예방의학회지 제42권 제6호(2009년 11월)
J Prev Med Public Health 2009;42(6):343-348
DOI: 10.3961/jpmph.2009.42.6.343

Special Issues

Issues in the Design of Molecular and Genetic Epidemiologic Studies

Jay H. Fowke

Division of Epidemiology, Vanderbilt University Medical Center, USA

The final decision of study design in molecular and genetic epidemiology is usually a compromise between the research study aims and a number of logistical and ethical barriers that may limit the feasibility of the study or the interpretation of results. Although biomarker measurements may improve exposure or disease assessments, it is necessary to address the possibility that biomarker measurement inserts additional sources of misclassification and confounding that may lead to inconsistencies across the research literature. Studies targeting multi-causal diseases and investigating gene-environment interactions must not only meet the needs of a traditional epidemiologic study but also the needs of the biomarker investigation. This paper is intended to highlight the major issues that need to be considered when developing an epidemiologic study utilizing biomarkers. These issues covers from molecular and genetic epidemiology (MGE) study designs including cross-sectional, cohort, case-control, clinical trials, nested case-control, and case-only studies to matching the study design to the MGE research goals. This review summarizes logistical barriers and the most common epidemiological study designs most relevant to MGE and describes the strengths and limitations of each approach in the context of common MGE research aims to meet specific MEG objectives.

J Prev Med Public Health 2009;42(6):343-348

Key words : Biomarker, Epidemiology, Study design

INTRODUCTION

At the heart of molecular and genetic epidemiology (MGE) is the idea that the measurement of a biomarker representing a hormonal, genetic, or other trait, or an external exposure, will lead to a better understanding of the relationship between an exposure and a disease of interest. Confounding, bias, and measurement error within a traditional epidemiologic framework limit our ability to identify risk factors for complex diseases such as cancer or diabetes. Use of biomarkers may improve the measurement of exposure and characterization of disease such that exposure - disease associations may be more easily observed [1,2]. Thus, MGE can be considered as extension of a broader epidemiologic paradigm that attempts to address the limitations of traditional epidemiologic methods by incorporating biological information.

Indeed, biological markers and genetic data have been utilized in epidemiologic research for some time [3]. What has changed recently, however, involves the rapid advances and reduced costs in biochemistry, genetics, and high-throughput technologies. Without the financial and logistical barriers that once made a MGE study exceptional, molecular or genetic marker analyses are now conducted within epidemiologic studies regardless of the original goals of the study. One of the central tenets of epidemiologic research is that results of one study should be reproduced or consistent with the results from another study. Although biomarker measurements may improve exposure or disease assessments, it is necessary to address the possibility that biomarker measurement inserts additional sources of misclassification and confounding that may lead to inconsistencies across the research literature. Studies targeting multi-causal diseases and investigating gene-environment interactions must not only meet the needs of a traditional epidemiologic study but also the needs of the biomarker investigation.

This paper is intended to highlight the major issues that need to be considered when developing an epidemiologic study utilizing biomarkers. After summarizing logistical barriers and the most common study designed employed in MGE, the application of these designs to meet specific MEG objectives is discussed.

ISSUES OF CONSENT AND FEASIBILITY FOR MGE STUDY DESIGNS

While issues of bias, measurement, and inference should dominate study design decisions, the final design is almost always adjusted in a manner to also maximize the feasibility of recruitment and biological specimen collection. Candidates for recruitment may be concerned about

Received : 2 October 2009, Accepted : 2 November 2009
Corresponding author : Jay H. Fowke (2525 West End Ave, 6th floor, Vanderbilt University Medical Center, Nashville, TN 37203-1783, USA , Tel : 615-936-2903, E-mail : jay.fowke@vanderbilt.edu)

confidentiality, particularly for genetic studies, and may chose not to consent unless their concerns are addressed. Once consent is obtained, blood or other biospecimens must be collected by trained staff and maintained at a temperature suitable for analytic preservation. If the protocol requires tissue procurement beyond standard blood collection, recruitment may be limited to candidates with disease or at high-risk for disease. Finally, biological samples must be stored for extended periods of time to support cohort or nested case-control studies.

MGE STUDY DESIGNS

The study designs most relevant to MGE research fall under the general categories of cross-sectional, cohort, and case-control. Additionally, there are several design variations from these general categories including clinical trials, nested case-control studies, and case-only studies (Table 1). Family linkage studies or case-sibling studies are used primarily for the study of highly penetrate genetic factors and will not be addressed.

I. Cross-Sectional Studies

A cross-sectional study is designed to collect data on a risk factor as well as clinical or disease status at the same point in time. This design has limited value in etiologic research because the temporal relationship between an exposure and outcome cannot be determined with confidence. Reverse causation may occur when a diagnosis potentially affects the exposure and the cause-effect relationship is unclear, as for example if cardiovascular disease (CVD) was associated with lower blood lipid levels because a CVD diagnosis led patients to change their diet. However, Hulka describes conducting early cross-sectional studies as critical step in MGE toward transition to an etiologic investigation using biological markers [4]. Subject selection and sample collection protocols may be optimized to characterize the biomarker in a target population with the disease or in a population at risk for disease. Variability in the biomarker between demographic or other factors may also be explored. The concept of a transitional study may also be extended beyond a simple cross-sectional design to collect multiple biological samples per participant in order to estimate variability in biomarker values within individuals.

II. Cohort Studies (Prospective and Retrospective)

The defining aspect of a cohort study is that exposure measurement and biological sample collection occurs prior to disease onset. Two or more exposure groups are defined from the study population using exposure estimates measured at baseline. Both these groups, the exposed and unexposed, are then followed over time until a suitable number of disease outcomes are identified. The statistical analyses characterize any differences in time to disease onset between the exposed and unexposed. Unlike a cross-sectional study, any observed exposure-disease association is unlikely to be the result of reverse causation. Excluding those cases identified soon after baseline (usually within the first year of follow-up) and thus more likely to have latent disease at baseline further strengthens the temporal relationship between exposure and disease.

Although retrospective cohort analyses using administrative or medical chart data is a common approach in occupational and pharmaco-epidemiology, few retrospective cohorts are adequately linked with a biospecimen repository suitable to support MGE. Most MGE cohort analyses are prospective, with well known examples including the Nurses' Health Study, the Health Professional's Follow-up Study, and the Shanghai Women's Health Study. These cohort studies tend to be quite large, often including many thousands of participants, and yet cohort initial analyses are often limited to the most common disease outcomes simply because of the time require to accumulate a sufficient number of disease outcomes for analysis. It also follows that the majority of effort to collect exposure information at baseline is targeted to the most common outcomes projected for that cohort, and detail data to characterize a particular exposure of interest, perhaps for a less common outcome, may be quite limited.

Under certain circumstances, large randomized clinical trials may be used like an observational cohort study for secondary prospective biomarker investigations. If the sample size is sufficient, a clinical trial provides a well characterized study population, often with extensive data collection for a target disease of interest and with active follow-up. Such analyses must carefully consider the potential effects of the intervention agent on the biomarker of interest, and it is not unusual for analyses to limit the investigation to the control arm of the trial. For example, Schenk et al. [5] analyzed data from the control arm of the Prostate Cancer Prevention Trial, a randomized trial of finasteride and prostate cancer prevention, and found blood adiponectin levels were inversely associated with incident benign prostatic hyperplasia. Extensive data collection at baseline and throughout follow-up permitted the investigators to exclude prevalent benign prostatic hyperplasia (BPH) cases and focus the analysis on those subjects with incident BPH.

III. Case-Control Studies

Case-control studies identify two groups; one with disease and one as a control without disease. Analyses determine if the prevalence of measured exposures are different between cases and controls. The case-control design may be the only feasible option with very rare outcomes, limited financial resources, or when extensive exposure assessments are required. Matching cases to controls by age or other parameters may improve analytic precision and

reduce required sample size.

Selecting the most appropriate control group can be challenging, but ideally controls are selected to represent the exposure pattern within the population that gave rise to the cases. Healthy community controls from the region where the cases reside may provide valid exposure estimates for comparison, but the recruitment of healthy community volunteers can be difficult and a low control recruitment rate may unknowingly lead to selective enrollment and induce bias in the analysis. Furthermore, biological sample collection becomes increasingly difficult as community volunteers are recruited from an increasingly large geographic region. Blood collection or collection of other perishable samples may not be feasible. An alternative approach recruits healthy controls identified from the hospital or institution used to identify cases. Hospital controls may be a legitimate alternative when cases are identified from only a few selected clinical centers or if the case series is believed to substantially differ from the surrounding community. Furthermore, clinical control recruitment usually supports a higher recruitment rate and biological sample collection is usually feasible. In practice, investigators may have laudable research aims, but the final MGE case-control protocol is usually a compromise between the research aims and the feasibility of recruiting and collecting biological samples.

IV. Other Study Designs

A nested case-control study is a powerful design that utilizes a case-control paradigm within a prospective cohort study. Cases from a prospective cohort are identified, then a random or matched sample of non-cases at the time the case was diagnosed is used as a control group. A case-cohort analysis is similar in that cases are identified from a cohort study, but controls include a random sample of the cohort. Nested case-control and case-cohort studies maintain the prospective nature of the investigation, as exposure assessment precedes disease, but these designs also increase efficiency and reduce cost because the biomarkers of interest are assayed only within a sample of the entire cohort study population.

Table 1. Study designs in molecular and genetic epidemiology

Design	Comment
Cross-sectional	Limited value in etiologic studies Highly valued in transitional studies to characterize biomarker in target population
Prospective cohort	Exposure and biological sample collection prior to disease Useful for common outcomes Expensive Limited exposure data
Randomized clinical trial	Well characterized study population Systematic follow-up and data collection Limited generalizability
Case-control	Feasible for rare diseases Potential for detailed exposure assessment Recall bias Reverse causality
Nested case-control and case-cohort	Maintains prospective design Efficient and less expensive
Case-only	Efficient for G x E interactions Requires assumptions that are difficult to test

A case-only design compares disease characteristics of cases with exposure to the disease characteristics of cases without exposure. This design is particularly well suited to the exploration of gene-exposure interactions and avoids the potential for selection bias that could develop if a control group substantively differs from the case series on one or more risk factors. However, the case-only design does not allow investigators to calculate a main effect of the exposure or the biomarker on the risk of disease, and therefore this design is less commonly used compared to a case-control design. Furthermore, the case-only design requires an assumption that the environmental exposure of interest is not associated with the biomarker or genetic variant.

MATCHING THE STUDY DESIGN TO THE MGE RESEARCH GOALS

I. Goals of Biomarkers in MGE

MGE studies typically have one or more of the following goals:

1. biomarker characterization
2. validation of another exposure assessment method
3. estimate internal dose of an exposure
4. identify genetic markers of disease susceptibility, including gene x exposure interactions associated with disease
5. investigate intermediate or transitional states between exposure and disease, including intermediate biomarkers used as outcomes within an intervention
6. identify markers of response to treatment
7. identify markers of disease sub-groups

There are no strict boundaries between these goals. For example, a biomarker may be considered an exposure marker or a disease marker depending on the intended use of the marker within the proposed study design. Similarly, estimates of internal dose and markers of treatment response share certain core ideas regarding the persistence and metabolism of an agent within the body. However, there is sufficient distinction across these goals such that study design issues are addressed for each.

II. Biomarker Characterization

Cross-sectional studies provide an efficient initial approach to characterize the prevalence of the marker within a target population. These studies may become quite complex and

increase in size if investigators wish to consider a range of age categories, race, gender, or other factors. Matching protocols may be needed to better compare biomarker values between specified groups while minimizing the effects of age or other known sources of systematic variation. It also may be feasible to explore genotype-phenotype associations and the functional activity of a genetic variant of interest using a cross-sectional design. Repeated biological sample collection at specified time intervals would allow investigators to partition variation in marker levels and determine variation in marker levels within individuals.

III. Validation Studies

Biomarkers are commonly used in studies designed to assess the reliability or validity of another measurement instrument. For example, food frequency questionnaires (FFQs) used to measure diet in large epidemiologic studies are susceptible to random and systematic reporting errors that may bias an analysis. Dietary biomarkers provide an alternative quantitative estimate of dietary intake, and error in diet biomarker values are assumed to be independent of questionnaire reporting biases derived from social desirability or other sources [6]. These studies may be relatively small (often 50-200 persons) with repeat biological sample collection over an extended time period to capture seasonal variation in diet and biomarker levels.

The correlation between a specific dietary biomarker and FFQ responses for that specific nutrient are often modest, often ranging from 0.1 to 0.5, and reflect the overall error in assessment and the separate errors between the questionnaire and biomarker methods of diet assessment [7]. In one of the most comprehensive of these studies, Subar and colleagues compared FFQ responses to levels of urinary doubly labeled water excretion as an unbiased biomarker of energy intake in a cross-section study of 484 men and women [8]. There were important differences in reporting between men and women, and with the type of FFQ utilized, that suggested substantial attenuation of relative risks investigating diet-disease associations in nutritional epidemiologic research

IV. Internalized Exposure and Biologically Effective Dose

A marker of exposure is any substance that reflects endogenous or exogenous exposure to the risk factor and that can be measured in a body tissue or fluid. This extends to the metabolites or adducts of endogenous or exogenous products. Indeed, many candidate risk factors, including many drugs and environmental carcinogens, are relatively inert and require metabolic activation after initial exposure in order to have an impact on disease. The internalized dose and persistence of a biologically activated agent within the body may be a weighted function of overall exposure, metabolism, and excretion. In this situation, it is possible that a biomarker provides a measure of exposure to an agent or risk factor of interest beyond what is possible through questionnaires or other assessment methods. Furthermore, exposure biomarkers could be a powerful tool in clinical trials to determine participant adherence to the intervention protocol or could be designed to complement otherwise limited exposure data that may be collected in a cohort study.

Reverse causation is a concern in case-control analyses unless it can be demonstrated that the exposure biomarker is stable despite disease and possible treatment. An exception to this concept may be the area of infectious disease epidemiology if the biomarker represents a highly acute exposure such as pretreatment viral infection or antibody titer, and there is believed to be minimal latency between exposure and disease onset.

Prospective cohort analyses better maintain the temporal relationship between the exposure biomarker and the outcome. Possible sources of bias are minimized as cases and non-cases derive from same baseline population and biological specimens are collected and handled in an identical manner. Nested case-control or case-cohort designs provide the opportunity to conduct a prospective analysis while minimizing financial cost, and matching between cases and controls with respect to age, time of follow-up, laboratory batch, or other factors may improve the analytic precision of the biomarker analyses. Whether a single biomarker measurement represents a persistent level within that study subject over an etiologically relevant time period needs to be addressed either with repeated biological sample collection within the cohort or a separate transitional study with repeated biomarker analysis.

V. Genetic Markers of Disease Susceptibility

A biological marker of susceptibility is an indicator of a person's likelihood of developing the disease or responding to a disease risk factor. This often involves genetic markers of inherited susceptibility that place one person at a different chance of developing disease compared to someone else. The development and analyses of such protocols lend themselves to ask which biological mechanisms and pathways are represented by these markers to advance disease progression.

Genetic epidemiologic studies identify disease susceptibility markers through a candidate gene approach or a genome-wide association study (GWAS). A candidate gene study uses prior knowledge of the disease to select a gene or panel of genes that may be hypothesized to be associated with disease. Genetic variants such as single nucleotide polymorphisms (SNPs) are selected within these gene regions to reflect some aspect of functional activity of the proteins hypothesized to be involved in the disease pathway. A GWAS approach, by contract, does not require any prior knowledge about how disease advances. Instead, the goal of GWAS is the discovery of new genetic regions that would not otherwise be considered based on existing

knowledge. Because multiple adjacent SNPs are strongly correlated (i.e., linkage disequilibrium), SNPs selected for GWAS include tagging SNPs that represent a larger panel of SNPs within a region of the DNA. Since SNPs in linkage disequilibrium may be transmitted as a unit across generations, a tagging SNP provides a marker of genetic variations within a specified DNA region, and a panel of tagging SNPs can be selected to capture genetic variation across the entire genome. With an appropriate study design, analysis of high-throughout scan for hundreds of thousands of these tagging SNPs (or more) per subject across a large number of study participants may identify a novel genome region associated with disease.

Candidate gene and GWA studies may employ either a cohort or a case-control design because it is assumed that germline genetic variants are stable within individuals. The major concern is usually regarding an inadequate number of cases, or a heterogenous case series, leading to an under-powered study. The combined distribution of the genetic variants with any environmental exposure of interest may further limit the investigation of gene x environment interactions. Similarly, associations between SNPs and disease in GWA studies tend to be fairly weak, with odds ratios usually less than 2.0, and GWA investigations require performing a very large number of statistical tests to select the most promising tagging SNPs and run the risk of a false positive finding. The sample size requirements of a study designed to identify a weak SNP-disease association while minimizing associations by chance alone has led to the formation of research consortia to pool data across prior research studies.

VI. Intermediate Markers of Disease Progression as an Outcome for a Trial

Any number of intermediate states between health and disease in cells or tissue may be identified, including high-grade prostatic intraepithelial neoplasia or colon adenomas as precursors to prostate or colon cancer, respectively. Any intermediary marker believed to be strongly associated with disease progression would be valuable as secondary end-points for clinical trials attempting to identify new disease prevention strategies [9,10], and a focused clinical trial with systematic follow-up may be the best choice to investigate the change in an intermediate marker associated with an exposure.

Observational studies may be used to identify specific exposures associated with the intermediate outcome and that might be developed as experimental agents for investigation in a clinical trial. However, investigation of risk factors for intermediate biomarker within an observational study design can be difficult. Often these preliminary markers of disease are asymptomatic and not easily identified without a laboratory or clinical test. Prospective cohort studies relying on participant reports or linkage with disease registries may miss a substantial number of outcomes. Participants may not know if they have a pre-clinical condition and disease registries do not necessarily record pre-clinical conditions. Alternatively, there may be opportunities within a retrospective cohort design to track the temporal relationships between an exposure, the intermediary biomarker, and clinical disease over time using medical chart or insurance data across a large sample of patients. Such studies are informative but may be limited to the investigation of drug effects or other clinical procedures that would be recorded in charts. Patient adherence to the drug of interest is rarely certain, and identification of the intermediate outcome relies on the quality of the administrative data and range of clinical decision-making processes that may lead to variation in diagnostic testing protocols necessary to detect the intermediate outcome.

Case-control investigation may be a feasible alternative to explore risk factors for intermediate disease states. Cases identification may require systematic medical record review for patients who tested positive after undergoing a diagnostic or screening test. Control selection becomes difficult if healthy persons that have also undergone the screening test of interest cannot be readily identified or recruited, or if those testing negative are subsequently diagnosed with an alternative disease or condition that would confound the interpretation between risk factor profiles and the risk of having the intermediate biomarker. This may require investigators to actively screen candidate controls from the community for latent conditions and to collect additional data measuring factors related to the utilization of the screening protocol that may differ between the cases and controls.

VII. Markers of Response to Treatment

Pharmacogenomics is a rapidly growing field and the foundation of individualized medicine and customized drug administration based on each patient's genetic profile. A pharmacoepidemiologic study is no different from other epidemiologic investigations except that the exposure of interest is usually a drug used to treat a clinical condition. Blood drug levels (an exposure biomarker) may be routinely monitored and subsequently adjusted, and indeed the final drug exposure is almost always associated with the underlying disease or drug side-effects. Separating the association between the drug and treatment outcome from the clinical indicators, and the many related factors that may lead to clinical intervention, is challenging. Heterogeneity across patients receiving the drug with regard to demographic, behavioral, or genetic traits also limit analyses of drug effects. Furthermore, it may be difficult to identify a patient control series that has not received the drug.

Genotyping within observational studies may allow investigators to begin to separate the effects of the drug from other indicator for drug. The proteins involved in drug absorption and metabolism are usually well-understood.

Genetic variation in these pathways leading to substantive differences in the bioavailability of the drug would separate patients receiving the drug into those receiving a high vs. low biologically available dose of the drug. Furthermore, genetic variation is more or less random within a population, creating a situation where the indicators of the drug would be randomized across the patients genetically predisposed for a high vs. low bioavailable drug exposure. Analyses of the interactions between the drug and genotype on clinical outcomes may clarify the effects of the drug on the outcome measure while controlling for indicators of drug use within the study design.

VIII. Disease Sub-Groups

Almost any complex disease may be partitioned into sub-groups for further analyses. Simple examples include stratification of breast cancer cases by estrogen receptor (ER)-positive vs. ER-negative status or prostate cancer cases by Gleason score. Increasingly sophisticated biological markers segregating specific sub-groups of disease continue to be developed and should be utilized within any study design to reveal the disease characteristics and pathways most susceptible to an exposure.

CONCLUSIONS

Although MGE has the potential to strengthen an epidemiologic protocol, frail study designs and bias may undermine the opportunity to produce consist results across investigations. Principles relevant for epidemiologic studies are necessary to apply to MGE. However, MGE study design also requires a multi-disciplinary team to support a strong biological rationale within an epidemiologic study framework. The stability of the biomarker within individuals and the ability to attain the required sample size are major obstacles to data interpretation, and transitional studies may be needed prior to a full investigation to understand the characteristics of the biomarker and to determine recruitment goals.

Suggested Readings

1. Hulka BS, Wilcosky TC, Griffith JD. *Biological Markers in Epidemiology*. New York: Oxford University Press; 1990.
2. Miller AB, Bartsch H, Boffetta P, Dragsted L, Vaino H. *Biomarkers in Cancer Prevention. IARC Scientific Publications No. 154*. Lyon: International Agency for Research on Cancer; 2001.
3. Schulte PA, Perera FP. *Molecular Epidemiology: Principles and Practices*. New York: Academic Press; 1993.
4. Rebbeck TR, Ambrosone CB, Shields PG. *Molecular Epidemiology: Applications in Cancer and Other Human Diseases*. London: Informa Healthcare; 2008.

ACKNOWLEDGEMENTS

I want to thank Dr. Wei Zheng for his comments and discussion throughout the process of developing this review.

REFERENCES

1. Rothman N, Stewart WF, Schulte PA. Incorporating biomarkers into cancer epidemiology: A matrix of biomarker and study design categories. *Cancer Epidemiol Biomarkers Prev* 1995; 4(4): 301-311.
2. Shields PG. Molecular epidemiology. *Prog Clin Biol Res* 1996; 395: 141-157.
3. Hulka BS, Wilcosky TC, Griffith JD. *Biological Markers in Epidemiology*. New York: Oxford University Press; 1990.
4. Hulka BS, Margolin BH. Methodological issues in epidemiologic studies using biologic markers. *Am J Epidemiol* 1992; 135(2): 200-209.
5. Schenk JM, Kristal AR, Neuhouser ML, Tangen CM, White E, Lin DW, et al. Serum adiponectin, C-peptide and leptin and risk of symptomatic benign prostatic hyperplasia: Results from the Prostate Cancer Prevention Trial. *Prostate* 2009; 69(12): 1303-1311.
6. Kaaks RJ. Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: Conceptual issues. *Am J Clin Nutr* 1997; 65(4 Suppl): 1232S-1239S.
7. Bingham SA, Day NE. Using biochemical markers to assess the validity of prospective dietary assessment methods and the effect of energy adjustment. *Am J Clin Nutr* 1997; 65(4 Suppl): 1130S-1137S.
8. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *Am J Epidemiol* 2003; 158(1): 1-13.
9. Marshall JR. High-grade prostatic intraepithelial neoplasia as an exposure biomarker for prostate cancer chemoprevention research. *IARC Sci Publ* 2001; 154: 191-198.
10. Owen RW. Biomarkers in colorectal cancer. *IARC Sci Publ* 2001; 154: 101-111.