Submitted to the Annals of Applied Statistics

TWO-PHASE SUBSAMPLING DESIGNS FOR GENOMIC RESEQUENCING STUDIES

By Thomas Lumley
University of Auckland
and

By Josée Dupuis
Boston University School of Public Health
and

By Kenneth M. Rice
University of Washington
and

By Maja Barbalic
University of Texas Health Science Center at Houston
and

By Joshua C. Bis
University of Washington
and

By L. Adrienne Cupples
Boston University School of Public Health
and

By Bruce M. Psaty
University of Washington
and

By Christopher J. O'Donnell
National Heart, Lung and Blood Institute
and

By Eric Boerwinkle
University of Texas Health Science Center at Houston

Targeted resequencing of DNA at specific genes or other genomic loci is now feasible for hundreds or thousands of samples, and costs for larger-scale resequencing are decreasing rapidly. For at least the

AMS 2000 subject classifications: Primary 62D05, 62P10; secondary 62J12
Keywords and phrases: case-cohort design, case-control design, sampling weights

next few years, resequencing will need to be confined to small subsets of the large samples on which genome-wide association studies have been recently been performed. This paper describes some strategies for subsampling an existing cohort for resequencing, and flexibly analysing the resulting data. We illustrate these strategies by describing the actual design and planned analyses for the example that motivated our research, the CHARGE-S resequencing study carried out by the CHARGE (Cohorts in Heart and Aging Research in Genomic Epidemiology) Consortium.

1. Introduction. The past few years have seen an explosion in the availability of genotype information. Genome-wide association studies have been performed with hundreds of thousands of genetic markers on sample sizes from tens to hundreds of thousands of people, and these have found many, mostly weak, associations between genetic variants and biological or clinical variables. The greatest likely benefit from these studies is in improving the understanding of biological processes in health and disease, rather than in direct prediction, but biological understanding is hampered by the fact that the association studies typically find a set of nearby genetic markers rather than the genetic variant or subset of variants that truly affects biological processes.

One approach to finding functional genetic variants near a set of markers is to determine the complete DNA sequence of a region of the genome. Analysis of complete sequence data should sharpen the association estimates compared with analysis only of marker data. More importantly, as statistical association is unlikely to be sufficient to narrow the association down to a single variant, having sequence data facilitates biological investigation of candidate variants, whether *in silico*, *in vitro*, or *in vivo*.

DNA resequencing is technologically feasible but very expensive at sample sizes sufficiently large for association studies. Although the costs are decreasing rapidly, over at least the next few years it will be necessary to resequence relatively small subsamples from the large samples that have been participated in genome-wide association studies. Two large resequencing projects in cardiovascular disease were funded by the US National Institutes of Health under the American Recovery and Reinvestment Act. In this paper we present the design and analysis principles for one of these studies, the CHARGE-S resequencing project, and provide some further study of this and alternative subsampling designs.

The goal of CHARGE-S is to find candidate functional variants at positions in the genome identified by genome-wide association studies involving the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Consortium, a group of five cohorts with a total of 40,000 participants(Psaty *et al.*, 2009). Because of the funding source, the resequencing was restricted to the three US member cohorts: the Atherosclerosis Risk in Communities (ARIC) study, the Cardiovascular Health Study (CHS), and the Framingham Heart Study (FHS). Sufficient funds were available to resequence approximately 60 genomic loci, in approximately 5000 people. The CHARGE consortium had more than 20 working groups that had found associations suitable for follow-up by resequencing.

2. Subsampling design.

Our aim in designing a subsampling procedure was to increase power for detecting associations, without introducing systematic bias and without unduly complicating the analysis. For all the participants eligible to be subsampled in CHARGE-S, information was available on all outcomes and other demographic variables, and on the genotypes measured in the genome-wide association study. In this section we consider how this information can be used in design.

2.1. *Subsampling based on outcome.*

The primary technique for increasing power at a fixed sample size is to sample from the extremes of the distribution of the outcome variable. For a rare event, oversampling those with the event in a case–control or case–cohort design can give almost the same power as would be available with complete data. Even for a continuous outcome variable, individuals at the extremes of the distribution should be more likely to carry genetic variants that affect the outcome. Increasing the sample size used in each comparison will also increase the powe, particularly for rare variants, which are hard to distinguish from genotyping error in small samples.

A popular design for a continuous outcome in genetic epidemiology is to sample the most extreme individuals from the upper and lower tails of the distribution and then to treat the outcome as a binary (high/low) variable in the subsample. This approach picks out the individuals most likely to carry functional variants and allows for simple exact analyses in small samples such as Fisher's exact test. For a moderate number of phenotypes this design may be the most powerful, and there is little loss of information in treating the outcome as binary (Guey *et al.*, 2011). Using such a design for a set of 20 outcome variables would require taking and analysing independent samples for each outcome. A total sample size of 5000 would then lead to samples of 125 from each tail of the distribution for each outcome.

A larger sample size for each comparison is possible at some cost in per-sample efficiency by sampling from just one tail of the outcome distribution and using a common reference sample across all outcomes. This is a generalization of the case–cohort design (Prentice, 1986) for survival data. With a total sample size of 5000, a reference sample of 1000 would allow samples of size 200 from one extreme of 20 outcomes. The increased sample size will more than compensate for the fact that the reference group is less extreme. The ARIC cohort has used the case–cohort design extensively in measuring genetic and biological risk markers (e.g. Lee *et al.*, 2008, 2007; Nambi *et al.*, 2008).

Sampling from a single tail reduces the ability to find functional variants that are concentrated in the other tail of the distribution. In the context of GWAS follow-up it is possible to be reasonably confident which tail of the distribution is of interest — if the minor alleles of the marker SNPs are associated with high values of the phenotype, the minor allele of the functional variant is also likely to be associated with high values and samples should be taken from the upper tail. In the context of an unbiased whole-exome or whole-genome search it might still be preferable to sample from only a single tail, which would give increased power for some functional variants and decreased power for others.

An additional benefit of the design using a common reference sample is that it is a proper probability sampling design, in which every individual has a known, non-zero probability of being included in the subsample. As a consequence, weighting methods from survey statistics can be used to estimate any population parameters that would be estimable with complete cohort data, including associations with other phenotypes that were measured at phase one but not targetted in the design. These weighting methods also allow the entire subsample to be used in the estimation of any parameter. With a total sample size of 5000 each analysis would include not only the 200 people sampled for extreme values of the outcome under study and the reference sample of 1000, but also the 3800 people sampled for extreme values of other outcomes. Section 4 includes simulations showing that the increased sample size does lead to increased power for the design with common reference sample under plausible assumptions about effects and correlations.

For some outcome variables there may be good measurements of non-genetic factors that influence the outcome, and rather than subsampling the extremes of the outcome it may be preferable to subsample the extremes of residuals after the influence of these variables has been removed. For example, blood pressure depends strongly on age and moderately strongly on sex, so residuals of blood pressure in a regression on age, age^2, and sex were used as the outcome for sampling.

2.2. *Subsampling based on genotype and outcome.* Since a major goal of the resequencing follow-up study is to find candidates for the functional variants responsible for associations in the genome-wide association study, it may be helpful to use the genetic markers as well as the outcome variable in choosing the subsample. One strategy is to attempt to increase the prevalence of the functional variant in the extreme subsample by oversampling carriers of the highest-signal genetic marker.

For example, Levy *et al.* (2009) found a marker in the gene *CYP17A* with minor allele frequency 10%, leading to lower blood pressure in carriers of the minor allele. If a subsample of 200 low blood pressure individuals was taken to follow up this marker we would expect to have 36 individuals with one copy of the marker and only two with two copies. Oversampling to obtain a more balanced distribution of the marker would likely increase the number of copies of the functional variant in the subsample. The price for this large number of variants is not only an increase in the variability in sampling probabilities but also a less extreme set of low blood pressure values. Additionally, the increased distortion in allele frequencies may complicate the process of quality control and data cleaning for the sequence data; experience on this is relatively limited and it is hard to assess the likely impact. The simulations in Section 4.3 explore the effect on power for finding a functional genetic variant in a limited set of scenarios.

3. Inference after two-phase sampling. The initial sampling of the cohorts followed by the subsampling of the individuals to be resequenced forms a two-phase sampling design. The phase-one data are all the variables available on the full cohorts, including the genome-wide SNP measurements; the phase-two data are the sequence variants. Since the two-phase sample is purely a cost-saving strategy, we wish to estimate the same parameters as would be estimated with sequence data on the full cohorts. Similarly, for testing the null hypothesis of no association between sequence variation and a phenotype the goal is to test the same hypothesis that would be tested with sequence data on the full cohorts. The constraints on estimation are stronger; an estimator of effect that is consistent for zero under the null hypothesis but not consistent for the true association under alternatives can still give a valid test. In section 3.1 we describe estimation methods based on sampling weights. In sections 3.2 and 3.3 we describe techniques that can give more powerful tests for association. In the following development we treat the original cohorts as if they were simple random samples from a data generating process, or superpopulation. Although treating the cohorts as simple random samples is an idealization, it is standard practice for CHS and ARIC. The Framingham study includes complex family structure that does need to be incorporated in the phase-one analysis; details of how this is done in each analysis are beyond the scope of the paper.

3.1. *Estimation.* In all the situations we consider, and much more generally, the estimators that would be used with complete data can be represented in terms of sums over the whole cohort of terms involving a single individual: either as a function of explicit sums or as the solution of an estimating equation that is a sum over individuals. For example, linear regression can be written either way. As a function of explicit sums over covariate vectors x_i and outcomes y_i in a cohort of size N the least squares estimator is

$$\hat{\beta} = \left(\sum_{i=1}^{N} x_i^T x_i\right)^{-1} \sum_{i=1}^{N} x_i y_i,$$

and it also solves the equation

$$\sum_{i=1}^{N} x_i(y_i - \beta^T x_i) = 0.$$

To extend these estimators to estimate the same parameters in a two-phase sample we replace the sum over individuals in the cohort by a weighted sum over individuals in the subsample, the Horvitz–Thompson estimator of a population total(Horvitz and Thompson, 1952). The Horvitz–Thompson estimator is unbiased and has been shown to be consistent and asymptotically Normal under various sets of asymptotics (Krewski and Rao, 1981; Hájek, 1964; Sen, 2009). These asymptotic properties extend to estimators solving weighted estimating equations, at least in simple cases; completely general asymptotic results are not yet available(Binder, 1983; Breslow and Chatterjee, 1999; Lin, 2000; Breslow and Wellner, 2008). Lin and Tang (2011) describes how to formulate many rare-variant tests as score tests, placing them in this framework.

We write R_i for the subsampling indicator for person i, and $\pi_i = E[R_i]$, where the expectation is conditional on all phase-one information. These probabilities must be known for individuals in the subsample; they typically are known for the whole cohort, since the subsampling is entirely under the control of the researchers. The pairwise sampling probabilities $\pi_{ij} = E[R_i R_j]$ must also be non-zero, and must be known at least for pairs of individuals in the subsample. Let

$$T_X = \sum_{i=1}^{N} X_i$$

be the sum over the whole cohort of a variable X. The Horvitz–Thompson estimator of T_X is then

$$\hat{T}_X = \sum_{i:R_i=1} \frac{1}{\pi_i} X_i$$

and an unbiased estimator of its variance around the true cohort total, conditional on all the phase one information is

$$\widehat{\mathrm{var}}\left[\hat{T}_X - T_X \mid \text{phase one}\right] = \sum_{i,j:R_iR_j=1} \frac{X_i}{\pi_i}\frac{X_j}{\pi_j} - \frac{X_iX_j}{\pi_{ij}}. \tag{1}$$

The same results hold if X_i is a function of observations on individual i and unknown parameters, such as a log-likelihood or an estimating function.(Binder, 1983)

The unconditional variance of $\hat{T}_X$ is the sum of the conditional finite-population sampling variance and the variance of T_X over repeated realizations of the cohort,

$$\mathrm{var}\left[\hat{T}_X\right] = \mathrm{var}\left[\sum_{i=1}^{N} X_i\right] + \mathrm{var}\left[\hat{T}_X - T_X \mid \text{phase one}\right]$$

and can be estimated by

$$\widehat{\mathrm{var}}\left[\hat{T}_X\right] = \frac{N^2}{n-1} \sum_{i:R_i=1} \frac{1}{\pi_i}(X_i - \bar{X})^2 + \sum_{i,j:R_iR_j=1} \frac{X_i}{\pi_i}\frac{X_j}{\pi_j} - \frac{X_iX_j}{\pi_{ij}}$$

where

$$\bar{X}_i = \frac{1}{N} \sum_{i:R_i=1} \frac{1}{\pi_i} X_i = \frac{\hat{T}_X}{N}.$$

Combining this approach with the delta method, if $J_i(\theta)$ is the efficient influence function for observation i in an estimator $\hat{\theta}$ based on complete cohort data, and $I_i(\theta)$ is the Fisher information from observation i then solving

$$\sum_{i:R_i=1} \frac{1}{\pi_i} J_i(\theta) = 0$$

gives a consistent, asymptotically Normal Horvitz-Thompson-type estimator $\hat{\theta}_{HT}$ based on the subsample, with variance estimated by

$$\widehat{\mathrm{var}}\left[\hat{\theta}_{HT}\right] = \left(\frac{N^2}{n-1} \sum_{i:R_i=1} \frac{1}{\pi_i} I_i(\hat{\theta}_{HT})\right)^{-1} + \sum_{i,j:R_iR_j=1} \frac{J_i(\hat{\theta}_{HT})}{\pi_i}\frac{J_j(\hat{\theta}_{HT})}{\pi_j} - \frac{J_i(\hat{\theta}_{HT})J_j(\hat{\theta}_{HT})}{\pi_{ij}}$$

Similarly, if $U_i(\theta)$ is the efficient score for a parameter θ, a score test of $\theta = \theta_0$ can be constructed using the weighted score statistic

$$\bar{U}(\theta_0) = \sum_{i:R_i=1} \frac{1}{\pi_i} U_i(\theta_0)$$

with variance estimated by

$$\widehat{\mathrm{var}}\left[\bar{U}(\theta_0)\right] = \frac{N^2}{n-1} \sum_{i:R_i=1} \frac{1}{\pi_i} I_i(\hat{\theta}_0) + \sum_{i,j:R_iR_j=1} \frac{U_i(\hat{\theta}_0)}{\pi_i} \frac{U_j(\hat{\theta}_0)}{\pi_j} - \frac{U_i(\hat{\theta}_0)U_j(\hat{\theta}_0)}{\pi_{ij}}.$$

The full generality of equation 1 is not needed in typical subsampling designs, which divide the cohort into strata and take a simple random sample without replacement from each stratum. Sampling probabilities π_i are then constant within a stratum and pairwise sampling probabilities π_{ij} depend only on the sampling fraction. We label strata by $h = 1, \ldots, H$ with cohort size N_h and subsample size n_h in stratum h, and write X_{hj} for the jth observation subsampled in the stratum. Equation 1 then reduces to

$$(2) \quad \widehat{\mathrm{var}}\left[\hat{T}_X - T_X \mid \text{phase one}\right] = \sum_{h=1}^{H} \left(\frac{N_h - n_h}{N_h} \times N_h^2 \times \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} (X_{hi} - \bar{X}_h)^2 \right)$$

where $\bar{X}_h$ is the subsample mean in stratum h. The formula for the contribution from each stratum differs from that under iid sampling only in the finite population correction factors $(N_h - n_h)/N_h$.

In the CHARGE-S study, a further feature of the analysis is that individual-level data are not shared between cohorts, so the analysis must be broken into cohort-specific computations followed by a meta-analysis or pooling stage where cohort-specific results are combined. This typically does not introduce any serious difficulty, since each sampling stratum is restricted to a single cohort. In the earlier genome-wide association studies the pooling stage simply involved taking a precision-weighted mean of point estimates from cohort-specific regression models. A score test can similarly be combined across cohorts by taking the precision-weighted sum of the efficient score from each cohort and dividing by its estimated variance.

3.2. *Testing.* If sampling is based on phenotype alone, and if a single phenotype Y is independent of the DNA sequence at a particular locus, then sampling based on this one phenotype does not affect the distribution of sequence variation, so the null hypothesis of no association can be tested without using sampling weights. This simplification does not occur for sampling based on multiple phenotypes, as is shown in the simulations in section 4. If the phenotypes are correlated and sequence variation is associated with one phenotype, the correlation may lead to spurious associations with other phenotypes in the subsample, a phenomenon related to Berkson's Fallacy(Berkson, 1946).

It is still possible to regain some of the precision from the unweighted analysis when sampling is based on multiple phenotypes, using the the idea of 'stabilized weights'(Robins, Hernán and Brumback, 2000) or 'adjusted weights'(Pfeffermann and Sverchkov, 1999). Let Y be the phenotype of interest. Under the null hypothesis that Y is independent of DNA sequence in the population, the sampling weights may be divided by any function of Y without inducing dependence between Y and DNA sequence in the weighted sample. The optimal function of Y to use is the one that minimizes the coefficient of variation of the weights, that is,

$$r(y) = \exp E\left[-R\log\pi|Y=y\right].$$

The optimal function is not available in practice, as it depends on the joint distribution of all phenotypes, but it can be estimated by one-dimensional smoothing. Testing can then be based on the same computations as in section 3.1 but using $(\hat{r}(Y_i)\pi_i)^{-1}$ in place of $\hat{\pi}_i^{-1}$ as the sampling weights.

If sampling were based only on Y this approach would recover the unweighted analysis, as $r(Y_i) = \pi_i$, so $(r(Y_i)\pi_i)^{-1} = 1$. With sampling based on multiple phenotypes, the variation in weights is reduced but not eliminated. The simulations in section 4 confirm that using stabilized weights leads to higher power without introducing bias when the null hypothesis is true, but that the gain in power is very small in the scenarios we consider.

3.3. *Unweighted testing.* When sampling is based only on phenotype, an unweighted test of association between a phenotype and DNA sequence is valid under the assumption that all other phenotypes have no association with DNA sequence at that locus. If the resequencing study is worth doing, this strong genetic null hypothesis must be implausible. However, genetic associations with complex traits have typically been very weak, so the strong genetic null hypothesis may be a reasonable approximation, and the unweighted tests may be close to being valid for plausible genetic effect sizes. Since the unweighted test often has greater power than the weighted test, a valid unweighted test would be very useful.

For the case of linear regression with no adjustment variables it is possible to examine the bias analytically. The score test for association is the test based on the covariance of genotype and phenotype. Let G be a genetic predictor variable and Y and Z be phenotypes. Assume for simplicity that X, Y and Z are standardized to have zero mean and unit variance in the whole cohort. We are interested in testing for association between G and Y, in the presence of association between G and Z. Our test statistic is based on

$$U = \frac{\sum_{i=1}^N R_i G_i Y_i}{\sum_{i=1}^N R_i}.$$

R depends on X only through Y and Z, so under the null hypothesis that X and Y are independent in the whole cohort, R is independent of X given Z, and RX is independent of RY given Z and R. Let $\pi_0 = E[R]$ be the overall sampling fraction, which is fixed by design. Then

$$\begin{aligned} E[U] &= \frac{1}{N\pi_0}E\left[RGY\right] \\ E\left[RGY\right] &= E\left[E\left[RGY|Z\right]\right] \\ &= E\left[E\left[G|Z\right]\cdot E\left[RY|Z\right]\right] \\ &= E\left[E\left[G|Z\right]\cdot E\left[E\left[R|Y,Z\right]\cdot Y|\,Z\right]\right]. \end{aligned}$$

The bias depends on the strength of association between genotype G and the other phenotype Z, between the two phenotypes Y and Z, and between the sampling probabilities and the two phenotypes. The correlation between the phenotypes may be high, certainly 0.5 is not out of the question. The association between G and Z is, by hypothesis, of the order of magnitude that we care about for genetic associations. So the only way that the unweighted test can be reliable is for $E[R|Y,Z]$ to be small or weakly correlated with Y and Z. If there were only two phenotypes, the sampling probabilities would be deterministic monotone functions of Y and Z, so the bias would be large, but as the number of phenotypes increases, the dependence on any given phenotype becomes weaker.

It is important to note that the unweighted test will not always be usefully more powerful, and intuition can be unreliable in this setting. A familiar example comes from the classical case–control design, where all cases of a rare disease are sampled together with approximately the same number of controls. Unweighted logistic regression is the standard analysis and is semiparametric efficient if the model is correctly specified (Prentice and Pyke, 1979; Breslow, Robins and Wellner, 2000). Weighted logistic regression using sampling weights would appear to be much less efficient, since very high weight is given to controls relative to cases. In fact, weighted logistic regression is fully efficient for saturated models, is locally efficient at the null hypothesis of zero covariate effects for any model, and is typically highly efficient for models with small numbers of discrete covariates (Lumley, 2010, chap 8). Substantial loss of efficiency does occur for continuous covariates having moderate to strong effects (Scott and Wild, 2002).

4. Simulations.

4.1. *Simulation set-up.* The code generating the simulated cohort populations is given in the Appendix. Each locus contains a marker SNP and a functional variant explaining the association at the marker. The marker has perfect specificity for the functional variant, but imperfect sensitivity; in the results presented here the functional variant is present in half the haplotypes containing the marker SNP. The results were qualitatively similar when we varied the sensitivity of the marker.

Each targeted locus also has rare variants. We assumed that these would be filtered using bioinformatic predictions of functional effects before analysis, and so the frequency of functional rare variants in the analysis could be quite high. The number functional rare variants were simulated from a compound Poisson–Poisson distribution with mean 0.25, and the number of non-functional rare variants from a compound Poisson–Poisson distribution with mean 0.5. The analysis used the total of the functional and non-functional numbers of variants as a mutation liability score. These simulations were targeted at the question of relative power and bias for different designs. They do not use a full population-genetic model for simulating haplotypes, and so may not give reliable estimates of absolute power.

Because bias from effects of correlated phenotypes was a concern for the case–cohort design and especially for the unweighted tests, the phenotypes were simulated to have moderately strong positive and negative correlations in a complex pattern. The iteration

$$x_{i+1} \leftarrow (x_i + \tau_i, \epsilon_{i1} - x_i + \tau_i, \epsilon_{i2})$$

creates 2^i phenotypes at stage i, where ϵ_{ij} are Normal or t_{15} and τ_i is a vector of scaling parameters. Our simulations used $\tau_i = 1$. The functional rare and common variants are simulated to have additive effects on the phenotype, with zero effects on half the phenotypes and non-zero effects on the other half. The correlation structure and pattern of null and strong non-null effects was chosen to be unfavorable for the unweighted tests, which tend to perform even better in other scenarios.

There were two major sampling designs. In the first design, with 20 phenotypes, binary extreme samples were taken as the highest and lowest 132 observations for a phenotype. In the generalized case–cohort design the cohort random sample of 1024 was taken first, and then individuals not already selected who had the 200 highest values for each phenotype. In the second design, with 4 phenotypes, the binary extreme samples were the 125 highest and lowest values of each phenotype. The generalized case–cohort design used a random reference cohort sample of 200 and four phenotype samples of 200, to give the same total sample size of 1000. The sampling probability was taken as 1 for individual in the phenotype samples. For those in the random reference cohort the probability was the empirical sampling probability: the number sampled who did not have extreme phenotype values divided by the number in the full cohort who did not have extreme phenotype values. Smoothing for stabilized weights used the supersmoother(Friedman, 1984).

In a design that samples both extremes of a phenotype there will be some overlap between samples for different phenotypes, allowing for either an increase in the sample size per phenotype or a reduction in total sequencing costs. The size of the overlap is sensitive to the size of the cohort and the correlation among phenotypes. We ignore this potential for overlap in the results presented here on sampling binary extremes.

4.2. *Results.* Figures 1, 2, 3, and 4 show boxplots of regression coefficients scaled by the simulation standard error for 4000 replications of a simulated study with 20 phenotypes. These boxplots show how the power and the control of Type I error vary across designs. Figures 1 and 2 are for a targeted variant; Figures 3 and 4 for rare variants. Figures 1 and 3 are for a Normal; Figures 2 and 4 for a heavy-tailed phenotype.

The first two pairs of boxes in each plot show a design than samples from both extremes of each phenotype. The remaining sets of boxes show weighted and unweighted

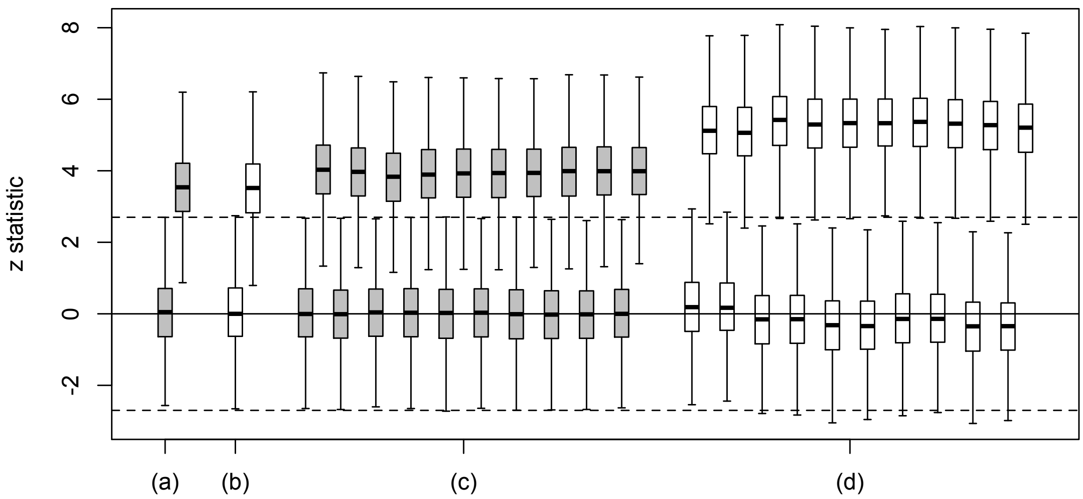

Fig 1. *Boxplots of z-statistics testing for effects of targeted variant on Normal phenotype: (a) binary extremes analysed as binary, (b) binary extremes analysed with quantitative phenotype, (c) case–cohort with weighted linear regression, (d) case–cohort with unweighted linear regression. The dashed lines are at* ±2.69*, the expected endpoints of a boxplot with Normal data*

tests in the generalized case-cohort design. The unweighted test in the generalized case-cohort design is always more powerful than the tests based on sampling two extremes, and the weighted test is more powerful in three of the four scenarios. Sampling from both extremes has lower relative power when the phenotype is heavy-tailed, as does the unweighted test. The unweighted test shows some deviation from the correct null distribution, but these are small compared to the effect size needed to induce the bias. Other simulations, not shown, confirm there is no bias under the strong null hypothesis of no genetic effects on any phenotype.

Figures 5 and 6 show boxplots of the regression coefficients from unweighted and weighted analyses of the generalized case–cohort design and from an analysis assuming sequence data is available for the whole cohort. The weighted estimates are unbiased for the whole-cohort estimates, and for the true simulation model. The unweighted estimates have small amounts of bias under the null, and for targeted variants where there is a true effect, but have substantial bias for rare variants where there is a true effect. These results are for a Normal phenotype; the same pattern of bias is seen with the heavy-tailed phenotype. Bias in the unweigted estimates is still small for functional variants when the functional variant allele frequency is reduced from 10% to 2.5% or to 1%.

Figure 7 shows results for a design with only four phenotypes. The generalized case–cohort design produces tests with lower power than sampling both extremes, though the loss of power is less for the heavy-tailed phenotype.

Figure 8 shows quantile–quantile plots of the null p-values. On the right is the weighted score test based on the two-phase design, which shows no deviations from the nominal

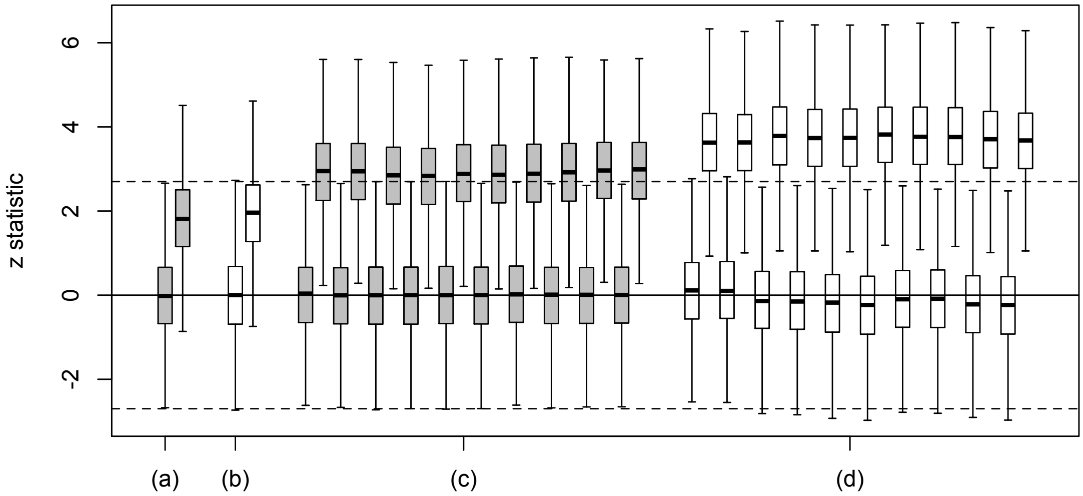

FIG 2. *Boxplots of z-statistics testing for effects of targeted variants on heavy-tailed phenotype: (a) binary extremes analysed as binary, (b) binary extremes analysed with quantitative phenotype, (c) case–cohort with weighted linear regression, (d) case–cohort with unweighted linear regression. The dashed lines are at* ±2.69*, the expected endpoints of a boxplot with Normal data*

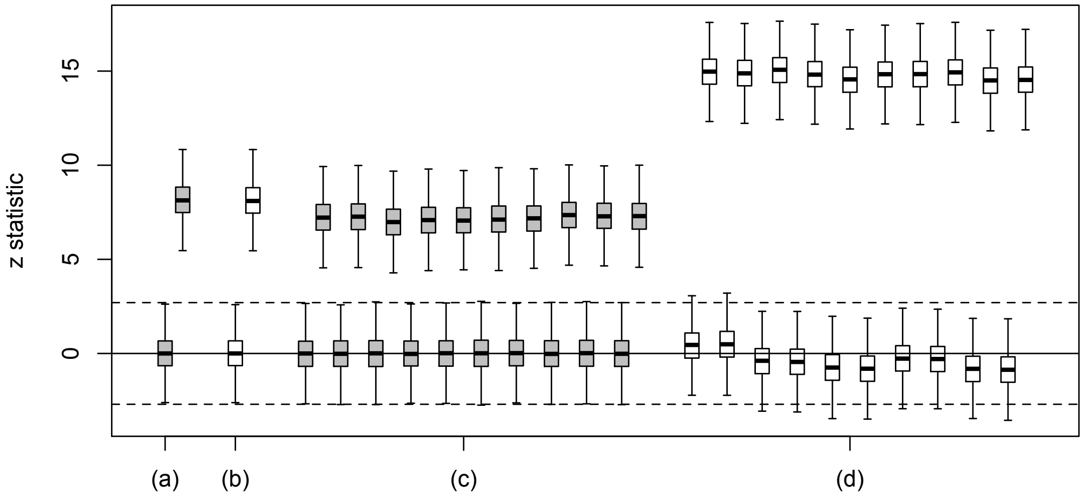

FIG 3. *Boxplots of z-statistics testing for effects of rare variants on Normal phenotype: (a) binary extremes analysed as binary, (b) binary extremes analysed with quantitative phenotype, (c) case–cohort with weighted linear regression, (d) case–cohort with unweighted linear regression. The dashed lines are at* ±2.69*, the expected endpoints of a boxplot with Normal data*

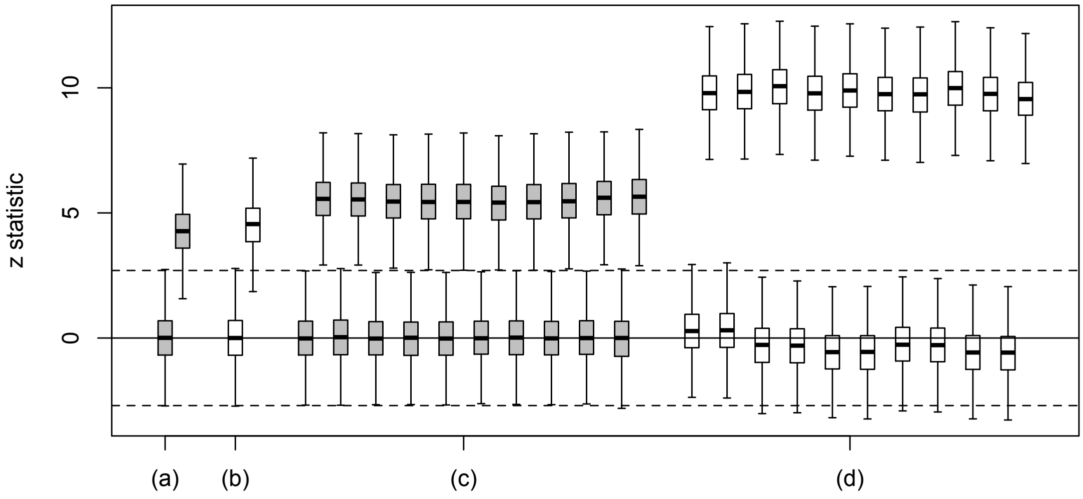

FIG 4. *Boxplots of z-statistics testing for effects of rare variants on heavy-tailed phenotype: (a) binary extremes analysed as binary, (b) binary extremes analysed with quantitative phenotype, (c) case–cohort with weighted linear regression, (d) case–cohort with unweighted linear regression. The dashed lines are at* ±2.69*, the expected endpoints of a boxplot with Normal data*

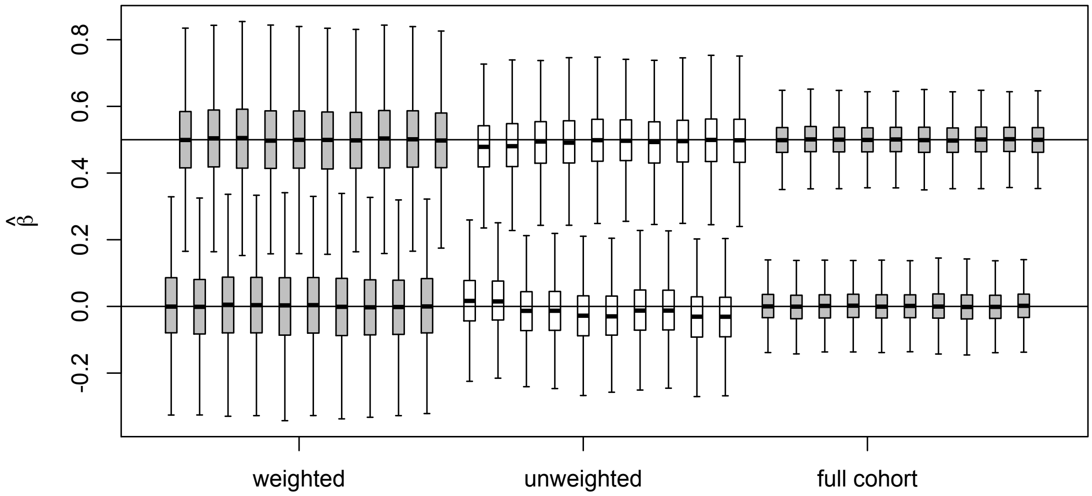

FIG 5. *Boxplots of $\hat{\beta}$ testing for effects of targeted variant on Normal phenotype: weighted estimates, unweighted estimates, estimates with sequence data on whole cohort*

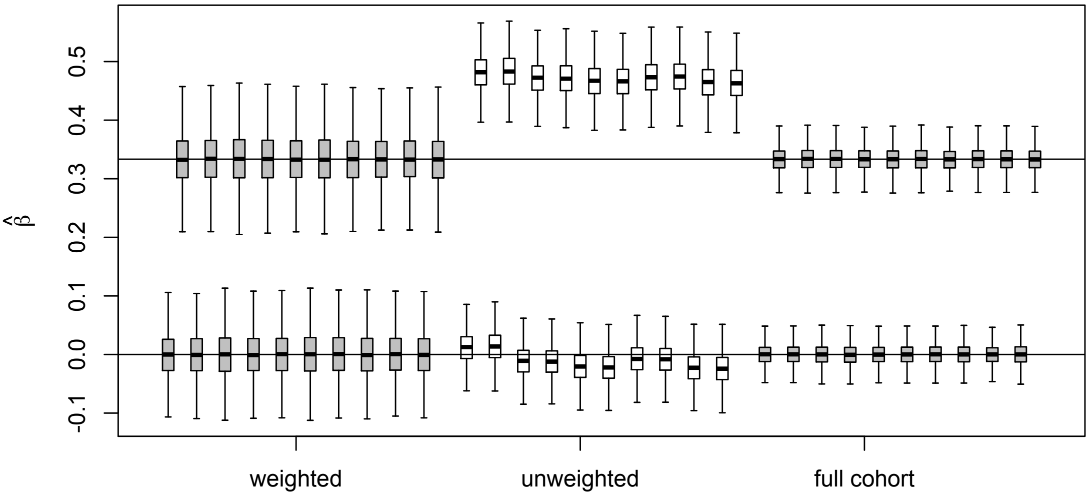

FIG 6. *Boxplots of $\hat{\beta}$ testing for effects of rare variants on Normal phenotype: weighted estimates, unweighted estimates, estimates with sequence data on whole cohort*

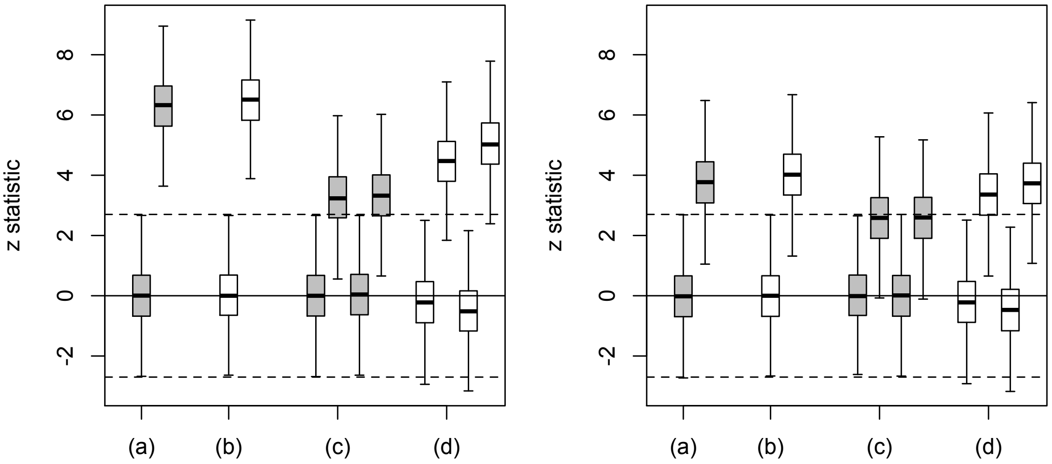

FIG 7. *Boxplots of z-statistics testing for effects of rare variants with four phenotypes. Left panel is Normal phenotype, right panel is heavy-tailed. (a) binary extremes analysed as binary, (b) binary extremes analysed with quantitative phenotype, (c) case–cohort with weighted linear regression, (d) case–cohort with unweighted linear regression. The dashed lines are at* ±2.69*, the expected endpoints of a boxplot with Normal data*

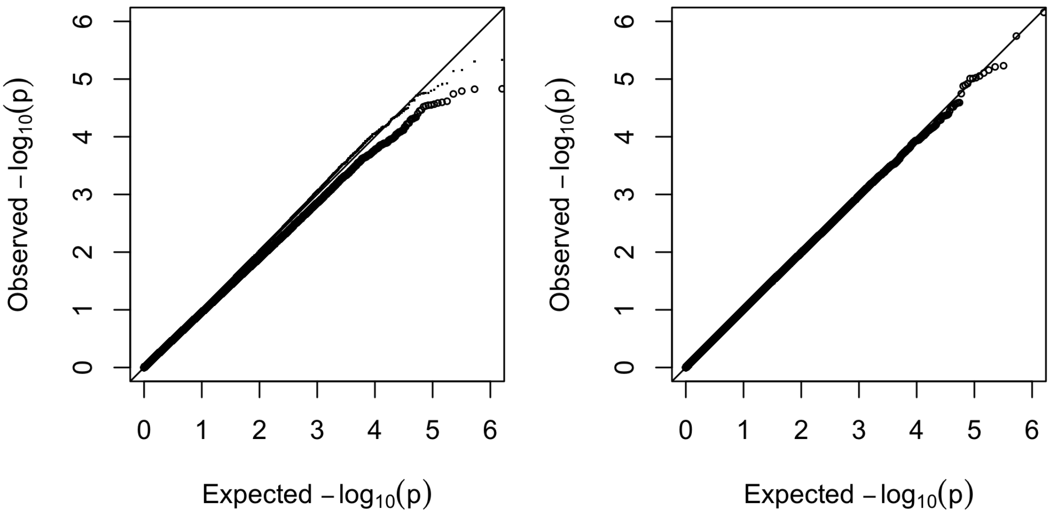

FIG 8. *Quantile-quantile plot of* 8×10^5 *p-values where the null hypothesis is true, on a logarithmic scale. Left panel is the unweighted test; circles indicate p-values using the (conservative) estimated standard error, dots indicate p-values using the true simulation standard error. Right panel is the weighted score test.*

distribution. On the left is the unweighted test. Despite the bias observed in the unweighted point estimates, the unweighted test is conservative, because the standard error is slightly conservative. We do not know whether the conservative standard error estimate is a general phenomenon, but as the second set of dots on the qq-plot shows, the test is not importantly anti-conservative even if the true simulation standard error is used.

Figures 9 compares the weighted and unweighted estimates to those using stabilized weights, for a targeted variant. The stabilized weights give unbiased estimates under the null, and biased estimates under the alternative. The net effect is an increase in power with no adverse effects on Type I error, but the increase in power is very small. The pattern with smaller numbers of phenotypes is similar; the gain in power is slightly greater for rare variants.

The weighted estimates are unbiased and the weighted tests have the correct level in all the scenarios we examined. The weighted tests and case–cohort design have more power in some situations and less power in other situations than a test and design based on sampling only from the extremes, with the case–cohort design performing better for larger numbers of phenotypes, for phenotypes with outliers, and for rare variants. The unweighted tests consistently have higher power than the weighted tests, and have higher power than a test based on sampling just the extremes. The unweighted tests always have correct Type I error when there are no true effects, but they may have unacceptable Type I error rates when the number of phenotypes is small and outliers are present.

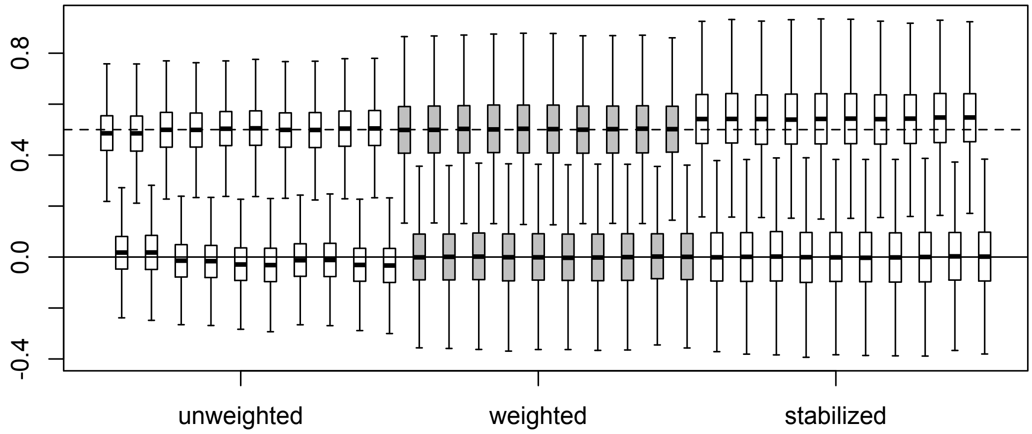

FIG 9. *Boxplots of regression coefficients for targeted variants, with 20 phenotypes: unweighted, weighted, and using stabilized weights. Dashed line indicates true regression coefficient.*

4.3. *Sampling on genotype.* As sampling on genotype makes sense only for the tail sample for a particular phenotype, not for the reference sample or the tail samples for other phenotypes, the most favorable scenario is a single phenotype and a single locus. We simulated cohorts of size 12000 with a Normal phenotype, and took a reference sample of 1000 and a phenotype tail sample of 200. Genotype data were simulated in the same way as in the previous section.

We compared the strategy of sampling the 200 highest values of the phenotype with two strategies using marker information. The first 'balanced' strategy sampled the highest values of the phenotype for each of the three values of the marker genotype, dividing the sample size as equally as possible. The second 'marker first' strategy sampled all individuals with two copies of the marker, then made up the remainder of the sample size of 200 using the highest phenotype values among those with one copy of the marker.

Since these sampling strategies induce an explicit relationship between phenotype and genotype, unweighted tests cannot be used. Weighted estimates were effectively unbiased, as expected, with all three sampling strategies. Table 1 shows simulation standard errors for point estimates from the three sampling strategies, for a range of minor allele frequencies. Even in the setting of a single phenotype and single locus, the reduction in standard errors from sampling based on genotype is modest except at the lowest minor allele frequencies. This is primarily because the reference sample makes up most of the data and will contribute more copies of a relatively common variant than are added by the targeted sample.

In a practical design with multiple phenotypes and multiple loci the benefit would be much smaller, because selection for any one phenotype and locus would be used for a smaller proportion of the data. In addition, it is not possible to use unweighted testing for the genotype-based designs. In the context of CHARGE-S, only 1.6% of

Table 1

Standard errors for effect of functional variant, comparing sampling based on phenotype alone to sampling based on phenotype and marker genotype, for a single phenotype and single locus. Monte Carlo error based on 5000 simulations is less than one unit in the last place quoted.

MAF		Sampling strategy		
Marker	Functional	Phenotype only	Balanced	Marker first
0.1	0.05	0.099	0.093	0.089
0.1	0.09	0.075	0.070	0.068
0.05	0.025	0.14	0.12	0.11
0.05	0.045	0.10	0.09	0.09
0.01	0.005	0.32	0.25	0.19
0.01	0.009	0.23	0.18	0.15

the samples would be targeted to each genotype:phenotype combination, leading to a negligible increase in power relative to a weighted analysis of a phenotype-only sampling design and a substantial loss in power relative to an unweighted analysis of a phenotype-only sampling design.

5. Example. The final CHARGE–S design used 20 extreme-phenotype samples. For phenotypes with more than three GWAS hits, all of which were in coding regions, it was more cost-effective to sequence the entire coding set of the genome (whole-exome sequencing) rather than designing capture sequences for targeted resequencing. Seven phenotypes used whole-exome sequencing, with a reference sample of 1000. Thirteen phentoypes used targeted sequencing, with a reference sample of 2000 that included the 1000 used for the whole-exome reference sample. Approximately 2 megabases of the genome was sequenced for the targeted loci; 18% of the sequencing for targets was in exons, where the whole-exome samples will also be available for use in the analysis.

Table 2 and 3 describe the phenotype samples for whole-exome and targeted sequencing. Phenotype working groups were given considerable flexibility in selecting samples, but for the majority of phenotypes, the sampling was 100 from ARIC, 50 from CHS, and 50 from FHS, which is roughly in proportion to cohort size. The main reason for varying from this plan was that a phenotype was not measured in one of the cohorts. Most phenotypes were sampled at one extreme; a few were sampled at both extremes.

6. Discussion. The strategy of taking a probability sample from the available full cohort data allows unbiased estimation and also allows sharing a common reference group between phenotypes. When the number of phenotypes is large, there are power gains compared to a strategy that samples both extremes of each phenotype. Importantly for the CHARGE Consortium, using a generalized case–cohort sample also allows estimation of the effects of sequence variants on the hundreds of other phenotypes available in the CHARGE cohorts. A large number of phenotypes also makes it much more challenging to follow an alternative strategy of maximum likelihood estimation which would require a joint probability model for twenty phenotypes.

Weighted estimation is straightforward and estimates the same parameters as would be estimated with complete cohort data, regardless of the sampling scheme. In some

Table 2

Selection of phenotype samples for targeted sequencing. 'Design N' is the number sampled specifically for this phenotype, 'Achieved N' adds in samples that were extreme on this phenotype but sampled for other phenotypes. 'Threshold' is the lowest or highest value in the sample, not the residual threshold used in selection. (F,C,A) indicates (Framingham, CHS, ARIC).

Phenotype	strategy	Design N	Achieved N	Threshold
		(F,C,A)	(F,C,A)	(F,C,A)
EKG PR interval	high resid	(50,50,100)	(64,50,109)	(188, 220,212)
EKG QRS interval	high resid	(50,50,100)	(59,72,114)	(110.5, 104,112)ms
Stroke	early	(50,70,80)	(42,98,173)	(?97.6,68.5,70)y
Blood pressure	low resid	(25,25,54)	(34,26,388)	(110/65, 128/58,124/63)
	high resid	(25,25,46)	(36,26,184)	(155/97,172/82,140/88)
Body Mass Index (BMI)	high	(50,50,100)	(65,51,135)	(35.43,32.5,39.0)
Fasting insulin	high	(50,50,100)	(66,61,182)	(33,31,29)
Bone mineral density by DEXA	low z-score	(100,100,0)	(109,103,)	(−1.6,−0.73,—)
Left ventricular diastolic diameter	high resid	(50,100,0)	(77,116,0)	(5.13,5.27,—)
C–reactive protein	high resid	(50,50,100)	(63,56,201)	(5.6,10.35,10.5)
Hematocrit	low resid	(50,50,100)	(59,72,260)	(41,37.9,36.9)
Retinal venule diameter	high	(0,34,166)	(0,44,220)	(—,216.9,231.8) CRVE
Carotid wall thickness	high	(50,50,100)	(63,63,276)	(0.915,1.44,1.01)mm
Pulmonary: $\mathrm{FEV}_1/\mathrm{FVC}$	low	(0,0,200)	(0,0,618)	(—,—,69.5)

Table 3

Selection of phenotype samples for whole-exome sequencing. 'Design N' is the number sampled specifically for this phenotype, 'Achieved N' adds in samples that were extreme on this phenotype but sampled for other phenotypes. 'Threshold' is the lowest or highest value in the sample, not the residual threshold used in selection. (F,C,A) indicates (Framingham, CHS, ARIC). The multivariate metabolic phenotype is the first principal component of a set of 'metabolic syndrome' variables.

Phenotype	strategy	Design N	Achieved N	Threshold
		(F,C,A)	(F,C,A)	(F,C,A)
Fibrinogen	high resid	(50,50,100)	(58,60,121)	(523, 436,448)
Menopause	early	(50,50,91)	(55,53,254)	(44,35,44)y
EKG QT interval	high resid	(50,50,100)	(57,64,109)	(461.5,384,382)ms
Fasting blood glucose	high but non-diabetic	(50,50,100)	(65,65,267)	(84,120,117)mg/dl
Waist:hip ratio	high	(50,50,100)	(63,68,560)	(0.90,1.03,0.98)
Multivariate metabolic	high and low	(50,50,100)	(61,81,116)	NA
Kidney function (eGFR)	low resid	(50,100,50)	(59,141,459)	(47.4,43.0,49.6)ml/min

circumstances an unweighted analysis will give an approximately valid test of the null hypothesis of no genetic association, and is often, but not always, usefully more powerful. When sampling is based on phenotype the unweighted analysis appears to perform better with large numbers of phenotypes, and to perform worse when the phenotype has outliers that are extreme for non-genetic reasons. In situations where the unweighted test does not perform well, the use of 'stabilized weights' appears to provide a useful gain in power. Since the justification for the unweighted tests rests on simulations, the tests should be used with caution in scenarios that vary significantly from the simulations. In particular, the behavior of the tests for p-value thresholds much lower than those planned for CHARGE-S has not been established.

Sampling based on the combination of phenotype and genotype at a marker makes it impossible to use unweighted tests, and does not provide much increase in power for weighted tests when the number of loci being studied is large; it may be more useful in studying gene:environment interactions. We did not investigate calibration of weights using the full-cohort marker data (Breslow *et al.*, 2009), because the plan for the CHARGE-S study is to use the unsampled cohort participants to replicate findings by genotyping promising variants.

DNA sequences contain many rare variants, so purely statistical methods will be inadequate. Biological information is needed to decide which variants are most promising for analysis. After associations are found, further biological investigation will be needed to decide which variants from a set that show association are actually functional. Increased precision and power in estimating associations, however, will make both these tasks easier.

Acknowledgements. The CHARGE-S sequencing project is funded by NHLBI award 1RC2HL102419-01 under the American Recovery and Reinvestment Act.

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022 and R01HL087641; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contracts HHSN268201100005C through HHSN268201100012C . The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

The CHS research reported in this article was supported by NHLBI contract N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, N01-85239, and grants HL075366, HL080295, HL087652, and HL105756 with additional contribution from the NINDS. Additional support from the NIA was provided through AG-023269, AG-15928, AG-20098, and AG-027058. A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.htm. DNA handling and genotyping was supported in part by National Center for Research Resources grant M01-RR00425 to the Cedars-Sinai General Clinical Research Center Genotyping core and National Institute of Diabetes and Digestive and Kidney Diseases grant DK063491 to the Southern California Diabetes Endocrinology

Research Center.

This work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195) and its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278), and by grants from the National Institute of Neurological Disorders and Stroke (NS17950; PAW) and the National Institute of Aging, (AG08122, AG16495; PAW). Analyses reflect intellectual input and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project.

References.

Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin* **2** 47–53.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51** 279–292.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms' tumor prognosis. *Applied Statistics* **48** 457–68.

Breslow, N. E., Robins, J. M. and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6** 447–455.

Breslow, N. E. and Wellner, J. A. (2008). A Z-theorem with estimated nuisance parameters and correction note for 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression'. *Scand J. Statist* **34** 186–192.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E. and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* **169** 1398–1405.

Friedman, J. H. (1984). A variable span scatterplot smoother. Technical Report No. 5, Laboratory for Compuational Statistics, Stanford University.

Guey, L. T., Kravic, J., Melander, O., Burtt, N. P., Laramie, J. M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., Nilsson, P., Almgren, P., Kathiresan, S., Groop, L., Seymour, A. B., Altshuler, D. and Voigh, B. F. (2011). Power in the Phenotypic Extremes: A Simulation Study of Power in Discovery and Replication of Rare Variants. *Genetic Epidemiology* **35** 236–246.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics* **35** 1491–1523.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

Krewski, D. and Rao, J. N. K. (1981). Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics* **9** pp. 1010-1019.

Lee, C. R., North, K. E., Bray, M. S., Couper, D. J., Heiss, G. and Zeldin, D. C. (2007). CYP2J2 and CYP2C8 polymorphisms and risk of cardiovascular events: the Atherosclerosis Risk in Communities (ARIC) study. *Pharmacogenet. Genomics* **17** 349–358.

Lee, C. R., North, K. E., Bray, M. S., Couper, D. J., Heiss, G. and Zeldin, D. C. (2008). Cyclooxygenase polymorphisms and risk of cardiovascular events: the Atherosclerosis Risk in Communities (ARIC) study. *Clin. Pharmacol. Ther.* **83** 52–60.

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., Morrison, A. C., Johnson, A. D., Aspelund, T., Aulchenko, Y., Lumley, T., Kottgen, A., Vasan, R. S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D. E., Mitchell, G. F., Mattace-Raso, F. U. S., Smith, A. V., Taylor, K., Scharpf, R. B., Hwang, S.-J., Sijbrands, E. J. G., Bis, J., Harris, T. B., Ganesh, S. K., O'Donnell, C. J., Hofman, A., Rotter, J. I., Coresh, J., Benjamin, E. J., Uitterlinden, A. G., Heiss, G., Fox, C. S., Witteman, J. C. M., Boerwinkle, E., Wang, T. J., Gudnason, V., Larson, M. G., Chakravarti, A., Psaty, B. M. and van Duijn, C. M. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics* **41** 677–687.

Lin, D. Y. (2000). On Fitting Cox's Proportional Hazards Models to Survey Data. *Biometrika* **87** 37–47.

Lin, D. Y. and Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American Journal of Human Genetics* **89** 354–367.

Lumley, T. (2010). *Complex surveys: a guide to analysis using R.* John Wiley and Sons, Hoboken, NJ.

Nambi, V., Hoogeveen, R. C., Chambless, L., Hu, Y., Bang, H., Coresh, J., Ni, H., Boerwinkle, E., Mosley, T., Sharrett, R., Folsom, A. R. and Ballantyne, C. M. (2008). Lipoprotein-Associated Phospholipase A2 and High-Sensitivity C-Reactive Protein Improve the Stratification of Ischemic Stroke Risk in the Atherosclerosis Risk in Communities (ARIC) Study. *Stroke.* Dec 18 (online).

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankyha, Series B* **61** 166-186.

Prentice, R. L. (1986). A Case-cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika* **73** 1–11.

Prentice, R. L. and Pyke, R. (1979). Logistic Disease Incidence Models and Case-control Studies. *Biometrika* **66** 403–412.

Psaty, B. M., O'Donnell, C. J., Gudnason, V., Lunetta, K. L., Folsom, A. R., Rotter, J. L., Uitterlinden, A. G., Harris, T. B., Witteman, J. C. M. and Boerwinkle, E. (2009). Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from five cohorts. *Circulation Cardiovascular Genetics* **2** 73–80.

Robins, J. M., Hernán, M. A. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

Scott, A. and Wild, C. (2002). On the Robustness of Weighted Methods for Fitting Models to Case-control Data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64** 207–219.

Sen, P. K. (2009). *Handbook of Statistics* **29A** Asymptotics in finite population sampling. Elsevier.

APPENDIX A: SIMULATION DETAILS

```
make.full.cohort<-function(samplesize=12000, phenotypes=16,maf=0.1,
spurious=0.1,tau=c(1,1,1,1,1),beta=rep(0.25,phenotypes),betarare=beta*2){

  pheno<-rt(samplesize,df=15)*tau[1]
  n<-1; i<-2;
  while(n<phenotypes){
    pheno<-cbind(pheno+rt(n*samplesize,df=15)*tau[i],
                 -pheno+rt(n*samplesize,df=15)*tau[i])
    n<-ncol(pheno)
    i<-i+1
  }

  commonvariant <- rbinom(samplesize, 2, maf)
  rarefvariants<-rpois(samplesize, rpois(samplesize,0.25))
  rarenvariants<-rpois(samplesize, rpois(samplesize,0.5))
  marker<- commonvariant+rbinom(samplesize, 2-commonvariant, spurious)

  pheno<-pheno[,1:phenotypes]+outer(commonvariant,beta)
  pheno<-pheno[,1:phenotypes]+outer(rarefvariants,betarare)
  data.frame(v1=commonvariant,vr=rarefvariants+rarenvariants,
            snp=marker, pheno=pheno)
}
```

Department of Statistics
The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand
E-mail: t.lumley@auckland.ac.nz

Department of Biostatistics
Boston University School of Public Health
E-mail: dupuis@bu.edu

Department of Biostatistics
University of Washington
Box 357232
Seattle, WA 98195-7232
E-mail: kenrice@uw.edu

Human Genetics Center
University of Texas Health Science Center
1200 Herman Pressler
Houston, TX, USA
E-mail: maja.barbalic@uth.tmc.edu

Cardiovascular Health Research Unit
University of Washington
E-mail: joshbis@uw.edu

Department of Biostatistics
Boston University School of Public Health
E-mail: adrienne@bu.edu

Cardiovascular Health Research Unit
University of Washington
E-mail: psaty@uw.edu

NHLBI Framingham Heart Study
Framingham
Massachusetts 01702, USA.
E-mail: odonnellc@nhlbi.nih.gov

Human Genetics Center
University of Texas Health Science Center
1200 Herman Pressler
Houston, TX, USA
E-mail: Eric.Boerwinkle@uth.tmc.edu