Association Studies in an Era of Too Much Information
Clinical Analysis of New Biomarker and Genetic Data

Joseph Loscalzo, MD, PhD

In the past century, the broad use of routine diagnostic assays in the clinical laboratory was a major advance that remains a core element of contemporary patient care. As quantitative, objective measurements, these routine clinical tests often provide unequivocal evidence for a diagnosis or help frame a therapeutic strategy. The development of specific laboratory tests is generally predicated on a plausible biological hypothesis grounded in (pre)clinical experimentation. As a result, the majority of routine laboratory tests are limited in number and sufficiently well understood by practitioners to be clinically useful.

In the past few years, the biomedical community has witnessed an increasing number of reports of associations between a wide range of newer molecular markers and disease phenotypes (pathophenotypes) or outcomes. These biomarkers—here broadly defined as molecular indicators of the presence of a disease, outcome, or response to therapy—including conventional laboratory measurements of proteins or enzymes, recently identified genetic polymorphisms (either singly [single nucleotide polymorphisms, or SNPs] or in combinations [haplotypes]), and plasma metabolites, peptides, or proteins, many with as yet unknown function. In contrast to the earlier era of biochemical testing, these contemporary biomarkers are generally not identified by hypothesis-driven, preclinical experimentation and the application of rational, reductionist principles; rather, they reflect the output of the brute-force forms of discovery science, viz., comparative total genome scanning, global proteomic analysis, and comprehensive metabolomic testing between populations with and without a well-defined disease phenotype. With the use of technically sophisticated molecular methods, the entire human genome or the plasma proteome, for example, is analyzed in some detail to identify changes in gene sequence or protein patterns, respectively, that are more or less prevalent in those with disease than in those without disease. The magnitude of the observed difference in prevalence of the biomarker between the 2 populations defines the strength of the association. Because of the extraordinary number of comparisons made between the 2 often large populations (eg, 500 000 SNPs in the human genome of 17 000 subjects in a recent genome-wide association study), modest differences in prevalence can achieve startlingly high statistical significance, even after adjustment for multiple comparisons.

It is instructive to review briefly the history of these associations between laboratory test results and disease risk because this history has particular relevance to cardiovascular medicine. The concept of a cardiovascular risk factor was first described in the Framingham Heart Study in 1961. Before that time, Framingham investigators had published a series of reports showing an association between hypercholesterolemia, hypertension, and left ventricular hypertrophy and the prevalence of coronary heart disease. In 1961, the investigators hypothesized and demonstrated that the presence of these risk factors in otherwise healthy individuals followed up for 6 years predicted incident coronary heart disease—ie, these factors put the subjects at risk for developing disease. In this way, the concept of disease risk factors was born, and the impetus for exploring the potential causative mechanisms underlying the association and for developing approaches to risk reduction took hold.

As the number of cardiovascular risk factors increased over the next 30 years, statisticians and epidemiologists recognized that they could not and should not be viewed in isolation: They often interact synergistically with one another to increase risk. In many cases, causative mechanisms are proposed to account for this associative synergy, eg, hypertension causes left ventricular hypertrophy, and smoking increases the oxidative modification of LDL, thereby enhancing its uptake by the macrophage scavenger receptor to promote foam cell formation. The growing number of cardiovascular risk factors and the increasing appreciation of their associative and, in some cases, causal synergies led investigators to attempt to ascertain the magnitude of the risk of developing disease that could be accounted for by them—ie, the attributable risk. Estimates of the cumulative attributable risk of the major cardiovascular risk factors taken together range from 50% to 75%.

The ability of conventional risk factors to define the preponderance of attributable risk was illustrated in a recent study by Wang and colleagues, who showed that the use of 10 contemporary biomarkers, most of which have possible causal links to coronary heart disease development, added only moderately to risk prediction models that used conventional risk factors alone. These investigators used the c statistic (the probability of concordance among individuals who can be compared) to classify risk, and they depicted risk prediction using receiver operator characteristic (ROC) curves for models with and without the 10 newer biomarkers added to conventional risk factors. Cook has criticized this...
analytical approach, arguing that the $c$ statistic, the area under the ROC curve, was developed to assess the utility of a laboratory test in discriminating diseased individuals from nondiseased individuals, not for prospective risk assessment. Predictive models rely not only on discrimination, which the $c$ statistic handles quite appropriately, but also calibration, which is a measure of the agreement between predicted probabilities and observed risk. Because the outcome has not yet occurred at the time of assessment of the predictors in a prognostic risk model, future disease development can only be estimated as a probability.\(^8\) For this reason, Cook provides some general guidelines for comparisons of models for risk prediction, which include the use of a sensitive determinant of the fit of the model, application of measures of both discrimination ($c$ statistic) and calibration (eg, Hosmer-Lemeshow statistic), and ascertainment of whether the new biomarker would lead to different treatment choices.

The importance of determining the attributable risk of conventional or newer biomarkers rests on defining the most effective targets for risk reduction. The synergism among conventional risk factors in promoting risk suggests that there may be synergistic benefit to the moderate, contemporaneous reduction of these risk factors that exceeds the benefit of reducing any single risk factor dramatically. This approach has yet to be broadly proven prospectively, but it remains a useful construct in developing cost-effective, safe, preventive strategies for large segments of the global population.\(^9\)\(^,\)\(^10\)

Despite estimates that the bulk of the attributable risk of disease is determined by conventional risk factors, the scientific cardiovascular community continues to search for a comprehensive, holistic set of the potential determinants of coronary heart disease. The reasons for this search are at one level quite logical and, in my opinion, justifiable from 2 complementary perspectives. First, there may be a key, as yet unrecognized determinant (or determinants) of risk that overwhelmingly influences disease expression by serving as a common mechanistic link to the conventional risk factors. If this common risk determinant can be identified and modified, it may provide a simpler target for preventive strategies than currently exists. Second, as the practice of molecular medicine becomes more personalized, defining the weighted balance of determinants of disease risk in subpopulations or communities, if not in individuals, will provide a means through which to target preventive prescriptions most effectively. These arguments are theoretical and may simply be rationalizations; however, they should be viewed as having merit until they are shown to be of no value in refining the modifiable attributable risk burden of a population or an individual.

A less compelling reason to continue the search for modifiable risk factors is that with the newer tools of whole-genome scans and comprehensive proteomic analyses, we have within our grasp the distinct possibility of identifying all possible intrinsic determinants of disease risk. This goal is, after all, the great promise of sequencing the human genome and all of its variants. Now that we have available to us the methods to do so, we are scientifically obligated to sort out these variants and their potential associations with human disease. Although this is a sound guiding principle, we must test it against the reality of practical utility so as to manage the expectations of the scientific community and the public thoughtfully and fairly.

Whereas I and many others believe that understanding the genome in precise molecular detail will ultimately give us truly unique insight into disease risk and pathogenesis, my review of the growing body of genome-wide association studies does not convince me that this goal is likely to be realized soon. For example, in the genome-wide association study of 7 common diseases in 17 000 subjects in the United Kingdom,\(^1\) an analysis of 500 000 SNPs in each individual identified some new and confirmed some previously recognized loci associated with coronary disease and diabetes, but failed to identify any SNP as being associated with hypertension below the predefined significance threshold. In addition, none of the genetic variants reported to be associated with hypertension in prior studies showed evidence for association in this analysis. There are several possible explanations that the authors offer for these results, which include the presence of fewer common risk alleles in hypertension of larger effect size than found in other complex phenotypes, phenotype misclassification bias, and the inability to detect genuine common susceptibility variants because of their poor tagging by the set of SNPs chosen.\(^1\) Very recent data on the complete diploid genome sequence of a single individual offer some support for this last point because DNA variation is more extensive than previously suggested, perhaps by a factor of 5 (ie, humans differ in DNA sequence by 0.5%, not 0.1%).\(^11\)

The United Kingdom study\(^1\) highlights some of the serious methodological, statistical, and interpretive challenges that face genome-wide association studies (and, by analogy, comprehensive proteomic and blood-based biomarker studies). First, are the markers chosen truly comprehensive, and, if not, what is the likelihood that a key, linked, polymorphic locus will fail to be detected? In the case of whole-genome studies, is the density of SNPs chosen sufficient to identify the likeliest disease-associated polymorphisms? Remember that although 500 000 SNPs may seem like a large number, there are 3.2 billion base pairs in the human genome, indicating that less than 0.02% of the genome is specifically assessed with this marker panel, which leaves a lot of genetic distance unaccounted for, on average, between SNPs. Genetic epidemiologists address this issue by noting the statistical association among groups of SNPs (ie, haplotypes): These evolutionary associations of polymorphisms within so-called haplotype blocks suggest that the identification of a few of the SNPs within the block can unambiguously identify all associated SNPs without the need to measure them directly. On the basis of this statistical assumption and the number of haplotype blocks estimated to compose the human genome, geneticists argue that a 500 000-SNP scan should provide coverage of $\approx$90% of the genome. The latest estimates suggest that there are $\approx$15 million SNPs within the human genome, and it is important to point out that some of these may affect phenotype independently of a haplotype association. This point, coupled with the recent study by Levy and colleagues\(^11\) on variability (which may influence the expression of disease pheno-
type), indicates that even denser genome scans will be necessary to optimize the likelihood of discerning associations with disease phenotypes.

Second, how does one define a statistically significant association given the large number of comparisons made? For example, for a chip that contains 500,000 SNPs, none of which has different frequencies in the 2 populations of individuals, one would expect 5000 (500,000/1000) of them to have a probability value ≤0.01 simply as a result of the play of chance. How, then, can one adjust for multiple testing and control the false-discovery rate? There are several methods to do so, including a sequential Bonferroni-type correction, but none has yet risen to the level of a global standard. As more large genome-wide association studies are published, their datasets should provide a rich source of information for statisticians to use to develop the optimal methodology to limit the false-discovery rate and minimize type I errors.

Third, how does one correct for selection bias? Most investigators argue that this correction can only be achieved by an independent validation sample; however, there may also be some merit to cross-validation and bootstrapping techniques applied to the initial sample to improve the locus-specific effect estimates. Even without selection bias, however, many investigators and editors demand independent validation to reduce the likelihood of finding spurious associations, as do we at Circulation.

Lastly, how does one determine the importance of a risk allele with a marginal odds ratio, albeit with striking statistical significance, determined in isolation? In the
genome-wide association study described above,\textsuperscript{1} the great majority of statistically significant risk alleles had odds ratios \(<2.0\). How can we make mechanistic sense of a collection of weak associations between SNPs or other biomarkers and a disease phenotype? This question is, perhaps, the most difficult, because it strikes at the heart of the genome-wide association studies of common, complex disease phenotypes.

Ernest Rutherford pointed out that “[a]ll science is either physics or stamp collecting.”\textsuperscript{13} Gathering huge datasets on genomic sequence variation puts us currently in the realm of collecting genetic stamps; the useful analysis of these data, however, will require a rigorous approach unlike any that has been applied thus far to these datasets. In complex traits such as hypertension or coronary heart disease, several of the loci with weak effects may code for genes that interact in common pathways to yield a synergistic mechanism of action in disease pathogenesis. Determining interactions among weakly associated alleles,\textsuperscript{14} therefore, will be a key exercise in sorting out the importance of their attributable risk and their potential as targets for risk reduction. Some of the methods that have been used to explore these gene–gene (ie, epistatic) interactions include multivariate linear regression,\textsuperscript{15} neural network analysis,\textsuperscript{16} significance analysis of function and expression framework methodology,\textsuperscript{17} and principal component analysis.\textsuperscript{18} The assessment of interactions between weakly associated alleles or between weakly associated alleles and environmental factors will require massive databases to have sufficient power to detect interactions and reasonable precision to estimate the magnitudes of the interaction effects. In addition, exploration of the potential mechanistic links between these weak risk alleles and disease expression requires not only the conventional reductionist approach, which builds on experiment-based knowledge of the increasing number of interlinked pathways, but also complex systems and network-based analyses in which multiple interacting disease determinants are assessed comprehensively in both static and dynamic quantitative models.\textsuperscript{15} These models must take into account not only interactions between genes but also those between genes or gene products and environmental factors.

It is a vast oversimplification, albeit a necessary one initially, to assume that all of the information that governs disease expression is determined by genetic sequence. Genes are expressed at different rates and at different levels in the steady state, and these levels cannot be predicted by simple DNA sequence. Furthermore, a host of epigenetic, posttranscriptional, and posttranslational events, many of which are a consequence of or influenced by genetically unpredictable environmental factors, govern the ultimate product of the genome, ie, the proteome, which itself can be modified posttranslationally by environmental and cellular metabolic factors. It is this postranslationally modified proteome that, in turn, truly defines phenotype. Viewed in this more complex and realistic way, the link between genotype and phenotype seems potentially rather remote, especially in the case of complex traits weakly linked to multiple alleles. (The Figure provides a brief summary of these points.)

Measuring these various postgenomic molecular events, cataloging them as we now do variations in genetic sequence, and sorting out how they interact with one another statically and dynamically will provide the ultimate analysis of the molecular determinants of disease phenotype. We are a long way from this goal, and, I fear, the hyperbole used to extol the importance of genomic sequence misleads the scientific community and lays public into a level of immediate expectation that is unrealistic.

This new era of discovery science holds great promise in defining not only the full panoply of genetic and molecular determinants of disease but also in identifying potential key targets for risk reduction. These targets can be defined for populations and for individuals, in the ideal case. We must, however, approach this exercise not only with intellectual rigor but also with humility, largely because, as G.K. Chesterton points out, life is a “trap for logicians” because “[i]t looks just a little more mathematical and regular than it is; its exactitude is obvious, but its inexactitude is hidden, its wildness lies in wait.”\textsuperscript{20}

**Acknowledgments**

The author thanks Drs Elliott Antman, Emelia Benjamin, Martin Larson, Jane Leopold, Joseph Vita, and Scott Weiss for helpful comments and discussions.

**Disclosures**

None.

**References**


**KEY WORDS:** Editorials | genetics | biomarkers
Association Studies in an Era of Too Much Information: Clinical Analysis of New Biomarker and Genetic Data
Joseph Loscalzo

Circulation. 2007;116:1866-1870
doi: 10.1161/CIRCULATIONAHA.107.741611

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/116/17/1866

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation 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 is online at:
http://circ.ahajournals.org/subscriptions/