In investigating lethal diseases like coronary heart disease (CHD) and major adverse events like myocardial infarction (MI) and death can sometimes seem a bit macabre. We are interested in understanding the events with the hope of preventing them; yet, to demonstrate effects, it is essential for a high rate of such unfortunate events to occur and to be observed. Fortunately, the increasing availability of big [event] data supports an unprecedented ability and power to explore genetic influences on primary and subsequent CHD events. Tempering the enthusiasm around this opportunity is the concern for biases that threaten the internal and external validity of such investigations.

Selection bias has been a well-acknowledged concern in genetic association. More recently, others have brought attention to and provided evidence for the influence of survival bias in cardiovascular genetic associations. In this issue of Circulation: Cardiovascular Genetics, Hu et al assert that a gap exists in our understanding of the extent to which these biases are expected to occur, specifically in the context of gene associations with subsequent CHD events. In their article titled, Impact of Selection Bias on Estimation of Subsequent Event Risk, they begin to address this contemporary problem by using in silico methods to estimate selection bias and survival bias effects on gene association with subsequent CHD events. Notably, they use the probability of subsequent MI after a first nonfatal MI as an exemplar case. The authors simulate these bias types from empirically derived population estimates of common candidate allele frequencies, MI event rates, and nongenetic risk exposure rates. Importantly, their empirical estimates for model parameters were chosen to best approximate the data anticipated among common cardiovascular genetics consortia data sets (here, the Genetics of Subsequent Coronary Heart Disease consortium). Among both time-to-event and logistic regression models, estimates were generated on 250 000 simulated subjects across 5000 iterations. For both prospective and case–control designs, results indicate that the estimated percentage of bias achieved was <10% for both selection bias and survival bias, unless the magnitude of genetic effect for a single variant was >2 (large effect size in common disease association). Additional sensitivity analyses revealed that the bias estimates were insensitive to variation in overall population disease prevalence, single nucleotide polymorphism minor allele frequency, and nuanced censoring parameters. Encouragingly, type 1 error rate and false discovery rate were stable across various models. The authors conclude that the effects of selection and survival bias on genetic associations for subsequent MI events are expected to be minimal; therefore, identification of common genetic variants for such events with either prospective or case–control design is methodologically valid.

The impact of these results should be considered in the context of some caveats dually acknowledged by the authors. These are simulated bias estimates, but carefully constructed with empirical parameters. The observed minimal effect of <10% is specific to genetic associations modeling subsequent (aka, recurrent) cardiac events. Concern over external validity about their use of marginal rather than conditional association is assuaged by the fact that the use of marginal estimates produces what the authors refer to as a worst-case scenario, such that incorporating covariates is presumed to diminish bias effects even further. Also, these analyses do not examine every distinct type of selection or survival bias that could be present in case–control designs. While providing power calculations helpful to future studies, the authors offer a cautionary note about their evidence of overprecision of confidence intervals that may fail to contain the true hazard, even with increasing sample size.

Among a meticulous and comprehensive set of in silico analyses presented by the authors, encouraging evidence emerges to reduce fears that genetic associations of subsequent CHD events are spuriously produced by selection or survival biases. Hu et al have provided us an informed starting point for evaluating the likelihood of such biases in models that demonstrate large genetic effects for common disease models (odds ratio or hazard ratio <2).
What is the value of such a threshold? In the case of survival bias specifically, estimations made by Hu et al.\(^5\) and Anderson et al.\(^4\) can be used to evaluate the expected amount of effect erosion among association models (change in hazard ratio or odds ratio likely to be observed in the presence of such bias). Furthermore, Kaplan–Meier curves can be generated for characterizing allele- and genotype-specific effects on survival that may indicate whether the source of bias is coming from loss or culling of the risk variant in the observed population.\(^2,3\)

Many sources of bias are often deemed unavoidable, but there are widely accepted means of reducing the likelihood of biases, such as through rigorous study design, randomization, and controlling for bias-related confounders. Statistical solutions for selection bias handling include treatment effects models, instrumental variables methods or propensity score methods. A more novel strategy to mitigate the undue influence of survival bias was recently presented by Chen et al.\(^7\) They demonstrated success in augmenting genetic data on patients experiencing fatal events from their immediate relatives. Their method even bolstered statistical power. Unfortunately, such an approach would present ethical and logistical challenges when working from population-based consortia data sets.

As healthcare advances increase the likelihood of surviving index events, and as subsequent event data are increasingly available through robustly powered cardiovascular genetics consortia, studies like those of Hu et al.\(^5\) increase our confidence to explore genetic effects on subsequent CHD end points. This work (and estimates herein) may be relevant to those researching other common disease phenotypes with similar prevalence and event rates. Further explorations into approaches for preventing and mitigating selection and survival biases will be important for refining gene associations and maximizing the use of secondary data to identify genetic underpinnings of risk for events.

**Disclosures**

None.

**References**

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9:356–369. doi: 10.1038/nrg2344.
2. Dungan JR, Qin X, Horne BD, Carlquist JF, Singh A, Hurdle M, et al. Case-only survival analysis reveals unique effects of genotype, sex, and coronary disease severity on survivorship. PLoS One. 2016;11:e0154856. doi: 10.1371/journal.pone.0154856.
3. Dungan JR, Hauser ER, Qin X, Kraus WE. The genetic basis for survivorship in coronary artery disease. Front Genet. 2013;4:191. doi: 10.3389/fgene.2013.00191.
4. Anderson CD, Nalls MA, Biffi A, Rost NS, Greenberg SM, Singleton AB, et al. The effect of survival bias on case-control genetic association studies of highly lethal diseases. Circ Cardiovasc Genet. 2011;4:188–196. doi: 10.1161/CIRCGENETICS.110.957928.
5. Hu Y-J, Schmidt AF, Dudbridge F, Holmes MV, Brophy JM, Tragante V, et al. Impact of selection bias on estimation of subsequent event risk. Circ Cardiovasc Genet. 2017;10:e001616. doi: 10.1161/CIRCGENETICS.116.001616.
6. Patel RS, Asselbergs FW. The GENIUS-CHD consortium. Eur Heart J. 2015;36:2674–2676.
7. Chen L, Weinberg CR, Chen J. Using family members to augment genetic case-control studies of a life-threatening disease. Stat Med. 2016;35:2815–2830. doi: 10.1002/sim.6888.

**Key Words:** Editorials • confidence intervals • genotype • probability • sample size • selection bias • survival analysis