The Effect of Survival Bias on Case-Control Genetic Association Studies of Highly Lethal Diseases

Christopher D. Anderson, MD*; Michael A. Nalls, PhD*; Alessandro Biffi, MD*; Natalia S. Rost, MD; Steven M. Greenberg, MD, PhD; Andrew B. Singleton, PhD; James F. Meschia, MD; Jonathan Rosand, MD, MSc

Background—Survival bias is the phenomenon by which individuals are excluded from analysis of a trait because of mortality related to the expression of that trait. In genetic association studies, variants increasing risk for disease onset as well as risk of disease-related mortality (lethality) could be difficult to detect in genetic association case-control designs, possibly leading to underestimation of a variant’s effect on disease risk.

Methods and Results—We modeled cohorts for 3 diseases of high lethality (intracerebral hemorrhage, ischemic stroke, and myocardial infarction) using existing longitudinal data. Based on these models, we simulated case-control genetic association studies for genetic risk factors of varying effect sizes, lethality, and minor allele frequencies. For each disease, erosion of detected effect size was larger for case-control studies of individuals of advanced age (age > 75 years) and/or variants with very high event-associated lethality (genotype relative risk for event-related death > 2.0). We found that survival bias results in no more than 20% effect size erosion for cohorts with mean age < 75 years, even for variants that double lethality risk. Furthermore, we found that increasing effect size erosion was accompanied by depletion of minor allele frequencies in the case population, yielding a “signature” of the presence of survival bias.

Conclusions—Our simulation provides formulas to allow estimation of effect size erosion given a variant’s odds ratio of disease, odds ratio of lethality, and minor allele frequencies. These formulas will add precision to power calculation and replication efforts for case-control genetic studies. Our approach requires validation using prospective data. (Circ Cardiovasc Genet. 2011;4:188-196.)

Key Words: stroke ▪ hemorrhage ▪ myocardial infarction ▪ genetics ▪ epidemiology

Genetic association tests using a case-control design depend on comprehensive ascertainment of cases to ensure an accurate representation of causal factors for analysis.1,2 Survival bias, a form of ascertainment bias caused by selective exclusion of individuals who have mortality related to the disease being studied, is difficult to predict, estimate, or control.3-5 Myocardial infarction (MI), acute ischemic stroke (AIS), and intracerebral hemorrhage (ICH) are prototypical diseases with high event-related mortality (lethality) that are affected by this phenomenon. For these conditions, prospective longitudinal studies of incident disease may offer advantages over cross-sectional studies capturing both prevalent and incident disease for the identification not only of those genetic variants that increase fatal outcome from disease but also those variants that influence both risk of disease as well as risk of disease-associated lethality. Cross-sectional studies are nonetheless commonly used because of their advantages in efficiency of disease ascertainment.

Editorial see p 108
Clinical Perspective on p 196

The presence of survival bias risks underascertainment of genotypes associated with high lethality, with consequent underestimation of disease risk effect sizes for those genotypes associated with both increased disease risk and disease-associated lethality.6 Although no variants have yet been identified that simultaneously confer modest risk of disease and substantial risk of lethality from such disease, such a scenario remains possible and could affect any comparison of longitudinal and case-control genetic association studies. For example, a recent genome-wide association study of ischemic stroke, based on a meta-analysis of longitudinal cohort data, demonstrated an association at the chromosome 12p13 locus.7

Received May 13, 2010; accepted January 14, 2011.
From the Center for Human Genetic Research (C.D.A., A.B., N.S.R., J.R.) and the Hemorrhagic Stroke Research Group (C.D.A., A.B., N.S.R., S.M.G., J.R.), Department of Neurology, Massachusetts General Hospital, Boston, MA; the Program in Medical and Population Genetics (C.D.A., A.B., N.S.R., J.R.), Broad Institute, Cambridge, MA; the Laboratory of Neurogenetics (M.A.N., A.B.S.), Intramural Research Program, National Institute on Aging, Bethesda, MD; and the Department of Neurology (J.F.M.), Mayo Clinic, Jacksonville, FL.
*Drs Anderson, Nalls, and Biffi contributed equally to this work.

The online-only Data Supplement is available at http://circgenetics.ahajournals.org/cgi/content/full/CIRCGENETICS.110.957928/DC1.
Correspondence to Jonathan Rosand, MD, Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge St, CPZN-6818, Boston, MA 02114; E-mail jrosand@partners.org
© 2011 American Heart Association, Inc.

Circ Cardiovasc Genet is available at http://circgenetics.ahajournals.org DOI: 10.1161/CIRCGENETICS.110.957928
An independent attempt at replication using a meta-analysis of case-control data failed to support this association.8 One potential explanation for failed replication could be that variants at 12p13 increase risk of stroke as well as stroke-related lethality, thereby limiting their detection in patients enrolled using cross-sectional designs. Replication efforts using such a study design would therefore have limited power to detect a true association with disease. Tools that allow calculation of effect size erosion caused by survival bias are necessary to determine if this explanation is plausible.

We hypothesized that by modeling diseases of high lethality based on data available from large population-based cohorts, we could simulate the effect of survival bias on genome-wide association studies of variants associated with both disease and lethality. Using AIS, ICH, and MI, we simulated case-control studies across a continuum of population ages, assuming a variety of minor allele frequencies (MAF), genotype relative risk (GRR) of disease (GRR-D), GRR of lethality (GRR-L), and recruitment rates of lethal cases. We then created formulas for estimating the erosion of effect size for case-control studies of each of these diseases as well as a formula for extrapolation of estimates to other diseases not modeled.

Methods

Literature Search for Data From Longitudinal Cohorts
A PubMed (http://www.ncbi.nlm.nih.gov) search was performed independently to identify articles reporting prevalence, incidence, and mortality of AIS, ICH, and MI. Search terms are shown in the Appendix. To be considered eligible for modeling, reported data had to refer to studies conducted after 1990 (to account for recent trends in incidence and mortality from vascular disease), with incidence and event-associated mortality detailed separately by age category. We also included only data from high-income industrialized countries to model scenarios consistent with populations currently enrolled in most genetic association studies.9–27 Data from stroke longitudinal cohorts that did not distinguish ischemic from hemorrhagic stroke were not considered eligible. We found only 1 article reporting data separately for ischemic stroke subtypes by Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria.14

Model Building for Disease Impact on the General Population
All simulations were performed using the R statistical package v2.10.0 (http://www.r-project.org). We modeled disease incidence rate within each age category by computing weighted (based on sample size) medians for incidence data from published longitudinal cohorts. Disease-associated mortality rate was computed similarly; all-cause mortality was modeled on the basis of summary World Health Organization epidemiological data. Mortality unrelated to vascular disease was modeled on the basis of summary World Health Organization data for industrialized countries (http://www.who.int/whosis/mort/download/en/index.html). We generated year-by-year disease incidence and mortality rates using the approx interpolation function in the R stats library.

We then proceeded to model the impact of each disease over time on longitudinal cohorts of 100 000 healthy individuals. At cohort inception, individuals were assumed to be age 45 years (mean), with a normal distribution and standard deviation of 7 years. Additional modeling using different standard deviations (range, 4 to 10 years) did not alter results significantly. This simulated longitudinal cohort was followed prospectively for 35 years. Each year, new disease cases (MI, ICH, or AIS) were registered based on incidence data from our literature search. Among these cases, a portion resulted in death as appropriate for their age category based on World Health Organization epidemiological data. Mortality unrelated to vascular disease was adjusted for cases surviving the disease-event according to published data9–27 to reflect the excess all-cause mortality in individuals expressing vascular disease phenotypes (online-only Supplement Tables S1 and S2). Cumulative disease incidence and disease-free survival are plotted in online-only Supplement Figure S1.

Genetic Risk Modeling
We simulated a single marker conferring increased risk for disease expression, assumed to be a common genetic variant in the population being studied. For the purposes of our simulation, only the minor allele was chosen for association with disease and lethality. Simulations were run for MAF of 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, and 0.45. The effect of this marker was expressed in terms of GRR-D, with different simulations accounting for values of 1.05, 1.10, 1.15, 1.20, 1.25, 1.30, 1.35, 1.40, 1.45, and 1.50. These values were selected to reflect effect sizes usually associated with common variants in genetic association studies.28 We also modeled GRR-L, expressed as the relative risk of death due to disease: simulated GRR-L included 1.05, 1.10, 1.15, 1.20, 1.25, 1.30, 1.35, 1.40, 1.45, and 1.50, with increments of 0.15, 1.70, 1.80, 2.0, 2.5, 3.0, 3.5, and 4.0. GRR-D was assessed after each year of longitudinal follow-up using a Cox proportional hazards model.

Simulation of Case-Control Genetic Association Studies
To quantify the impact of survival bias on case ascertainment, we simulated a case-control study conducted within the modeled population each year (for MI, ICH, and AIS). Samples ranged in mean age from 45 to 80 years (SD = 7 years). For each year of cohort follow-up, 1000 cases and 1000 control subjects were enrolled through simple random sampling with replacement from survivors, and effect size (expressed as odds ratio [OR] for each copy of the risk allele) was computed using logistic regression analysis. All combinations of prespecified genetic parameters (MAF, GRR, and GRR-L) were simulated independently. Results were expressed in terms of erosion of effect size caused by survival bias, computed as the ratio between the observed OR and the underlying “true” GRR. Observed MAFs for cases and control subjects were also recorded for each simulation and stored. To model the effect of lethal case recruitment, we performed separate simulations assuming 0% through 40% ascertainment of lethal cases in our case-control studies (0% and 40% lethal case ascertainment results reported).

Effect Size Erosion Estimation Formulas
To provide a disease-specific formula to estimate the impact of survival bias on observed OR in case-control studies for MI, ICH, and AIS, we used results from all simulated scenarios (9 MAF values×10 GRR values×18 GRR-L values=1620 simulations) to regress the erosion of effect size due to survival bias (dependent variable, expressed as OR/GRR and log-transformed for normality). Predictors included mean age of enrolled subjects, MAF of the genetic marker in the study population, risk for phenotype expression (GRR-D), and risk of phenotype-associated mortality (GRR-L).

Furthermore, we computed a general (non–disease-specific) formula based on user-specified disease incidence and mortality rates for the age range of interest. Specified annual incidence rates for formula computation ranged from 0.0001 to 0.03, and specified disease-associated mortality ranged from 0.01 to 0.40. All genetic parameter (MAF, GRR-D and GRR-L) ranges were identical to those simulated for AIS, ICH, and MI. The resulting formula (based again on linear regression) allows for quantification of effect size erosion due to survival bias in diseases other than MI, AIS and ICH, provided that age-specific incidence and mortality data are available to investigators.
Sensitivity Analyses
To verify the consistency of our simulation approach, we conducted several sensitivity analyses. (1) We performed separate simulations of disease incidence and mortality in men and women, based on literature data to assess sex-specific effects. (2) We modeled earlier ages of disease onset (range of simulation values: 2 to 10 years earlier than published data) associated with the causal genetic marker of interest. This was done to simulate the effect of several known genetic risk variants acting on both disease risk and age of onset.26 (3) We simulated a genetic-confounded risk of non-disease-related increase in mortality (range of simulated values, 1.05 to 2.0) to our model(s). This analysis allowed us to evaluate the impact of survival bias on discovery efforts for variants acting on both disease risk and accelerated aging/biological frailty (eg, APOE in Alzheimer disease). (4) We performed additional modeling of vascular disease recurrence and recurrence-associated mortality (based on literature data). This analysis aimed at assessing whether the concomitant effect of the variant of interest on first-ever disease manifestation, disease recurrence, and risk of disease-related death (at all time points, independent of all-cause mortality) modified the impact of survival bias. (5) We compared results obtained from interpolating incidence and mortality data on a year-by-year basis against the direct use of published data by age-category. (6) We performed separate analyses using both the lowest and highest published incidence and mortality figures for each age category in each disease in order to account for the variability in the published estimates for these figures. (7) In simulation of longitudinal cohort studies, we varied the number of healthy individuals at enrollment to values of 50 000, 200 000, 500 000, and 1 000 000. (8) In simulation of case-control studies for each disease state, we varied the number of cases and control subjects sampled from the survivor population for each year at values of 500/500, 1000/1000, 2500/2500, and 5000/5000.

Results

Literature Search and Construction of Disease Models
A PubMed (http://www.ncbi.nlm.nih.gov/pubmed) search identified 10 articles for AIS, 7 articles for ICH, and 6 articles for MI, providing incidence and mortality data by age category, presenting data obtained since 1990.9–27 On the basis of this search, we constructed models for each disease and used an interpolation algorithm to simulate data for each year (Figure 1). We then constructed longitudinal and cross-sectional studies of genetic variants influencing both disease incidence and lethality, across a range of MAF, GRR-D, and GRR-L (Figure 2). Hazard ratios for disease (HR-D) in the simulated longitudinal study were compared with the OR for disease (OR-D) from the simulated case-control study for each year of follow-up in the inception cohort of 100 000 individuals.

Measurement of the Effects of Survival Bias
We tabulated the output of the longitudinal (HR-D) and case-control (OR-D) genetic simulations and reported results as the erosion of apparent effect size obtained using a case-control design (ie, the percentage of HR-D that was lost to detection due to survival bias in calculation of an OR-D using a case-control study design) (Figures 3, 4, and 5). For all diseases, effect size erosion increased with increasing mean ages of enrolled individuals and increasing variant-confounded GRR-L. These findings reflect the close association between baseline disease lethality and extent of ascertainment lost as the result of survival bias. Case-control studies with higher rates of mortality among carriers of the causal variant, whether because of lethal effects of the variant itself or because of the higher disease mortality among older individuals, had larger effect size erosions (Figures 3 through 5).

As an illustration of this phenomenon, for a genetic variant that simultaneously increased the risk of ischemic stroke (GRR-D range, 1.05 to 2.0) and doubled the risk of death from that of stroke (GRR-L = 2.0), the effect size erosion was substantially <20%, assuming a cohort of 65 to 75 years old and no enrollment whatsoever of lethal cases (Figure 3). By comparison, a variant increasing risk for ICH in a similar fashion would be subject to an average erosion in observed effect size of approximately 40%, with the possibility of reversal of observed effect in some situations as cases possessing the risk allele become selectively excluded to such a degree that the risk allele begins to achieve a higher frequency in control subjects compared with cases. This can be seen in situations where the effect size erosion surpasses 100% (Figure 4).

Erosion in effect size estimates decreased substantially for all diseases, with just 40% enrollment of lethal cases, and decreased even further for more conservative estimates of GRR-L. In ICH, the condition with the highest lethality, enrollment of 40% of lethal cases reduced the erosion in effect size estimate (on average) to <20% and prevented reversal of effect from being observed up to GRR-L of 2.0 (Figure 4).

Simulation of Observed MAF
Using the genetic case-control study simulations by year for each disease, we calculated the observed MAF for the simulated variant in cases and compared this with the specified population MAF at simulation outset for MAF 0.10 to 0.30, assuming no population stratification. For this analysis, GRR-D of 1.05 to 1.5 and 0% to 40% ascertainment of lethal cases were modeled (results shown, GRR-D of 1.3 with lethal-case ascertainment of 0% and 40%). For all population MAFs tested, we observed a GRR-L and age-dependent decrease in observed cases of MAF. This MAF depletion was predictably less severe for 40% lethal case ascertainment than for 0% ascertainment (online-only Supplement Figures S2 through S4). The severity of depletion of MAF was strongly dependent on GRR-L for all simulated values of GRR-D and lethal case ascertainment (all P<0.05). As a result, depletion in observed MAF in cases compared with the population MAF could be suggestive of the presence of survival bias.

Formulas for Estimation of Effect Size Erosion in Case-Control Genetic Studies
On the basis of these simulations, we used linear regression to identify a formula for each disease that approximates the percent erosion of effect size estimate for a genetic variant, with a priori assumptions of cohort mean age (range, 40 to 80), and the variant’s MAF (range, 0.05 to 0.45), GRR-D (range, 1.05 to 1.50), and GRR-L (range, 1.05 to 4.0). Note that because ICH has a lower annual incidence than AIS or
Figure 1. AIS indicates acute ischemic stroke; ICH, intracerebral hemorrhage; and MI, myocardial infarction. Bars represent the incidence and mortality estimates for each disease for the age ranges specified, according to the published reports for each cohort. Lines represent the interpolated incidence and mortality curves used to create the disease models for simulation of longitudinal and case-control study designs.
MI, the MAF exerts a proportionally larger influence on effect size erosion than in the other 2 disease states.

\[
\% \text{ effect size erosion in AIS} = \exp \left( 0.08(\text{AGE}) - 0.13(\text{GRR-D}) + 0.67(\text{MAF}) + 1.43(\text{GRR-L}) - 11 \right)
\]

\[
\% \text{ effect size erosion in ICH} = \exp \left( 0.034(\text{AGE}) + 0.085(\text{GRR-D}) + 1.06(\text{MAF}) + 1.53(\text{GRR-L}) - 6.8 \right)
\]

\[
\% \text{ effect size erosion in MI} = \exp \left( 0.062(\text{AGE}) - 0.069(\text{GRR-D}) + 0.94(\text{MAF}) + 1.50(\text{GRR-L}) - 8.9 \right)
\]

To allow extension of our simulation results to other highly lethal phenotypes, we again used a linear regression to create a general case formula to estimate effect size erosion. This formula requires the variant’s MAF, GRR-D, and GRR-L (same ranges as above) as input, in addition to disease annual incidence (Φ; range, 0.0001 to 0.03) and disease-related mortality (Ω; range, 0.01 to 0.40) estimates for the age range of included subjects.

\[
\% \text{ effect size erosion} = \exp \left( 8.73(\Omega) \times 1.4(\Phi) + 0.067(\text{GRR-D}) + 1.02(\text{MAF}) + 1.53(\text{GRR-L}) - 7.25 \right)
\]

Application of the general formula to AIS, ICH, and MI using incidence and mortality data from the published literature returned estimates of effect size erosion caused by survival bias within 5% of those provided by disease-specific formulas presented above.

We offer the following as a theoretical application of our formulas. We can construct a scenario of two genetic studies of ICH, one in a population-based longitudinal cohort and another in a hospital-based case-control cohort with a mean age of 67 years. Let us suppose that the longitudinal study identified a novel variant with an HR of 1.7 for risk of ICH and an OR of 2.0 for ICH-related mortality, with a MAF of 0.15 in the general population. Turning to the hospital-based case-control study, let us assume that researchers were unable to consent and enroll any lethal ICH cases. Using these data, we can apply our formula for effect size erosion in ICH to show that the case-control researchers should expect a 33% erosion in their effect size estimate for the same variant, yielding an expected OR of 1.14. If the case-control study is underpowered to detect an OR of this size, it may fail to confirm the ICH risk association for the variant identified in the population-based study. Of note, if the case-control researchers were able to enroll 40% of lethal cases from the population, the variant’s observed OR for ICH risk would rise to 1.46, and power to detect the variant would increase.

**Discussion**

Our results demonstrate that failure to enroll lethal cases can distort the measured effect sizes for genetic variants affecting both disease incidence and lethality. However, the effect size erosion is <20% in ischemic stroke, and, while more substantial in extremely high mortality diseases such as ICH and MI, it can be reduced to <20% if ascertainment of at least a proportion (40%) of lethal cases can be ensured. Variants that increase disease risk while more than doubling the risk for disease-related mortality (lethality) and case-control genetic association studies performed on cohorts >75 years of age do result in more substantial erosion in effect size estimates for disease risk, but these situations are unlikely to be encountered in the real-world of complex disease genetics.

Incomplete ascertainment of individuals with a variant increasing risk of both disease and lethality leads to an observed MAF in cases that is lower than the true value. With increasing lethality risk and cohort age, we have shown that this observed case MAF can drop below the MAF of the population from which the cohort is drawn. In such a situation, the variant would spuriously appear to be protective...
rather than harmful because it is seen less often in cases than in control subjects. This apparent reversal of effect would appear in the presence of an imbalance between alleles, as homozygotes would not been seen as often as expected because of their dramatically increased risk of lethality. As a result, the variant would not be in Hardy-Weinberg equilibrium, which would provide a warning that this apparent reversal of effect was not a true result. We have shown a

Figure 3. Effect size erosion in case-control design is plotted according to varying cohort ages and genetic relative risk of lethality (GRR-L) for acute ischemic stroke. The median (thick line) and range (box plot with whiskers) for each GRR-L simulation reflects the ranges of minor allele frequency and genetic relative risk of disease. Top panel assumes 0% ascertainment of lethal cases; bottom panel assumes 40% ascertainment of lethal cases.

Figure 4. Effect size erosion in case-control design is plotted according to varying cohort ages and genetic relative risk of lethality (GRR-L) for intracerebral hemorrhage. The median (thick line) and range (box plot with whiskers) for each GRR-L simulation reflects the ranges of minor allele frequency and genetic relative risk of disease. Top panel assumes 0% ascertainment of lethal cases; bottom panel assumes 40% ascertainment of lethal cases.
significant correlation between the GRR-L and the degree of observed MAF deviation from the population value. When comparing results obtained from case-control studies with those obtained from prospective longitudinal cohort studies, this signature of depleted MAF in cases may prove useful for detecting the presence of survival bias. MAF depletion can disappear with improved lethal case enrollment. Forty percent enrollment of lethal cases results in no MAF depletion below the population MAF, even for cohorts of advanced age and diseases of high mortality such as ICH and MI, as long as the GRR-L is relatively low (1.3). For higher GRR-L values, MAF depletion can remain substantial even with 40% lethal case ascertainment.

The ascertainment rate of individuals with event-related lethality is of critical import in the estimation of power erosion caused by survival bias. For our simulations, we first estimated a 0% ascertainment rate for lethal cases, which is quite conservative for at least some highly lethal phenotypes. We have shown in our analysis that increasing the ascertainment rate for lethal cases to 40% substantially reduces the effect size erosion caused by survival bias.

Researchers wishing to apply our formulas to their populations should substitute the nonascertainment rate caused by disease lethality (if known) for the absolute lethality rate whenever possible to avoid overestimation of effect size erosion for their cohort.

Using our simulation results, we have constructed disease-specific formulas for the estimation of effect size erosion as well as a general case formula that can be used for other highly lethal diseases of known incidence and mortality. These formulas will allow designers of future case-control genetic analyses of diseases of high lethality to predict the erosion of true effect size for their study, allowing for more accurate ad hoc power calculations. These formulas will also be of utility to investigators seeking to replicate the findings of prior genetic studies, as the modeling of survival bias in addition to Winner’s Curse should allow for very precise power calculations. Thus, in the case of attempted validation of genetic association with chromosome 12p13 in ischemic stroke, the original longitudinal study identified an association at rs12425791, with OR of 1.39. After correction for Winner’s Curse using WINNER v1.1 (http://csg.sph.umich.edu/boehnke/winner), this association decreases to OR of 1.25. Applying our model for ischemic stroke, using the known population MAF of the variant in both the longitudinal discovery and case-control replication cohorts (0.20), and assuming that the variant doubles the risk of stroke lethality (GRR-L = 2.0), we find that the expected OR in the case-control replication effort is 1.14. This case-control study had 0.999 power to detect this effect size. The lack of replication at rs12425791 in populations ascertained using a case-control design is therefore unlikely to have arisen due to failure to ascertain sufficient numbers of fatal cases of stroke. Of note, no evidence exists that the 12p13 locus is associated with such a dramatically increased risk of disease-associated lethality, but the use of such a substantial estimate of lethality provides a conservative estimate of effect size erosion in replication.

Our study has limitations. Even longitudinal studies are not exempt from the possibility of survival bias because individuals are often not enrolled and genotyped until later in life, when some potentially lethal variants could have already
been depleted from the population. However, this is likely to increase the similarity of results between longitudinal and case-control studies, and therefore our simulation, by assuming perfect ascertainment in longitudinal cohorts, represents the most conservative comparison between the two. Our simulations only apply to common genetic variants; therefore, additional methods will be required to model survival bias in experiments identifying rare variants or analyses using other functional units (ie, exon or gene-based units instead of single-nucleotide polymorphisms). In such analyses, deviations from population MAFs would be unreliable due to the inherent imprecision in MAF estimates for rare variants. This issue is unlikely to have affected recent replication efforts because most genetic association studies to date have not been designed to detect rare or private variants. An additional limitation is that the models for MI, ICH, and AIs used for our simulations were built using incidence and mortality data for these diseases from longitudinal studies performed in 1990 to the present.21–27 If disease incidence and overall mortality decrease, these formulas will need to be updated. Similarly, our formulas were constructed from incidence and mortality data from high-income industrialized societies, limiting their generalizability. Experiments performed in underdeveloped countries or following groups with restricted access to health care could underestimate the effects of survival bias caused by increased overall mortality and event-related lethality.21,32,33 Likewise, our formulas were constructed using a priori ranges for MAF, GRR-D, and GRR-L. Values falling outside of these ranges would necessitate reconstruction of the formulas. Another potential limitation is that specific variants could predominantly exert a lethal effect on specific age groups (ie, increasing the lethality of late-onset stroke). This effect has not been modeled in our simulation and thus could lead to incorrect estimates of effect size erosion for certain age groups. Finally, our study has been performed entirely by simulation. Validation of our estimates of effect size erosion in comparison with real-world longitudinal and case-control cohorts is needed if our results are to be considered wholly accurate.

Survival bias represents a substantial concern in the design of case-control genetic association studies of disease with high fatality rates. Our simulation shows that the impact of survival bias is relatively small for most cohort ages and variant lethalities and is further minimized by proactive enrollment of lethal cases. The tools provided by this study will be useful in calculation of power for future genetic association studies of common variants. Future studies are needed to validate and extend our findings to other diseases and to adapt our approach to other functional units of genetic analysis.

Appendix

PubMed (http://www.pubmed.org) search terms for population-based incidence and mortality were Acute Ischemic Stroke: Stroke; Ischemic Stroke; Acute Ischemic Stroke; Cerebral Infarction; Cerebrovascular: Incidence; Prevalence; Mortality; Lethality; Epidemiology; Population; Longitudinal; Registry; Intracerebral Hemorrhage: ICH; Hemorrhage; Hemorrhagic; Stroke; Intracerebral Hemorrhage; Cerebrovascular; Incidence; Prevalence; Mortality; Lethality; Epidemiology; Population; Longitudinal; Registry; Myocardial Infarction: Myocardial Infarction; Acute Myocardial Infarction; MI; Cardiovascular; Incidence; Prevalence; Mortality; Lethality; Epidemiology; Population; Longitudinal; Registry.

Sources of Funding
This work was supported by funding from the American Heart Association/Bugher Foundation Centers for Stroke Prevention Research (0775011N, 0755984T), Deane Institute for Integrative Study of Atrial Fibrillation and Stroke, and the National Institute for Neurologic Disorders and Stroke (R01 NS063925, R01 NS059727, P50 NS051343, K23NS064052-01A1). This research was also supported in part by the Intramural Research Program of the National Institute of Health, National Institute on Aging, under Annual Report No. Z01-AG000954-06:2007.

Disclosures
None.

References
CLINICAL PERSPECTIVE

Genetic association studies seek to establish links between specific genetic variants and phenotypes such as quantifiable traits or disease states. These studies use the same recruitment formats and are subject to many of the same types of bias that can invalidate the results of standard epidemiological studies. Survival bias, the phenomenon by which individuals are excluded from analysis of a trait because they die as a result of the trait’s expression, has been suggested as a potential explanation for the lack of case-control replication of results obtained using longitudinal cohorts. A genetic variant that increases risk of both incidence and mortality of a disease could be found in a longitudinal cohort but could remain undetected in case-controls because individuals with that variant die before study recruitment. We modeled 3 cardiovascular diseases of high lethality: myocardial infarction, acute ischemic stroke, and intracerebral hemorrhage, using existing data from population-based studies. We simulated genetic variants that modify baseline risk for both incidence and mortality of these diseases and constructed longitudinal and case-control studies to test the observed associations between our genetic variants and disease incidence. Our results suggest that the erosion of effect size caused by survival bias is small and is calculable using formulas we provide. Furthermore, study designs that can ensure 40% recruitment of lethal cases can lessen the percentage of effect size erosion caused by survival bias. These results should help genetic researchers compare the results of genetic studies using differing recruitment frameworks.
The Effect of Survival Bias on Case-Control Genetic Association Studies of Highly Lethal Diseases
Christopher D. Anderson, Michael A. Nalls, Alessandro Biffi, Natalia S. Rost, Steven M. Greenberg, Andrew B. Singleton, James F. Meschia and Jonathan Rosand

_Circ Cardiovasc Genet._ 2011;4:188-196; originally published online February 3, 2011; doi: 10.1161/CIRCGENETICS.110.957928

_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/4/2/188

Data Supplement (unedited) at:
http://circgenetics.ahajournals.org/content/suppl/2011/02/03/CIRCGENETICS.110.957928.DC1

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/
The Effect of Survival Bias on Case-Control Genetic Association Studies of Highly Lethal Diseases

Anderson, CD et al.

MS ID#:  CIRCCVG/2010/957928

SUPPLEMENTAL MATERIALS
**Supplemental Table S1.** Absolute and Excess Mortality Rates per 100,000 per year following acute events in AIS, ICH and MI survivors

<table>
<thead>
<tr>
<th>Age Category</th>
<th>All cause-mortality in the general population</th>
<th>All cause excess mortality in AIS / ICH survivors</th>
<th>All cause absolute mortality in AIS / ICH Survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
</tr>
<tr>
<td>45 to 54</td>
<td>547.8</td>
<td>319.9</td>
<td>+21</td>
</tr>
<tr>
<td>55 to 64</td>
<td>1131.0</td>
<td>698.5</td>
<td>+35</td>
</tr>
<tr>
<td>65 to 74</td>
<td>2612.2</td>
<td>1736.3</td>
<td>+55</td>
</tr>
<tr>
<td>over 75</td>
<td>8339.5</td>
<td>7240.6</td>
<td>+85</td>
</tr>
</tbody>
</table>

**Supplemental Table S1 Legend.** AIS = acute ischemic stroke. ICH = intracerebral hemorrhage. MI = myocardial infarction. All cause mortality reflects non-AIS and non-ICH total mortality in AIS and ICH survivors.
### Supplemental Table S2. Absolute and Excess Mortality Rates per 100,000 per year following acute events in MI survivors

<table>
<thead>
<tr>
<th>Age Category</th>
<th>All cause-mortality in the general population</th>
<th>All cause excess mortality in MI survivors</th>
<th>All cause absolute mortality in MI survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
<td>Males</td>
</tr>
<tr>
<td>45 to 54</td>
<td>547.8</td>
<td>319.9</td>
<td>+17</td>
</tr>
<tr>
<td>55 to 64</td>
<td>1131.0</td>
<td>698.5</td>
<td>+31</td>
</tr>
<tr>
<td>65 to 74</td>
<td>2612.2</td>
<td>1736.3</td>
<td>+48</td>
</tr>
<tr>
<td>over 75</td>
<td>8339.5</td>
<td>7240.6</td>
<td>+78</td>
</tr>
</tbody>
</table>

### Supplemental Table S2 Legend.

AIS = acute ischemic stroke. ICH = intracerebral hemorrhage. MI = myocardial infarction. All cause mortality reflects non-MI total mortality in MI survivors.
Supplemental Figure S1. Disease simulation results: cumulative-incidence and disease-free survival

AIS

Cumulative incidence (per 1,000)

Cohort Median Age

Cohort Median Age

ICH

Cumulative incidence (per 1,000)

Cohort Median Age

Cohort Median Age

MI

Cumulative incidence (per 1,000)

Cohort Median Age

Cohort Median Age

% Disease-free Survivors

% Disease-free Survivors

% Disease-free Survivors

% Disease-free Survivors
Supplemental Figure S2. Observed depletion in minor allele frequency with increasing age and relative risk of lethality in case-control simulation of acute ischemic stroke
Supplemental Figure S3. Observed depletion in minor allele frequency with increasing age and relative risk of lethality in case-control simulation of intracerebral hemorrhage
**Supplemental Figure S4.** Observed depletion in minor allele frequency with increasing age and relative risk of lethality in case-control simulation of myocardial infarction
**Supplemental Figure S1 Legend.** Cumulative-incidence per 100,000 individuals (panels A,C,E) and percent of individuals enrolled at baseline surviving free of disease (panels B,D,F) under the simulation model presented in the Methods for acute ischemic stroke (panels A,B), intracerebral hemorrhage (panels C,D), and myocardial infarction (panels E,F). Estimates are calculated from the cohort after a simulated 35 year period from inception.

**Supplemental Figure S2 Legend.** RR(L) = genetic relative risk of lethality for the assumed genetic variant. Observed case minor allele frequencies (MAF) in a case-control framework are plotted for varying mean cohort ages and genetic relative risk of lethality. Separate simulations are shown for starting population MAF of 0.10, 0.20, and 0.30, as represented by the thin grey horizontal line on each plot. The top 3 graphs assume a 0% ascertainment of lethal cases, while the bottom 3 graphs assume a 40% ascertainment of lethal cases. Apparent reversal of effect is seen when the observed case MAF is below the population MAF (grey line).

**Supplemental Figure S3 Legend.** RR(L) = genetic relative risk of lethality for the assumed genetic variant. Observed case minor allele frequencies (MAF) in a case-control framework are plotted for varying mean cohort ages and genetic relative risk of lethality. Separate simulations are shown for starting population MAF of 0.10, 0.20, and 0.30, as represented by the thin grey horizontal line on each plot. The top 3 graphs assume a 0% ascertainment of lethal cases, while the bottom 3 graphs assume a 40% ascertainment of lethal cases. For graphs in which high RR(L) curves are absent, the genetic variant’s MAF rapidly has rapidly decreased to zero, resulting in a monomorphic variant. Apparent reversal of effect is seen when the observed case MAF is below the population MAF (grey line).

**Supplemental Figure S4 Legend.** RR(L) = genetic relative risk of lethality for the assumed genetic variant. Observed case minor allele frequencies (MAF) in case-control are plotted for varying mean
cohort ages and genetic relative risk of lethality. Separate simulations are shown for starting population MAF of 0.10, 0.20, and 0.30, as represented by the thin grey horizontal line on each plot. The top 3 graphs assume a 0% ascertainment of lethal cases, while the bottom 3 graphs assume a 40% ascertainment of lethal cases. For graphs in which high RR(L) curves are absent, the genetic variant’s MAF rapidly has rapidly decreased to zero, resulting in a monomorphic variant. Apparent reversal of effect is seen when the observed case MAF falls below the population MAF (grey line).
PubMed (http://www.pubmed.org) search terms for population-based incidence and mortality:

Acute Ischemic Stroke
Stroke
Ischemic Stroke
Acute Ischemic Stroke
Cerebral Infarction
Cerebrovascular
Incidence
Prevalence
Mortality
Lethality
Epidemiology
Population
Longitudinal
Registry

Intracerebral Hemorrhage
ICH
Hemorrhage
Hemorrhagic
Stroke
Intracerebral Hemorrhage
Cerebrovascular
Incidence
Prevalence
Mortality
Lethality
Epidemiology
Population
Longitudinal
Registry

Myocardial Infarction
Myocardial Infarction
Acute Myocardial Infarction
MI
Cardiovascular
Incidence
Prevalence
Mortality
Lethality
Epidemiology
Population
Longitudinal
Registry