Analytical selection bias

Sometimes our study sample does not make up a stratum of a sampling collider, but we inadvertently create such a stratum when we restrict the analysis to part of the sample. The most common trap is, perhaps, the studying of an early causal link in a chain of cause and effect—after excluding those who had reached a later effect.

Suppose, for example, that we are interested in the effect of fibrinogen (a coagulation protein) on atherosclerosis, which is a known cause of clinical cardiovascular disease. According to our background knowledge, fibrinogen causes cardiovascular disease by mechanisms that do not pass through atherosclerosis because it has a pro-thrombotic effect (Figure 7–4, top part). Evidently, cardiovascular disease status plays the role of a collider on a path that connects fibrinogen and atherosclerosis:

fibrinogen→coagulation→**cardiovascular disease status**←atherosclerosis

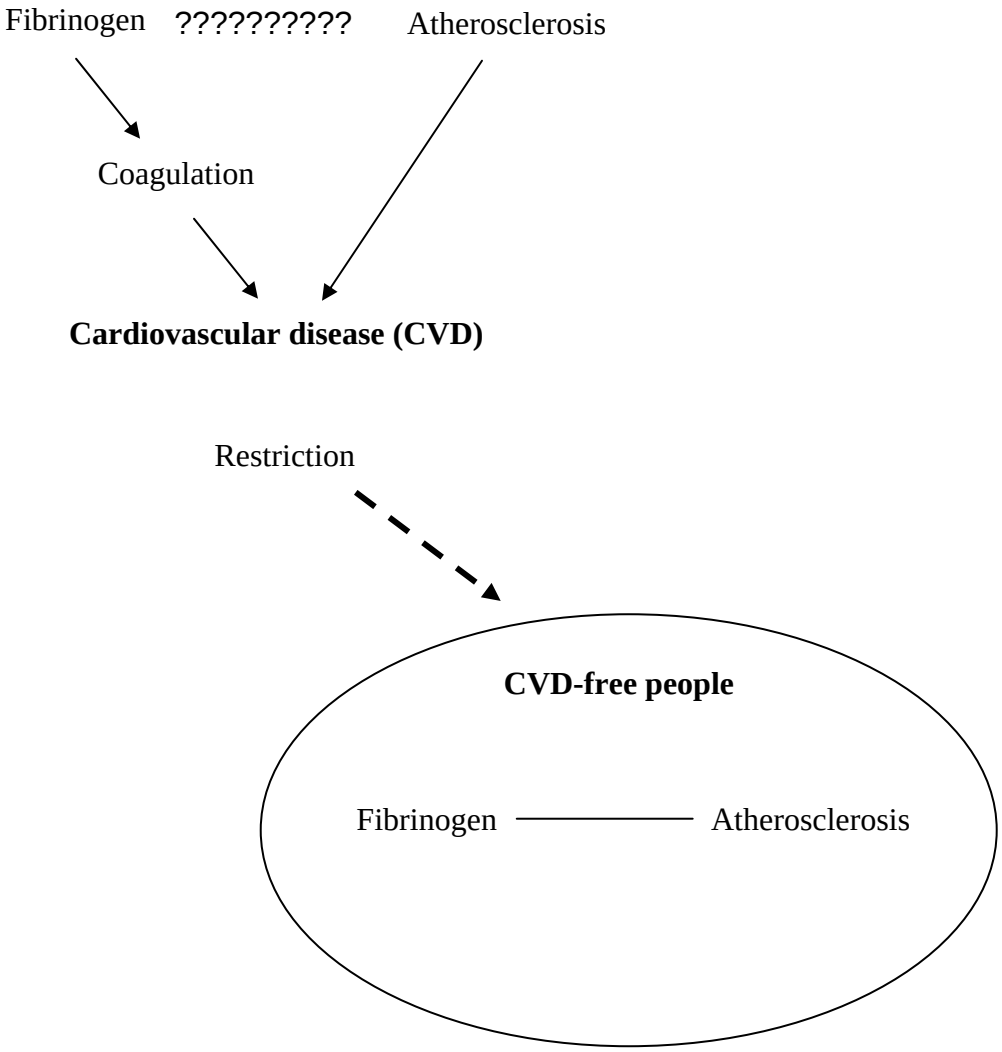

Figure 7–4. A causal diagram showing why the association between fibrinogen and atherosclerosis is affected by selection bias when the analysis is restricted to those who are free of cardiovascular disease. The variable cardiovascular disease status is a collider.

Unaware of the pitfall of conditioning on a collider, many researchers (including me) have studied the cross-sectional associations between atherosclerosis and risk factors for cardiovascular disease after excluding people who had already developed the disease. From the standpoint of causal diagrams, it might have been an unfortunate methodological mistake. Restricting the analysis to one stratum of the variable "cardiovascular disease status" is a form of selection bias because we condition on a collider and thereby induce or alter the association between atherosclerosis and its putative cause (Figure 7–4, bottom part). A similar kind of bias is lurking in cross-sectional studies of risk factors for sub-clinical disease that have recruited only people who did not develop clinical disease.

The decision to condition on a late effect (by restriction) is often driven by fear of estimating a reversed causal pathway. For example, if clinical cardiovascular disease can somehow change the plasma concentration of fibrinogen, we are facing the following diagram:

atherosclerosis→cardiovascular disease status→fibrinogen

which implies a marginal association between atherosclerosis and fibrinogen due to reversed causality of no immediate interest. That association may be blocked by conditioning on cardiovascular disease status (for example, by restricting the analysis to disease-free people). Unfortunately, in a cross-sectional sample it is impossible to distinguish between the path above (reversed causality) and the path below (an intermediary collider) which was depicted in Figure7–4:

atherosclerosis→cardiovascular disease status←fibrinogen

That means that we can't tell which evil is smaller: restricting the sample to disease-free people or analyzing the entire sample? In some examples, however, the mechanism of reverse causality is nonsensical and, therefore, restriction is unquestionably wrong. For instance, cardiovascular disease status cannot (yet) affect one's genotype or one's sex group. To sum up, we may cause more harm than good by naïvely assuming that a study of an early link (A→B) in a causal chain (A→B→C) would benefit from excluding people who developed a later effect (C) in that chain.

*

The idea of selection bias applies to several other situations: 1) selection of controls for a case-control study; 2) selection of prevalent cases rather than incidence cases; 3) and losses to follow up in a cohort study. In all three situations it is possible to explain how the sampling of controls or cases (in a case-control study) or a special type of losses to follow up (in a cohort study) amounted to conditioning on a collider. Examples will be provided in chapter 21.