Study-Design Issues in the Development of the University of Southern California Consortium’s Colorectal Cancer Family Registry

Robert W. Haile, Kimberly D. Siegmund, W. James Gauderman, Duncan C. Thomas

The University of Southern California Consortium is a participating center in the National Cancer Institute’s Collaborative Family Registry for Colorectal Cancer Studies (CFRCCS). Because data collection takes time, money, and effort, all of which are in short supply, we first defined our research objectives and then attempted to design our registry to enable us to address these objectives in an efficient manner. We decided on a family-based design, and our objectives are to characterize cloned genes that are generally accepted causes of colorectal cancer, to assess putative candidate genes, to map new genes, and to conduct prevention trials in high-risk subjects. For the gene characterization objectives, our primary aim is to estimate gene frequency and penetrance, with a secondary aim to investigate factors that may affect penetrance (allele-specific effects plus gene–gene and gene–environment interactions). We describe a multiple-stage design to select families into the registry. After a family is selected into the registry, we collect questionnaire data and blood samples on selected subjects only, and we tailor data collection decisions to each family (given who is affected and who is available) to optimize power per unit effort and cost. We also discuss practical decisions faced by our registry, including 1) defining a reference period for use in questionnaires; 2) deciding whether or not to establish cell lines and, if so, on whom; and 3) determining which cases should be tested for microsatellite instability. Finally, we address the appropriate use of data derived from high-risk clinics, within more broadly defined, population-based research. [Monogr Natl Cancer Inst 1999;26:89–93]

The University of Southern California (USC) Consortium is a participating center in the National Cancer Institute’s Collaborative Family Registry for Colorectal Cancer Studies (CFRCCS). Other centers in the CFRCCS are the Australasian Registry, encompassing centers in Australia and New Zealand; the Fred Hutchinson Cancer Center in Seattle; the Hawaii Cancer Research Center; the University of Toronto; and the Mayo Clinic. Member institutions of the USC Consortium are the Cleveland Clinic Foundation (CCF), Dartmouth, University of Arizona, University of Colorado, University of Minnesota, University of North Carolina, and the University of Southern California. Comments here pertain to deliberations undertaken by the USC Consortium in the development of its registry and will be limited to issues of study design (another manuscript, separate from this monograph, is in preparation to discuss practical, logistical, and ethical issues addressed by our consortium).

Our consortium comprises one high-risk clinic, the CCF, and six population-based registries that cover the states of Arizona, Colorado, Minnesota, and New Hampshire, 33 counties in North Carolina, and the County of Los Angeles in Southern California. Combined, the registries ascertain more than 6000 cases of colorectal cancer per year, including approximately 500 African-American, 100 Asian, 400 Latina/o, and 5000 non-Hispanic white case patients per year. From these registries, we plan to screen 5000 case patients with colorectal cancer for a family history of colorectal cancer (50% of white case patients diagnosed at or above 50 years of age in 1996, plus 100% of case patients diagnosed under 50 years of age and all African-American, Asian, and Latina/o case patients over a 4-year period, from 1996 through 1999. The CCF has 382 families with familial adenomatous polyposis (3250 individuals), 100 families with hereditary nonpolyposis colorectal cancer (HNPPC) (900 subjects), and 250 families with “HNPPC-like”/non-Amsterdam criteria (2250 subjects).

Because data collection takes time, money, and effort, all of which are in limited supply, we first defined our research objectives and then attempted to design our registry to enable us to address these objectives in a very efficient manner. Broadly defined, our research objectives are to characterize cloned genes that are generally accepted causes of colorectal cancer, such as the mismatch repair genes, to assess putative candidate genes, to map new genes that cause colorectal cancer, and to conduct prevention trials in high-risk subjects. For accepted and putative cancer genes, our primary aims are to estimate gene frequency and penetrance and to investigate factors that affect penetrance (allele-specific effects plus gene–gene and gene–environment interactions).

Design: Population-Based Registries

Aims

With these objectives in mind, Drs. Thomas, Gauderman, and Siegmund at USC entered discussions and analyses with Dr. Haile to determine an optimal design for the registry. For reasons described in another paper (J), we favored the use of a family-based design to meet our objectives (no unrelated, population-based control subjects). We were aware of another design option that includes control families but did not consider it further because it was our intuition that such control families do not provide enough additional information to warrant the increased costs (in terms of fieldwork and laboratory work) of including them, but this belief merits further analysis and consideration.
Within the general design framework (a family-based study in which the families are ascertained through a case patient), the researchers addressed questions of what proportions of single-case versus multiple-case families to include in the registry and, within families, on which subjects to collect risk-factor questionnaires, blood samples, tumor blocks, and other data.

Detailed results of deliberations are presented in an accompanying paper by Siegmund et al. (2). We discuss here some of the actual issues we dealt with in applying a multistage design to a “real life” circumstance, with reality-based constraints and practical considerations. The first question addressed was what types of families (defined by family history of colorectal cancer) should be sampled in what proportions to achieve an optimal design. In general, if one were primarily interested in estimating gene frequency or penetrance accurately, results suggest including a substantial number of single-case families because the variance of the estimator is greater in these families (a lower proportion carries the high-risk allele or mutation). In contrast, if one were primarily interested in gene–covariate interactions or mapping new genes, one would want to start with families with the most case patients and then ascertain families with fewer and fewer case patients, as the “heavily loaded” families are exhausted. It was clear that no single sampling strategy was optimum for all objectives. At this point, the principal investigators of each center in the consortium, along with Dr. Thomas, had to prioritize the research objectives. There was a strong consensus that obtaining accurate estimates of gene frequency and penetrance should be our highest priority (it should be noted that all P.I.s are epidemiologists and take a population-based perspective of this research). There was also a strong interest in retaining a reasonable ability to investigate gene–environment interactions. Mapping of new genes was considered the lowest priority for this registry because we believed other efforts were in a better position to find new genes and that we should concentrate on our strength of population-based research on cloned genes. With these priorities in mind, a multistage sampling scheme was recommended, as described in detail by Siegmund et al. (2). We comment on relevant issues here.

**Multistage Sampling Protocol**

The first step in the registry is to screen 5000 case patients with colorectal cancer for a family history of cancer. Given the large number of case patients to screen, it was imperative that we design a screening questionnaire that was quick and relatively accurate, understandable that we could not afford to verify all reported cancers at this stage of the study. To keep the questionnaire short, we decided to ask probands only about their siblings and parents and to ask only about colorectal cancer. For purposes of selecting families for our second-stage sample, any classification strategy that was applied equally to all case patients would not create any biases. Therefore, we could safely ignore other cancers and cancer history in second-degree relatives. It was also our feeling, not based on formal simulations, that including children of probands at this stage was not worth the effort, given the expected low yield of case patients. We also decided not to ask about polyps because information on polyp status is usually highly selected by family history, access to care, and other variables that are difficult to quantify. A key point here is that it does not matter if the information obtained at this stage is “accurate” in the multistage sampling paradigm suggested for our registry, provided one uses the measured, not the “true,” family history information to compute the sampling fractions as weights in the analysis. A sensitivity analysis of our design showed that underreporting and overreporting of family history (7% and 27%, respectively) resulted in only a slight loss in statistical efficiency that was partially overcome by a larger sample size and increased total costs (Siegmund KD: unpublished data). Therefore, even if having more accurate information initially would lead to a statistically more efficient sampling strategy, these inaccuracies would not produce any bias if analyzed correctly.

The screening survey will identify with some error nuclear families with one, two, three, or more cases of colorectal cancer. As described by Siegmund et al. (2), we decided to employ a sampling scheme that would have us include 16% of single-case, 32% of two-case, 48% of three-case families, and so on, for further data collection. This sampling scheme would yield a set of families that provides reasonable power to estimate gene frequency and penetrance [see Table 4, (2)] for a reasonable range of these parameters and also adequate power to estimate gene–environment interactions for reasonable ranges of effects, particularly when analyses can include case-cousin pairs in addition to case-sib pairs and case-sib triplets.

Given that we now had a sampling scheme for families, the next set of questions was on which subjects within families to collect questionnaire data and blood samples. Again, this selection procedure is described in detail by Siegmund et al. (2). We were guided by the principle that we should tailor data-collection decisions to each family (given who is affected and who is available) to optimize power for unit effort and cost. As one example, if an unaffected sibling were available, we would want to obtain both a blood sample and a questionnaire from that subject, who would serve as a control in case–control comparisons. If no such control were available, but parents of the proband were available, we would obtain a blood sample from them as they are informative for some genetic analyses, but we would not obtain risk-factor questionnaires because they would not be comparable to the case patients’ questionnaire, given the generational difference and secular trends in selected exposures of interest. For example, assume a case patient was diagnosed in 1999 at 40 years of age and she had a mother who was 70 years. Asking about history of oral contraceptive use during her 20s would be problematic because oral contraceptives were not widely available when the mother was in her 20s (the exposure prevalence in the source population of the case patient would be much higher than in the source population of the mother) (3–5). (We note here that gene–environment interactions may, in principle, be examined in parent–child trios with exposure data only on the proband, not the parents. This requires the assumption that the gene and environmental factor of interest are independently distributed conditional on the parents’ genotypes, or equivalently within families, not across families as in the case-only design.) In addition to sibling controls, we would consider the use of cousin controls for estimating main effects and gene–environment interactions. This theme of tailoring our data collection to maintain validity and enhance efficiency has guided many of our decisions, as described below.

**Practical Decisions**

We have addressed other study-design issues that have grown out of the basic decisions described above. We briefly address...
three here to illustrate the types of decisions one may face in developing a registry: 1) defining the reference period in questionnaires when asking about exposures or practices at one point in time (such as diet) as opposed to collecting a “lifetime” exposure history; 2) deciding whether or not cell lines should be established and, if so, on whom; and 3) determining which cases should be tested for microsatellite instability (MSI) or replication error (RER) phenotype.

Reference Period

Defining an appropriate reference period was problematic. A reference period is necessary to set an upper limit for obtaining information on “lifetime” exposures up to the reference period (e.g., before the diagnosis of cancer in a case patient), and it is also necessary when we are obtaining a report of behaviors or exposures that refer to one period of time (e.g., one’s typical diet the year before diagnosis). Ideally, we want to set a reference period that is comparable for case patients and control subjects, that minimizes bias, and that does not compromise accuracy of recall. One distinction we face in defining “time” of exposure is between calendar time (a concern if there are secular trends in exposure or incidence) and “biologic” time, typically measured by age (a concern if there are critical ages of exposure and because we want case patients and control subjects to have similar periods at risk). An additional factor in family studies is that diagnosis of one case patient in a family may possibly cause other relatives in that family to alter their exposures if they believe those exposures may affect risk of disease (e.g., changing one’s smoking behavior or diet in response to a family member’s diagnosis of lung cancer or colon cancer). This family history may also influence physicians to alter advice they give to that family (e.g., possibly suggesting the use of oral contraceptives to a woman with a family history of ovarian cancer in the belief it will reduce her risk of ovarian cancer). In our opinion, there is no simple and completely correct definition of the reference period. For “lifetime” exposures, one may collect “complete” data up to the present and then treat the calculation of appropriate individual-specific reference dates as an analysis issue. For some exposures, such as diet, this option is not feasible.

The option recommended by the Epidemiology Working Group for all centers of the CFRCCS was to use “2 years ago” as the reference period for all subjects. The arguments for doing this were that it was a simple standard for all subjects (cases and controls) and avoided referring to the date of diagnosis, which some investigators believe would reduce the potential for recall bias. There are arguments against it.

One argument is that it is naive to believe that the case patients are not using their date of diagnosis as a reference period in their minds because it is such an important event to them. Therefore, we ought to use the date of diagnosis because it provides a more natural reference period than “2 years ago” that has an arbitrary quality about it, especially for case patients. This arbitrary quality may make it difficult for all subjects to relate to it, thereby compromising the accuracy of responses. A second argument is that, within families, one subject may complete a questionnaire in the current year, while another subject, for a host of reasons (divorce, ill health, and changing jobs are some reasons we have encountered), may delay completing the questionnaire for a number of years; therefore, the reference period now refers to a different calendar time.

An alternative is to define the reference period as the year before the diagnosis for case patients. Some investigators worry about recall bias because this reference period means more to a case patient than to a control subject. Also, within multiple-case families, case patients may have been diagnosed many years apart, even within sibships, so case patients are left referring to different periods of time (either calendar time or age, depending on how the reference period is defined). Even if case patients are asked to refer to a period of time before diagnosis, the situation of what to do with control subjects is unresolved. The most natural period for them is to recall events, such as diet, within the past year.

Some researchers worry about the lack of comparability with case patients if they are referring to a period a few years ago, before diagnosis, and control subjects are referring to the current time, especially for exposures with strong secular trends. Setting the reference period for control subjects based on the date of diagnosis for the case patient in the family is another option. Here, we may be asking control subjects about exposures for a period of time that may seem arbitrary and removed from the present.

As we stated at the beginning, no simple solution is available that satisfies all investigators. The CFRCCS has decided to use the reference period of 2 years ago, with the option of stretching this reference period to 3 years to avoid asking case patients about exposures after their diagnosis.

Cell Lines

Another question that we face is on which subjects to establish cell lines that offer, in principle, an infinite source of DNA and RNA. Because characterizing cloned genes, as opposed to finding new ones, is a higher priority, we will limit comments to research with cloned genes. One purpose of the registry is to evaluate possible candidate genes, with association studies of main effects and possibly gene–gene and gene–environment interactions. We usually use a polymerase chain reaction-based technique for genotyping the polymorphism of interest, so we would probably not run out of DNA (provided we collect enough in the beginning) for this purpose; therefore, establishing and maintaining cell lines would probably not be justified for this type of research.

Another major purpose of the registry is to characterize, at the population level, relatively rare susceptibility genes that have already been identified as causes of colorectal cancer (such as the mismatch repair genes) in terms of gene frequency, penetrance, and gene–covariate interactions. With this purpose in mind, we face the prospect of screening large genes for relatively rare mutations. This process can use up a lot of DNA to identify and confirm the nature of the sequence variation. It is for this type of research that cell lines may be justified. We see two extreme options and various scenarios in between the extremes. One extreme is to establish a cell line only on the proband in the family. If we actually ran out of DNA on other family members but had it on the proband, we could still provide model-based estimates of gene frequency and penetrance. One could also explore gene–environment interactions if we were prepared to assume gene–environment independence within families, as discussed above. At the other extreme, we could establish cell lines for “all” family members who would contribute useful information to our planned analyses (frequency, penetrance, and interactions).
obvious disadvantages are the substantially increased labor and costs.

At least two less extreme options are worth considering. One is to establish cell lines only for the proband and require the initial search for mutations to use DNA (or RNA) only from the proband. Once a mutation is identified, more efficient assays can be used to test for that mutation in all other family members. Valid case–control comparisons would, however, have to be restricted to mutations that could be detected using the same methodology for cases and controls. The other option is to tailor the establishment of cell lines on a family-by-family basis, depending on the structure of the family and the intended analyses, which is similar to what we are planning to do for collection of blood samples and questionnaires.

**Tumor Blocks**

Another issue that we are addressing is the collection of tumor blocks and the testing for MSI, also referred to as RER. This activity becomes an issue because the collection and processing of tumor blocks takes time and money. A consensus at the CFRCCS is that we should determine MSI status for all cases (when isolated cases are entered into the registry) or for all families in the CFRCCS. If germline mutations in a mismatch repair gene were the only interest, perhaps collecting tumor blocks from one case patient with colorectal cancer per family and testing for MSI would suffice as a relatively accurate classification of that family with respect to MSI status (for purposes of detecting germline mutations that are responsible for the MSI). This approach would not work, however, for classifying cases in which the MSI is caused by alternative (nongermline) events, such as hypermethylation of the promoter region of hMLH1, that may or may not track in families. Collecting tumor blocks from other case patients with colorectal cancer (in addition to the proband) would enable us to investigate the prevalence of MSI (by all causes), sharing of MSI status among cases, as well as sharing of other molecular events, which could help elucidate modifier genes and pathogenetic mechanisms associated with colorectal cancers with MSI. Collecting tumor blocks from other sites of cancer would enable us to extend the investigations to these cancers as well, shedding light on common mechanisms and helping us to define what cancers are actually associated with HNPCC and other syndromes. Currently, our consortium has decided to try to obtain tumor blocks for all reported cancers in family members we deem potentially informative for planned analyses, but we are tracking the costs and may amend this decision in the future.

In summary, with respect to our population-based research, we started by defining our objectives. The reality that we cannot address all objectives equally well forced us to set priorities. Within these priorities, we tailor decisions about sampling and data collection to the family structure (which relatives are affected and which ones are likely to be available for data collection) in an attempt to maximize our power for planned analyses per unit cost and effort while retaining a valid design.

**HIGH-RISK CLINIC**

A final, major issue we are addressing is the appropriate use of data derived from high-risk clinics, such as those available from the CCF, a member institution in our consortium. In our opinion, such data may be quite useful for multiple purposes. First, research focused and applied to high-risk families is important, per se, because we need more information on how to best serve members of these families. For example, our consortium is interested in prevention trials specifically in high-risk subjects. Two means of defining high risk are a positive family history or actual genetic data that identify a subject as carrying a mutation in a gene of interest. High-risk registries generally provide a higher yield of subjects with a positive family history and carriers of mutations than a sample from the general population. Results from such trials would be most directly applicable to other, similarly defined subjects.

A broader question is whether results obtained from studies of subjects identified through high-risk clinics (with ascertainment schemes characterized at this meeting as “indescribable”) can be used to estimate meaningful population parameters. We start by considering the generalizability of results from prevention trials. Here the main question would be the prevalence of effect modifiers and whether the prevalence of actual modifiers differs to an important degree between the high-risk families and other subjects to whom one may want to generalize the results. The problem is that we often do not know what these modifiers are and, therefore, have to make judgments about generalizability without this knowledge.

Another question is whether we can derive valid estimates of gene frequency and penetrance from these types of families. In principle, one may be able to do so if the correct analysis is performed and the model underlying the analysis is correctly specified. It may be instructive here to consider the results published to date about the estimated penetrance of BRCA1. In the analysis by Easton et al. (6), penetrance was estimated to be 82% by age 70 years, based primarily on data from highly selected high-risk families. In the analysis of BRCA1 mutations in Ashkenazi Jewish subjects by Streuwing et al. (7), penetrance was estimated to be about 56% by age 70 years. Results from a population-based study in Australia, which is part of the Collaborative Family Registry for Breast Cancer Studies, suggest that the penetrance of protein-truncating mutations is around 40% (8). Because penetrance drives so many decisions about how to manage carriers, it is important to resolve or to understand the basis of these different estimates.

It seems that at least two reasons are likely for the estimate of 85% by Easton et al. (6). It is possible that there is heterogeneity in the penetrance of BRCA1 mutations, defined by modifying factors, either genetic or environmental. High-risk families may be selected for modifying factors that increase penetrance. In this case, the estimates by Easton et al. may be correct for the types of families included in this analysis. The estimates by Streuwing et al. (7) may be lower because the specific mutations studied, which are more common in Ashkenazi Jews, may be less penetrant than other mutations or because the particular Ashkenazi Jewish populations studied may have other protective modifying factors. It is also conceivable that the analyses failed to adequately take into account residual familial correlations, which could bias penetrance upward. These two explanations are not mutually exclusive. In fact, it seems plausible that there may be some important modifying factors and that these factors may correlate within families. Given the experience with BRCA1 to date, it would seem prudent not to rely on estimates of gene frequency and penetrance derived from variously ascertained high-risk families until we better understand determinants of penetrance and how these determinants are associated with ascertainment.

92 Journal of the National Cancer Institute Monographs No. 26, 1999
This brings us to a consideration of determinants of penetrance. Can findings regarding gene–gene and gene–environment interactions derived from variously ascertained high-risk families be generalized to some broader population? Here again, we must consider the question of modifiers, in this case, modifiers of the interaction. We must ask how strong are these possible modifiers and how likely are they to be differentially distributed in a substantial manner. Because we are often operating in ignorance of the answers to these questions, it would seem unwise to generalize results. In our consortium, we plan to investigate gene–covariate interactions in data from the CCF (because the yield of carriers is expected to be higher than a population-based series of families) as a source of leads that may be followed up in our population-based families. In this manner, we can begin to build an empirical basis for seeing what interaction effects are similar or different between variously ascertained high-risk families and population-based series of families. In summary, high-risk families derived primarily from clinical settings are certainly a prevalent and potentially valuable resource, so we need careful thought about their proper use and integration into the population-based research many of us are conducting.

REFERENCES