Bias and confounding are the twin banes of epidemiology, but of the 2, bias is worse. Confounding can often be detected and at least partly corrected by appropriate statistical analyses, but if bias is built into the design of a study, it can be difficult to detect, and even more difficult, if not impossible, to correct. Different types of epidemiological studies are prone to different types of biases. One type of particular concern in case-control studies is survival bias, which can arise when risk factors for a disease that is often fatal are measured only in the patients who survive it. An exposure that is more frequent in the survivors may appear to be related to the disease when it is actually associated with survival. Alternatively, if an exposure is related both to disease and its lethality, its association with the disease may be diminished by survival bias.

The latter possibility motivated the study by Anderson et al this issue of Circulation: Cardiovascular Genetics. In an extensive series of simulations, they demonstrate that, under certain circumstances, even substantial levels of survival bias will not completely obscure the effects of genetic variants that increase both the risk of developing a disease and the risk of dying of it. To show this, they used published estimates of age-specific incidence and mortality for acute ischemic stroke, intracerebral hemorrhage, and myocardial infarction, and of age-specific all-cause mortality, to simulate disease incidence and survival in a longitudinal cohort. Within this framework, they constructed a variety of genetic models in which specific genotypes conferred varying levels of risk for disease-specific incidence and mortality and observed the consequences of missing different proportions of fatal cases for each of the 3 diseases when selecting samples of cases and controls from the full cohort at different ages.

They report that for patients younger than 75, even failure to include a single lethal case in a study would, under a range of conditions, reduce the estimate of the genotype relative risk for disease by <20%, even for genotypes that doubled the risk of dying. The impressive amount of work they used to establish this included simulations to test the sensitivity of their findings to different values of age-specific disease incidence and mortality, different sex-specific incidence and mortality rates, different recurrence risks of disease and death, and different sample sizes, among other factors. None of these differences changed the estimates from their primary analyses by >5%.

Where and when these findings might apply to actual studies remains questionable. One motivation for the work was to investigate the potential effects of genetic variants that increase the risk of dying of a disease more than they increase the risk of developing it. Whether such variants exist, or how common they might be, is unclear, although a few reports exist. For instance, in patients with type 2 diabetes mellitus, homozygosity for the B1 allele of the TaqIB polymorphism of the CETP gene that encodes the cholesteryl ester transfer protein was associated with a slightly higher hazard ratio for sudden death (1.46; 95% confidence interval: 1.01 to 2.12) than for incident myocardial infarction (1.33; 95% confidence interval: 0.87 to 2.05) even after adjusting for multiple risk factors. The difference, if real, is quite small, but both values are within the range of those tested in the simulations of Anderson and his colleagues. Another CETP variant was reported to be associated with cardiovascular mortality, but not with nonfatal myocardial infarction or stroke, in a cohort of subjects with coronary artery disease. Both of these results were from prospective studies; it would be interesting to see what attempts to replicate these findings in case-control studies would find.

Even if such variants do exist, a well-designed case-control study of them would not be compromised if case ascertainment were complete, or nearly so, and the investigators were careful to avoid mixing prevalent and incident cases. It is a misconception that case-control studies, because they involve existing cases, necessarily involve prevalent cases—a case-control study and a cross-sectional study are 2 different things, occasional references to “cross-sectional case-control studies” notwithstanding. If it can be determined that disease onset in an existing case occurred at a specific time or within a specific interval covered by the study, that case qualifies as an incident case. In a well-designed case-control study, epidemiologists will go to great lengths to try to establish the time of disease onset and include only incident cases in analyses relating risk factors to disease. Although it may be true, as stated on page 188, that “Cross-sectional studies are nonetheless commonly used because of their advantages in efficiency of disease ascertainment,” conscientious epidemiologists will make every effort to distinguish incident and prevalent cases and not mix them in their analyses. The dangers of mixing incident and prevalent cases in analyses are emphasized repeatedly in epidemiological methods courses.

In one regard, case-control studies of genetic risk factors for chronic diseases have an advantage over those of other
risk factors, in that the exposure is virtually guaranteed to precede the disease, something the use of only incident cases in nongenetic case-control studies attempts to ensure. But DNA must be available for both cases and controls, and if cases of a frequently fatal disease are identified through medical records or registries, DNA samples from fatal cases would often be unavailable, and the danger of selectively excluding a specific class of patients from a genetic case-control study would be quite real. But there is more than one type of case-control study, and not all types would be equally susceptible to this problem. Anderson et al specifically designed their simulations with genome-wide association studies in mind, and in case-cohort or nested case-control studies derived from the sorts of large prospective cohorts frequently used for genome-wide association studies, neither lack of available DNA nor the inability to distinguish prevalent and incident cases are likely to be insuperable problems. Anderson et al suggest that effect erosion from survival bias may explain the recent failure of a meta-analysis of (mainly) case-control studies to verify associations between polymorphisms on chromosome 12p13 and stroke that were found in analyses involving 4 prospective cohorts. This may be the case, but given the dismayingly high likelihood that any newly reported positive associations will turn out to be false, more prosaic explanations may be responsible.

That survival bias has not yet been proved to have affected genetic association studies does not mean that the work of Anderson et al should be ignored. They quantified the extent to which survival bias may reduce the frequency of a risk allele in cases relative to controls, and suggest that this could be useful as a signature of the possible presence of survival bias. In general, this would show itself as a failure of genotype frequencies in cases to match Hardy-Weinberg expectations. This could occur for reasons other than survival bias, but is certainly something any investigator would want to check in any case-control study of genetic risk factors for a disease.

None of the considerations mentioned here are intended to discount the findings of Anderson et al. The fact remains, however, that the problems they sought to investigate and quantify are really problems of poor study design, not inevitable flaws of genetic case-control studies of frequently fatal diseases. That incomplete case ascertainment and failure to distinguish incident and prevalent cases are problems that have plagued many case-control studies in the past does not invalidate case-control studies, just poorly designed ones.

In a sense, the findings of Anderson et al might be considered reassuring because they show that under a range of conditions, a poorly designed genetic case-control study may still yield at least partly valid results. But one hopes that no one will take much comfort in this. For one thing, as the authors of the study readily concede, their results are based entirely on simulations, and there can be no guarantee how broadly applicable their findings might be in realistic situations. Simulations can never account for every factor that might affect complex biological systems and pathological processes. In addition, Anderson et al were more concerned with the possibility that survival bias could lead to underestimating the effect of a mutation on disease risk, rather than on mortality risk. But failure to detect and genotype lethal cases could make a mutation look more innocuous than it is, even if a detectable association with nonlethal incident disease remained. Furthermore, incomplete case ascertainment could compromise inferences related to nongenetic factors, if, as is usually the case in epidemiological studies, investigators attempted to assess multiple risk factors in a single case-control study.

Thus, one hopes that no one will interpret the finding that, in many circumstances, failure to ascertain substantial numbers of fatal cases may not seriously affect the findings of a genetic case-control study to mean that one can get away with careless study design. In any case-control study, every possible effort must be made to ensure complete, or at least unbiased, ascertainment of cases and to avoid mixing incident and prevalent cases in analyses.

When it comes to designing epidemiological studies, knowing how much one can get away with can be a great comfort because no epidemiological study, however perfectly designed, can ever be perfectly executed. But such comfort can be a dangerous thing if it leads to overconfidence and carelessness in designing a study, or if lessons valid in one context are misapplied in another where they do not hold. Poor epidemiological study design can wreck a study in many ways, even though one may be lucky enough to get away with it now and then.

Disclosures

None.

References


The Folly of Being Comforted
D. Michael Hallman

doi: 10.1161/CIRCGENETICS.111.959783
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circgenetics.ahajournals.org/content/4/2/108

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Genetics can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Genetics is online at:
http://circgenetics.ahajournals.org/subscriptions/