CHAPTER 14.

Population-based study designs in molecular epidemiology

Montserrat García-Closas, Roel Vermeulen, David Cox, Qing Lan, Neil E. Caporaso, and Nathaniel Rothman

Summary

This chapter will discuss design considerations for epidemiological studies that use biomarkers in the framework of etiologic investigations. The main focus will be on describing the incorporation of biomarkers into the main epidemiologic study designs, including cross-sectional or short-term longitudinal designs to characterize biomarkers, and prospective cohort and case–control studies to evaluate biomarker-disease associations. The advantages and limitations of each design will be presented, and the impact of study design on the feasibility of different approaches to exposure assessment and biospecimen collection and processing will be discussed.

Introduction

There is a wealth of existing and emerging opportunities to apply a vast array of new biomarker discovery technologies, such as genome-wide scans of common genetic variants, mRNA and microRNA expression arrays, proteomics, metabolomics and adductomics, to further our understanding of the etiology of a broad range of diseases (1–9). These approaches are allowing investigators to explore biologic responses to exogenous and endogenous exposures, evaluate potential modification of those responses by variants in essentially the entire genome, and define disease processes at the chromosomal, DNA, RNA and protein levels. At the same time, most biomarkers analysed by these technologies can still be classified into the classic biomarker categories defined more than 20 years ago (Figure 14.1), which include biomarkers of exposure, intermediate endpoints (e.g. biomarkers of early biologic effect), disease and susceptibility (10–17). Biomarkers in epidemiological studies can also be used to evaluate behavioural characteristics that affect the likelihood of exposure, such as tobacco smoking, as well as clinical behaviour and progression of disease. The use of biomarkers associated with exposure, disease development, and clinical progression within the same overall design is of increasing interest and has recently been termed ‘integrative epidemiology’ (18,19).

Figure 14.1. A continuum of biomarker categories reflecting the carcinogenic process resulting from xenobiotic exposures

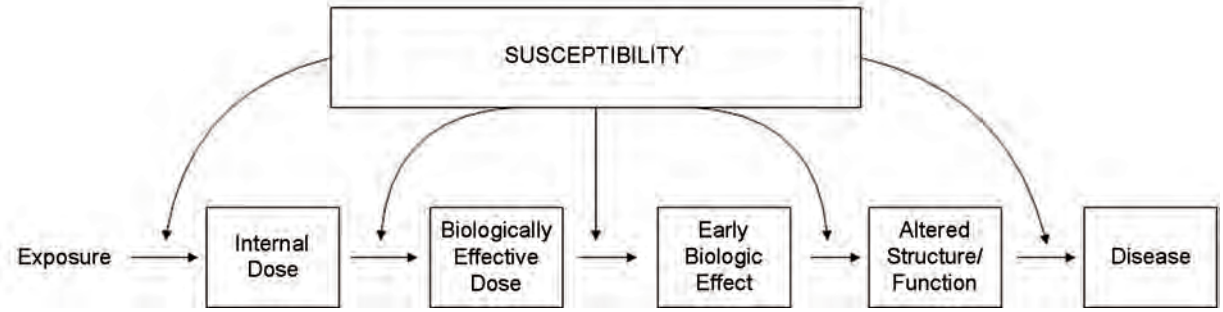

Figure compiled from (10) and (21, copyright © 2008, Informa Healthcare. Reproduced with permission of Informa Healthcare).

Regardless of the appellation used to describe the use of biomarkers in epidemiologic research, be it "molecular epidemiology," "integrative epidemiology," or the more limited "genetic epidemiology," the successful application of new and established biomarker technologies still depends on integrating them into the appropriate study design with careful attention to the time-tested principles of the epidemiologic method (16,20–24). Basic principles in vetting new biomarkers and technologies in pilot or transitional studies apply now more than ever (25–28). Understanding how to collect, process and store biologic samples (see Chapter 3), and the factors that influence biomarker levels, with particular attention to within- and between-person variation for non-fixed biomarkers (see Chapter 9), are key concerns. Testing for and optimizing laboratory accuracy and precision are also critical to the successful use of biomarkers in epidemiology studies (see Chapter 8). Finally, selecting the most appropriate, effective, and logistically feasible study design to use a given biomarker technology that answers a particular research question remains of paramount importance.

The focus of this chapter is on design considerations for epidemiological studies that use biomarkers, primarily in the context of etiologic research, including cross-sectional or short-term longitudinal designs to characterize biomarkers, and prospective cohort and case–control studies to evaluate biomarker–disease associations. A description of the general principles of study design (29–31) is outside the scope of this chapter. Instead, the focus is on describing the incorporation of biomarkers into the main epidemiologic study designs, pointing out the advantages and limitations of each, and showing how study design affects the feasibility of different approaches to both exposure assessment and biospecimen collection and processing.

Study designs in molecular epidemiology

Cross-sectional and short-term longitudinal studies with biomarker endpoints

In epidemiological terms, a cross-sectional study refers to a study design in which all of the information refers to the same point in time. As such, these studies provide a 'snapshot' of the population status with respect to exposure variables and intermediate endpoints, and, in some instances, disease at a specific point in time. Short-term longitudinal biomarker studies are studies in which subjects are prospectively followed for a short period of time (usually a few weeks to up to a year). Investigations are usually performed on healthy subjects exposed to particular exogenous or endogenous agents where the biomarker is treated as the outcome variable. These studies generally focus on exposure and intermediate endpoint biomarkers, and sometimes evaluate genetic and other modifiers of the exposure–endpoint relationship.

Questions addressed by cross-sectional and short-term longitudinal studies

Cross-sectional and/or short-term longitudinal studies are often used as follows:

1) To answer questions about whether or not a given population has been exposed to a particular compound, the level of exposure, the range of the exposure, and the external and internal determinants of the exposure. For instance, recent studies on haemoglobin adducts of acrylamide have shown that exposure to this toxic chemical

is widespread in the general population, due to dietary and lifestyle habits (32,33).

2) To evaluate intermediate biologic effects from a wide range of exposures in the diet and environment, as well as from lifestyle factors (e.g. obesity and reproductive status). This design can be used to provide mechanistic insight into well established exposure–disease relationships, and to supplement suggestive but inconclusive evidence on the possible adverse health effects of an exposure. For instance, studies have used haematological endpoints to investigate the effects of benzene on the blood forming system at low levels of exposure (34,35). These studies have found decreased levels in peripheral blood cell counts at exposures below 1 ppm, indicating that at low levels of exposure, perturbations in the blood forming system can be detected. These results hinted at the possibility of increased risk of leukaemia at low levels of benzene exposure, given the putative link between benzene poisoning (a severe form of haematotoxicity) and increased risk for leukaemia.

3) To evaluate whether or not there are early biologic perturbations caused by new exposures, or recent changes in lifestyle factors that have not been present long enough to have been evaluated for their association with disease. For example, there is considerable public health concern about the increased use of nanoparticles in both research and manufacturing operations (36). Various initial research studies and evaluations have demonstrated greater biological activity of nanoparticles compared with larger particles of the same material, and significant potential toxicity has been observed in laboratory animals exposed to some types of nanoparticles. However, given their recent introduction into commerce, the time between first exposure and the occurrence of any chronic health effect is most likely too short. In this particular example, the assessment of preclinical indicators of disease (e.g. markers of pulmonary inflammation) in asymptomatic individuals would be of importance to identify potential adverse health effects at an early stage.

4) To study changes in exposure and/or intermediate endpoints to determine the effectiveness of intervention studies. For instance, the effect of exercise and weight loss interventions on serum levels of four biomarkers related to knee osteoarthritis (cartilage oligometric matrix protein (COMP), hyaluronan, antigenic keratin sulfate, and transforming growth factor-β-1 (*TGF-β1*)), and clinical outcome measures (e.g. medial joint space, pain) were examined (37). Intervention programmes indeed resulted in changes in COMP (which was associated with decreased knee pain) and *TGF-β1*.

Cross-sectional and short-term longitudinal studies using exposure markers

Biomarkers of exposure measure the level of an external agent, its metabolic by-products in either the free state or bound to macromolecules, or the specific immunologic response it elicits. In addition, exposure biomarkers measure endogenously produced compounds, which may be influenced directly or indirectly by external factors (e.g. hormones), as well as by genetic factors. The first epidemiological evaluation of potential biomarkers of exposure generally occurs in cross-sectional studies in the general population, or in subgroups with specific, well characterized exposure and lifestyle patterns. Sometimes a biomarker of exposure can be used only in cross-sectional studies to determine if a population is exposed to an agent of concern, or used as an independent marker of exposure in studies evaluating intermediate biomarker endpoints.

The applicability of exposure biomarkers in cross-sectional studies depends on certain intrinsic features related to the marker itself (e.g. half-life, variability, and specificity of the marker) and the exposure pattern (see Chapter 9). The first requirements for successful application of an exposure marker are that the assay is reliable and accurate (see Chapter 8), the marker is detectable in human populations, and important effect modifiers (e.g. nutrition and demographic variables) and kinetics are known (20). Second, the timing of sample collection in combination with the biological half-life of a biomarker of exposure is key, as this determines the exposure time window that a marker of exposure reflects. The time of collection may be critical if, as is often the case in cross-sectional studies, only one sample per subject can be obtained on a given occasion, and if the exposure is of brief duration, highly variable in time, or has a distinct exposure pattern (e.g. diurnal variation in certain endogenous markers, such as hormones) (38). Chronic, near-constant exposures pose fewer problems. However, most biomarkers of internal dose generally provide information about recent exposures (hours to days), with the exception of markers of persistent pesticides, dioxins, polychlorobiphenyls, certain metals, and serological markers related to infectious agents, which can reflect exposures received many years before.

Cross-sectional and short-term longitudinal studies using intermediate endpoints

Intermediate biomarkers directly or indirectly represent events on the continuum between exposure and disease, and can provide important mechanistic insight into the pathogenesis of disease. As such, they complement classic epidemiological studies that use disease endpoints. For instance, the use of intermediate biomarkers in cross-sectional studies can provide initial clues about the disease potential of new exposures years before a disease develops (10,15,39–41).

One group of intermediate biomarkers, biomarkers of early biologic effect (10), generally measure early biologic changes that reflect early, and generally non-persistent, effects. Examples of early biologic effect biomarkers include: measures of cellular toxicity; chromosomal alterations; DNA, RNA and protein expression; and early non-neoplastic alterations in cell function (e.g. altered DNA repair, altered immune function). Generally, early biologic effect markers are measured in substances such as blood and blood components (red blood cells, white blood cells, DNA, RNA, plasma, sera, urine) because they are easily accessible, and, in some instances, it is reasonable to assume that they can serve as surrogates for other organs. Early biological effect markers also can be measured in other accessible tissues such as skin, cervical and colon biopsies, epithelial cells from surface tissue scrapings or sputum samples, exfoliated urothelial cells in urine, colonic cells in feces, and epithelial cells in breast nipple aspirates. Other early effect markers include measures of circulating biologically active compounds in plasma that may have epigenetic effects on disease development (e.g. hormones, growth factors, cytokines).

For maximum utility, an intermediate biomarker must be shown to be predictive of disease occurrence, preferably in prospective cohort studies (40) or potentially in carefully designed case–control studies. The criteria for validating intermediate biomarkers have focused on the calculation of the etiologic fraction of the intermediate endpoint, which varies from 0 to 1 (40,41). For intermediate endpoints with etiologic fractions that are close to 1.0, either positive or negative results in cross-sectional studies of an exposure–intermediate endpoint relationship are particularly informative. For intermediate endpoints linked to risk of developing disease but with a substantially lower etiologic fraction, the interpretation must be more circumspect. Specifically, a positive association between an exposure and an intermediate biomarker is informative, but a null association does not rule out that the exposure is associated with adverse health outcomes, as the exposure may act through a mechanism not reflected by the particular endpoint under study.

One of the most well known examples of a validated intermediate marker is low-density lipoprotein (LDL) cholesterol. Epidemiological studies have shown that elevated LDL cholesterol is one of the major causes of coronary heart disease (CHD). Given its high predictiveness, risk management/intervention programmes are focused on lowering and identifying factors that would reduce LDL levels and thus the risk for CHD (42). Unfortunately, there are very few examples like LDL. For instance, chromosomal aberrations in peripheral blood lymphocytes have been extensively used as the classic biomarker of early genotoxic effects in cross-sectional studies of populations exposed to a wide variety of potential carcinogens (43–45). Several cohort studies have reported that the prevalence of chromosomal aberrations in peripheral lymphocytes can predict subsequent risk of cancer (46–51). The predictive performance of this biomarker was shown to be similar irrespective of whether the subjects had been smokers or occupationally exposed to carcinogenic agents (52). In contrast, such associations were not observed for the sister chromatid exchange assay, another biomarker of genotoxicity also measured in peripheral lymphocytes (49–51).

Interpretation of results from cross-sectional studies using intermediate endpoints is, as indicated before, premised on the assumption that the intermediate endpoints reflect biological changes considered relevant to disease development. This may be based on *in vitro* and animal models or on previous observations that the biomarker is altered in human populations exposed to known toxicants. However, these studies are not capable, in and of themselves, of directly establishing or refuting a causal relationship between a given exposure or a given level of exposure and risk for developing diseases. Results of studies using most intermediate biomarkers as outcome measures are only suggestive; a biomarker may be overly sensitive (i.e. it may respond to low levels of chemical exposures that are below the disease threshold, if one exists), be insensitive, reflect phenomena that are irrelevant to the disease process, or fail to reflect important processes involved in the pathogenesis of disease.

Variance in biomarker response

The applicability of exposure and intermediate endpoint markers in cross-sectional and semi-longitudinal studies depends to a large extent on the variability in biomarker response between persons and over time (see Chapters 8 and 9). If a biomarker response is highly variable over time within a person, then it is clear that a single measurement of such a marker would be a poor estimate of the average marker level of a certain individual. However, even if the variance over time is small, and thus a reasonable estimate of the individual's average marker response, the applicability in epidemiological studies might be limited if the variance between individuals is small as well. In the end, the applicability of a marker in epidemiological research depends on the relative level between the interindividual and intraindividual variability in marker response. A useful measure in this regard is the intraclass correlation coefficient (ICC), which can be defined as the interindividual variance divided by the sum of the interindividual and intraindividual variance; in other words, it represents the fraction of the total variance that can be attributed to differences between individuals. Short-term longitudinal studies are ideal to collect information needed to estimate this key parameter. Chapter 9 provides a more detailed description of methods used to quantify biomarker variability and its impact on biomarker-disease associations.

Strengths and limitations

A distinct advantage of cross-sectional and short-term longitudinal studies is that detailed and accurate information can be collected on current exposure patterns, potential confounders, and effect modifiers. Furthermore, they can take advantage of a wide range of potential analytic (molecular) approaches, particularly those that require cell culturing and extensive processing within an often short period of time after collection (e.g. RNA, protein stabilization).

Cross-sectional and short-term longitudinal biomarker studies can collect very accurate information on the dose–response relationship between external or internal exposures and intermediate endpoints; these detailed exposure status data should be exploited to the fullest. As most biologic markers of exposure reflect exposures over the previous several days to months, this information must be collected over the etiologically relevant time period. For example, in a study on haematologic, cytogenetic and molecular endpoints among workers exposed to benzene, measurements were collected for over a year before determination of the biological endpoints to unequivocally assess individual exposures (53). Furthermore, given the increasing interest in identifying potential gene-environment interactions in chronic diseases, accurate measurement of the environment becomes very important. Simulation studies have shown that even a modest amount of nondifferential exposure misclassification can dramatically attenuate the estimate of the interaction parameter and increase sample size requirements (54). As such, cross-sectional and semi-longitudinal studies could have a distinct advantage in elucidating gene-environment interactions.

Summary and future directions

Cross-sectional and semi-longitudinal study designs have been successfully applied to: answer questions about whether or not a given population has been exposed to a particular compound; evaluate intermediate biologic effects from a wide range of exposures in the diet and environment; evaluate whether or not there are early biologic perturbations caused by new exposures or recent changes in lifestyle factors that have not been present long enough to have been evaluated for their association with disease; and to study changes in exposure and/or intermediate endpoints to determine the effectiveness of intervention studies. However, as indicated previously, the interpretation and therefore the usefulness of these studies depend heavily on the validity of the markers measured. The availability of numerous prospective cohort studies with stored blood specimens should enhance the ability to rapidly test the relationship between a wide variety of early biologic effect markers, using both standard and emerging technologies (55,56), and disease risk. Such studies could ultimately produce a novel endpoint to evaluate the disease potential and mechanisms of action of various risk factors.

Prospective cohort studies

In contrast to cross-sectional studies where biomarkers are the outcome variable, in prospective cohort and case–control studies the risk of disease is the outcome of interest. Prospective cohort studies collect exposure information and biological specimens from a group of healthy subjects who are then followed-up to identify those who develop disease. Establishing a cohort study is initially very costly and time-consuming, as large populations must be recruited and followed-up long enough to identify sufficient numbers of cases with the disease of interest.

Although power is limited by the overall cohort size and frequency of the outcome, in the long run the cohort design becomes more cost-efficient, since it can study multiple disease endpoints and provide a well defined population that can be easily sampled for efficiency (57–60). This section describes key features of cohort studies, particularly with regards to the use of biomarkers. Table 14.1 summarizes the strengths and limitations of this study design, as compared to the case–control designs described later in this chapter.

Participation in the study

Subjects in a cohort study often have distinct characteristics compared to their population of origin, by design or because of the motivation and level of commitment required to be included in such studies. Collection of biospecimens to measure biomarkers can have adverse effects on participation rates, even when collection procedures require minimally invasive procedures (e.g. buccal swab or oral rinse as opposed to blood collection). Cohort studies that collect questionnaires and biological samples at baseline or before disease onset can avoid selection biases, as long as specimens for each participant remain available and follow-up is complete. However, subjects with biological specimens collected after the cohort has been formed might have different characteristics from the rest of the cohort. Increasingly, concerns over privacy have also affected the willingness of participants to take part in some research studies.

Exposure assessment, timing of exposure, and misclassification

A major strength of the cohort design is that the sequence between exposure assessment and outcome is the same as the causal pathway: exposures are measured before disease diagnosis. This is particularly important for biomarkers that are directly or indirectly affected by the disease process (61), with the caveat that undiagnosed or preclinical disease may alter levels of specific biomarkers measured on specimens collected close to the date of diagnosis. Screening for disease at baseline (e.g. requiring recent colonoscopy for samples used in studies of colorectal cancer), or excluding cases diagnosed in the first few years after sample collection (lag analyses), can limit the effect of preclinical disease on biomarker levels.

Environmental data collected through questionnaires is less prone to recall biases (i.e. differential recall between cases and controls) than in case–control studies, thus facilitating the assessment of biomarker-environment interactions, such as gene-environment interactions (62–66). However, prospective studies often have a lower level of detail on specific exposures than case–control studies focusing on one or a few related diseases, due to the need to collect at least minimal data on the multiple exposures relevant to multiple outcomes. Therefore, although cohort studies can minimize the occurrence of differential misclassification, nondifferential misclassification of exposure might be larger than in alternative study designs.

Cohort studies with extended follow-up provide a wide range of time periods between biomarker collection and diagnosis of the outcome. This can be used to evaluate hypotheses relating to latent periods between the exposures of interest and the outcome. Theoretically, cohort studies have the advantage of collecting serial biological samples over time to evaluate biomarkers that vary in time. However, logistical and cost constrains often result in large studies collecting a single biological sample at one point in time. This results in diminishing the value of evaluating the relevant time window of exposure for disease causation, and studying markers with substantial seasonal or day-to-day variations, such as short-term exposure markers.

Chapter 9 describes important considerations in data analysis and inference related to the timing of exposure assessment. Below is a summary of these considerations in the context of cohort studies.

Misclassification due to random within-person variation. Most biomarkers vary from time to time within the same person. This variation could be due to multiple factors, including diurnal (e.g. melatonin) or monthly (e.g. estrogens) cycles, seasonal variation (e.g. vitamin D), recent dietary or supplement intake (e.g. vitamin C), as well as from the specificity of the assay used to measure the biomarker. If this variation is random and nondifferential between cases and controls, the bias will tend to attenuate measures of association.

When within-person variability for a particular biomarker is random, the correlation between single measurements in a population and the average of multiple measurements can be used to gauge the extent of attenuation in the association measure, or be used explicitly to correct the attenuation of relative risk estimates due to nondifferential misclassification (67). This information can be obtained from a representative subsample of

the cohort in which the biomarkers are measured at two or more distinct time points. It is important that the subjects with repeated measurements represent the larger cohort, so that the correlation used to correct risk estimates can be generalized to the entire cohort. However, true random sampling is often difficult to perform in large cohort studies, due to geographic dispersion and lower-than-optimal participation rates in more burdensome subsampling studies.

Time integration. The most common conceptual timeframes for exposure data in the epidemiology of chronic disease in cohort studies are long-term average measurements, as the induction time of most chronic diseases (e.g. cardiovascular disease, diabetes or cancer) is thought to be in the order of years or decades. Therefore, biomarkers would optimally represent cumulative exposure over relatively long periods of time, such as months or years. Some biomarkers may be able to integrate exposure time, which might also depend on sampling, processing, and storage protocols. For example, concentrations of many nutrients are less susceptible to short-term fluctuations in erythrocytes than in plasma or serum. Concentrations in adipose tissue, which is often more difficult to acquire, reflect even more long-term exposure history. It is therefore important to balance feasibility of sample collection with implications for time-integration of the biomarkers of interest.

Multiple biomarker levels. Obtaining multiple samples over time can increase the time-integration of exposures and biomarker measurements. Multiple biomarker measurements can be used in several ways, including averaging measurements or comparing subjects with consistently high versus consistently low levels. If within-person variation in biomarker measurements is assumed to be random, methods exist to estimate the number of replicate measurements required to estimate the 'true' mean value of a biomarker within a specified range of error. On the other hand, if variation is not random, due to changes in behaviour or secular trends, multiple measurements can be used to estimate exposure error. From a practical point of view, collecting multiple biological specimens from large numbers of subjects in cohort studies increases the cost of sample collection and storage, as well as the burden on study subjects, and thus might not always be feasible or recommendable.

Inference from biomarker/disease associations. The association between a biomarker and disease may also be influenced by the point in time in which the biomarker was assayed with respect to where it influences disease on the causal pathway. In the case of cancer, early events on the causal pathway ('initiators') may need to be distinguished from later events ('promoters'). Therefore, it is important that any biomarker related to exposures that are either initiators or promoters be measured during a time period when the exposure is most likely to exert its influence on disease. For example, initiating exposures should most likely be measured many years, possibly decades, before cancer diagnosis. In contrast, exposures considered to be promoters should be measured more closely to the time of diagnosis. If a biomarker of exposure is not measured at the etiologically relevant time, the association between the exposure and disease could be attenuated or not observed at all. The problem is that the true latency between exposure and disease diagnosis is often not known. The within-person variability in the biomarker of exposure is also important to consider in respect to the optimum time point for measuring it with respect to disease risk. If the within-person variability is low, then careful timing of exposure measurement is not necessary. However, if there is large within-person variability in the biomarker, then measurements should be made as close to the time of predicted maximum effect as possible.

Considerations in biospecimen collection, processing and storage

The collection and storage of large numbers of samples needed for cohort studies using biological specimens is very complex (see Chapter 3 for considerations in sample collection, processing, and storage). Given that samples are often used years after collection, optimal biospecimen collection, processing and storage protocols that will allow the performance of a wide range of assays in the future are critical (68,69). Therefore, validation studies aimed at optimizing sample handling and storage protocols, according to the impact of these procedures on the stability of samples and biomarker measurements, are strongly recommended (69–72). Validation studies can assess considerations such as the influence of time of collection to arrival at the processing laboratory (e.g. blood collection tubes with different preservatives, anti-coagulants or clot accelerators, temperature during shipping, impact of time between collection to processing), processing protocols (e.g. isolation of serum for proteomic analyses (73)), and long-term storage (e.g. freezing temperature,

impact of thaw/freeze cycles) on biomarker measurements.

Of paramount consideration is limiting the loss of information due to exhaustion of archived samples. The problem of sample exhaustion is most evident in prospective cohorts examining incident disease, as the amount of sample collected before diagnosis is by definition finite. Moreover, due to the advantages of the prospective cohort design, interest in the utilization of biological specimens could be great among the scientific community. Therefore, it is important for investigators to try to minimize the volume of sample used for measuring any one biomarker, either by reducing the volume of sample used or by maximizing the number of assays that can be made at any one time (multiplexing). Also beneficial is the formation of an advisory board to aid investigators in evaluating both internal use and external requests for access to precious prospective samples. While these considerations are obvious for those participants who are diagnosed with disease, biological samples from healthy or control subjects should also be carefully preserved, as these subjects may become cases in the future.

Despite advances in technology, such as whole-genome amplification to increase the amount of available DNA for assays, many biomarkers still require large amounts of biological samples. This limits the number of measurements that can be carried out on the limited resource of biological samples from cohort studies. Collecting additional specimens is often difficult in cohort studies, as members move, are lost to follow-up, or do not wish to go through the further inconvenience of providing an additional sample.

Statistical power

The major weakness of cohort studies is that even for common diseases the number of cases is limited by the cohort size and follow-up time. Even a very large cohort may not acquire enough cases of rare diseases to achieve adequate statistical power after long follow-up periods. Considering that cohorts using biological samples tend to be small or subsets of larger questionnaire-based cohorts, this is a particular problem for studies using biomarkers. Recently, a movement to form consortia of cohorts, such as the Cohort Consortium to study causes of cancer (http://epi.grants.cancer.gov/Consortia/cohort.html), has begun to address the problem of statistical power by coordinating biomarker measurements and analyses. Many consortia have been formed to support genome-wide association studies of many diseases (updated information on new publications from these efforts can be found at http://www.genome.gov/gwastudies/).

Sampling designs

When a cohort is chosen at random from the general population, the exposures in the cohort will be representative of the exposures in the general population. If the hypotheses to be tested rely on participants having either rare or extreme exposures in the general population, then oversampling these people or restricting the cohort to certain exposed groups would increase efficiency by increasing the prevalence of these exposures in the cohort.

For many large cohort studies, it is not feasible to assay all participants for a given biomarker. Therefore, with a few exceptions (such as assays that can only be performed on fresh samples), some selection of cases and controls will be necessary. This can be attained by using sampling designs in which only samples from cases and a random subset of non-cases are analysed, thus considerably reducing laboratory requirements and cost (60).

Nested case–control

The nested case–control study is an efficient sampling scheme that includes all cases identified in the cohort up to a particular point in time, and a random sample of subjects free of disease at the time of the case diagnosis. Increasing the case-to-control ratio to two or three controls per case can easily increase the efficiency of nested case–control studies. Optimally, controls should be selected for each case from the pool of participants that have not developed the disease at the time the case was diagnosed (risk set sampling). Alternatively, controls may be selected from all of the participants at baseline who were not diagnosed with disease throughout follow-up. Simulation studies have shown that as long as the proportion of the baseline cohort that acquires disease is low (e.g. less than 5%), the bias introduced by violating risk set sampling is minimal.

Case–cohort

A case–cohort design includes a random sample of the cohort population at the onset of the study and all cases identified in the cohort, up to a particular point in time (74). This design allows for the evaluation of several disease endpoints using the same comparison group (referred to as a subcohort). It may reduce the amount of laboratory work by assaying a subcohort of

subjects at baseline, and then adding case information as cases accrue. While there are statistical considerations that must be taken into account when analysing case–cohort studies, these are now included in most statistical packages. Of greater concern in case–cohort studies are problems more unique to biomarker studies. For example, if the biomarker being assayed degrades over time or if there is substantial laboratory drift in measurement, then cases assayed at varying time periods after baseline (when controls were assayed) can lead to bias. Additionally, laboratory personal are less easily blinded to case or control status, which can also lead to bias. These factors limit the utility of the case–cohort design in biomarker studies. Another limitation of this design is that since the same disease-free subjects are repeatedly used as controls for different disease endpoints, depletion of samples from this group can become an issue.

Sample comparability

The methods by which biological samples are collected, handled, and stored can influence the measurement of many biomarkers of exposure (see Chapter 3). Therefore, to have valid biomarker studies, case and control samples must be handled in the same way. For prospective studies, it is also important to consider the length and type of storage, as some biomarkers may degrade over time even under ideal conditions. Thus, it is important to match cases and controls on the method of sample collection, duration of storage, as well as other factors that may be related to the biomarker of interest, such as fasting status or season of collection. Additionally, batch-to-batch variation in assay measurement should also be considered. This can and should be minimized by assaying matched cases and controls at the same time, regardless of the study design.

Screening cohorts

Prospective cohort studies are sometimes designed within screening cohorts. In this design, screening failures lead to missing prevalent cases among cohort participants that are misclassified as controls (75). Although repeated screening reduces misclassification of subjects, cases discovered in follow-up cannot be distinguished from prevalent cases missed by the initial screening or incident disease. However, the degree of misclassification of prevalent and incident cases can be assessed by analyses of time to diagnosis or pathological characteristics. Intensive screening may also uncover a reservoir of latent disease that would not otherwise become clinically relevant, and that might differ from disease detected through clinical symptoms (76,77).

Resources and infrastructure

The vastly greater size of cohorts compared to other designs, such as case–control studies, and the time period required for the cohort to mature, mean that a substantially greater initial investment is required to establish the cohort. For cohort studies that incorporate biological materials, the infrastructure to support biospecimens' databases, freezers, and processing require a correspondingly greater effort and cost. While all studies with biospecimens must consider the risk of untoward events (e.g. freezer failure), the anticipated long useful life of the samples from cohorts requires special emphasis on quality control and security issues (e.g. backup generators, monitoring, distributing samples among different freezers). In the next few years, however, the cost-per-case for studies fielded from a cohort will offer economies in comparison to fielding a new case–control study (57).

Summary and future directions

Prospective cohort studies provide invaluable resources to study biomarkers of risk, particularly those that can be affected by disease processes. Multiple prospective cohort studies are currently being followed-up for disease incidence with basic risk factor information from questionnaires and stored blood components, including white blood cells that can be used as a source of DNA. At the completion of ongoing collections, current studies will have stored DNA samples on over two million individuals (16). These studies will provide very large numbers of cases of the more common cancer sites (breast, lung, prostate, and colon) to evaluate genetic markers of susceptibility; biomarkers in serum or plasma, such as hormone levels; chemical carcinogen levels; and proteomic patterns. Most cohort studies do not have cryopreserved blood samples, as the procedure is very expensive and logistically challenging in large studies. Also, cohort studies often have a limited capability to collect detailed disease information or biological specimens to facilitate disease classification, as well as to follow-up cases for disease progression and survival studies. New cohort studies based on large institutions, such as health maintenance organizations (HMOs), could enable access to clinical records with more detailed disease information, archived biological specimens, and easier

follow-up of cases for treatment response and survival. Caucasian populations in wealthier countries are overrepresented in studies of most diseases, and the recent establishment of consortia in other populations, such as the Asia Cohort Consortium (http://www.asiacohort.org/), will be critical to study disease across geographically and ethnically diverse populations that might have different exposures to environmental risk factors and frequencies of susceptibility alleles.

Case–control studies

Case-control studies are conceptualized as a retrospective sampling of cases and controls from an underlying prospective cohort, referred to as the source population (29,31). The case–control design has been a mainstay of molecular epidemiology studies due to its well known traditional strengths including depth and focus of questionnaire information, biologically intensive specimen collection, potential to enrol large numbers of cases rapidly, and ability to target rare diseases that occur in small numbers in prospective cohort studies.

Types of case–control designs

Case–control studies can be hospital- or population-based depending on how the cases and controls are identified (Table 14.1). A major concern of case–control studies is proper case and control selection. Proper controls are representative of the study base from which the cases arise (29). Identifying either a random sample from the general population or the source population for cases presenting at a particular hospital(s) may be difficult. Population-based studies attempt to identify all cases occurring in a pre-defined population during a specified period of time, and controls are a random sample of the source population where the cases came from. On the other hand, cases and controls in hospital-based studies are identified among subjects admitted to or seen in clinics associated with specific hospitals. As in the population-based design, the distribution of exposures in the control group should represent that from the source population of the cases. However, the source population is often more difficult to define in hospital-based studies.

Molecular epidemiology studies often use the hospital-based case–control design, as the hospital setting facilitates the enrolment of subjects, thus enhancing response rates, as well as the collection and processing of biological specimens. Enrolment of subjects is also made easier by having in-person contact with study participants by doctors, nurses or interviewers, which usually results in higher participation rates (78). Because study subjects are generally less geographically distributed than those in population-based or cohort studies, rapid shipment of specimens to central laboratories for more elaborate processing protocols, such as cyropreservation of lymphocytes, is made possible. Rapid ascertainment of cases through the hospitals also facilitates the collection of specimens from cases before treatment, thus avoiding the potential influence of treatment on some biomarker measurements.

Potential for selection bias is one of the most important limitations of case–control designs. The impact of selection bias in hospital-based studies is not only related to the reasons for non-participation, but also to diseases in the control population. An example of this is selection of controls admitted to the hospital for other diseases that might themselves introduce bias if they are related to the genetic or environmental exposures under study, particularly when evaluating gene–environment interactions or joint effects (79). Further potential for selection bias occurs if cases or controls are less likely to participate because of problems in the collection of biospecimens. Since the source population for cohorts is explicit, selection bias is less of a problem as long as follow-up rates are high (61). Low participation rates in case–control studies, and particularly refusals related to providing biological specimens, can bias results, especially when cases are less likely to participate than controls and selection is related to the biomarker of interest. Low participation rates additionally threaten the population-based nature of the study, undermining its use for estimating absolute and attributable risk (29). Use of non-intensive biospecimen collection protocols can increase participation rates—for instance, the collection of buccal cells as a source of DNA instead of the more invasive phlebotomy (80).

Single disease

Case–control studies are generally limited to one disease outcome (or a few related diseases), but are unconstrained by the rarity of the disease, while cohort studies (including full cohort, nested case–control, or case–cohort studies) may identify multiple disease endpoints. The focus of a case–control study on one disease entity permits more detailed documentation of disease information and detailed diagnostic procedures not routinely collected in clinical practice, such as specialized imaging and access to pathologic tissue and other

Table 14.1. Advantages and limitations of prospective cohort and case-control designs in molecular epidemiology relevant to the collection of biological specimens and data interpretation

biological specimens for application of novel biomarkers of disease. Obtaining disease-related data in cohorts entails mounting an effort that is generally less efficient and more costly. The advantage of cohorts' ability to examine multiple outcomes may be somewhat limited by resources and logistics, limited exposure information, the diverse approaches to documenting disease incidence or mortality, and the rarity of some outcomes.

Costs for a series of case–control studies of different diseases can sometimes be reduced by sharing a single control group. When different diseases require different exposures, the partial questionnaire design may offer reduction in the burden to respondents, thereby potentially increasing participation (81). Even if these options are not feasible, using the same infrastructure for control selection for repeated studies can reduce costs.

Exposure assessment and misclassification

Exposure assessment through questionnaires in case–control studies of a single disease or multiple diseases sharing risk factors (e.g. breast, ovarian, and endometrial cancer) can be more detailed and focused than prospective cohort studies that often study multiple unrelated diseases. However, studies that rely on

retrospective exposure assessment may be affected by disease or its treatment. Also, questionnaire responses subject to rumination by respondents are susceptible to bias from differential misclassification. Biomarkers (except germ-line genetic variation) and responses to questionnaires may change as a consequence of the early disease process or diagnosis itself. Differential errors or recall bias from questionnaire information collected in case–control studies are possible, and their extent should be evaluated in the context of specific exposures and populations under study. Similarly, levels of certain biomarkers measured after diagnosis can be influenced by the disease process or treatment, and must be considered and evaluated to the extent possible for each biomarker of interest. Differences in biomarker levels among cases diagnosed at different stages of the disease can help evaluate whether differences in biomarker levels between cases and controls reflect an influence of the disease on the biomarker rather than the contrary.

The applicability of exposure biomarkers in case–control studies depends on certain intrinsic features related to the marker itself (e.g. half-life, variability, specificity) and the exposure time window that a marker of exposure reflects in relation to the biologically relevant time of exposure and timing of sample collection. Methods to evaluate these key biomarker features before their use in case–control and other epidemiological studies are described in the previous section and in Chapter 8. The time of collection may be critical if the exposure is of brief duration, is highly variable in time, or has a distinct exposure pattern (e.g. diurnal variation for certain endogenous markers, such as hormones). However, chronic, near-constant exposures pose fewer problems. Ideally, the biomarker should persist over time and not be affected by disease status in case–control studies. However, most biomarkers of internal dose generally provide information about recent exposures (hours to days), with the exception of markers such as persistent pesticides, dioxins, polychlorinated biphenyls, certain metals, and serological markers related to infectious agents, which can reflect exposures received many years before. If the pattern of exposure being measured is relatively continuous, short-term markers may be applicable in case–control studies of patients with early disease, so that disease bias would be less likely. However, short-term markers have generally limited use in case–control studies, as they are less likely to reflect usual patterns, and the disease or its treatment might influence its absorption, metabolism, storage, and excretion.

Biomarkers of susceptibility in case–control studies

The approaches to studying genetic susceptibility factors for disease have evolved very quickly over the last several years, owing to advances in genotyping technologies, substantial reductions in genotyping costs, and improvements in the annotation of common genetic variation, namely, the most common type of variant, the single nucleotide polymorphism (SNP). The principles and quality control approaches for the use of genetic makers in epidemiological studies are described in Chapter 6. Because inherited genetic markers measured at the DNA level are stable over time, the timing of measurement before disease diagnosis is irrelevant. In addition, it is highly likely that most genetic markers are not related to factors influencing the likelihood of participation in a study, and therefore selection bias in case–control studies is less of a concern for studying the main effect of genetic risk factors. Indeed, the robustness of genetic associations with disease for different study designs has been demonstrated in findings from consortia of studies that have shown remarkably consistent estimates of relative risk across studies of different design (82,83). Because genetic markers might influence disease progression, incomplete ascertainment of cases in case–control studies can introduce survival bias, particularly for cancers associated with high morbidity and mortality rates, such as pancreatic and ovarian cancers. This is a particular concern for population-based studies, unless a very rapid ascertainment system is implemented that enrols cases as close as possible to the time of diagnosis.

Susceptibility biomarkers can also be measured at the functional/phenotypic level (e.g. metabolic phenotypes, DNA repair capacity) (16). While genotypic measures are considerably easier to study than phenotypic measures, since they are stable over time and much less prone to analytical measurement error, phenotypic measures are likely to be closer to the disease process and can integrate the influences of multiple genetic and post-transcriptional influences on protein expression and function (84). Therefore, in spite of the advantages in measuring genotypic changes, when complex combinations of genetic variants and/or important post-transcriptional events determine a substantial portion of interindividual variation in a particular biologic process, phenotypic assays may be the only means to capture important variation in the population.

For example, several studies have assessed the role of DNA repair capacity (DRC) regarding cancer risk by using *in vitro* phenotypic assays mostly on circulating lymphocytes (e.g. mutagen sensitivity, host cell reactivation assay). These studies have shown differences in DRC between cases and controls; however, interpretation of these results must account for study design limitations, such as use of lymphocytes to infer DRC in target tissues, the possible impact of disease status on assay results, and confounding by unmeasured risk factors that influence the assay (85–87). The application of functional assays in multiple, large-scale epidemiological studies will require development of less costly and labour-intensive assays. In the future, assays that assess non-clonal mutations in DNA, through the analysis of DNA isolated from circulating white blood cells, may capture some of the same information as the above functional assays and have wider application because of greater logistic ease.

Considerations in biospecimen collection, processing and storage

A case–control study in a relatively small geographic region, or a defined set of hospitals, can permit efficient collection of medical records or specimens (e.g. blood, urine, surgical tissue and other pathologic material) along with supporting documentation. Hospital-based case–control studies or population-based studies served by a small number of hospitals can have direct contact with patients in a hospital setting, thus offering advantages for the collection of different types of specimens or elaborate processing protocols (e.g. cryopreserving lymphocytes and Epstein-Barr Virus transformation to ensure large quantities of DNA), since resources for collection, processing and storage are often available in diagnostic hospitals. This offers the potential for conducting functional assays that require live cells, such as mutagen sensitivity (85), which in general are not methodologically feasible in cohort studies. Pre-treatment specimens, critical for evaluation of biologic markers that could be affected by treatment, such as chemotherapy, can be obtained through rapid identification systems that recruit cases right at the time of diagnosis.

Biomarker measurements can be very sensitive to differences in handling of samples (e.g. fasting status at blood collection or time between collection and processing of specimens). Therefore, it is important that samples from cases and controls be collected during the same timeframe and use identical protocols to avoid differential biases. Ideally, the nursing and laboratory staff should be blinded with respect to the case–control status of the subjects. However, because the differences in handling samples between cases and controls are not always avoidable, it is important to record key information such as date and time of collection, processing and storage problems, time since last meal, current medication, and current tobacco and alcohol use to be able to account for the influence of these variables at the data analysis stage. This information can also be used to match cases and controls selected for specific biomarker measurements in a subset of the study population. This will ensure efficient adjustment for these extraneous factors during data analysis.

Biomarkers measured in samples collected from subjects during a hospital stay might not reflect measurements from samples collected outside the hospital, as habits and exposures change during hospitalization (e.g. dietary habits, medication used and physical activity). Therefore, even if cases and controls are selected through a hospital-based design, collection of specimens after the patients return home and are no longer taking medications for the conditions that brought them to the hospital should be considered, if feasible. On the other hand, specimens to measure biomarkers that are influenced by long-term effects of treatment should be collected before treatment is started at the hospital, within logistic limitations.

Case–control studies might also allow more detailed characterization of disease through the use of biomarkers, such as the presence of eosinophils in sputum to identify eosinophilic and non-eosinophilic asthma, typing of viruses in infectious diseases, or molecular characterization of tumours in cancer. This more detailed classification of disease permits the analysis of genetic and environmental risk factors and clinical outcomes by biologically important disease subtypes. These analyses can lead to improvements in risk assessment by identifying diseases with distinct risk profiles. In addition, identifying subclasses of disease of different etiology can aid in understanding the pathogenic pathways to disease, as well as developing targeted prevention programmes (e.g. use of hormonal chemoprevention for women at high risk of estrogen-receptor positive breast tumours). Review of medical records can be used to obtain information on disease characteristics determined for clinical practice, such as histological tumour type and tumour grade in cancer patients. However, more

detailed characterization of disease might require large collections of biological specimens to determine disease biomarkers, which is facilitated in hospital-based studies.

Follow-up of cases to determine clinical outcomes

The prospective collection of clinical information from cases enrolled in case–control studies (e.g. treatment, recurrence of disease, and survival) greatly increases the value of these studies, since critical questions on the relationship between biomarkers and disease progression can be addressed in well characterized populations (see Chapter 4). Designing a survival study within a case–control study is easier to do at the beginning of the case–control study rather than later after subject enrolment is completed. Given the value that such studies have for carrying out translational research in a very efficient manner, consideration should be given to implementing this type of study whenever possible. The collection of clinical information is facilitated in hospital-based studies when cases are diagnosed in a relatively small number of hospitals, and in stable populations where patients are likely to be followed-up in the diagnostic hospitals or associated clinics.

Information on clinical outcomes can be obtained through active follow-up of the cases, in which patients are contacted individually through the course of their treatment and medical follow-up, or through passive follow-up by extracting information from medical records. Passive follow-up is less costly; however, it is often limited by difficulties in obtaining detailed information on treatment from medical records, or by loss to follow-up in populations where patients change cities or hospitals. Use of database resources, such as death registries in populations where cases are diagnosed, can be helpful in determining survival from cases lost to follow-up.

The case–control method in relation to other epidemiological designs

Existing cohort studies and their consortial groups that have or are collecting blood samples will accrue large numbers of cases with common diseases over the coming years. Appropriately, questions are being raised about the utility of carrying out new case–control studies, either population- or hospital-based, to study the main effects of common polymorphisms and their interaction with environmental exposures. Designers of a new case–control study will need to show that it offers benefits that cannot be obtained from existing cohorts. Below are some considerations when planning to carry out a new case–control study in contrast with performing nested studies within existing cohorts:

1) Disease incidence. A key advantage of case–control studies is the ability to enrol large numbers of cases with less common diseases in a relatively short period of time. Given the need for large sample sizes (up to several thousand cases and controls) to investigate weak to moderate associations, such as main effect for common susceptibility loci, as well as gene–environment interaction (88), and to explore data subsets, it is only feasible to collect enough cases of the more common diseases in most cohort studies. However, pooling efforts across cohorts or case–control studies, such as consortia of studies of specific tumour sites (http://epi.grants.cancer.gov/Consortia/tablelist.html), are critical to attain very large sample sizes.

2) Inclusion of diverse population groups. Case–control studies can focus on enrolling a narrow range of ethnic, racial, age or socioeconomic levels that are particularly interesting or important but not adequately represented in existing cohort studies.

3) Specialized specimen collection and processing protocols. Case–control studies can use labour- and technology-intensive biological collection, processing and storage protocols that would not be logistically feasible or cost-efficient in a large prospective cohort study.

4) Depth of exposure data. Case–control studies can collect more detailed and broader information about exposure from both interviews and records than is feasible in a cohort study. This is particularly important when there is concern about a specific type of exposure that is not generally assessed at all or in adequate detail in the typical cohort questionnaire (which usually focuses on diet and general lifestyle factors). Examples could include occupational and environmental exposures requiring complete occupational and residential histories, respectively. Cohorts have an inherent limitation in that their aim is to study multiple endpoints, and thus they collect less extensive data on exposures relevant to any one particular disease, although the opportunity to return to participants at later time points may partially ameliorate this point. Case–control studies can more readily focus on new exposures of concern for particular diseases, tailoring methods to optimally capture target data. In contrast, cohort studies will have instruments in place that will inevitably lack precision or entirely miss new exposures.

Summary and future directions

Case–control studies play a critical role in molecular epidemiologic research, particularly for biomarkers that are unlikely to have disease bias, such as DNA-based markers of genetic susceptibility. They can rapidly enrol large numbers of cases, even with rare conditions, in multicentre studies and by combining across studies in consortia. In addition, case–control studies can apply detailed diagnostic procedures, including specialized imaging approaches not routinely used in the usual healthcare setting, and state-of-the-art molecular analyses when tissue samples are collected. Given that a substantial number of rapidly developing new "omic" technologies can be readily applied to the case–control setting, this design should continue to be a core component of research programmes on the etiology of chronic diseases.

Case-only and other study designs

Case-only studies

Studies including subjects with the disease of interest without a control population (for instance, case series or clinical trials), are often used to evaluate questions related to disease treatment and progression, including secondary effects of treatment. These designs are also well suited to evaluate the influence of genetic and environmental risk factors on disease for disease progression and response to treatment, and can be very valuable to evaluate etiological questions, such as gene–gene and gene–environment interactions (89,90), and etiologic heterogeneity for different disease subtypes. An advantage of these designs is their ability to obtain extensive information on the disease to allow a more accurate definition of disease and refined classification of complex diseases, such as cancer, diabetes or hypertension, into entities more biologically or etiologically homogeneous among groups. By having direct access to patients, biological specimens, and clinical records, case series studies may be able to define diseases or preclinical conditions based on molecular events driving biological processes rather than clinical symptoms. For instance, cancers can be classified according to pathological and molecular characteristics of tumours, infectious diseases such as hepatitis can be classified according to the causal virus, and asthma can be more precisely defined according to pathophysiologic mechanisms (91).

The case-only design, however, has limitations when evaluating etiological questions, most notably related to the inability to directly estimate risk for disease. Although the case-only design can be used to estimate multiplicative interactions between risk factors under certain assumptions, it is susceptible to misinterpretation of the interaction parameter (92), is highly dependent on the assumption of independence between the exposure and the genotype under study (93), and it cannot be used to estimate additive interactions. The degree of etiologic heterogeneity in case-series studies can be quantified by the ratio of the relative risk for the effect of exposure on one disease subtype to the relative risk for another subtype. This parameter is equivalent to the relative risk for the association between exposure and disease subtype (94). However, case-only studies are limited to the estimation of the ratio of relative risk, and cannot be used to obtain estimates of the relative risk for different disease types. It should be noted that the relative risk from a case-only design would underestimate the relative risk derived in a case–control design when the exposure of interest is associated with more than one disease type.

Another potential limitation of the case-series design is the generalizability of findings, since this design can include highly selected cohorts of patients to address specific treatment protocols, such as in clinical trials. In etiological studies, it is always reassuring to observe associations between established factors and disease risk in a particular study population; however this cannot be observed in case series. Identification of cases through well characterized population-based registries, or evaluation of established associations between disease characteristics and clinical outcomes or risk factors, could address some of these limitations.

Other designs

Alternative study designs have been proposed to address some of the limitations of the classical epidemiological designs. For instance, the two-phase sampling design can be used to improve efficiency and reduce the cost of measuring biomarkers in large epidemiological studies (95). The first phase of this design could be a case–control or cohort study with basic exposure information and no biomarker measurements. In a second phase, more elaborate exposure information and/or determination of biomarkers (with collection of biological specimens if these were not collected in the first phase) is carried out in an informative sample of individuals defined by disease and exposure (e.g. subjects with extreme or

uncommon exposures). Multiple statistical methods, such as simple conditional likelihood (96) or estimated-score (97), have been developed to analyse data from two-sampling designs. Another example is the use of the kin-cohort design as a more efficient alternative to case–control or cohort studies, when the goal is to estimate age-specific penetrance for rare inherited mutations in the general population (98,99). In this design, relatives of selected individuals with genetic testing form a retrospective cohort that is followed from birth to onset of disease or censoring.

Concluding remarks

The field of molecular epidemiology has undergone a transformational change with the incorporation of powerful genomic technology. Further, important advances are being made in the development of new approaches in exposure assessment (http://www.gei.nih.gov/exposurebiology). At the same time, large and high-quality case–control studies of many diseases have been established with detailed exposure data and stored biological specimens, previously established cohorts are being followed-up, and new cohort studies with biological samples are being established in developing as well as developed countries. The confluence of extraordinary technology and the availability of large epidemiologic studies should ultimately lead to new insights into the etiology of many important diseases and help to facilitate effective prevention, screening and treatment. However, this will only be achieved if molecular epidemiologists adhere to the fundamental epidemiologic principles of careful study design, vigilant quality control, thoughtful data analysis, cautious interpretation of results, and well powered replication of important findings.

Acknowledgements

This chapter has been adapted and updated from a book chapter on *Application of biomarkers in cancer epidemiology* (16) and a chapter on *Design considerations in molecular epidemiology* (21). We thank Elizabeth Azzato, David Hunter, Maria Teresa Landi, and Sholom Wacholder for their valuable comments to the chapter.

References

1. Aardema MJ, MacGregor JT (2002). Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. *Mutat Res,* 499:13–25. PMID:11804602

2. Wang W, Zhou H, Lin H *et al.* (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem,* 75:4818–4826.doi:10.1021/ac026468x PMID:14674459

3. Hanash S (2003). Disease proteomics. *Nature*, 422:226–232.doi:10.1038/nature01514 PMID:12634796

4. Baak JP, Path FR, Hermsen MA *et al.* (2003). Genomics and proteomics in cancer. *Eur J Cancer,* 39:1199–1215.doi:10.1016/S0959-8049(03)00265-X PMID:12763207

5. Sellers TA, Yates JR (2003). Review of proteomics with applications to genetic epidemiology. *Genet Epidemiol,* 24:83–98. doi:10.1002/gepi.10226 PMID:12548670

6. Staudt LM (2003). Molecular diagnosis of the hematologic cancers. *N Engl J Med,* 348:1777–1785.doi:10.1056/NEJMra020067 PMID:12724484

7. Strausberg RL, Simpson AJ, Wooster R (2003). Sequence-based cancer genomics: progress, lessons and opportunities. *Nat Rev Genet,* 4:409–418.doi:10.1038/nrg1085 PMID:12776211

8. Smith MT, Rappaport SM (2009). Building exposure biology centers to put the E into "G x E" interaction studies. *Environ Health Perspect,* 117:A334–A335. PMID:19672377

9. Schembri F, Sridhar S, Perdomo C *et al.* (2009). MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proc Natl Acad Sci USA*, 106:2319–2324.doi:10.1073/pnas.0806383106 PMID:19168627

10. Committee on Biological Markers of the National Research Council (1987). Biological markers in environmental health research. *Environ Health Perspect,* 74:3–9. PMID:3691432

11. Rothman N, Wacholder S, Caporaso NE *et al.* (2001). The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta,* 1471:C1–C10. PMID:11342183

12. Schulte PA (1987). Methodologic issues in the use of biologic markers in epidemiologic research. *Am J Epidemiol,* 126:1006–1016. PMID:3318408

13. Perera FP (1987). Molecular cancer epidemiology: a new tool in cancer prevention. *J Natl Cancer Inst,* 78:887–898. PMID:3471998

14. Perera FP (2000). Molecular epidemiology: on the path to prevention? *J Natl Cancer Inst,* 92:602–612.doi:10.1093/jnci/92.8.602 PMID:10772677

15. Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publication; 1997.

16. García-Closas M, Vermeulen R, Sherman ME *et al.* Application of biomarkers in cancer epidemiology. In: Schottenfeld D, Fraumeni JF Jr, editors. Cancer epidemiology and prevention. 3rd ed. New York (NY): Oxford University Press; 2006. p. 70–88.

17. Perera FP, Weinstein IB (1982). Molecular epidemiology and carcinogen-DNA adduct detection: new approaches to studies of human cancer causation. *J Chronic Dis,* 35:581–600.doi:10.1016/0021-9681(82)90078-9 PMID:6282919

18. Spitz MR, Wu X, Mills G (2005). Integrative epidemiology: from risk assessment to outcome prediction. *J Clin Oncol,* 23:267–275.doi:10.1200/JCO.2005.05.122 PMID:15637390

19. Caporaso NE (2007). Integrative study designs–next step in the evolution of molecular epidemiology? *Cancer Epidemiol Biomarkers Prev,* 16:365–366.doi:10.1158/1055-9965.EPI-07-0142 PMID:17372231

20. Rothman N, Stewart WF, Schulte PA (1995). Incorporating biomarkers into cancer epidemiology: a matrix of biomarker and study design categories. *Cancer Epidemiol Biomarkers Prev,* 4:301–311. PMID:7655323

21. García-Closas M, Lan Q, Rothman N. Design considerations in molecular epidemiology. In: Rebbeck T, Ambrosone C, Shields P, editors. Molecular epidemiology: applications in cancer and other human diseases. New York (NY): Informa Healthcare; 2008. p. 1–18.

22. Ransohoff DF (2009). Promises and limitations of biomarkers. *Recent Results Cancer Res,* 181:55–59.doi:10.1007/978-3-540-69297-3_6 PMID:19213557

23. Pepe MS, Feng Z, Janes H *et al.* (2008). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst,* 100:1432–1438.doi:10.1093/jnci/djn326 PMID:18840817

24. Ransohoff DF (2007). How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol,* 60:1205–1219.doi:10.1016/j.jclinepi.2007.04.020 PMID:17998073

25. Schulte PA, Perera FP. Transitional studies. In: Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publications; 1997. p. 19–29.

26. Rothman N (1995). Genetic susceptibility biomarkers in studies of occupational and environmental cancer: methodologic issues. *Toxicol Lett,* 77:221–225.doi:10.1016/0378-4274(95)03298-3 PMID:7618141

27. Hulka BS, Margolin BH (1992). Methodological issues in epidemiologic studies using biologic markers. *Am J Epidemiol,* 135:200–209. PMID:1536135

28. Hulka BS (1991). ASPO Distinguished Achievement Award Lecture. Epidemiological studies using biological markers: issues for epidemiologists. *Cancer Epidemiol Biomarkers Prev,* 1:13–19. PMID:1845163

29. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992). Selection of controls in case-control studies. I. Principles. *Am J Epidemiol,* 135:1019–1028. PMID:1595688

30. Breslow NE, Day NE. Design considerations. In: Breslow NE, Day NE, editors. Statistical methods in cancer research. Vol 2. The design and analysis of cohort studies. Lyon: IARC Scientific Publication; 1987. p. 272–315.

31. Rothman KJ, Greenland S, editors. Modern epidemiology. Philadelphia (PA): Lippincott-Raven; 1998.

32. Paulsson B, Larsen KO, Törnqvist M (2006). Hemoglobin adducts in the assessment of potential occupational exposure to acrylamides – three case studies. *Scand J Work Environ Health,* 32:154–159. PMID:16680386

33. Wirfält E, Paulsson B, Törnqvist M *et al.* (2008). Associations between estimated acrylamide intakes, and hemoglobin AA adducts in a sample from the Malmö Diet and Cancer cohort. *Eur J Clin Nutr,* 62:314–323. doi:10.1038/sj.ejcn.1602704 PMID:17356560

34. Qu Q, Shore R, Li G *et al.* (2002). Hematological changes among Chinese workers with a broad range of benzene exposures. *Am J Ind Med,* 42:275–285. doi:10.1002/ajim.10121 PMID:12271475

35. Lan Q, Zhang L, Li G *et al.* (2004). Hematotoxicity in workers exposed to low levels of benzene. *Science*, 306:1774–1776. doi:10.1126/science.1102443 PMID:15576619

36. Schulte PA, Geraci C, Zumwalde R *et al.* (2008). Occupational risk management of engineered nanoparticles. *J Occup Environ Hyg,* 5:239–249.doi:10.1080/15459620801907840 PMID:18260001

37. Chua SD Jr, Messier SP, Legault C *et al.* (2008). Effect of an exercise and dietary intervention on serum biomarkers in overweight and obese adults with osteoarthritis of the knee. *Osteoarthritis Cartilage,* 16:1047–1053.doi:10.1016/j.joca.2008.02.002 PMID:18359648

38. Rejnmark L, Vestergaard P, Heickendorff L *et al.* (2001). Loop diuretics alter the diurnal rhythm of endogenous parathyroid hormone secretion. A randomized-controlled study on the effects of loop- and thiazide-diuretics on the diurnal rhythms of calcitropic hormones and biochemical bone markers in postmenopausal women. *Eur J Clin Invest,* 31:764–772.doi:10.1046/j.1365-2362.2001.00883.x PMID:11589718

39. Schulte PA, Rothman N, Schottenfeld D. Design considerations in molecular epidemiology. In: Schulte PA, Perera FP. Molecular epidemiology: principles and practices. San Diego (CA): Academic Press, Inc.;1993. p. 159–98.

40. Schatzkin A, Freedman LS, Schiffman MH, Dawsey SM (1990). Validation of intermediate end points in cancer research. *J Natl Cancer Inst,* 82:1746–1752.doi:10.1093/jnci/82.22.1746 PMID:2231769

41. Schatzkin A, Gail M (2002). The promise and peril of surrogate end points in cancer research. *Nat Rev Cancer,* 2:19–27.doi:10.1038/nrc702 PMID:11902582

42. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001). Executive summary of the third report of The National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *JAMA*, 285:2486–2497.doi:10.1001/jama.285.19.2486 PMID:11368702

43. Tucker JD, Eastmond DA, Littlefield LG. Cytogenetic end-points as biological dosimeters and predictors of risk in epidemiological studies. In: Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publication; 1997. p. 185–200.

44. Zhang L, Eastmond DA, Smith MT (2002). The nature of chromosomal aberrations detected in humans exposed to benzene. *Crit Rev Toxicol,* 32:1–42.doi:10.1080/20024091064165 PMID:11846214

45. Zhang L, Rothman N, Wang Y *et al.* (1999). Benzene increases aneuploidy in the lymphocytes of exposed workers: a comparison of data obtained by fluorescence in situ hybridization in interphase and metaphase cells. *Environ Mol Mutagen,* 34:260–268.doi:10.1002/(SICI)1098-2280(1999)34:4<260::AID-EM6>3.0.CO;2-P PMID:10618174

46. Boffetta P, van der Hel O, Norppa H *et al.* (2007). Chromosomal aberrations and cancer risk: results of a cohort study from Central Europe. *Am J Epidemiol,* 165:36–43.doi:10.1093/aje/kwj367 PMID:17071846

47. Bonassi S, Norppa H, Ceppi M *et al.* (2008). Chromosomal aberration frequency in lymphocytes predicts the risk of cancer: results from a pooled cohort study of 22 358 subjects in 11 countries. *Carcinogenesis*, 29:1178–1183.doi:10.1093/carcin/bgn075 PMID:18356148

48. Smerhovsky Z, Landa K, Rössner P *et al.* (2001). Risk of cancer in an occupationally exposed cohort with increased level of chromosomal aberrations. *Environ Health Perspect,* 109:41–45.doi:10.1289/ehp.0110941 PMID:11171523

49. Liou SH, Lung JC, Chen YH *et al.* (1999). Increased chromosome-type chromosome aberration frequencies as biomarkers of cancer risk in a blackfoot endemic area. *Cancer Res,* 59:1481–1484. PMID:10197617

50. Bonassi S, Abbondandolo A, Camurri L *et al.* (1995). Are chromosome aberrations in circulating lymphocytes predictive of future cancer onset in humans? Preliminary results of an Italian cohort study. *Cancer Genet Cytogenet,* 79:133–135.doi:10.1016/0165-4608(94)00131-T PMID:7889505

51. Hagmar L, Brøgger A, Hansteen IL *et al.* (1994). Cancer risk in humans predicted by increased levels of chromosomal aberrations in lymphocytes: Nordic study group on the health risk of chromosome damage. *Cancer Res,* 54:2919–2922. PMID:8187078

52. Bonassi S, Hagmar L, Strömberg U *et al.*; European Study Group on Cytogenetic Biomarkers and Health (2000). Chromosomal aberrations in lymphocytes predict human cancer independently of exposure to carcinogens. *Cancer Res,* 60:1619–1625. PMID:10749131

53. Vermeulen R, Li G, Lan Q *et al.* (2004). Detailed exposure assessment for a molecular epidemiology study of benzene in two shoe factories in China. *Ann Occup Hyg,* 48:105–116.doi:10.1093/annhyg/meh005 PMID:14990432

54. García-Closas M, Rothman N, Lubin J (1999). Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev,* 8:1043–1050. PMID:10613335

55. Nicholson JK, Wilson ID (2003). Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat Rev Drug Discov,* 2:668–676.doi:10.1038/nrd1157 PMID:12904817

56. Merrick BA, Tomer KB (2003). Toxicoproteomics: a parallel approach to identifying biomarkers. *Environ Health Perspect,* 111:A578–A579.doi:10.1289/ehp.111-a578 PMID:12940285

57. Potter JD. Logistics and design issues in the use of biological specimens in observational epidemiology. In: Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publications; 1997. p. 31–37.

58. Rundle AG, Vineis P, Ahsan H (2005). Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiol Biomarkers Prev,* 14:1899–1907.doi:10.1158/1055-9965.EPI-04-0860 PMID:16103435

59. Prentice RL (1995). Design issues in cohort studies. *Stat Methods Med Res,* 4:273–292.doi:10.1177/096228029500400402 PMID:8745127

60. Wacholder S (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology,* 2:155–158.doi:10.1097/00001648-199103000-00013 PMID:1932316

61. Hunter DJ. Methodological issues in the use of biological markers in cancer epidemiology: cohort studies. In: Toniolo P, Boffetta P, Shuker DEG *et al.*, editors. Application of biomarkers in cancer epidemiology. Lyon: IARC Scientific Publications; 1997. p. 39–46.

62. Banks E, Meade T (2002). Study of genes and environmental factors in complex diseases. *Lancet*, 359:1156–1157, author reply 1157.doi:10.1016/S0140-6736(02)08140-0 PMID:11943294

63. Burton P, McCarthy M, Elliott P (2002). Study of genes and environmental factors in complex diseases. *Lancet*, 359:1155–1156, author reply 1157.doi:10.1016/S0140-6736(02)08138-2 PMID:11943293

64. Clayton D, McKeigue PM (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, 358:1356–1360.doi:10.1016/S0140-6736(01)06418-2 PMID:11684236

65. Wacholder S, García-Closas M, Rothman N (2002). Study of genes and environmental factors in complex diseases. *Lancet*, 359:1155, author reply 1157.doi:10.1016/S0140-6736(02)08137-0 PMID:11943292

66. García-Closas M, Thompson WD, Robins JM (1998). Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol,* 147:426–433. PMID:9525528

67. Rosner B, Spiegelman D, Willett WC (1992). Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Am J Epidemiol,* 136:1400–1413. PMID:1488967

68. Holland NT, Smith MT, Eskenazi B, Bastaki M (2003). Biological sample collection and processing for molecular epidemiological studies. *Mutat Res,* 543:217–234.doi:10.1016/S1383-5742(02)00090-X PMID:12787814

69. Tworoger SS, Hankinson SE (2006). Collection, processing, and storage of biological samples in epidemiologic studies: sex hormones, carotenoids, inflammatory markers, and proteomics as examples. *Cancer Epidemiol Biomarkers Prev,* 15:1578–1581.doi:10.1158/1055-9965.EPI-06-0629 PMID:16985015

70. Peakman TC, Elliott P (2008). The UK Biobank sample handling and storage validation studies. *Int J Epidemiol,* 37 Suppl 1;i2–i6.doi:10.1093/ije/dyn019 PMID:18381389

71. Jackson C, Best N, Elliott P (2008). UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. *Int J Epidemiol,* 37 Suppl 1;i16–i22. doi:10.1093/ije/dym280 PMID:18381388

72. Elliott P, Peakman TC; UK Biobank (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol,* 37:234–244.doi:10.1093/ije/dym276 PMID:18381398

73. Fu Q, Garnham CP, Elliott ST *et al.* (2005). A robust, streamlined, and reproducible method for proteomic analysis of serum by delipidation, albumin and IgG depletion, and two-dimensional gel electrophoresis. *Proteomics*, 5:2656–2664.doi:10.1002/pmic.200402048 PMID:15924293

74. Prentice RL (1986). On the design of synthetic case-control studies. *Biometrics*, 42:301–310.doi:10.2307/2531051 PMID:3741972

75. Franco EL (2000). Statistical issues in human papillomavirus testing and screening. *Clin Lab Med,* 20:345–367. PMID:10863644

76. Welch HG, Black WC (1997). Using autopsy series to estimate the disease "reservoir" for ductal carcinoma in situ of the breast: how much more breast cancer can we find? *Ann Intern Med,* 127:1023–1028. PMID:9412284

77. Morrison AS. Screening. In: Rothman KJ, Greenland S, editors. Modern epidemiology. 2nd ed. Philadelphia (PA): Lippincott-Raven Publishers; 1998. p. 499–518.

78. Morton LM, Cahill J, Hartge P (2006). Reporting participation in epidemiologic studies: a survey of practice. *Am J Epidemiol,* 163:197–203.doi:10.1093/aje/kwj036 PMID:16339049

79. Wacholder S, Chatterjee N, Hartge P (2002). Joint effect of genes and environment distorted by selection biases: implications for hospital-based case-control studies. *Cancer Epidemiol Biomarkers Prev,* 11:885–889. PMID:12223433

80. Lum A, Le Marchand L (1998). A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol Biomarkers Prev,* 7:719–724. PMID:9718225

81. Wacholder S, Benichou J, Heineman EF *et al.* (1994). Attributable risk: advantages of a broad definition of exposure. *Am J Epidemiol,* 140:303–309. PMID:8059765

82. Cox A, Dunning AM, García-Closas M *et al.*; Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer; Breast Cancer Association Consortium (2007). A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet,* 39:352–358.doi:10.1038/ng1981 PMID:17293864

83. Easton DF, Pooley KA, Dunning AM *et al.*; SEARCH collaborators; kConFab; AOCS Management Group (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447:1087–1093.doi:10.1038/nature05887 PMID:17529967

84. Ahsan H, Rundle AG (2003). Measures of genotype versus gene products: promise and pitfalls in cancer prevention. *Carcinogenesis*, 24:1429–1434.doi:10.1093/carcin/bgg104 PMID:12819189

85. Wu X, Gu J, Spitz MR (2007). Mutagen sensitivity: a genetic predisposition factor for cancer. *Cancer Res,* 67:3493–3495. doi:10.1158/0008-5472.CAN-06-4137 PMID:17440053

86. Berwick M, Vineis P (2000). Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *J Natl Cancer Inst,* 92:874–897.doi:10.1093/jnci/92.11.874 PMID:10841823

87. Spitz MR, Wei Q, Dong Q *et al.* (2003). Genetic susceptibility to lung cancer: the role of DNA damage and repair. *Cancer Epidemiol Biomarkers Prev,* 12:689–698. PMID:12917198

88. García-Closas M, Lubin JH (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol,* 149:689–692. PMID:10206617

89. Yang Q, Khoury MJ, Sun F, Flanders WD (1999). Case-only design to measure gene-gene interaction. *Epidemiology*, 10:167–170. doi:10.1097/00001648-199903000-00014 PMID:10069253

90. Khoury MJ, Flanders WD (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls. *Am J Epidemiol,* 144:207–213. PMID:8686689

91. Bel EH (2004). Clinical phenotypes of asthma. *Curr Opin Pulm Med,* 10:44–50. doi:10.1097/00063198-200401000-00008 PMID:14749605

92. Schmidt S, Schaid DJ (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am J Epidemiol,* 150:878–885. PMID:10522659

93. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001). Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol*, 154:687–693. doi:10.1093/aje/154.8.687 PMID:11590080

94. Begg CB, Zhang ZF (1994). Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev*, 3:173–175. PMID:8049640

95. White JE (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol*, 115:119–128. PMID:7055123

96. Cain KC, Breslow NE (1988). Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol*, 128:1198–1206. PMID:3195561

97. Chatterjee N, Chen Y, Breslow N (2003). A pseudoscore estimator for regression problems for two phase sampling. *J Am Stat Assoc,* 98:158–168 doi:10.1198/016214503388619184.

98. Wacholder S, Hartge P, Struewing JP *et al.* (1998). The kin-cohort study for estimating penetrance. *Am J Epidemiol,* 148:623–630. doi:10.1093/aje/148.7.623 PMID:9778168

99. Chatterjee N, Shih J, Hartge P *et al.* (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype and family history data from the Washington Ashkenazi Study. *Genet Epidemiol,* 21:123–138.doi:10.1002/gepi.1022 PMID:11507721