Mendelian Randomization: How It Can—and Cannot—Help Confirm Causal Relations between Nutrition and Cancer

Arthur Schatzkin,1 Christian C. Abnet,1 Amanda J. Cross,1 Marc Gunter,2 Ruth Pfeiffer,1 Mitchell Gail,1 Unhee Lim1 and George Davey-Smith3

Abstract
Observational epidemiologic studies of nutrition and cancer have faced formidable methodologic obstacles, including dietary measurement error and confounding. We consider whether Mendelian randomization can help surmount these obstacles. The Mendelian randomization strategy, building on both the accuracy of genotyping and the random assortment of alleles at meiosis, involves searching for an association between a nutritional exposure—mimicking gene variant (a type of “instrumental variable”) and cancer outcome. Necessary assumptions are that the gene is independent of cancer, given the exposure, and also independent of potential confounders. An allelic variant can serve as a proxy for diet and other nutritional factors through its effects on either metabolic processes or consumption behavior. Such a genetic proxy is measured with little error and usually is not confounded by nongenetic characteristics. Examples of potentially informative genes include LCT (lactase), ALDH2 (aldehyde dehydrogenase), and HFE (hemochromatosis), proxies, respectively, for dairy product intake, alcoholic beverage drinking, and serum iron levels. We show that use of these and other genes in Mendelian randomization studies of nutrition and cancer may be more complicated than previously recognized and discuss factors that can invalidate the instrumental variable assumptions or cloud the interpretation of these studies. Sample size requirements for Mendelian randomization studies of nutrition and cancer are shown to be potentially daunting; strong genetic proxies for exposure are necessary to make such studies feasible. We conclude that Mendelian randomization is not universally applicable, but, under the right conditions, can complement evidence for causal associations from conventional epidemiologic studies.

It has been long and widely believed that nutrition plays an important role in the development of cancer. Nevertheless, although much effort has been devoted to clarifying the links between nutrition and malignant disease, and some consensus has been reached for at least a few nutrition-related factors, many important questions remain unresolved (1). The scientific community has been debating how to move the field forward and generate more definitive, and credible, public health recommendations (2). In that light, we discuss whether Mendelian randomization, a relatively new research strategy combining genomics and epidemiologic methods, can help us make progress in this area.

Nutrition and Cancer—the Problem

For several centuries, writers have maintained, based on nonsystematic observation of human experience, that nutrition influences the development of cancer (5). Animal studies have provided strong evidence that nutritional interventions can modify carcinogenesis, even among genetically altered rodents (6–9). Human metabolic (feeding) studies show that nutritional interventions can modulate putatively cancer-related intermediate end points, such as fecal bile acid concentrations (10) and blood hormone levels (11). Ecologic data—international correlations (12, 13), time trends (14), and migration studies (15, 16)—are consistent with dietary causation:
what people eat clearly varies among countries, has changed substantially in certain countries over the past few decades, and changes with migration and acculturation. These three types of evidence, although suggestive, are hardly definitive: rodents are not people; modulation of an intermediate end point does not necessarily equate to a change in cancer outcome (17); and ecologic findings may well be confounded by other lifestyle factors causally linked to carcinogenesis.

Epidemiologic case-control studies suggested increased risks of cancer at several anatomic sites in association with dietary fat and meat, and reduced risks for intakes of, for example, fruits, vegetables, dietary fiber, and certain micronutrients. Several of these findings, however, have not been clearly or consistently confirmed in cohort studies, which are largely free of the recall and selection biases that potentially compromise case-control investigations (18–20).

Even clinical trials have produced consternation. A few trials have yielded modest findings: calcium supplementation, for example, decreased colorectal adenoma recurrence (21), and a combination of selenium, vitamin E, and β-carotene reduced both total and gastric cancer incidence (22). However, the protection of β-carotene against lung cancer, suggested in observational epidemiologic studies, was not seen in clinical trials (23). Polyp trials did not support the protective associations reported in earlier observational studies (24) for fat, fiber, or fruits and vegetables (25, 26). The Women’s Health Initiative recently reported a marginal effect ($P = 0.07$) of a low dietary fat pattern on invasive breast cancer, which has been interpreted as suggesting both a protective and a null effect (27, 28).

**Why the Uncertainty in Nutrition and Cancer?**

In observational epidemiology, a consensus has emerged in favor of prospective cohort studies of diet and cancer (20). Nevertheless, even with the establishment of many such cohort studies, other persisting methodologic problems may account for some of the inconsistency and uncertainty in the field.

**Measurement error**

Diet is notoriously difficult to measure. A lively debate, recently intensified by results from biomarker-based methodologic studies (29), has arisen over the accuracy and suitability of the food frequency questionnaire, the instrument used typically in large-scale epidemiologic studies of diet and cancer. It has been argued that dietary measurement error may be causing epidemiologists to miss important nutrition-cancer relations or even report false protective or deleterious associations (30–32).

**Confounding**

In an epidemiologic study, individuals who consume, say, a low-fat, high-fiber diet may differ from their high-fat, low-fiber counterparts not only in what they eat but also in a variety of lifestyle and even biological characteristics that truly affect the development of cancer. For a potent risk factor, such as heavy cigarette smoking, which engenders relative risks for lung cancer in the range of 10 to 20, the likelihood that some other characteristic of heavy smokers could cause the entire association is slim. In fact, such a confounding factor would need to have a prevalence ten times greater in smokers than in nonsmokers to explain a relative risk of 10 between heavy smoking and lung cancer (33). The more modest relative risks of 1.5 to 2.0 often found for nutritional exposures vis-à-vis cancer could more plausibly be generated by confounding, particularly when several important confounders may all distort an exposure-outcome association. Randomized trials largely avoid confounding because those persons randomized to the intervention and control groups are likely to have similar distributions of known and unknown confounders. Although trials have their own methodologic limitations—problems with adherence and diminished power to detect true associations, dose selection difficulties, potentially inadequate intervention and follow-up time, and use of pre-cancerous lesions rather than frank cancer end points (e.g., in polyp trials)—the disparate results from trials and observational epidemiologic studies have raised the possibility that confounding in the observational studies has yielded misleading results for several nutrition-chronic disease associations (34, 35).

**The Mendelian Randomization Strategy**

Mendelian randomization is a term for a research strategy that uses genomic information in an epidemiologic context to cast light on the etiologic role of nutritional and other environmental exposures (3, 4, 36). Mendelian randomization has the potential to address, at least in part, the problems of measurement error and confounding.

Mendelian randomization in nutritional epidemiology is generally premised on the fact that genes involved in substrate metabolism or receptor function are polymorphic, such that different allelic variants lead to reduced enzyme activity or altered receptor function and, in various ways, can therefore mimic or serve as proxies for different degrees of exposure to a given dietary factor. In other words, for a specific polymorphic gene, allelic variant $A$ produces a physiologic-biochemical state that is biologically equivalent to one level of “exposure,” whereas alternative allele $B$ produces a state indicative of a qualitatively different degree of “exposure,” which could be either higher or lower than that reflected by allele $A$.

In essence, we are using genotypic information to obtain an unbiased assessment of exposure to nutritional factors. It is an approach complementary to asking people what they eat, via a self-report instrument, or measuring nutrient levels in blood. Both the self-report and blood levels are imperfect measures of dietary intake. Self-reported dietary data are prone to both random and systematic errors, which cause true relations between diet and cancer to be attenuated or, in some circumstances, inflated (32). Most nutrient levels in blood, so-called concentration biomarkers (37), are subject to the influence of various personal behaviors and biological characteristics (e.g., smoking) and can engender bias in diet-cancer associations. Moreover, substantial confounding of any association of the dietary measure (questionnaire or biomarker) and cancer outcome may occur.

The use of genotypic information to enhance causal inference for nutritional exposures in relation to cancer is an example of the “instrumental variables” approach that has been commonly used in econometrics and social science and more recently discussed in an epidemiologic context (38, 39). An instrumental variable is one that is associated with an outcome...
only through its association with an intermediate variable—in this case, the nutritional exposure of interest—and is independent of potential confounders. This is illustrated in Fig. 1, a directed acyclic graph in which Z is the genotype, X is a nutritional exposure, Y is cancer outcome, and C, which is associated with both X and Y, represents one or more potential confounders of an observed association between X and Y. The assumptions underlying the use of an instrumental variable—and the Mendelian randomization strategy—are as follows:

(a) The instrumental variable (genotype) Z is associated with the nutritional exposure of interest (X).

(b) The genotype (Z) is independent of any variable (C) that potentially confounds the association between X and Y.

(c) The association between genotype (Z) and cancer (Y) exists only because the genotype is associated with the nutritional exposure (X); that is, Z is independent of outcome Y given X. This is indicated by the dotted line from Z to Y.

General theoretical and specific statistical aspects of the instrumental variables approach, as well as its potential application in Mendelian randomization studies, have been more extensively discussed elsewhere (39).

Both case-control and cohort designs are applicable to Mendelian randomization studies. We noted earlier that case-control studies of nutritional factors in relation to cancer are subject to recall, reverse causation, and other “retrospective” biases. Such biases are not likely to compromise case-control studies of a gene variant in relation to malignant disease. Moreover, case-control studies offer particular advantages in accruing a large number of cases in a relatively short time.

How Does Mendelian Randomization Deal with Potential Bias in Observational Epidemiologic Studies?

Measurement error

Genotyping errors exist but are small compared with the apparent error in self-reported dietary assessment. Having established that an allelic variant truly reflects an enzyme deficiency and that it mimics, for example, low exposure to a given nutrient, an investigator can have a high degree of confidence in the truth of the statement that an individual is homozygous for that variant and, therefore, has lower exposure to the nutritional factor than someone who does not carry the variant.

Confounding

According to Mendel’s second law, alleles are distributed randomly during meiosis, without regard to potentially confounding characteristics (3). Whereas a particular nutritional exposure (e.g., high fat or low folate intake) may be correlated with a variety of other nutritional, lifestyle, socioeconomic, or biological characteristics, a nutritional exposure—mimicking allelotype generally does not display correlation with such potential confounding factors (Z is independent of C in Fig. 1). For example, smoking is a potential confounder of associations between alcohol and a number of cancers because smoking is correlated with alcohol consumption and is a causal factor for multiple malignancies. Smoking behavior, however, is not correlated with ALDH2 genotype (40, 41), although the genotype is strongly associated with alcohol intake and therefore will not confound the ALDH2-cancer relation. Thus, analysis of ALDH2 can complement findings from more conventional multivariable-adjusted prospective analyses of dietary factors and cancer. A recent report from the British Women’s Heart and Health Study shows substantial pairwise correlation among 96 behavioral, socioeconomic, and physiologic factors, whereas the pairwise correlations between 23 genetic factors and the 96 nongenetic factors were no greater than what would be expected by chance (42).

How Do Genes Give Us Information on Nutritional Exposure?

(a) Gene affects exposure propensity

The LCT gene encodes the lactase enzyme, which is central to the metabolism of lactose in dairy products. Lactase activity is high in infants but is generally down-regulated during adulthood. A dominantly inherited trait, however, can result in lactase persistence throughout adult life, which is common in people with a northern European heritage. Lactase persistence has been linked to two genetic variants: a C-to-T change ~14 kb upstream of LCT and a G-to-A change 22 kb upstream of LCT (43). Lactase nonpersistent individuals have difficulty metabolizing lactose and, after consuming dairy products, often have symptoms of bloating, abdominal pain, and diarrhea. As a result, individuals with lactase nonpersistence

![Fig. 1. A directed acyclic graph depicting how the instrumental variables approach is used in Mendelian randomization. Specifically, Z is the genotype, X is a nutritional exposure, Y is cancer outcome, and C, which is associated with both X and Y, is a potential confounder of an observed association between X and Y. A key assumption underlying the use of instrumental variables in the Mendelian randomization setting is that the association between genotype (Z) and cancer (Y) exists only because the genotype is associated with the nutritional exposure (X); that is, Z is independent of outcome Y given X. This is indicated by the dotted line from Z to Y.](https://www.aacrjournals.org/doi/fig)
tend to consume less lactose-containing dairy products and, therefore, the variant associated with lactase nonpersistence can be a proxy for low exposure to such foods (44). As an example of how this gene can be used in the Mendelian randomization context, a recent study showed that the CC genotype in postmenopausal women is associated with low dietary intake of calcium from milk, lower bone mineral density at the hip and spine, and an increased risk of nonvertebral fractures (45). With regard to malignant disease, if, for example, an inverse association were found between the lactase nonpersistence variant and prostate cancer, this would suggest that low, as opposed to high, dairy food consumption protects against the disease, thereby helping to clarify a recent nutrition-cancer controversy (Fig. 2; ref. 46).

Acetaldehyde is the first metabolite of ethanol. Aldehyde dehydrogenase (ALDH2) is the enzyme primarily responsible for the elimination of acetaldehyde. ALDH2 is functionally polymorphic in some populations, and an individual’s genotype at this locus plays a major role in acetaldehyde levels after consumption of alcohol (47). The ALDH2*2 allele results from a single point mutation in ALDH2*1 (the wild-type allele) and encodes for a protein unable to metabolize acetaldehyde after alcohol consumption. Blood acetaldehyde levels are 18 times higher in persons homozygous for ALDH2*2, and 5 times higher in those heterozygous for the allele, compared with levels in wild-type (*1*1) homozygotes. ALDH2*2*2 homozygotes experience nausea, flushing, drowsiness, headache, and other dysphoric symptoms after drinking alcohol. A number of studies have found a protective association between the homozygote ALDH2*2*2 variant and esophageal squamous cell carcinoma, which implicates alcohol in pathogenesis (40).

The situation is less straightforward, however, for heterozygosity. Heterozygotes, as a group, are less likely to drink and have lower esophageal cancer risk than those with wild-type ALDH2. However, heterozygotes who do drink, compared with those with wild-type ALDH2, are at higher risk of esophageal cancer at the same level of alcohol consumption—a phenomenon consistent with an enhanced carcinogenic effect of longer exposure to acetaldehyde (41, 48). Because heterozygotes are more likely to drink than ALDH2*2*2 homozygotes, heterozygosity is an ambiguous proxy for drinking and a questionable genetic instrument for Mendelian randomization studies of alcohol and cancer.

A large body of evidence now implicates alcohol consumption, even at modest levels, in the etiology of breast cancer in women (49). Because the relative risks for 1 to 2 drinks per day are around 1.2 to 1.3, concern persists that such risks reflect confounding. Mendelian randomization is a potentially valuable analytic tool for this hypothesis. If the modest association between alcohol and breast cancer indicates causality, then we would expect to see an inverse relation between ALDH2*2*2 and breast malignancy. The ALDH2-breast cancer association merits examination in large studies in Asian and other populations with adequate prevalence of ALDH2*2*2. A similar case can be made for Mendelian randomization studies of ALDH2 and colorectal cancer in both men and women (50).

(b) Gene determines metabolic state reflecting altered exposure

High serum iron has been proposed as a causal factor for certain malignancies (51). HFE is a gene involved in iron absorption; it is associated with the iron overload condition of hereditary hemochromatosis (52). A transition mutation in HFE (845G to A) leads to a cystosine-tyrosine amino acid change at position 282 and is known as mutation C282Y. A second common transition mutation (187C to G) in codon 63, mutation H63D, results in a histidine-to-aspartic acid substitution. These HFE mutations may serve as genetic instrumental variables for high serum iron, as reflected, for example, in serum ferritin levels (Fig. 3). Note that we are not equating HFE status with iron intake. That is because the connection between intake and serum level is complex, involving absorptive and metabolic factors (including such potentially confounding exposures as smoking) that are not reflected by the HFE gene variant.

A recent population-based case-control study of 475 cases and 833 controls reported a significant 40% increased risk of colorectal cancer in individuals with HFE mutations. The risk was greatest in those who consumed high levels of iron (53). This direct association between HFE gene mutations and cancer suggests that high serum iron is causal in colorectal carcinogenesis. Further studies of HFE variants versus colorectal and possibly other cancers are needed.
Methylenetetrahydrofolate reductase (MTHFR) is a gene with potential as an instrument for Mendelian randomization (3). MTHFR encodes an enzyme that irreversibly converts 5,10-methylenetetrahydrofolate (derived from dietary folate) to 5-methyltetrahydrofolate, which is used to convert homocysteine to methionine and to facilitate methylation reactions. A polymorphism in the MTHFR gene that yields a C-to-T substitution at base 677, MTHFR 677C>T, leads to reduced enzyme activity (54). The homozygous genotype 677TT has been shown to be associated with elevated homocysteine levels.

Nevertheless, Mendelian randomization studies of folate and MTHFR in relation to cancer may be problematic because of a nutrient-gene interaction: the effect of 677TT on blood levels of folate and homocysteine seems to depend on the level of folate intake. If folate intake is relatively low, then 677TT is associated with low blood folate, high blood homocysteine (55), and, with respect to consequences for malignant transformation, reduced global methylation of DNA (56). 677TT in the presence of high folate intake, however, does not result in low blood folate, high blood homocysteine (57–59), or reduced tissue methylation (56). Mechanistic evidence for such an interaction is provided by experiments showing that the instability of the variant MTHFR enzyme structure was offset, in part, by the availability of high folate (60, 61).

In other words, whether or not 677TT is a good proxy for reduced folate (or elevated homocysteine) levels depends on knowing true folate intake. In many epidemiologic studies of cancer, there is likely a range of folate intake among study participants; with supplement use and fortification of foods, a substantial proportion of any study population likely has high intake levels. Although we can use conventional tools to assess whether a participant’s folate intake is high or low, we then face the measurement error difficulties we were hoping to avoid with Mendelian randomization. It is questionable, therefore, whether MTHFR 677TT can serve as a reasonably unambiguous genetic instrument for studies of folate and cancer.

Thus far, we have discussed metabolically active genetic proxies for specific dietary factors. Mendelian randomization may also be valuable in the investigation of more complex nutrition-related exposures. For example, hyperinsulinemia has been proposed to partly underlie the relation of obesity with colorectal cancer (62). Associations between insulin (or C-peptide) and colorectal cancer may be subject to confounding, and prospective data on this relation are both limited and inconsistent (63–66). Circulating insulin levels have been shown to correlate with a variable number of tandem repeat (VNTR) polymorphisms located in the 5′ region of the insulin gene (INS): carriers of the “class III” allele of the INS 5′-VNTR have higher insulin levels than noncarriers (67). Another insulin gene polymorphism, INS IVS1-6 A>T, is in tight linkage disequilibrium with the INS 5′-VNTR and carriage of the 5′-VNTR class III and IVS1-6-T alleles are highly correlated. Given that the INS IVS1-6-T allele may therefore serve as an unconfounded marker for hyperinsulinemia, the detection of a positive association between this allele and colorectal cancer risk would support a functional role for insulin in colorectal carcinogenesis. A recent study, for example, suggested that carriers of the INS IVS1-6-T/T genotype were 30% more likely to have colorectal adenomas than those homozygous for the INS IVS1-6-A allele (68). Although these data implicate a causal role for insulin in colorectal neoplasia, replication is needed given the many genes involved in insulin resistance and the potential for false-positive results.

**Complexities of the Mendelian Randomization Strategy**

**Inadequate prevalence of exposure-mimicking allelic variant**

A reasonable argument can be made that ALDH2 allelic variation reflects alcohol consumption: for the homozygous wild-type drinking is unimpeded, and for the homozygous mutant ALDH2*2*2 drinking is largely precluded. Variation in the lactase gene corresponds to dairy product consumption. In many populations, however, neither the ALDH2*2*2 homozygote nor variation in lactase persistence occurs with sufficient frequency to allow epidemiologic comparisons. Use of Mendelian randomization will thus be confined to studies in populations with sufficient prevalence of the altered exposure–reflecting variants.

---

**Z (HFE mutation)**

**Serum ferritin level**

**Y (cancer)**

**C (confounders)**

Fig. 3. A directed acyclic graph depicting how the HFE gene can be used as a proxy (instrumental variable) for serum ferritin in a Mendelian randomization study of cancer.
Limited availability of genetic proxies for nutritional exposures

In spite of rapid advances in genomics over the past several years, understanding of the functional effects of allelic variation remains limited for many genes involved in nutritional physiology. Although the relation of vitamin D status (reflecting both dietary and sunlight exposure) to cancer, for example, has considerable public health importance, the evidence supporting a protective role for vitamin D in humans is not definitive (69). Several polymorphisms of the vitamin D receptor have been identified, including TaqI, BsmI, ApaI, and FokI. Although rare mutations of the VDR gene lead to the autosomal recessive hereditary vitamin-D–resistant rickets, the functional activity of the more common polymorphisms is less clear (70).

Undoubtedly, for a number of genes related to nutrient and food consumption, via metabolic pathways, receptor integrity, or transport activity, allelic variants that are potentially useful for Mendelian randomization have simply not yet been identified. Moreover, certain important nutritional exposures (dietary fiber could be an example) may not have genetic variant proxies suitable for Mendelian randomization studies.

Genome-wide association studies (71) may reveal additional genes that “light up” vis-versa cancer and which, after suitable functional studies are carried out, may be appropriately exposure-mimicking and thereby suitable for Mendelian randomization investigations. Genome-wide association studies targeted to nutrition-related intermediate end points such as obesity or nutrient (such as vitamin D) levels—scans of obese or low nutrient “cases” versus nonobese or high nutrient “controls”—can contribute to this functional research. However, to avoid chance associations without causal significance, replicating studies and statistically accounting for multiple comparisons are crucial (72). Genome-wide association studies may turn out to be a boon for Mendelian randomization studies of nutrition and cancer, but false positivity needs to be guarded against in this research strategy.

Genetic complexities of exposure inference: linkage disequilibrium and pleiotropy

Human genetics is highly complex (73) and the (desirable) inference of exposure from genotype data is often problematic. Several well-established genetic phenomena potentially compromise the exposure assumptions necessary for a successful Mendelian randomization approach. Here we discuss linkage disequilibrium and pleiotropy; canalization, which may have limited applicability to the later life events important in carcinogenesis, has been discussed elsewhere (3, 74).

Linkage disequilibrium. It is possible that a polymorphism under investigation (as a proxy for nutrient exposure) is in linkage disequilibrium with another polymorphic locus that might influence disease outcomes through a different mechanism. There is evidence, for example, that different polymorphisms affecting alcohol metabolism are in linkage disequilibrium (75). The effect of linkage disequilibrium on the Mendelian randomization approach depends on the structure of relations among the two “linked” loci, the exposure of interest, and the cancer outcome. If (in Fig. 1) the allelic variant proxy (Z) is in linkage disequilibrium with a second genetic variant, which is associated with the nutritional exposure (X) but is conditionally independent of cancer Y given X, then Z is still a valid proxy for X, and examination of the Z-Y relation is informative with respect to the relation between X and Y. If, however, the second genetic variant is associated with the outcome Y through a different mechanism than that suspected for allelic variant Z, then confounding is introduced into the Z-Y relation and one of the assumptions underlying Mendelian randomization is violated.

Pleiotropy. Many genes have more than one function; that is, there can be a one-to-many relation of genes to phenotypes. Suppose a gene truly reflects a nutritional exposure but also has tissue, cellular, or molecular consequences that influence carcinogenesis independently of that nutritional exposure-related pathway. In that scenario, an association between the genetic polymorphism and cancer is not clearly interpretable as the nutritional exposure effect. We note that the ALDH2 system metabolizes alcohols in general, not just ethanol (76). Given that retinol is another alcohol of potential importance in cancer causation, this further complicates the interpretation of ALDH2-cancer associations.

Dose imprecision

Conventional dietary assessment, measurement error aside, does quantify intake in terms of grams or servings of a given food or nutrient, enabling an investigator to derive a dose-response relation for a dietary factor versus cancer. Classification of study participants into “high” and “low” intake status on the basis of their allelotype lacks such dose precision. Extending this classification to include heterozygotes as proxies for “intermediate” intake would enhance the dose-response information. As we saw for ALDH2, however, it can be problematic whether heterozygosity truly reflects such “intermediate” exposure.

Statistical Power and Sample Size

In discussing statistical power and sample size implications of the Mendelian randomization approach, we present two examples featuring relatively strong and weak exposure-gene associations (“instruments”). More details on the statistical approach underlying the sample size computations can be found in Appendix 1.

(a) Strong instrument: alcohol intake and colorectal cancer in Asian men

Epidemiologic studies have generally shown a direct association between alcohol intake and colorectal cancer, with relative risks per 100 g of ethanol intake/wk of approximately 1.2 (50). A Mendelian randomization approach could help us evaluate whether this association is causal or confounded by unknown or poorly measured lifestyle or biological factors. Allelic variants of the ALDH2 gene have been reported to be directly associated with alcohol intake; one study reported mean intakes (in units of 100 g of ethanol intake/wk) of 2.1303, 0.7969, and 0.2604, respectively, for the 1/1, 1/2, and 2/2 genotypes (77). Thus, the correlation between the gene and the exposure in this example is very high, r = −0.78 (letting 1/1 be the wild-type genotype).

To address sample size requirements, we assume (a) the following values of alcohol intake (in units of 100 g ethanol/wk): 0 (for nondrinkers), 0.2604, 0.7969, and 2.1303; (b) genotype prevalences (in a Japanese population) of 0.57, 0.37, and 0.06
for 1*I, 1*2, and 2*2 genotypes (77); (c) colorectal cancer incidence of 49.3/100,000; and (d) a colorectal cancer relative risk of 1.21 per 100 g of ethanol/wk. Detecting such a relative risk with 80% power at a two-sided α level of 0.05, in the absence of confounders, would require 73 cases and 73 controls. Using these parameters and assuming that, given alcohol consumption, colorectal cancer does not depend on the gene, we calculate that 472 cases and 472 controls would be needed if the gene, rather than alcohol intake, is included in the logistic regression model (78).

(b) Weak instrument: body mass index and postmenopausal breast cancer

A direct association between body mass index (BMI) and postmenopausal breast cancer is well established, with relative risks for BMI in the obese range (≥30), compared with normal weight (≤25), in the range of 1.3 to 1.5. Allelic variants of the FTO gene have been recently reported to be directly associated with BMI in some populations, such that each additional copy of the rs9939669 A alleles was associated with a BMI increase of 0.4 kg/m² (79). The per A allele odds ratio for obesity in a meta-analysis was 1.31, which corresponds to an odds ratio of 1.72 for the homozygous variant compared with homozygous wild-type individuals.

We base our calculations on the following assumptions: (a) “obesity” is coded as zero for women with BMI <30, and 1 for women with BMI ≥30; (b) among women 20 years of age or older, the prevalence of obesity is estimated to be 33% (80); (c) the T-allele frequency in the general population was assumed to be 0.61 (79); (d) the gene-obesity odds ratio was 1.31 per A-allele (79); (e) breast cancer incidence is 127.8 per 100,000 women; and (f) the relative risk for breast cancer associated with a BMI in the obese range (≥30), compared with the normal, is 1.5 (81). To detect such a relative risk with 80% power at a two-sided α level of 0.05, in the absence of confounders, would require 396 cases and 396 controls. If, however, the gene is used in a logistic regression model instead of obesity, using the above parameters and assuming that, given obesity, breast cancer does not depend on the gene, the required numbers of cases and controls would be 48,910 each! This dramatic increase, which makes this Mendelian randomization study infeasible, is explained by the relatively low correlation (r = 0.12) between genotypes and obesity. Even if the odds ratio for obesity and breast cancer were 2 instead of 1.5, the required sample size would still be daunting: 16,006 cases and 16,006 controls.

Therefore, although the strength of the nutritional exposure-cancer association is important, sample size requirements for Mendelian randomization studies are especially sensitive to the strength of the gene-exposure association. Given that most nutritional exposure-cancer associations are likely to be modest, the genetic association with the nutritional exposure must be fairly strong for Mendelian randomization to be a practical research strategy, even for study consortia. Investigators need to carry out sample size calculations to determine whether a given gene-exposure (Z-X) association is strong enough to permit a Mendelian randomization study. It is conceivable that a combination of two or more gene variants would show a stronger association with a given nutritional exposures than either variant alone, such that the combination would serve as an adequate genetic proxy in the Mendelian randomization context.

What Mendelian Randomization Is Not

We have argued that Mendelian randomization can make a distinctive contribution to the epidemiology of nutrition and cancer. We should be clear, however, on what Mendelian randomization is not.

(a) Mendelian randomization is not fundamentally about discovering how genetic variation influences human carcinogenesis. In Mendelian randomization, genotype is used strictly as a proxy for nutritional (or, more generally, environmental) exposure.

(b) Mendelian randomization is not a strategy for detecting genes that confer a higher risk of cancer and therefore can contribute to a screening tool for clinical trial recruitment or public health practice.

(c) Mendelian randomization does not address the microprocesses through which that exposure influences carcinogenesis. It is an epidemiologic tool, not a molecular or physiologic inquiry into underlying mechanism.

Nutrition-Gene Interactions

Investigators have proposed extending the Mendelian randomization strategy beyond the examination of “main effect” associations between exposure-mimicking genes and cancer to include the search for nutrition-gene interactions. An example is a recent study of consumption of isothiocyanate-rich cruciferous vegetables in relation to lung cancer among participants with active and inactive isothiocyanate-metabolizing enzymes (the null variants for glutathione-S-transferase enzymes GSTM1 and GSTT1). The finding of a protective association between cruciferous vegetables and lung cancer in only those with the inactive isothiocyanate-metabolizing enzymes has been argued to constitute strong Mendelian randomization evidence for a causal protective role of crucifers and isothiocyanate in this malignancy (82).

The isothiocyanate example, however, still faces some of the methodologic difficulties of traditional nutritional epidemiologic studies. Measurement error could differ across genetically defined categories if allelic variation affects behavior, which, in turn, influences the accuracy of self-report. Although this genetic influence on questionnaire response is more likely to be an issue for genes clearly affecting exposure propensity such as ALDH2 or lactase deficiency, we cannot definitely rule out effects on behavior (and dietary reporting) of other enzymes involved in the metabolism of foods and nutrients. Nor can we ignore possible confounding in this diet versus cancer-within-genotype scenario. Because the activity status of metabolizing enzymes such as glutathione-S-transferase may well condition the ultimate exposure of lung tissue to potential confounding carcinogens, especially those in cigarette smoke that are also found in the diet (both procarcinogenic and anticarcinogenic), the anticonfounding virtues of random allele (and enzyme status) assignment do not obviate...
potential confounding within genotype category. Because active and inactive variants of the glutathione-S-transferase enzyme may differentially affect exposure of target tissue to tobacco carcinogens or their metabolites, differential confounding by smoking could at least partially explain cruciferous vegetable-lung cancer findings that differ between genotype categories.

This is not to say that diet-gene interaction studies are uninformative. In fact, it can be reasonably argued that the examination of diet-gene interactions is inferentially superior to diet-diet interactions, which involve at least two (potentially confounded) variables measured with (potentially correlated) error. Nevertheless, the diet-gene studies do not fully escape the problems of measurement error and potentially correlated error. Nevertheless, the diet-gene studies do not fully escape the problems of measurement error and confounding as do the direct exposure-mimicking gene versus cancer studies, and therefore do not provide the same degree of evidentiary support as the “classic” Mendelian randomization strategy.

Conclusion

The promise of finding foods and other nutrition-related factors that are clearly causally related to cancer at various sites—offering possibilities of both primary and secondary prevention—remains tantalizing. Large-scale observational epidemiologic studies of diet and cancer are a critical tool in diet-and-cancer research, but methodologic difficulties, notably confounding, as a result of the clustering of behavioral, demographic, and physiologic characteristics, and dietary measurement error, hamper progress in these investigations. The Mendelian randomization strategy, by which genes reflecting altered dietary exposure are examined for association with cancer, can, with some serious caveats, help to overcome these methodologic difficulties and provide evidence to complement the findings from traditional observational studies. Suppose that similar nutrition-cancer associations emerge from both conventional epidemiologic investigations and Mendelian randomization studies. For that association not to be causal, one would have to argue (rather unconvincingly) that the nutritional exposure-cancer findings from the conventional epidemiologic study are confounded at the same time that the gene-cancer findings are biased by one of the Mendelian randomization limitations discussed earlier.

Mendelian randomization is hardly a panacea. This strategy will neither substitute for continued efforts to improve dietary assessment in epidemiologic studies nor replace the randomized clinical trial with its avoidance of confounding. However, it is not yet established that we will get much better at assessing diet in observational studies or that incorporating new instruments will make a qualitative difference, although there is promise in this regard (83). Nor can clinical trials ethically or feasibly address all questions, leaving aside the expense and logistical complexity of such undertakings. In the end, it is the totality of evidence in the nutrition and cancer field that will lead to clear understanding and effective prevention. Under the right conditions, including especially the availability of a genetic instrument that is both strongly associated with the nutritional exposure and (given the exposure) independent of cancer outcome, Mendelian randomization can contribute to that totality.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Appendix 1. Sample Sizes for Mendelian Randomization Case-Control Studies

This section follows closely the computations presented by Pfeiffer and Gail (84) for genetic association studies.

We assume a random sample of R cases and S unrelated controls. The disease outcome is denoted by \( Y \), with \( Y = 1 \) for diseased and \( Y = 0 \) for healthy subjects. The nutritional exposure is given by \( X \), and we assume that the probability of disease in the population follows the model

\[
P(Y = 1 | X) = \frac{\exp[\mu + \beta X]}{1 + \exp[\mu + \beta X]}
\]

Instead of \( X \), we assess the association of \( Y \) with a biallelic marker, with genotypes \( aa \), \( aA \), and \( AA \). We define random variable \( M = 0, 1, 2 \), which corresponds to the number of \( A \) alleles in the marker genotype. Let \( Z(M = i) = Z_i \) be a score associated with marker genotype \( M = i \), for \( i = 0, 1, 2 \), with \( Z_0 = 0 \), \( Z_1 = k \), and \( Z_2 = 1 \). For a dominant genetic model \( k = 1 \), for a recessive genetic model \( k = 0 \), and for an additive genetic model \( k = 1/2 \).

The case-control data can then be summarized in a \( 2 \times 3 \) table, where the columns correspond to genotype, \( M \), and the rows to disease status, \( Y \); see Table 1.

The score test for testing for a trend in proportions is \( U = Z' (1 - \phi) r - \phi s \), where \( \phi = \mu / N \) is the proportion of cases in the case-control study, and \( N = R + S \). The vector of scores is \( Z' = (Z_0, Z_1, Z_2) \), and the genotype counts for cases and controls, \( r' = (r_0, r_1, r_2) \), and \( s' = (s_0, s_1, s_2) \) follow independent multinomial distributions with indices \( R \) and \( S \) and respective probabilities \( \phi' = (p_0, p_1, p_2) \) and \( q' = (q_0, q_1, q_2) \), where \( p_i = P(M = i | Y = 1) \) and \( q_i = P(M = i | Y = 0) \). Alternatively, \( U \) can be written as

\[
U = \sum_{i=0}^{2} \frac{Z_i}{N} [s_{r_i} - r_{s_i}]
\]

The variance of \( U \) is \( V = \text{var}(U) = (1 - \phi)^2 RZ'SZ + \phi^2 S^2 Z'Z \), where \( \Sigma_p \) denotes the correlation matrix for the genotype counts for the cases with \( (\Sigma_{pi})_{ij} = p_i (1 - p_i) \) and \( (\Sigma_{jj}) = -pp_i \), and \( \Sigma_q \) is defined analogously for the controls. Under the null hypothesis, \( H_0 \), that \( p_i = q_i, i = 0, 1, 2 \), a valid estimate of \( V \) is the pooled variance estimate, obtained by using \( \Sigma_p = \Sigma_q = \Sigma \).

Table 1. Summary case-control data for sample size calculations

<table>
<thead>
<tr>
<th>Marker genotype</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( aa )</td>
<td>( aA )</td>
</tr>
<tr>
<td>Cases ( r_0 )</td>
<td>( r_1 )</td>
</tr>
<tr>
<td>Controls ( s_0 )</td>
<td>( s_1 )</td>
</tr>
<tr>
<td>Total counts ( n_0 )</td>
<td>( n_1 )</td>
</tr>
</tbody>
</table>
with estimates \( \hat{p} = \hat{q} = n/N, \) where \( n = (n_0, n_1, n_2) \) is the vector of total counts in Table 1. To be explicit,

\[
V_\text{p} = \text{var}(U) = N \phi(1 - \phi) Z_1^2 \Sigma Z_j \quad \text{and} \quad V_\text{p} = \frac{RS}{N} \left( \sum Z_i^2 \frac{n_i}{N} - \frac{1}{n_1 N} \right) - \frac{1}{n_1 N} \sum_{i,j} Z_i Z_j \frac{n_i n_j}{N^2} \frac{n_i}{N}
\]

For an alternative hypothesis, \( H_1, \) in which \( p_i \neq q_i \) for some \( i = 0, 1, 2, \) the asymptotic power of the two-sided trend test \( |V_\text{p}^{1/2}| > z_{\alpha/2} \) can be expressed in terms of \( E_{\text{null}}/N, \) \( \sigma_i^2 = \text{var}(U) \) and the limit of \( V_\text{p}^{1/2} \) under \( H_1 \) denoted by \( \sigma^2 \) as

\[
P \left( |V_\text{p}^{1/2}| > z_{\alpha/2} \right) = 1 - \Phi \left( z_{\alpha/2} \sigma_i / \sigma_1 - \sqrt{N E_{\text{null}}/U} / \sigma_1 \right) + \Phi \left( -z_{\alpha/2} \sigma_i / \sigma_1 - \sqrt{N E_{\text{null}}/U} / \sigma_1 \right) \quad \text{(B)}
\]

where \( \Phi \) stands for the standard normal distribution function and \( z_{\alpha/2} = \Phi^{-1}(\alpha/2). \)

### Computation of the Moments of the Test Statistic under the Alternative

Taking expectations of \( U \) under the alternative yields

\[
E_{\text{null}} U = \sum_i \pi_i \phi_i = \frac{1}{N} \sum_i \pi_i \phi_i,
\]

and

\[
\lim_{n \to \infty} E_{\text{null}} V_\text{p} = \phi(1 - \phi) Z_1^2 (\phi_1 + (1 - \phi) q_1) (1 - \phi q_1) - \phi(1 - \phi) \sum Z_j (\phi_j + (1 - \phi) q_j) (1 - \phi q_j).
\]

The calculation of \( p_i = P(M = i, Y = 1) \) and \( q_i = P(M = i, Y = 0) \) depends on the extent of correlation between the genetic locus and the true exposure and on the strength of association between the disease and the exposure.

If \( M \) has no effect on the probability of disease given \( X, \) that is, \( P(Y = 11 X, M) = P(Y = 1 | X), \) and assuming that the joint distribution of the marker, \( M, \) and the nutritional exposure, \( X, f(MX), \) is known, the probabilities are

\[
p_i = P(M = i | Y = 1) = \frac{P(M = i, Y = 1)}{P(Y = 1)} = \int P(Y = 1 | X = x) dF(x)
\]

When the exposure \( X \) is discrete, the integral is replaced by a sum. The calculations for the \( q_i \)’s for the controls are analogous, with \( P(Y = 1 | X) \) replaced by \( 1 - P(Y = 1 | X)). \)

The power and sample size for the score test for trend therefore depend on the allele frequencies at the marker, the effect size for the true exposure, and, through the joint distribution of \( X \) and \( M, \) the amount of correlation between the nutritional exposure, \( X, \) and the gene used in the study.

### References


Mendelian Randomization: How It Can—and Cannot—Help Confirm Causal Relations between Nutrition and Cancer


Updated version
Access the most recent version of this article at:
doi:10.1158/1940-6207.CAPR-08-0070

Cited articles
This article cites 77 articles, 13 of which you can access for free at:
http://cancerpreventionresearch.aacrjournals.org/content/2/2/104.full#ref-list-1

Citing articles
This article has been cited by 4 HighWire-hosted articles. Access the articles at:
http://cancerpreventionresearch.aacrjournals.org/content/2/2/104.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.