Potential for Revealing Individual-Level Information in Genome-wide Association Studies

Thomas Lumley; Kenneth Rice

JAMA. 2010;303(7):659-660 (doi:10.1001/jama.2010.120)

http://jama.ama-assn.org/cgi/content/full/303/7/659

Potential for Revealing Individual-Level Information in Genome-wide Association Studies

Thomas Lumley, PhD
Kenneth Rice, PhD

GENOME-WIDE ASSOCIATION STUDIES[1] GENERATE large volumes of results. While the strongest signals are the focus of most reports, full online publication of thousands or millions of association results has been encouraged.[2] These aggregate results may be valuable for future scientific research,[3] but analysts have recently shown[4-7] that the aggregate results actually may reveal information about participants. Specifically, if study participants' genetic information is available, large-scale reporting of population-level variant-disease associations enables easy reconstruction of individuals' disease states.

That individual-level information can be obtained from aggregate association results may be surprising. In fact, high precision individual data may be achieved. For example, in a study of 1000 cases and 1000 controls, reporting separate variant × disease counts for 5000 variants could enable anyone who knew a study participant's genotype to determine his or her disease status with 99% sensitivity and specificity.[6,7] Aggregate regression results also may reveal information; reporting the default (additive model) odds ratios for 10 000 variants and disease gives the same prediction accuracy.

The phenomenon is not limited to binary disease states (ie, either having or not having a disease). In a typical recent genome-wide association study,[8] associations between left ventricular mass and 2.5 million genetic variants were studied in 12 612 individuals. A full report of these associations would provide 2.5 million regression estimates. With data from 1 variant, the corresponding regression estimate can be used to give a very weak prediction of a participant's left ventricular mass; for example, multiplying the regression coefficient by a person's number of copies of the variant gives a weak prediction of how far his or her disease state is from the sample average. With 2.5 million variants, the average of these predictions yields a very precise determination of an individual's left ventricular mass.

The FIGURE illustrates the phenomenon, giving within-sample predictions of left ventricular mass based on just 35 000 variants. The correlation between predicted and measured left ventricular mass is 0.86, a value typically seen in test-retest variability of left ventricular mass measurement. In other words, for predicting the original measurement, using genotypes and aggregate results performs at least as well as obtaining another actual echocardiogram.

Because the clinical predictive ability of common genetic variants has been disappointing,[1] it may seem paradoxical that individual outcomes can be determined so well from a set of association results. However, the targets of prediction are different. Clinicians want to predict disease in new patients; the genetic data are being used to reconstruct the original disease status used in the published regression estimates. In other applications of predictive models, this distinction between in-sample and out-of-sample prediction is well known and has motivated bias-correcting techniques such as replication, cross-validation, and resampling.[9]

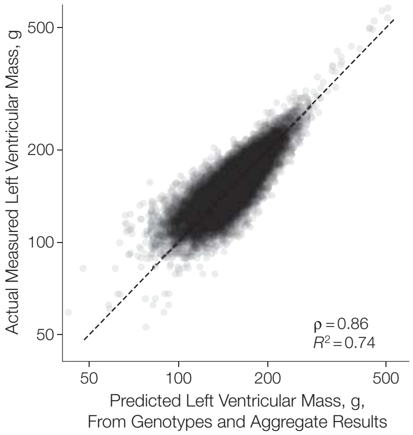
The ability to infer the disease states of genome-wide association studies' participants raises important issues of consent. Even if participants have agreed to the release of their genetic data, typically they will not have consented to the release of sensitive disease information. In this situation, publishing aggregate results for thousands of variants could be seen as breaching the limits of consent. Problems also arise when genetic data are not public. If a participant has used 1 of the increasing number of commercial genotyping services,[10] naive publishing of aggregate results could disclose his or her disease information to that third party.

Given these concerns, publishing complete genome-wide aggregate results is not safe. Because genotypes at a few thousand independent variants appear to be the minimum data required for accurate prediction, reporting only the highly significant results from most genome-wide association studies will typically not disclose disease information. However, the compromise position of publishing all associations that reach intermediate levels of significance (such as $P<10^{-3}$) will often allow unacceptably accurate predictions.

In the post–genome era, traditional boundaries between individual and aggregate data have become blurred. Although further work is required to identify situations in which individual information will be disclosed, as an interim measure, the current authors recommend that genome-wide

Author Affiliations: Department of Biostatistics, University of Washington, Seattle.
Corresponding Author: Kenneth Rice, PhD, Department of Biostatistics, University of Washington, F-600 Health Sciences Bldg, Box 357232, Seattle, WA 98195-7232 (kenrice@u.washington.edu).

Figure. Predicted and Observed Measurements of Left Ventricular Mass



To construct the predictions based on a hypothetical cohort of 12 612 individuals, additive models were fitted for each of 35 000 variants. The product of each regression coefficient and number of copies of each variant gives subject × variant predictors. Averaging these for each subject gives an overall prediction score. Finally, predictions of left ventricular mass are constructed by scaling these prediction scores to match the 25% and 75% percentiles of the observed left ventricular mass measurements. Thirty-five thousand independent variants were used, each with minor allele frequency 20%. Greater accuracy could be obtained with more variants. In line with the findings of Visscher and Hill,[7] these results are not sensitive to the assumed minor allele frequency. The dashed diagonal line indicates perfect prediction.

association studies ordinarily publish no more than 500 regression results. This will typically be sufficient to describe novel signals, attempted replications of prior findings, and suggestive associations with biologically motivated candidate genes. In the longer term, all those who contribute to research will need to agree on new standards for release of potentially sensitive data, through direct or indirect means. These contributors include study participants, clinicians, investigators, journal editors, and funding agencies.

Financial Disclosures: None reported.

Funding/Support: This research was supported in part by grants U01 HL080295 and R01 HL087652 from the National Heart, Lung, and Blood Institute.

Role of the Sponsor: The National Heart, Lung, and Blood Institute had no role in the preparation, review, or approval of this commentary.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

Additional Contributions: We thank Bruce M. Psaty, MD, University of Washington Cardiovascular Heath Research Unit, Seattle, for his helpful comments. Dr Psaty received no remuneration.

REFERENCES

1. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299(11):1335-1344.

2. Hunter DJ, Kraft P. Drinking from the fire hose—statistical issues in genome-wide association studies. *N Engl J Med*. 2007;357(5):436-439.

3. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. *BMC Med Genet*. 2009;10:6.

4. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*. 2008;4(8):e1000167.

5. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. *Nat Genet*. 2009;41(9):965-967.

6. Jacobs KB, Yeager M, Wacholder S, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat Genet*. 2009;41(11):1253-1257.

7. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis [published online October 2, 2009]. *PLoS Genet*. 2009;5(10):e1000628. doi:10.1371/journal.pgen.1000628.

8. Vasan RS, Glazer NL, Felix JF, et al. Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *JAMA*. 2009;302(2):168-178.

9. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York, NY: Springer-Verlag; 2001.

10. McGuire AL, Burke W. An unwelcome side effect of direct-to-consumer personal genome testing: raiding the medical commons. *JAMA*. 2008;300(22):2669-2671.