THE INTERFACE OF EPIDEMIOLOGIC METHODS

Interface of Genetics and Epidemiology

Terri H. Beaty¹ and Muin J. Khoury²

INTRODUCTION

Genetic epidemiology is a hybrid discipline with a relatively short history of less than 50 years. The first mention of "epidemiologic genetics" can be found in the 1954 book by Neel and Schull (1), who suggested that the two parent disciplines must consider each other's strategies when studying common diseases such as cancer, diabetes, and heart disease that reflect joint effects of both genes and environmental factors. The hybrid discipline of genetic epidemiology emerged slowly over the next two decades, but the pace of interaction between genetics and epidemiology has accelerated recently.

Epidemiology, the methodological science of public health, is the older of the two parent disciplines and has its roots in use of vital records or observational data to identify causal agents and/or mechanisms of disease transmission during the "bacteriologic revolution" of the 19th century. Epidemiology's history of systematic study design for observational data sparked development of numerous statistical approaches to test hypotheses and make causal inferences (2). The broader field of public health builds its combined intervention and health policy efforts upon sound epidemiologic findings. Savitz et al. (3) argued that epidemiology as a discipline must strive to maintain its scientific objectivity and independence from public health policy initiatives.

Modern genetics (the biologic science of inheritance) can be traced to Gregor Mendel, an Austrian monk whose plant breeding experiments in the middle of the 19th century produced the underlying paradigm for genetics (4). The value of Mendel's experimental work was not fully appreciated until the beginning of the 20th century, however, when cell biologists recognized that eukaryotic chromosomes (and the genes they carried) followed regular patterns of segregation during meiosis. At the turn of the century, several workers concurrently argued that these chromosomes could be the physical units of inheritance described in the abstract by Mendel (5).

Genetic epidemiology began to emerge when both geneticists and epidemiologists realized that the etiology of the major chronic diseases cannot be understood without simultaneously considering both genes and environmental factors. Common chronic diseases such as cancer, heart disease, and diabetes are "multifactorial" or "complex." Their etiology reflects the interaction of genes and environmental risk factors, and substantial etiologic heterogeneity is to be expected. As people from both disciplines began to work together, however, differences in terminology were an impediment to understanding the strengths and limitations of one another's tools and study designs. For example, geneticists have long recognized that a heterogeneous population structure (i.e., a population is composed of genetically distinct subgroups) can lead to spurious statistical associations between a genetic marker and a disease. They use the term "population stratification" to describe this phenomenon of heterogeneity within a sample of cases and controls. Epidemiologists likewise recognized that whenever a factor (e.g., membership in a subgroup) is associated with both the disease outcome and the risk factor (e.g., an allele at the genetic marker), erroneous inferences can result from statistical associations, but they call this occurrence "confounding." Thus, it is not surprising that some confusion remains to this day about how to interpret any statistical association between a disease and a genetic marker.

There is a spectrum of causality from "single-gene" or "Mendelian" disorders (where genes are the sole cause of disease) to genes that determine disease only under selected environmental exposures (e.g., glucose-6-phosphate dehydrogenase or G6PD deficiency causes hemolytic anemia only when there is exposure...
to antimalarial drugs or other agents) to the complex
diseases (where genes contribute to the pathogenesis
as one of many risk factors). Even though Mendelian
diseases are relatively rare, they constitute a major
public health burden in the aggregate. Furthermore, in
selected populations, individual Mendelian diseases
can be common enough to represent major public
health burdens (e.g., the beta-thalassemias in many
Mediterranean populations). The focus of medical
genetics on rare diseases often has been at odds with
epidemiology's focus on common diseases, but as
common allelic variants associated with increased risk
are discovered, both disciplines will be forced to con-
sider the broader public health impact of genetic risk
factors. For both Mendelian and complex diseases,
however, identification of interacting environmental
factors can serve as the basis for public health inter-
vention since environmental factors generally are eas-
ier to modify (e.g., newborn screening for phenylke-
tonuria allows dietary intervention among susceptible
homozygotes for this Mendelian disorder).

SCOPE OF GENETIC EPIDEMIOLOGY

King et al. (6) defined three essential questions in
genetic epidemiology: 1) Does the disease cluster in
families? 2) Is this clustering due to genes or shared
environments? and 3) What is the best model of inheritance? These essential ques-
tions can be extended as shown in table 1, in which
appropriate study designs or analytical approaches are
also noted.

As the 21st century begins, the search for genes that
correlate to the etiology of complex diseases is an
important research area in genetic epidemiology. Since
multiple genes are likely to contribute to any one dis-
ease (some only in certain subgroups and some only
under environmental exposure), it becomes critical to
deal with etiologic heterogeneity that may exist in any
set of families as well as across data sets from differ-
ent populations. A variety of epidemiologic study
designs can be used to test for such heterogeneity, and
this may become the prime activity for the field of
genetic epidemiology in the near future.

Because replication is important in the process of
accumulating scientific evidence to define gene action,
genetic epidemiology also will be forced to deal with
inconsistency among studies, a dilemma common to
all of epidemiology (7). Adapting epidemiologic prin-
ciples to classic genetic study designs to test for con-
sistency across different studies will raise further
unique challenges, as samples drawn from different
populations have different genetic backgrounds. Combining these approaches to incorporate informa-
tion on environmental factors will be necessary to put
the role of genes into a broader public health perspec-
tive, and the final question in table 1 can be viewed as
the application phase of genetic epidemiology, often
called public health genetics. Without a solid research
foundation for the first five questions (table 1), how-
ever, public health intervention strategies for the sixth
question will be doomed to fail.

STUDY DESIGNS FOR GENE DISCOVERY

Discovery of genes involved in complex diseases
commonly relies on linkage analysis, a statistical pro-
cedure originally developed for Mendelian diseases.
Elston (8) argued that showing evidence for linkage
(or cosegregation) between a genetic marker and a
complex phenotype is the highest possible level of sta-
tistical proof that the phenotype is under genetic con-

| Table 1. Central questions in genetic epidemiology and analytical designs to address them |
|------------------------------------------|---------------------------------------------------------------|
| Question | Analytical design |
| Does the disease cluster in families? | Case-control studies, familial aggregation/correlation studies |
| Is this clustering due to genes or shared environments? | Heritability, path analysis, variance components, family studies of environmental risk factors and genetic factors |
| What is the best model of inheritance? | Segregation analysis |
| Can we locate causal gene(s) for a complex disease? | Linkage analysis and mapping, linkage disequilibrium tests to discover causal genes |
| How does this gene influence the level of disease risk? | Tests for heterogeneity and interaction between genes and the environment, integration of genetic markers into epidemiologic studies |
| How can susceptibility genes or genetic risk factors be incorporated in public health interventions? | Screening, education, clinical trials, etc. |
trol. Lander and Schork (9) divided the available statistical methods for identifying genes underlying complex diseases into two groups: linkage methods and allele-sharing methods. The former tests for cosegregation between alleles at an observed marker locus and an unobserved trait locus by using family data; the latter tests for excess sharing of marker alleles among relatives that should reflect linkage between the marker locus and an unobserved trait locus.

Linkage methods include primarily likelihood-based methods to estimate genetic distance (i.e., the recombination fraction) and provide some formal test statistic for the null hypothesis of independent assortment. Commonly, the log odds of linkage versus no linkage or lod score is used. Likelihood-based methods have the advantage of statistical rigor, but they were developed primarily for Mendelian diseases and require strong assumptions about the model of inheritance (which are seldom justifiable for complex traits for which the true model of inheritance is rarely if ever known). Still, there are circumstances in which even approximately correct models can be used productively. For example, mapping of genes for breast cancer and prostate cancer has yielded important scientific knowledge from highly selected samples of families, although many questions remain about the true role of genes in these common cancers.

Allele-sharing methods, sometimes called "model free" or "nonparametric," generally do not require full knowledge of the model of inheritance and may be more appropriate for complex diseases. However, they focus only on testing the hypothesis that the distribution of marker alleles shared in pairs (or sets) of relatives is no different than expected under the null hypothesis of independent assortment between the marker and a putative trait locus. Rejection of this null hypothesis should reflect linkage, but it is often impossible to estimate the map position itself by using allele-sharing methods alone. Typically, genetic distance between the marker and the unobserved trait locus is confounded by other parameters such as the variance attributable to the trait locus. Still, evidence that genes underlie complex diseases such as non-insulin-dependent diabetes is beginning to accumulate in some populations (10, 11), although proving causality remains a substantial challenge.

A third group of statistical tools relies on tests of linkage disequilibrium; these tests are more indirect and require more cautious interpretation. Linkage disequilibrium is defined as an excess of certain combinations of a marker allele with a presumed "high-risk" allele at the unobserved disease locus, which can be used to map Mendelian disorders (12). However, the transitory nature of disequilibrium (which will slowly fade toward multilocus Hardy-Weinberg equilibrium) makes it beneficial to select the population to be examined carefully. For example, "founder populations," such as Finland, where a relatively small group of founders produce a much larger group of descendants in relative isolation from continuous genetic admixture (which can itself create disequilibrium), are more likely to yield evidence of the strong linkage disequilibrium necessary to map disease genes in case-control designs.

Classic epidemiologic designs such as the case-control design can be used to test for linkage disequilibrium; however, it is important to understand the limitations of this approach. In case-control samples, the null hypothesis essentially is that the frequencies of alleles or genotypes are the same in both groups. Rejection of this null hypothesis could be explained by the following:

1. A simple type I error.
2. A marker allele in the causal pathway of the disease (although less likely for anonymous genetic markers, this explanation is very likely for expressed variants such as the null allele PiZ that leads to deficiency of alphal antitrypsin, which in turn contributes to the pathogenesis of emphysema).
3. Confounding due to population stratification in which a spurious statistical association is created by heterogeneity in the sample, with both disease and allele frequency differing between unrecognized subgroups.
4. Linkage disequilibrium between the observed marker locus and an unobserved disease locus in which one marker allele occurs more frequently with a high-risk allele.

Thus, a statistically significant association between a marker allele and a disease in a case-control study may or may not reflect linkage disequilibrium.

An extension of the classic case-unrelated-control study design is the case-parental-control or case-parent-trio design, in which matched controls are artificially constructed from marker alleles (or genotypes) not transmitted to the observed case (13). This study design has the advantage of sidestepping the problem of confounding because the controls cannot be genetically different from the cases; they are merely meiotic permutations that did not occur. The roots of this design go back to the matched case-control design of the "haplotype relative risk" approach (14, 15), and the design creates a combined test of linkage and disequilibrium due to linkage (16). Thus, it offers a narrower range of possible biologic explanations for a statistically significant result (17). Recently, this approach
has been extended to accommodate missing parental genotypes and to rely on sib controls to reconstruct parental genotypes (18, 19). It is promising for detecting linkage in the presence of disequilibrium and thus should help identify genes that play a role in complex diseases. However, issues of heterogeneity within samples and among populations, as well as issues of gene-environment interaction, still must be addressed. Of course, this approach will be useless if there is linkage but no disequilibrium.

**INTERFACE OF GENETICS AND EPIDEMIOLOGY IN THE 21ST CENTURY**

As the discovery of genes associated with the risk of a wide range of human diseases accelerates in the post–Human Genome Project era, genetic risk factor information will be integrated routinely into epidemiologic studies (20, 21). Because of the emerging importance of gene-environment interaction in explaining the etiology of diseases, the traditional epidemiologic paradigm of the $2 \times 2$ table (two disease categories by two exposure categories) may be replaced by a $2 \times 4$ table (two disease categories by four combinations of risk factors and genotypes) as the basic unit of analysis (21). Shpilberg et al. noted that “sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street pump (20, p. 638).” In fact, integration of genetics into epidemiologic research has been occurring for some time. Over the last decade, molecular epidemiology began using DNA along with biochemical measures of exposure in epidemiologic studies, especially in the area of environmental carcinogenesis (20). When the Human Genome Project has been completed, epidemiologic researchers will have an overwhelming array of information on genetic variation at thousands of loci that can be used in cohort, case-control, and cross-sectional studies, but there must be some guidelines and priorities on how to incorporate this new information and how to interpret the results. More complex methods will be needed to evaluate the role of multiple genes, along with possibly multiple environmental factors.

Since most discoveries of genetic variants influencing disease risk are based on studies of high-risk families or highly selected subgroups (e.g., BRCA1 and BRCA2 mutations in breast and ovarian cancer), population-based epidemiologic studies will be needed to quantify and confirm the impact of these genes on the risk of disease in a broader population. Further epidemiologic studies will be required to identify and

<table>
<thead>
<tr>
<th>TABLE 2. Categories and examples of human genome epidemiology studies*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study type</strong></td>
</tr>
<tr>
<td>Assess the prevalence of gene variants in different populations</td>
</tr>
<tr>
<td>Assess the magnitude of disease risk associated with gene variants in different populations</td>
</tr>
<tr>
<td>Assess the contribution of gene variants to occurrence of the disease in different populations</td>
</tr>
<tr>
<td>Assess the magnitude of disease risk associated with gene-gene and gene-environment interactions in different populations</td>
</tr>
<tr>
<td>Assess the validity and utility of genetic tests in different populations</td>
</tr>
<tr>
<td>Evaluate the determinants and impact of using genetic tests and services in different populations</td>
</tr>
</tbody>
</table>


*Epidemiol Rev* Vol. 22, No. 1, 2000
quantify the impact of any modifiable risk factors that could interact with "high-risk" alleles. Such studies should help medical and public health professionals target interventions. Epidemiologic studies are also required for clinically validating new genetic tests, monitoring use of genetic tests in different populations, and determining the safety and effectiveness of genetic tests and services in various populations (22). To effectively translate discoveries into public health action, however, epidemiologists must work closely with clinical genetics, behavioral and social sciences, health services research, and communication sciences professionals while also balancing thorny legal and ethical issues. The scientific approaches of genetic epidemiology will be essential for better understanding of disease etiology and for developing valid diagnostic tools. Results from such collaborations are needed now to create both medical and public health policy. For example, issues are still being debated regarding population-based genetic testing for breast cancer in relation to BRCA1 mutations (23), for Alzheimer’s disease in relation to the E4 allele at the apolipoprotein E locus (24), and for iron overload and risk of liver disease in relation to the hemochromatosis gene (25). Given the paucity of population-based data regarding the frequency of these genetic variants and their true disease risks (which may depend on interaction with environmental factors), it is not yet clear how appropriate health policy for use of genetic tests can be developed.

Khoury and Dorman (26) coined the term “human genome epidemiology” (HuGE) to denote an evolving field of systematically using epidemiologic methods in population-based studies to measure the impact of genetic variation on health and disease. This work can be viewed as the intersection between genetic epidemiology and molecular epidemiology. The spectrum of topics addressed by investigators working on human genome epidemiology, as well as selected examples, is shown in table 2 (adapted from Khoury (27)) and ranges from simple population-based research on allele frequency to evaluation of genetic tests and services. Ultimately, application of genetic epidemiologic research to a population setting will provide the raw data for rational development of public health policy. In 1998, the Human Genome Epidemiology Network (HuGE Net) was established to promote collaboration in disseminating peer-reviewed epidemiologic information on human genes and to develop an updated and accessible knowledge base on the World Wide Web (28).

The World Wide Web as a medium for information exchange is likely to continue, and use of genetic databases on the Internet is already well established (29). As massive quantities of information on genetic variation in humans become available, there will be an increasing need for systematic review, synthesis, and dissemination. Such information will be crucial in driving policy decisions regarding use of genetic information in public health. Genetic epidemiologists should join forces with experts in informatics to develop, organize, and disseminate the rapidly growing information on the human genome and to delineate how this information can best be integrated into the practice of public health in the 21st century.

REFERENCES