Causal Models for Investigating Complex Disease: I. A Primer

Ann M. Madsen\textsuperscript{a} Susan E. Hodge\textsuperscript{b–d} Ruth Ottman\textsuperscript{a, c, e}

\textsuperscript{a}Department of Epidemiology and \textsuperscript{b}Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, \textsuperscript{c}Division of Epidemiology, New York State Psychiatric Institute, \textsuperscript{d}Department of Psychiatry, College of Physicians and Surgeons, and \textsuperscript{e}G.H. Sergievsky Center and Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, N.Y., USA

Abstract

Background/Aims: To illustrate the utility of causal models for research in genetic epidemiology and statistical genetics. Causal models are increasingly applied in risk factor epidemiology, economics, and health policy, but seldom used in statistical genetics or genetic epidemiology. Unlike the statistical models usually used in genetic epidemiology, causal models are explicitly formulated in terms of cause and effect relationships occurring at the individual level. Methods: We describe two causal models, the sufficient component cause model and the potential outcomes model, and show how key concepts in genetic epidemiology, including penetrance, phenocopies, genetic heterogeneity, etiologic heterogeneity, gene-gene interaction, and gene-environment interaction, can be framed in terms of these causal models. We also illustrate how potential outcomes models can provide insight into the potential for confounding and bias in the measurement of causal effects in genetic studies. Results: Our analysis illustrates how causal models can elucidate the relationships among underlying causal mechanisms and measures obtained from statistical analysis of observed data. Conclusion: Causal models can enhance research aimed at identifying causal genes.

Introduction

Causal models provide a basis for understanding the relationships among underlying causal mechanisms and measures obtained from statistical analysis of observed data. While these models are increasingly applied in classical epidemiology, economics, and health policy, they are seldom used in statistical genetics or genetic epidemiology\cite{1–3}. In this paper, we describe two causal models, the sufficient component cause (SCC) model and the potential outcomes (PO) model, and discuss their relevance to important concepts in research on genetically complex disorders. Our goals are to make causal model-based methods more accessible to genetic researchers and to illustrate their utility to the field.

Research in human genetics has relied heavily on the use of statistical models to draw inferences about the genetic architecture of complex disorders (e.g. the number
of genes that contribute to disease risk, the presence of epistasis or genetic heterogeneity, etc.) [4–11]. The choice of appropriate models has sometimes been controversial, in part because of uncertainty about the mathematical relationships that best correspond to underlying biological mechanisms (e.g. [9, 12]). Causal models can help to resolve some of these controversies. Unlike the statistical models usually used in genetic epidemiology, causal models are explicitly formulated in terms of cause and effect relationships occurring at the individual level. This allows derivation of predictions about mathematical relationships (e.g. between disease and genetic variant frequency in a sample), under a transparent set of assumptions about underlying cause and effect relationships. The second paper of this two-part series illustrates the use of a causal model for deriving such predictions [13].

While one can think about causation in many ways, epidemiologists researching the causes of disease have widely adopted a ‘counterfactual’ framework [14, 15]. According to this framework, a cause (X) of an effect (Y) meets the condition, if ‘X had not occurred Y would not have occurred’ (at least not when and how it did). The occurrence of X is factual. The condition, ‘if X had not occurred’ is counterfactual. Thus, a genotype is a cause of disease, in the counterfactual sense, if an individual has the disease and the genotype and if that individual would not have developed the disease (at least not when or how he/she did) if he/she had not had the genotype. Two prominent approaches to causal modeling used in epidemiology, the SCC model and the PO model, reflect counterfactual reasoning.

**SCC Model**

In a seminal paper, Rothman [16] described the SCC model, a clear and succinct model for understanding and describing the possible relations among insufficient and unnecessary causes of disease. Rothman’s paper closely paralleled the work of contemporaneous scholars of the philosophy of causation, in particular, Mackie’s work regarding INUS causes (insufficient but non-redundant part of unnecessary but sufficient conditions) [17, 18]. The introduction of the SCC model reflected a transition in the focus of epidemiologic research from infectious diseases, characterized by a one-to-one relationship between a germ and a disease, to chronic diseases, characterized by many causes, none of which alone is necessary or sufficient [19]. This parallels the transition in the focus of human genetics research from monogenetic disorders to genetically complex disorders with multiple unnecessary and insufficient genetic and environmental causes.

To construct a model of disease causation, Rothman defines a sufficient cause as a ‘set of minimal conditions and events that inevitably produce disease’ [20], and a component cause as one of the minimal conditions or events included in a sufficient cause. Component and sufficient causes are illustrated in ‘causal pie’ diagrams, a heuristic that conveys how a factor can be a cause of disease while neither necessary nor sufficient. Figure 1a depicts a SCC model of a hypothetical disease. In the figure, A is a cause of Disease Z, but is insufficient – it leads to Disease Z only in the presence of unknown factor(s) or stochastic event(s) represented by U 1. Together, A and U 1 complete a sufficient cause of Disease Z. When an individual has both A and U 1, Sufficient Cause I is completed and Disease Z occurs. A and U 1 are each component causes of Sufficient Cause I. U 1 represents the ‘casual partner’ of A and vice versa [21].

The model of disease causation in figure 1a also demonstrates lack of necessity. Sufficient Cause II, consisting of component causes B and U 2, depicts another set of conditions/exposures sufficient to cause Disease Z. In the figure, no single component cause is necessary or sufficient for disease to occur. However, SCC models do not prohibit necessary and sufficient causes. For example, an SCC model of Down syndrome would consist of a single necessary and sufficient cause: extra genetic material from a critical region of chromosome 21. In contrast, in a SCC model of AIDS, HIV infection would be a necessary but insufficient cause, since no case of AIDS occurs without HIV infection, but not all infected individuals develop AIDS. In this SCC model, HIV infection would be included in every causal pie, but each pie would also include causal partner(s) as additional component causes. If the causal partners of HIV differ among individuals, the causal model would include as many pies as there are ways for an individual to acquire AIDS.

Figure 1b depicts a general SCC model for a complex genetic disease with two sufficient causes. Genotype G, included in Sufficient Cause I, is an unnecessary and insufficient cause of Disease Z. Other genetic or environmental factors, represented in the schematic as U 1, are also required in order for G to cause disease in the absence of other causes. Some individuals may have the same Disease Z, but not carry G because they have Sufficient Cause II, which has only a single component cause, X, representing all sufficient causes that do not involve G.

The SCC model in figure 1b does not specify the mode of inheritance of the susceptibility genotype. For a SCC model of a dominant genetic disease, G would be defined by the presence of at least one copy of the causal allele, whereas for a recessive mod-
el, G would be defined by the presence of two copies of the causal allele. Models of intermediate (or co-dominant) inheritance could also be depicted using a SCC model (e.g. by specifying one sufficient cause for individuals with two causal alleles and a different sufficient cause for those with one causal allele) [11].

As an example of a SCC model for a complex disease, consider the application of the model depicted in figure 1b to the causes of breast cancer. Here, G could represent BRCA1 mutations and U1 the ‘causal partners’ that when combined with BRCA1 mutations lead to breast cancer. X would represent all other sufficient causes of disease that do not include BRCA1 mutations, such as BRCA2 mutations and their causal partners, radiation exposure and its causal partners, etc. Of course, a causal model could alternatively be constructed from the viewpoint of BRCA2 mutations, in which case BRCA1 mutations would be subsumed under X.

Below, we use SCC models to illustrate important concepts in statistical genetics and genetic epidemiology: penetrance, phenocopies, genetic heterogeneity, etiologic heterogeneity, gene-gene interaction, and gene-environment interaction. Table 1 shows SCC models corresponding to these terms and gives examples of statistical models commonly used to represent them. For simplicity, we assume a dichotomous disease outcome and a general risk genotype G (without specifying mode of inheritance). The diagrams representing the ‘simple’ SCC models illustrate the minimal number of component and sufficient causes to convey the concept. The diagrams representing the ‘complex’ models give examples in which the causes in the simple model are neither necessary nor sufficient. However, even the ‘complex’ SCC models are simplified, in the sense that they include only one concept at a time, rather than combinations of concepts such as genetic heterogeneity and gene-gene interaction. SCC models can comprise any number of component and sufficient causes to incorporate known or hypothesized information.

<table>
<thead>
<tr>
<th>Genetic term</th>
<th>Statistical model</th>
<th>Causal model**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penetrance</td>
<td>P[Disease</td>
<td>G]</td>
</tr>
<tr>
<td>Phenocopy</td>
<td>P[Disease</td>
<td>G]</td>
</tr>
<tr>
<td>Genetic Heterogeneity</td>
<td>e.g., Statistical heterogeneity model $p_{ij} = \alpha_i + \beta_j - \alpha_i \beta_j$</td>
<td>Simple</td>
</tr>
<tr>
<td></td>
<td>e.g., Additive model $p_{ij} = \alpha_i + \beta_j$</td>
<td>Complex</td>
</tr>
<tr>
<td>Etiologic Heterogeneity</td>
<td>e.g., Statistical heterogeneity model $p_{ik} = \alpha_i + \gamma_k - \alpha_i \gamma_k$</td>
<td>Simple</td>
</tr>
<tr>
<td></td>
<td>e.g., Additive model $p_{ik} = \alpha_i + \gamma_k$</td>
<td>Complex</td>
</tr>
<tr>
<td>Gene-Gene Interaction</td>
<td>e.g., Multiplicative model $p_{ij} = \alpha_i \beta_j$</td>
<td>Simple</td>
</tr>
<tr>
<td></td>
<td>e.g., Departure from additive or multiplicative model on specified scale</td>
<td>Complex</td>
</tr>
<tr>
<td>Gene-Environment Interaction</td>
<td>e.g., Multiplicative model $p_{ik} = \alpha_i \gamma_k$</td>
<td>Simple</td>
</tr>
<tr>
<td></td>
<td>e.g., Departure from additive or multiplicative model on specified scale</td>
<td>Complex</td>
</tr>
</tbody>
</table>

* G = Susceptibility genotype at locus G; H = susceptibility genotype at locus H; E = environmental exposure; $\alpha_i$ = effect of genotype G; $\beta_j$ = effect of genotype H; $\gamma_k$ = effect of exposure k for environmental exposure E; $p_{ij}$ = penetrance in individuals with genotype G and genotype H; $p_{ik}$ = penetrance in individuals with genotype G and genotype H; $p_{ik}$ = penetrance in individuals with genotype G and exposure E.

** The ‘complex’ SCC models illustrate only one concept at a time, rather than combinations of complexities, such as genetic heterogeneity and gene-gene interaction.
Penetrance

Penetrance is usually defined as 'the conditional probability of observing [a] phenotype given [a] specified genotype' [22]. This definition, based on the observable joint distribution of the disease and genotype in a sample, conveys a statistical concept rather than a causal concept. That is, by this definition penetrance refers to the probability of a disease occurring, but does not take account of what caused it to occur. We refer to this meaning as 'statistical penetrance'. As a causal concept, penetrance would be defined differently, as 'the proportion of individuals in whom the genotype caused the disease'. This would be modeled in the SCC framework as 'the proportion of individuals in whom a sufficient cause containing the genotype is completed'. We refer to this quantity as the 'causal penetrance', or 'proportion susceptible to the genotype' [23]. As discussed below, this quantity cannot be determined from a joint distribution unless the causal partners of the genotype of interest are known and measured in the population.

SCC models make apparent and explicit the conditions under which these quantities differ. Consider the SCC model depicted in figure 1b, from which we derive the statistical penetrance and causal penetrance for the causal genotype(s) at locus G. By genotype G, we mean all allele combinations at the risk locus that lead to disease when co-occurring with U1. G encompasses the complement to G, i.e. all combinations of alleles at the locus that do not partner with U1 to cause disease.

The statistical penetrance of G is reflected by the joint distribution of G and disease. Causal models describe the factors that cause this joint distribution. In this example, individuals who have G but not U1 might have disease caused by X; these individuals contribute to the statistical penetrance of G, but not to the causal penetrance (proportion susceptible to G).

Thus, if we assume G, U1, and X are independently distributed in the population, so that P[U1 | G] = P[U1], then

\[ P(\text{Disease} | G) = P[X] + (1 - P[X])P[U1] = P[U1] + (1 - P[U1])P[X]. \] (1)

That is, the statistical penetrance of G is equal to the probability that an individual with G has either the causal partners of G or other factors that cause disease independently of G. (As shown in the Appendix, the expression for statistical penetrance in equation 1 can easily be generalized to the situation where G, U1, and X are not independent.) In contrast, the causal penetrance of G consists only of the probability that a sufficient cause containing G is completed, which is equal to the probability that an individual has the causal partners of G alone, i.e. P[U1]. The statistical penetrance of G equals the causal penetrance of G only when P[X] = 0. The right-hand side of equation 1 makes these distinctions clear.

The SCC model depicted in figure 1b does not specify a sufficient cause of Disease Z containing G. This implies that the causal penetrance of G equals zero (i.e. the absence of the genotype is not a cause of disease for any individual). In contrast, the statistical penetrance of G, i.e. the probability of disease given G, does not necessarily equal zero, because individuals with G will have Disease Z when they have X. Thus, P[Disease | G] = P[X].

For an example of how the causal model depicted in figure 1b could describe the causes of a common disease, consider Alzheimer’s disease, with a prevalence of 13% in individuals aged 65 or older [24]. Again, assume G, U1, and X are independent. Suppose a common genetic variant, G, with 2% frequency in the population, raises risk for the disorder, and P[U1] = 0.25, so that 25% of individuals who carry the variant develop the disease in the absence of other causes (i.e. 25% are susceptible to the effects of the variant). The total prevalence, P[D], is made up of two components: those who have disease because of X, and those who do not have X but do have G and its causal partners U1:

\[ P[D] = P[X] + (1 - P[X]) \cdot P[G] \cdot P[U1]. \] (2)

From equation 2, we can solve for P[X]:

\[ P[X] = \frac{P[D] - P[G]P[U1]}{1 - P[G]P[U1]} = \frac{0.13 - (0.02)(0.25)}{1 - (0.02)(0.25)} = 0.126. \] (3)

Equation 3 illustrates that if the overall disease prevalence is 13%, then the frequency of other causes of disease, X, must also be close to 13%, because most people do not carry the genetic variant. Under these assumptions, 'statistical penetrance' (from equation 1) is equal to 0.25 + (1 - 0.25)(0.126) = 0.34. This is notably higher than the causal penetrance of 0.25. Statistical penetrance and causal penetrance are equal only if the disease involves only one sufficient cause [25].

Whether statistical penetrance or causal penetrance is of interest depends on the research question. The concept of causal penetrance facilitates consideration of the potential impact of preventive interventions on disease. For example, based on the parameters from the SCC model of Alzheimer’s disease described above, if the causal effect of the gene could be eliminated (i.e. P[U1] = 0), disease risk in carriers of the genetic variant would be reduced to P[X] = 0.126 (from equation 1), compared with the original estimate of 0.34. This estimate could not be obtained if the different components of the causal model were not considered.

One can always observe the statistical penetrance P[D | G] for a given sample. In contrast, the causal penetrance, P[U1], cannot be determined from observational data, but instead reflects a characteristic of the underlying model of disease causation. Any causal model can be realized in many ways, depending on the frequencies of the component causes of disease, but each realization gives rise to a single observable joint distribution of genes and disease [26].

Although penetrance is often considered a characteristic of a genotype, as evidenced by the expression ‘highly penetrant gene’, neither definition of penetrance given above suggests penetrance is a fixed characteristic of a genotype alone. Unless the biological effect of a genotype (or genetic variant) is sufficient to cause disease by itself, penetrance of the genotype will depend on the prevalence of other factors that influence susceptibility. Hence, as noted by Bateman in 1957 [27], penetrance can vary across populations as a function of the conditions in which the phenotype occurs. According to a SCC model, this is because both definitions of penetrance reflect the distribution of causes other than the genotype in the sample: U1 (the causal partners of G) and X (causes that do not involve G) in the case of statistical penetrance, and only U1 in the case of causal penetrance.

In genetic research, the term ‘penetrance’ refers to what we call ‘statistical penetrance’, and hence we suggest this term be defined as the ‘proportion of individuals with disease among those who
carry a specific genotype within a defined population'. Likewise, we recommend that the term 'causal penetrance' be used to refer to the proportion of individuals susceptible to G within a defined population – i.e. the proportion of individuals with G who would have the disease in the absence of other causes, i.e. $P(\text{Disease} \mid G, X) = P(U_X)$ [25].

**Phenocopies**

Ott defines phenocopies as 'individuals who are affected not owing to genetic distribution (at the locus under study)' [22]. In terms of a SCC model, phenocopies are defined as individuals with disease caused by sufficient causes that do not include the genotype of interest (e.g. individuals with disease due to X in fig. 1b). As discussed in the penetrance section above, according to this model an increase in X leads to an increase in the statistical penetrance of the causal genotype G (unless penetrance of G is 100%). However, in statistical genetics (e.g. in modeling penetrance for linkage analysis), an increase in phenocopies is often modeled as an increase in the probability of disease only among individuals without the putatively causal genotypes. Use of the SCC model suggests that to account for phenocopies in linkage analysis, an alternative approach for modeling penetrance might be preferable.

**Genetic and Etiologic Heterogeneity**

The statistical meaning of heterogeneity derives from the notion of statistical independence, i.e. lack of interaction, meaning that the interaction coefficient in a regression model equals zero [7]. In statistical genetics, lack of interaction is typically modeled by assuming an additive relationship between the statistical penetrances of two genotypes [5, 7, 8]. Examples of the types of additive models commonly assumed are shown in table 1. These models are similar, but not identical, to Risch's [5] parameterization, as described in detail in the second paper of this series [13]. The biological interpretation of such models is unclear.

In contrast, biological definitions of genetic and etiologic heterogeneity pertain to ideas about independence underling biologic pathways of disease causation. Genetic heterogeneity can be defined as the production of identical or similar phenotypes by different genetic mechanisms. If the X in figure 1b represents, in part, a genetic factor that causes disease in addition to G, the SCC model in figure 1b depicts 'causal genetic heterogeneity'. 'Etiologic heterogeneity' often refers to a non-genetic cause of disease in addition to the genotype of interest, i.e. if X consists entirely of non-genetic factor(s) that cause disease in addition to G, the SCC depicts 'causal etiologic heterogeneity'.

Researchers have considered whether statistical models applied to observational data can be used to infer when a disease is characterized by underlying causal genetic heterogeneity and which study designs are best suited for such studies (e.g. [5, 6, 28]). In the second paper in this series, we illustrate how to use a SCC model as the basis for determining the expected joint distribution of genotypes and disease under specific assumptions about causal relationships between genotypes and disease [13].

**Gene-Gene Interaction (Epistasis)**

The terms 'epistasis' and 'gene-gene interaction' are used interchangeably in the literature. Here, we use gene-gene interaction. As a statistical concept, interaction denotes a departure from independence of effects, e.g. significant interaction terms in a regression model. We refer to this concept as 'statistical gene-gene interaction'. In statistical genetics, researchers model gene-gene interaction for binary outcomes employing a variety of models [10]. For example, a multiplicative relationship between the statistical penetrances of two genotypes is often used to model two genes acting in non-independent biologic pathways to influence disease [4, 7, 8]. Again, as with additive models, the biologic interpretation of this model is unclear.

We use the expression 'causal gene-gene interaction' to describe two or more genotypes in a single sufficient cause of disease. This is consistent with Cordell's 'qualitative epistasis' [7] and Moore and Williams's 'biologic epistasis' [29]. A SCC model of causal interaction between genotypes G at one locus and H at another locus would depict both genotypes as component causes in a single sufficient cause of a disease as shown in the 'simple' example of causal gene-gene interaction in table 1. For a complex disease, one might anticipate that even G and H together could be insufficient to cause disease, and could require a causal partner when they co-occur. Additionally, one would anticipate causes of disease that do not include either G or H. These complexities are shown in the 'complex' example of causal gene-gene interaction in table 1. However, even this example does not include all of the possible sufficient causes that may be involved in a complex disease. Other causes may involve G or H alone (with causal partners specific to each genotype); this situation would include both causal gene-gene interaction and causal genetic heterogeneity.

Conclusions derived from multiplicative models necessarily depend on the scale of measurement; for example, transforming the multiplicative model to a logarithmic scale results in an additive model [7, 8, 30]. While the distinction between a causal concept of interaction and a statistical concept of interaction has been recognized in genetics, the field has not settled on a method for reconciling the two concepts [6–8, 29]. As with statistical heterogeneity, causal models can be used to clarify the conditions under which multiplicative models accurately represent the joint distribution of genotypes and disease.

**Gene-Environment Interaction**

Like statistical gene-gene interaction, statistical gene-environment interaction is defined as a departure from independent effects of the genotype and environmental exposure, and modeled in various ways. Perhaps owing to the focus on gene-environment interaction in risk factor epidemiology, researchers increasingly model gene-environment interaction as departure from additivity when the goal is to infer an underlying biologic mechanism [31, 32]. This practice reflects the recognition, derived from causal models, that greater-than-additive risks for two factors together indicate 'causal synergy' [18].

**Summary: SCC Model**

The SCC model provides a unified framework for unambiguously illustrating the causal meanings ascribed to genetic terms, such as penetrance, phenocopies, genetic heterogeneity, etiologic heterogeneity, gene-gene interaction, and gene-environment interaction. As illustrated in table 1, these causal meanings differ from the statistical meanings ascribed to these terms. The joint distribution of causes and disease realized for a SCC model depends on the frequency of the component causes in the population. Whether (or for which study designs and analytic methods)

58 Hum Hered 2011;72:54–62 Madsen/Hodge/Ottman
the underlying component causes and their causal model relationships (i.e. same or separate sufficient causes) can be inferred from an observed joint distribution is of considerable interest in statistical genetics. Deriving the joint distribution from SCC models offers the opportunity to address such questions.

In epidemiology, a causal model perspective has led to the recommendation that ‘causal synergy’, i.e. causal interaction, be modeled as statistical interaction on an additive scale [18, 20]. Statistical geneticists should establish whether this conclusion is generally true for causal gene–gene and gene-environment interaction considering dominance, recessivity, penetrance, phenocopies, etc. Also, questions about the relative efficiency or validity of various study designs or analytic methods for diseases characterized by causal interaction or causal heterogeneity should, in principle, be addressed by first determining the expected joint distributions from realizations of SCC models and then evaluating which study design is most powerful assuming a given mechanism.

PO Model

While the SCC model and corresponding heuristic are useful for considering the structure of complexity that underlies disease, the PO model is useful for considering the measurement of effects of causes on disease. Explicitly based on counterfactual reasoning, the PO model recognizes that to determine the effect of a genotype on disease, the ideal experiment would involve contrasting the disease status of an individual carrying the susceptibility genotype and the disease status of the same individual not carrying that genotype. However, this ideal ‘counterfactual contrast’ is unobservable; an individual cannot be observed under both the carrier and non-carrier conditions. The PO framework ‘asserts that individuals have potential outcomes in all states, even though they can actually only be observed in one state’ [33]. Neyman expressed the notion of potential outcomes in 1923 in the context of a randomized trial [34], and Rubin [35] developed a more theoretically complete expression of the model that can be applied to observational studies.

**Notation**

Rubin developed the PO notation to convey the comparison required for causal effect estimation using counterfactual reasoning (i.e. by imagining an individual in two different states). Because of the historical context of randomized trials, the terms ‘treatment’ and ‘control’ are used to convey these two states, so that with a genetic study, the ‘treatment’ would be a genotype. Of interest are the effects of specific putatively causal genotypes, in comparison with the putatively non-causal genotypes. Holland [36] summarizes a general notation for PO models. Below, we describe this notation in the context of estimating the causal effects of a binary genotype on a dichotomous outcome:

- **U** = population of individuals (‘universe U’; note that this is not the same as **U**₁ and **U**₂ defined above as ‘causal partners’ in the SCC model)
- **u** = individual subject in **U**
- **t** = treatment condition (i.e. susceptibility genotype carrier)
- **c** = control condition (i.e. non-carrier of susceptibility genotype)
- **Y(u)** = disease response in individual **u** (1 if diseased, 0 if not diseased)
- **Y_t(u)** = disease response in individual **u** under treatment condition (i.e. carrying susceptibility genotype)
- **Y_c(u)** = disease response in individual **u** under control condition (i.e. not carrying susceptibility genotype)
- **Y_i(u_j)** = disease response under condition **i** in individuals observed to have condition **j**, where **i** and **j** can equal either treatment or control conditions (i.e. genotype carriers or non-carriers).

**Measures of Effect**

The causal effect of a specific genotype on the disease outcome for a single individual is defined as the difference between the disease outcome for that individual if he/she carried the genotype and the disease outcome under the counterfactual condition that he/she did not carry the genotype. Using the above notation:

\[
\text{True causal effect} = Y_t(u) - Y_c(u)
\]  

(4)

The true causal effect is unobservable because an individual cannot be observed in both states; he/she can only carry the genotype or not carry the genotype [36].

The average causal effect (ACE) of a genotype in a population is:

\[
\text{ACE (risk difference)} = E[Y_t(u) - Y_c(u)] = E[Y_t(u)] - E[Y_c(u)].
\]  

(5)

Like the variables in equation 4, the variables in equation 5 are unobservable and therefore not practically useful for determining the effect of a genotype on an outcome. However, formulating the effect measure as a population average suggests equation 6, consisting of variables observable in a population.

\[
\text{Estimated ACE (risk difference)} = E[Y_t(u_t)] - E[Y_t(u_c)]
\]  

(6)

The difference between equations 5 and 6 is the replacement of **u**, the individuals in the population, with **u_t** and **u_c**, the subsets of individuals actually observed under the ‘treatment’ condition (i.e. having the genotype) and ‘control’ condition (i.e. not having the genotype). Hence the estimated ACE in equation 6 contrasts the observed proportion of individuals with the genotype who have the disease versus the observed proportion of individuals without the genotype who have the disease, whereas the true ACE in equation 5 contrasts the proportion of individuals who would have disease if they had the genotype with the proportion of these same individuals who would have disease if they had not had the genotype. Equation 6 can be used to estimate the ACE using data observed in a population **U**, consisting of **u_t** individuals who are carriers of a susceptibility genotype and **u_c** individuals who do not carry the genotype. Equation 6 is the usual risk difference statistic calculated in a cohort study.

Relating the observable risk difference to the counterfactual contrast makes transparent the threats to valid estimation of the causal effects of a genotype. One such threat occurs when the estimated ACE (equation 6) is not a valid estimate of the true ACE (equation 5). This occurs when **Y_t(u_t) ≠ Y_t(u_c)**, i.e. when the average outcome in the treatment state would have been different if the treatment had been applied to the control group. In genetic terms, this inequality would imply that the average disease out-
come in individuals with the causal genotype would have been different if the individuals in the population who did not have the genotype had instead had the genotype. Similarly, the estimated ACE will differ from the true ACE when \( Y_c(u_t) \neq Y_c(u_l) \), i.e. when the average outcome in individuals without the causal genotype would have been different if the individuals who were observed to be genotype carriers had instead been non-carriers. These threats can be understood better using the concept of response types, as described below.

**Response Types**

Greenland and Robins [37] use ‘response types’ to clarify threats to validity when estimating the effect of a binary treatment and binary outcome. Response types are the potential outcomes an individual would demonstrate if one could observe him/her in both the carrier and non-carrier states. To illustrate, assume that a disease is described by the SCC model in figure 2. The response types with respect to genotype G and a binary outcome are given in table 2.

G can cause, prevent, or have no effect on disease. The pattern in line 1 of table 2 indicates an individual who would have disease when a carrier of G and also would have disease when not a carrier of G. We designate this response type as ‘inevitable’ to convey the concept that disease is inevitable in individuals of this type. (In the literature, the term ‘doomed’ is often used for this response type [37]; we avoid this term to eliminate any negative connotation.) A proportion \( P_1 \) of individuals in the population has this response type. An individual who has disease, whether or not a G carrier, must have been exposed to a completed sufficient cause of disease that does not include G. In figure 2, this is represented by Sufficient Cause III, which does not include G as a component cause.

Line 2 of table 2 describes the response type denoted ‘causal’, consisting of individuals who would have disease if they carried G but not if they were non-carriers. Given the SCC model in figure 2, this pattern might arise among individuals exposed to \( U_1 \), the causal partner of G in Sufficient Cause I, and negative for \( X \) (otherwise they would be ‘inevitable’). Line 3 describes the response type denoted ‘preventive’. Individuals with this response type would not have disease if they carried G and would have disease if they did not carry G, so that G prevents their disease. This pattern might arise among individuals exposed to \( U_2 \) but not exposed to \( X \). Finally, line 4 describes the response type denoted ‘immune’, in which individuals will never have disease, whether they are carriers or non-carriers. An individual demonstrating this response type given the SCC model in figure 2 must not have been exposed to \( X, U_1 \), or \( U_2 \).

Viewing response types through SCC models can be nuanced, and some of these subtleties have been addressed previously [26, 38]. For this reason, a frequent assumption is ‘monotonicity’, i.e. that a genotype cannot both cause and prevent disease.

**Applications of the PO Model**

The PO model provides insight into the potential for confounding and bias in the measurement of causal effects in genetic studies. Valid estimation of the causal effect of a putative genetic cause depends in part on the distribution of response types in the study sample [37]. Specifically, the study sample must meet an ‘exchangeability’ criterion, i.e. response types must be distributed equally among carriers and non-carriers of the putative causal genotype. Lack of exchangeability, or unequal proportions of response types among carriers and non-carriers of the putative causal genotype, can lead to confounding.

In risk factor epidemiology, the PO framework informs study design and method development for non-randomized (observational) studies, and this framework should similarly inform association studies of complex genetic diseases. Randomizing with regard to treatment theoretically allows one to detect the causal effect of the treatment; thus, study design and analytic methods should be approached as an attempt to emulate a randomized study. To estimate the causal effect of a genotype on disease, one can imagine randomizing a cohort with regard to the genotype. For the SCC model in figure 2, for example, randomization would eliminate bias in the distribution of \( X \) among carriers and non-carriers. If \( X \) is known, one might match genotype carriers and non-carriers with respect to \( X \) to avoid differences in the distribution of response types due to chance.

In summary, the PO model makes explicit the assumptions required for the measured effect to estimate the causal effect of a genotype, and illustrates the conditions under which an observed effect measure may not reflect the true causal effect. The model also serves as the basis for designing studies robust to confounding (e.g. due to population stratification) and analytic methods to control for confounding statistically.

---

**Table 2.** The response types for a single susceptibility genotype G and a binary disease outcome (adapted from [37])

<table>
<thead>
<tr>
<th>Line</th>
<th>Disease status if G carrier</th>
<th>Disease status if G non-carrier</th>
<th>Description</th>
<th>Proportion of population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>inevitable</td>
<td>( P_1 )</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>–</td>
<td>causal</td>
<td>( P_2 )</td>
</tr>
<tr>
<td>3</td>
<td>–</td>
<td>+</td>
<td>preventive</td>
<td>( P_3 )</td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>–</td>
<td>immune</td>
<td>( P_4 )</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** A hypothetical SCC model giving rise to the potential disease outcomes given exposure to G.
Discussion

Causal models facilitate the articulation of the causal meaning of important concepts such as causal effect measures, causal interaction, and causal heterogeneity, and help to establish the validity (or lack thereof) of different statistics for addressing causal questions. Causal models also allow researchers to determine the correspondence between the unobservable underlying causal mechanism (i.e. causal model) and the observable joint distribution. By articulating the causal meaning of a concept, e.g. causal gene–environment interaction, one can evaluate whether realizations of this model are described by a particular statistical model. Using a causal framework to think about complex genetic diseases should help clarify current problems in gene mapping studies, suggest causal model-based analytic approaches, and identify valid statistical models. Causal models can inform evaluation of methods for controlling for population stratification, and can help to understand the reasons for failure to replicate results in an association study. The causal model framework makes apparent that the magnitude of a genotype’s effect on disease risk is affected by differences between families or populations in the distribution of its causal partners and of causal factors external to it. In fact, the prevalence of the genotype’s causal partners and the other sufficient causes of disease in the population determine the magnitude of the effect rather than any innate characteristic of the genotype itself [20]. Therefore, a different effect measure may be obtained in a replication study in a population with a different distribution of other causes, and a failed replication in a sample different from the initial study may not mean that the initial finder causes, and a failed replication in a sample different study in a population with a different distribution of other effect measure may be obtained in a replication genes

A causal model framework also facilitates evaluation of the utility of particular study designs or analytic methods in genetic epidemiology. Use of causal models for this purpose would involve simulating the joint distribution of disease and component causes for a known causal model and evaluating the distribution using different sampling or analytic approaches. This contrasts with the current practice of simulating data without beginning from causal assumptions. For example, one study simulated data using multiplicative odds ratios to investigate one- and two-stage approaches to GWAS for interacting genes [40]. An alternative approach would have been to simulate data based on a SCC model rather than assuming multiplicative odds ratios, which may not actually correspond to biologic interaction.

In summary, causal models provide the basis for innovative statistical methods that improve the validity of tests of causal hypotheses. Here, we have summarized the principles of these methods and demonstrated their applicability to complex genetic diseases. We hope this paper will facilitate the application of causal model-based methods to research aimed at identifying causal genes.

Acknowledgments

A.M.M. was supported in part by NIMH grant T32-MH065213. This work was also supported by NIH grants R01-NS043472, R01-NS036319, R01-NS053998, and RC2-NS070344 (to R.O.), and NIMH grant R01-MH048858 (to S.E.H.). We are grateful to Sharon Schwartz, PhD, for instructive insights and critical comments.

Appendix

Assuming the causal model depicted in figure 1b, the formula for statistical penetrance can be generalized as shown below, for the situation where G, U1, and X are not all mutually independent. If D denotes ‘disease’,

\[ P[D \mid G] = P[D \mid X,G]P[X \mid G] + P[D \mid \bar{X},G]P[\bar{X} \mid G]. \]  

(A1)

Then \( P[D \mid X,G] = 1, P[D \mid \bar{X},G] = P[U_1 \mid \bar{X},G] \) and \( P[\bar{X} \mid G] = 1 - P[X \mid G] \), so equation A1 becomes:

\[ P[D \mid G] = P[X \mid G] + (1 - P[X \mid G])P[U_1 \mid \bar{X},G] = P[U_1 \mid \bar{X},G] + (1 - P[U_1 \mid \bar{X},G])P[X \mid G]. \]  

(A2)

The right-hand side of equation A2 is analogous to the right-hand side of equation 1, with \( P[U_1 \mid \bar{X},G] \) replacing \( P[U_1] \) and \( P[X \mid G] \) replacing \( P[X] \). Equation A2 gives the expression for statistical penetrance in the more general situation where all the factors are not independent.

References


Phillips CV, Goodman KJ: Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. Emerg Themes Epidemiol 2006;3:5.


