

Cohort Studies for Characterizing Measured Genes

Bryan Langholz, Nathaniel Rothman, Sholom Wacholder, Duncan C. Thomas

We describe the advantages of using established cohort studies that have collected blood samples to investigate the role of genes in the etiology of cancer. These studies include the cost-efficiency and reliability of nested case-control substudies from the cohort for exploration of gene-disease associations and gene-environment interactions as well as gene penetrance. Also, the cohort may serve as a well-defined "mini-population" from which to study population stratification and molecular markers of ethnicity. We conclude that cohort studies can play a significant role in assessing the role of genetic markers for common tumors or multiple cancer sites. [Monogr Natl Cancer Inst 1999;26:39-42]

Many open questions exist about the importance of genetics in causing cancer and about the degree to which genetic variation can explain the variation in cancer risk across different populations. Estimates of familial risk and penetrance for putative major genes can be obtained from family studies via segregation analysis (1,2) or, once localized, via joint segregation and linkage analysis (3). Once the gene has been cloned, it is of interest to assess the role of the gene in cancer etiology in the general population. In some circumstances, family studies can be used to estimate penetrance (4). But, to assess relative and attributable risk in the general population, population case-control studies have been the main approach used over the past decade to evaluate the effects of common genetic polymorphisms on cancer risk (5). Case-control studies are particularly appealing when large numbers of cases need to be rapidly accrued or when time-consuming and resource-intensive exposure assessment is required. In studies of genetic factors, these advantages of the case-control design are not reduced by potential information biases, which may occur more commonly for certain exposures (i.e., dietary) assessed in case-control studies than in prospective cohort studies.

Cohort studies may also be used to evaluate genetic effects and may have some important advantages over case-control studies. Even if a case-control study might be more appropriate for answering a question about multiple exposures obtained from questionnaires about a single outcome quickly and efficiently, the cohort study allows one to study multiple outcomes with only laboratory costs hindering collection of information on many loci. Furthermore, although case-control designs will necessarily be the method of choice for studying very rare diseases, cohort studies currently under way are of sufficient size to yield enough cases to efficiently test genetic markers of common tumors, of multiple cancer sites, and, potentially, of important noncancer outcomes that can be assessed accurately (6-8).

Table 1 lists many of the world's major prospective, relatively general cohort studies that have collected comprehensive food-frequency questionnaires and have or are collecting blood samples on at least 10 000 adults. Each of these cohort studies also has, to a variable extent, questionnaire data on other lifestyle and demographic characteristics, including race, ethnicity, and family history. At the completion of ongoing collections,

blood samples will be stored on about one million individuals in these studies. Although samples from every study listed may not necessarily be available for investigations of genetic factors, this table does demonstrate the feasibility of collecting and storing biologic samples from subjects in large epidemiologic cohort studies.

Even within a cohort, one would capitalize on the efficiency of the case-control approach to investigate gene-disease associations. The cohort would serve as a study base for case-control studies in which only the blood from case and sampled control subjects would be assayed for genotype. In this paper, we discuss how this approach can be used to advantage in the study of gene-disease associations as well as how cohort studies may be used to address other hypotheses about the genetic etiology of cancer.

STUDY METHODS

The starting point for studies of genetic factors within a cohort study is, of course, a relatively large number of individuals for whom blood or other biologic material has been collected and stored. This cohort is followed over time for occurrence of disease. Once enough cases of disease have occurred, case-control studies of genetic factors could be performed, with controls appropriately selected from the cohort members. These methods include nested case-control [e.g., (9-11)] and case-cohort [e.g., (12,13)] designs. (To simplify the terminology, we will only refer to nested case-control studies, but our discussion applies to case-cohort designs as well.) The biologic material from the subjects in this case-control study would be analyzed to determine genotype, and association studies of the genetic factor would proceed. The feasibility of this approach is based on the ability to store and retrieve biologic samples for a large number of people to reach the required number of cases but to genotype a much smaller number of subjects. We consider some potential advantages of this approach, in particular, in comparison to population-based, case-control studies.

ADVANTAGES OF NESTED CASE-CONTROL STUDIES FROM AN ESTABLISHED COHORT

Nested Case-Control Studies Can Be Done Quickly

Enrolling subjects to participate in the cohort study; building the baseline database; collecting, cataloging, and storing biologic materials; and developing follow-up procedures to update disease event and time-varying information takes considerable

Affiliations of authors: B. Langholz, D. C. Thomas, Department of Preventive Medicine, University of Southern California, Los Angeles; N. Rothman, S. Wacholder, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Correspondence to: Bryan Langholz, Ph.D., Department of Preventive Medicine, School of Medicine, University of Southern California, 1540 Alcazar St. CHP-220, Los Angeles, CA 90033 (e-mail: langholz@hsc.usc.edu).

See "Note" following "References."

© Oxford University Press

Table 1. Selected cohort studies with blood sample collections and food-frequency questionnaire data*

| Study† | Year blood collection began | No. of subjects with blood samples |
|--|-----------------------------|------------------------------------|
| NYU Women's Health Study, United States | 1985 | 14 000 |
| Northern Sweden Health and Diseases Study, Sweden | 1985 | ~43 000 |
| ATBC, Finland | 1985 | 29 000 |
| Health Professional's Follow-up Study, United States | 1986 | 18 000 |
| ORDET, Italy | 1987 | 11 000 |
| ARIC, United States | 1987 | 16 000 |
| Nurses' Health Study, United States | 1989 | ~33 000 |
| Washington County, Maryland, United States | 1989 | 33 000 |
| Melbourne Collaborative Cohort Study, Australia | 1990 | 42 000 |
| JPHC, Japan | 1990 | 49 000 |
| Nurses' Health Study II, United States | 1991 | 30 000 |
| Women's Health Study, United States | 1992 | 27 000 |
| Women's Health Initiative, United States | 1993 | 164 000 |
| EPIC, Europe | 1993 | 350 000 |
| PLCO Study, United States‡ | 1994 | ~65 000 |
| Shanghai Women's Health Study, China§ | 1997 | ~55 000 |
| CPS-II Lifelink, United States¶ | 1998 | ~40 000 |

*Modified from Willett W. Nutritional epidemiology. 2nd ed. New York (NY): Oxford University Press; 1998. p. 486-7.

†NYU = New York University; ATBC = Alpha-Tocopherol Beta-Carotene Cancer Prevention Study Group; ORDET = Hormones and Diet in the Etiology of Breast Tumors Study; ARIC = Atherosclerosis Risk in Communities Study; JPHC = Japan Public Center-based prospective study on cancer and cardiovascular diseases; EPIC = European Prospective Investigation into Cancer and Nutrition; PLCO = Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; and CPS-II Lifelink = American Cancer Society Cancer Prevention Study-II Lifelink Cohort.

‡Hayes R: personal communication (planned number, still enrolling).

§Zheng W, Chow WH: personal communication (planned number, still enrolling).

¶Thun M: personal communication (planned number, still enrolling).

time, money, and effort. But once the cohort study resource is established and a sufficient number of cases have occurred, a study of genetic factors can proceed much more quickly and efficiently than a population-based study. At this stage, the registry has already ascertained cases, and control selection from within the cohort is a statistical and administrative activity. The main effort at this point is the laboratory work needed to determine genotypes. In contrast, a population-based, case-control study requires ascertainment and enrollment of diseased subjects, as well as blood collection, before reaching the laboratory analysis stage. Thus, once the cohort study is established and matured, the additional cost of a nested case-control study is relatively small, and, perhaps more important, the time needed to do the studies can be much shorter. Given this advantage, nested case-control studies provide a timely and reliable way to verify gene-disease associations found in other studies.

Multiple Studies From the Cohort Are Comparable

The cohort resource permits case-control studies of different disease outcomes, all from the same study base. The ability to efficiently study multiple tumors might compensate for their poorer efficiency for a single hypothesis. In addition, the outcomes can be reliably compared, if case ascertainment is comparable. For this purpose, a case-cohort study design should be

considered because controls can be shared among the various outcomes (14).

Nested Case-Control Studies Are Relatively Free of Selection Bias

It can be difficult to get controls in population-based, case-control designs that adequately represent the base population, and seldom is there any way to assess their representativeness. In a cohort study, the cohort is the base population, so random sampling from it is straightforward. Even if one cannot obtain genotypes on all of them, one can at least characterize the losses in terms of baseline characteristics.

Gene by Environment Studies

Gene-environment interactions can be explored in cohort studies with the use of relatively unbiased exposure data assessed biologically or by questionnaire that can potentially be collected at more than one point in time. Under these circumstances, the nested case-control study would avoid recall or other information biases that can occur when "environment" data are retrospectively obtained by interview, as is usually the situation with population-based, case-control studies. We note that misclassification of exposure status can seriously bias the assessment of gene-environment interactions and substantially reduce power (15). If nested case-control study subjects need to be contacted to obtain additional exposure information, contact information will be available to the study investigators. Also, participation rates will be high because cohort members have already agreed to participate in the main cohort study. Of course, interview data collected retrospectively will be subject to the same information biases as a population-based study and, to the extent that cases have died or refuse to participate, selection biases. Also, there can be added selection bias from loss to follow-up if censoring is informative.

Penetrance Estimation

Because the study population is enumerated, the penetrance of the genes can be reliably estimated from the nested case-control study, accounting for risk factors measured on the case-control subjects (16). In this regard, such studies could be definitive about penetrance for populations that are reasonably represented, in terms of unmeasured risk factors, by the cohort. Absolute risk can be estimated from population based, case-control studies but will be subject to much more error because of unidentified cases and because of inaccuracies in the estimation of numbers in the underlying population (17).

Penetrance estimation from family-based studies requires strong assumptions to extrapolate the prevalence of the gene from the included families to the general population. Although methods can, in principle, correct for ascertainment (18), these methods may still lead to upwardly biased estimates if the penetrance is not the same in all families (19).

Cost-Efficient Sampling Schemes

As described above, nested case-control studies from a cohort are a cost-efficient study design in the same way that case-control studies from a population are cost-efficient. To understand the relative contribution of individuals with respect to information about gene characterization from the cohort, it is useful to think of the cross-classification of four states on the basis of presence or absence of the gene and disease status (20).

The relative size of these cells is proportional to the amount of information contributed by individuals in these cells. Thus, if the disease is rare, the information contributed by each diseased subject is far greater than from a nondiseased subject, so that virtually all of the information about the gene–disease association may be captured with the use of information from a small fraction of the nondiseased subjects. This, of course, is the basis for case–control study designs. Disease status information that is available on all cohort members is used in sampling from the cohort, so that costly information (genotype) needs only to be collected on a fraction of the entire cohort to capture nearly all of the information in the cohort about gene–disease association. If the gene is rare, then the individuals with gene-positive cells will carry relatively large amounts of information. Although, of course, genotype status is not known for the cohort subjects, some factor that is correlated with the presence of the gene, such as family history, or a positive value from an inexpensive lab assay, may be available. Just as disease status information may be exploited to increase the information per sampled subject, gene-related information can be used to increase the information from each sampled subject and to decrease the sample size, and thus the cost, of the study. Counter-matching and other two-stage designs that exploit gene-related information on cohort members may reduce the number of subjects who must be genotyped, relative to standard case–control studies (21–25).

OTHER USES FOR COHORT STUDIES

We have thus far focused on case–control substudies within the cohort. But an established cohort study can be useful in other study designs and also can be used to address issues other than gene–disease associations.

Opportunities to Expand the Usefulness of the Cohort by Adding Information From Family Members

In addition to exploiting the information available on the cohort members themselves, it may be feasible in some circumstances to use them as probands for family-based, cohort or case–control studies [e.g., (26,27)]. For example, one might use the cases from the cohort and a random or matched sample of unaffected controls as probands in the kin-cohort design (19). In this approach, only the cohort members would require genotyping. The only additional data to be obtained would be the subjects' family histories in first-degree relatives, which might be already available from the probands' baseline questionnaires or might be obtained by follow-up questionnaires to the probands, without the need to formally enroll their relatives. These designs could also be extended to include more distant family members or to obtain genotypes from selected relatives. Multistage sampling designs might be useful in this context (24).

Conceptually, these family-based designs are no different from those discussed elsewhere in this workshop—sampling probands from the general population. However, given the need to involve family members, the availability of already collected family history information on a baseline questionnaire could considerably simplify their implementation, particularly for designs restricted to multiple-case families. For example, in the Department of Preventive Medicine, University of Southern California, Los Angeles, a sibling case–control study of breast and ovarian cancers nested is currently being conducted within a multiethnic cohort study in Los Angeles and Hawaii. All sibships with at least two cases (in either the sibship or their par-

ents) and at least one control are being enrolled, this information being readily available from the baseline family-history questionnaire. In addition, being able to characterize the representativeness of the case series vis-à-vis the larger cohort would help address concerns about possible selection biases in such designs.

Opportunities to Explore Population Stratification Bias

Some investigators (26,28,29) have voiced concerns about the potential for bias in case–control studies with unrelated controls because of population stratification. Although diabetes in the Pima Indians provides one notable example of such bias (30), the extent of the problem in general is still unresolved. Cohort studies that use unrelated individuals are, in principle, subject to exactly the same concerns. However, baseline information on ethnicity and other risk factors that are sometimes available on the entire cohort provides opportunities to examine the potential severity of this problem in ways that may not be feasible when sampling from the general population without a well-defined sampling frame. In other words, by treating the cohort as a “mini-population,” one could contrast the results from alternative case–control designs (e.g., using unrelated and family member controls) in terms of their ability to correctly estimate the parameter that would have been obtained by using the full cohort and to examine the role of the available ethnic or other baseline data to account for methodologic differences. The cohort results could then be used to infer the potential effect of population stratification in other studies on the basis of the ethnicity and other risk factors in the study population and the type of study design used.

Control of Ethnic Stratification With the Use of Molecular Markers

A relatively unexplored approach to the ethnic stratification problem entails the use of polymorphic markers to infer ethnicity. Shriver et al. (31) have proposed an approach to inferring the ethnic affiliation of individuals with the use of a panel of markers whose allele frequencies vary substantially between ethnic groups. By using a similar panel with a maximum likelihood approach to discrimination, Shriver et al. demonstrated reasonable separation in the likelihood scores between European Caucasians and African-Americans but less discrimination for Hispanics. To date, we are not aware of any applications of this approach to ethnic stratification in epidemiologic studies of candidate gene associations. However, Pritchard and Rosenberg (32) discuss the potential of this approach for detecting population stratification and provide guidelines as to the number of unlinked markers that would be needed. Although this idea could be implemented in a population based, case–control study, the approach might be most easily implemented in a cohort study by first picking a few ethnically matched controls per case and, as part of the genetic typing, ascertain the alleles for markers of “ethnic origin.” A further analysis of cases with controls of the most closely matching ethnicity would provide an indication of the extent that population stratification might explain an observed gene–disease association.

DISCUSSION

Genetic studies within a cohort study are feasible when blood or other genetically analyzable material is available on study members and enough cases of disease are (or will be) available to ensure statistical precision. We have suggested that case–

control studies within the cohort would be the design of choice for gene-association studies because blood from only a small fraction of the cohort members would need to be genotyped. We have outlined a number of benefits of this approach, including low marginal cost and high reliability of such studies. Also, they could be used to investigate familial clustering of disease and population stratification issues.

Beside the restriction to relatively common diseases and genes, some other potential limitations exist to using cohort studies for genetic studies. Cohorts are often not representative of the "general population." This lack of representation may be because the cohort is formed precisely because it is a "high-exposure" group, such as in many occupational cohorts. Or subjects are enrolled from a convenient administrative entity, such as unions, or religious or professional organizations. Furthermore, those who choose to participate and, in particular, agree to provide a blood sample can ultimately be a rather small and selected proportion of those approached, thus, perhaps, limiting the generalizability of the estimated associations and risks. Cohort study reliability depends on many data-quality issues, such as the quality of questionnaire and other information gathered in the field, completeness of follow-up, and outcome ascertainment. In particular, if loss to follow-up is related to presence of the gene, bias can arise. If loss to follow-up is significant and differential across a gene-related factor (e.g., race or ethnicity), some protection from such bias would be obtained by stratification of the analysis by those factors because the follow-up will be less differential within these strata.

Mounting and maintaining a cohort study is a major endeavor. But, once established, the cohort study can play an important role by quickly and definitively verifying gene-disease associations and, potentially, gene-environment interactions as well as establishing the penetrance of the gene.

REFERENCES

- (1) Schwartz AG, King MC, Satariano WA, Swanson GM. Risk of breast cancer to relatives of young breast cancer patients. *J Natl Cancer Inst* 1985;75:665-8.
- (2) Elston RC, Rao DC. Statistical modeling and analysis in human genetics. *Annu Rev Biophys Bioeng* 1978;7:253-86.
- (3) Easton DF, Ford D, Bishop DT. Breast and ovarian cancer incidence in BRCA1-mutation carriers. *Am J Hum Genet* 1995;56:265-71.
- (4) Gail MH, Pee D, Benichou J, Carroll R. Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotype-proband designs. *Genet Epidemiol* 1999;16:15-39.
- (5) Caporaso N, Rothman N, Wacholder S. Case-control studies of common alleles and environmental factors. *Monogr Natl Cancer Inst* 1999;26:25-30.
- (6) Rothman N, Stewart W, Schulte PA. Incorporating biomarkers into cancer epidemiology: a matrix of biomarker and study design categories. *Cancer Epidemiol Biomarkers Prev* 1995;4:301-11.
- (7) Munoz A, Grange SJ. Methodological issues for biomarkers and intermediate outcomes in cohort studies. *Epidemiol Rev* 1998;20:29-42.
- (8) Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology*. New York (NY): Oxford University Press; 1993.
- (9) Mantel N. Synthetic retrospective studies and related topics. *Biometrics* 1973;29:479-86.
- (10) Lubin J, Gail M. Biased selection of controls for case-control analysis of cohort studies. *Biometrics* 1984;40:63-75.
- (11) Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci* 1996;11:35-53.
- (12) Kupper L, McMichael A, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *J Am Stat Soc* 1975;70:524-8.
- (13) Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986;73:1-11.
- (14) Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology* 1991;2:155-8.
- (15) Rothman N, Garcia-Closas M, Stewart WS, Lubin J. The impact of misclassification in case-control studies of gene-environment interactions. In: Vineis P, Malats N, Lang M, d'Errico A, Caporaso N, Cuzick J, et al., editors. *Metabolic polymorphisms and susceptibility to cancer*. Lyon (France): International Agency for Research on Cancer (IARC); 1999.
- (16) Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics* 1997;53:767-74.
- (17) Benichou J, Gail M. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* 1990;51:182-94.
- (18) Risch N. Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. *Am J Hum Genet* 1984;36:363-86.
- (19) Wacholder S, Hartge P, Struewing JP, Pee D, McAdams M, Brody L, et al. The kin-cohort study for estimating penetrance. *Am J Epidemiol* 1998;148:623-30.
- (20) Wacholder S. Design issues in case-control studies. *Stat Meth Med Res* 1995;4:293-309.
- (21) Langholz B, Borgan O. Counter-matching: a stratified nested case-control sampling method. *Biometrika* 1995;82:69-79.
- (22) Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect* 1994;102:47-51.
- (23) Goldstein AM, Andrieu N. Detection of interaction involving identified genes: available study designs. *Monogr Natl Cancer Inst* 1999;26:49-54.
- (24) Siegmund KD, Whittemore AS, Thomas DC. Multistage sampling for disease family registries. *Monogr Natl Cancer Inst* 1999;26:43-8.
- (25) Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort studies. *Lifetime Data Anal*. In press 1999.
- (26) Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. *Monogr Natl Cancer Inst* 1999;26:31-7.
- (27) Gail MH, Pee D, Carroll R. Kin-cohort designs for gene characterization. *Monogr Natl Cancer Inst* 1999;26:55-60.
- (28) Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037-48.
- (29) Witte JS, Gauderman WJ, Elston RC, Thomas DC. Bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999;149:693-705.
- (30) Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988;43:520-6.
- (31) Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet* 1997;60:957-64.
- (32) Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;65:220-8.

NOTE

Supported by Public Health Service grants CA42949 and CA52862 (B. Langholz and D. C. Thomas) from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.