cohorts or administrative databases and are not eligible for long-term follow-up. However, in this issue of *Canadian Journal of Cardiology*, Banack and colleagues[12] present a potential

Table 2. Population from Table 1 with selection bias (ie, excluding the 50 patients with high risk-factor burdens and the 25 patients with low risk-factor burdens who died en route to hospitals)

	MACE	No MACE	Total	Risk of MACE
High risk-factor burden	120 − 50 = 70	880	950	70/950 x 100% = 7.4%
Low risk-factor burden	100 − 25 = 75	900	975	75/975 x 100% = 7.7%
Total	145	1780	1925	

Risk ratio (RR) = 70/950 ÷ 75/975 = 0.96.
MACE, major adverse cardiac event.