



AJE-sponsored Society For Epidemiologic Research Symposium

The Role of Epidemiology in the Era of Molecular Epidemiology and Genomics: Summary of the 2013 AJE-sponsored Society of Epidemiologic Research Symposium

Lewis H. Kuller*, Michael B. Bracken, Shuji Ogino, Ross L. Prentice, and Russell P. Tracy

* Correspondence to Dr. Lewis H. Kuller, Graduate School of Public Health, University of Pittsburgh, 130 North Bellefield Avenue, Room 550, Pittsburgh, PA 15213 (e-mail: kullerl@edc.pitt.edu).

Initially submitted September 10, 2013; accepted for publication September 10, 2013.

On June 20, 2013, the *American Journal of Epidemiology* sponsored a symposium at the Society for Epidemiologic Research's 46th Annual Meeting in Boston, Massachusetts, entitled, "What Is the Role of Epidemiology in the Era of Molecular Biology and Genomics?" The future of epidemiology depends on innovation in generating interesting and important testable hypotheses that are relevant to population health. These new strategies will depend on new technology, both in measurement of agents and environment and in the fields of pathophysiology and outcomes, such as cellular epidemiology and molecular pathology. The populations to be studied, sample sizes, and study designs should be selected based on the hypotheses to be tested and include case-control, cohort, and clinical trials. Developing large mega cohorts without attention to specific hypotheses is inefficient, will fail to address many associations with high-quality data, and may well produce spurious results.

immunology; pathology; study design

On June 20, 2013, the *American Journal of Epidemiology* sponsored a symposium at the Society for Epidemiologic Research's 46th Annual Meeting in Boston, Massachusetts, entitled, "What Is the Role of Epidemiology in the Era of Molecular Biology and Genomics?" The symposium was based on a series of important and controversial articles in the *Journal* on the future and importance of epidemiologic research. Two main issues evolved from these papers. The first is whether the future of epidemiology requires big data collections utilizing electronic medical records and other new information technology for both recruitment and follow up of participants to identify "new" risk factors and gene-environment interactions. The second is whether the relevancy of epidemiology is uncertain in the era of metabolomics, genomics, proteomics, and a focus on personalized medicine and individualized risk prediction in contrast to population epidemiologic research. The symposium was chaired by Lewis H. Kuller and included 4 speakers: Michael B. Bracken, Shuji Ogino, Ross L. Prentice, and Russell P. Tracy.

Epidemiologic research is dependent on the development of new technologies for improved measurement of 1) the host, such as genomics; 2) the agent and environment, for example, new technologies for quantifying nutrition, energy expenditures, exercise, environmental exposures, and the totality of

exposures that have an impact on disease processes; and 3) phenotypes, deep phenotyping, and molecular pathology. A unique contribution of epidemiology is selection of populations at risk, study designs, and analytical approaches to test specific hypotheses related to etiology and ultimately the treatment and prevention of disease. Epidemiology is considered one of the basic fields of preventive medicine and is perhaps very successful in studying epidemics; in the last half century or so, epidemiology has also emerged as a major tool for defining differences in disease distributions among and within populations and for the identification of the impacts of various risk factors, both environmental and genetic. Continuing this work will require more precise and accurate phenotyping resulting from increasing the specificity of disease of interest and host characteristics.

Dr. Prentice made several key observations at the beginning of his presentation that have very important implications for the future of epidemiologic studies. First, the risk and distribution of chronic diseases in the population is determined primarily by lifestyle variables. The following factors support this observation: 1) There are very substantial variations in disease distributions among populations; 2) these variations are almost certainly due to differences in exposures to etiological determinants of disease; and 3) migrant populations tend to assume

the risk of their new environment within relatively few generations. If the variations in disease among populations were primarily determined by genetics, we would not observe these marked changes in rates of disease among populations that migrated to new environments.

Our current prediction models do not allow us to determine an individual's risk of diseases with long incubation periods with even modest precision. We can predict the risks for the appropriate group of individuals defined by similar characteristics, such as the Framingham risk score, but we have little ability to predict any single individual's risk within the groups, a phenomenon that highlights the differences between population and personalized medicine prediction. Diet-specific nutrients, total energy intake, and energy expenditure (exercise) are very likely to continue to be important determinants of many chronic diseases, such as cancer, cardiovascular disease, diabetes, and obesity, and perhaps even some inflammatory autoimmune diseases and age-related diseases.

Although there are some genetic differences among people that are important factors in chronic complex diseases (e.g., changes in lipid genes, such as *PCSK9*, that affect cholesterol levels and therefore atherosclerosis or variation in coagulation genes, such as factor V Leiden, that affect coagulation status and therefore thrombosis), many genetic variants are important determinants of host susceptibility that act primarily as mediators secondary to the exposures to lifestyle or environmental agents. In other words, the genome responds to different lifestyle and environmental exposures. Variations in the distributions of specific susceptibility genes in the population have evolved because of differences in exposure to environmental agents. For example, the much higher prevalence of apolipoprotein L in blacks could possibly be associated with trypanosomiasis, and malaria may influence the prevalence of sickle cell anemia. Some of these evolving genetic attributes previously provided success against infectious agents (i.e., inflammatory responses provided innate and adaptive immunity) but may now be deleterious in a changing environment.

PHENOTYPES

Drs. Tracy and Oginio both stressed the importance of deep phenotyping of participants in epidemiologic studies. Dr. Tracy focused on cellular epidemiology, the effects of specific cellular responses, such as the innate and adaptive immune responses as related to the epidemiology of cardiovascular disease, and aging. Dr. Oginio noted the importance of molecular pathology in understanding the interrelationships between risk factors, various therapeutic agents in cancer etiology, and prevention. The limitations of current laboratory epidemiology studies include the lack of specificity of measurements related to the specific tissues of interest. An example would be measuring cytokines in the blood to estimate a complex, cell-based inflammatory response in a specific tissue, such as in skeletal joints in rheumatoid arthritis or in coronary plaques in atherosclerosis. In addition, the limited sensitivity of measurements in many easily obtainable biological specimens, such as serum or plasma samples, restricts these studies. The low levels of analytes in these samples result in substantial laboratory and within-person variability that, combined with the lack of tissue specificity, may produce biased results. Many

acute-phase proteins and coagulation factors are produced in the liver. Liver injury and/or infection may result in increases or decreases in the production of these mediators independent of their association with a specific disease. The double-edged sword of lack of specificity for biological processes and locations and common pathophysiology of many chronic diseases often results in collinearity among variables, with numerous variables showing significant associations and low relative risk; this is often seen in studies of cytokines, inflammatory markers in the blood, etc.

Mendelian randomization studies have become a popular genetic epidemiology tool to evaluate the specificity of associations, but the effect of specific single nucleotide polymorphisms on phenotype level is often so small as to preclude or limit the mendelian randomization approach. Still, for some associations, this is an important technique for genetic epidemiology.

Cellular epidemiology has a focus on studying specific cell types and activities, such as lymphocytic function, in an attempt to increase specificity for human biology. Cellular epidemiology approaches have the potential for dynamic interrogation of biological systems in epidemiologic studies in response to various agents and environmental exposures, such as the response of specific T-cell types to environmental challenges. Collection of specimens for such deep phenotyping in cellular epidemiology is not as simple as drawing blood and separating plasma, serum, and cells. Collaboration between epidemiologists, field centers, and laboratories is key. The Multi-Ethnic Study of Atherosclerosis (MESA), for example, has overcome many logistical problems to measure T-helper cell bias and its relationship to measures of subclinical atherosclerosis and determine that Th1 bias was associated with greater coronary artery calcium. Th1 bias was positively correlated with titers of cytomegalovirus exposure. Epidemiologic approaches are the only way to determine whether cytomegalovirus or other herpes viruses may be a driving force in T-cell immunosenescence and in the balance with aging of adaptive and innate immunity. Because it is associated with increased inflammation and disease burden, viral burden over time may be an important determinant of inflammatory response and may contribute to aging and the chronic diseases of aging, especially in association with specific genetic attributes.

Cellular epidemiology studies provide a unique opportunity to further understand the relationships of potential agents of disease, such as viruses, chemicals, dietary factors, and others, with immune responses at the cellular level, genetic host susceptibility, and disease development. An example is the interrelationship of inflammatory immune responses and risk of thrombosis, which may be a key to precipitation of heart attack and stroke. These are, however, costly studies that require careful attention to specific hypothesis testing and collaboration between epidemiologists and laboratory scientists.

Molecular pathological epidemiology focuses on the molecular heterogeneity of diseases defined by clinical characteristics. The molecular heterogeneity of diseases such as colon and breast cancer is due in large part to somatic mutations in specific malignant and premalignant tissues, which allows the identification of specific clones of cells with unique genetic and epigenetic characteristics. There are different environmental and lifestyle exposures associated with one or another molecular

type of colon cancer. Pooling all colon cancer cases as one “disease” may result in missing important etiological associations or in substantially reducing the estimated relative risks. A specific exposure, for example, cigarette smoking, obesity, or a specific nutrient in the diet, may be a risk factor for a molecularly specific disease subtype. Epidemiologic studies, such as the Nurses’ Health Study (NHS), have analyzed specific molecular characteristics of both benign and malignant colon tumors using formalin-fixed paraffin-embedded tissue specimens. Such studies are expensive but clearly represent a newer approach in epidemiology and clinical trials. For example let’s assume that for colon cancer, type A accounts for 20%, type B accounts for 20%, and type C accounts for 60%. It may be possible that a dietary attribute would increase type A colon cancer by 4-fold but would only increase overall colon cancer by 1.6-fold if the other 80% of colon cancers were unaffected. If we had the ability to distinguish type A colon cancer from the others, such estimates could be determined and our view of that dietary attribute would change dramatically. Similarly, clinical trials that do not include better definitions of phenotypes by means of, for example, deep phenotyping using molecular pathological epidemiology, may have overall null results despite substantial benefits for predesignated phenotypic subgroups.

MEASUREMENT OF EXPOSURES

Nutritional epidemiologic studies have not been very successful in identifying specific determinants of many diseases, especially cancer. The most likely problem Dr. Prentice noted is the weakness of the tools to measure specific nutrients, especially within relatively homogenous populations.

In recent years, large cohort studies have become the mainstay of nutritional epidemiologic research, but they have been substantially limited by confounding of diet by other variables, measurement biases, and probably most important, issues in the accuracy of dietary assessment for specific nutrients and foods. It is important to do well-designed, controlled dietary feeding studies to develop biomarkers for additional components of diet, as was done in the past in diet-heart research, to help unravel some of these issues. You can then utilize new biomarker technology to relate specific dietary intakes with biochemical and metabolomic variables. This new information can then be used to recalibrate population-level data on reported dietary intake of specific nutrients. For example, in the Women’s Health Initiative (WHI), measurement of total energy intake over 2 weeks using doubly labeled water protocols with repeated evaluation over a 6-month period was used to estimate total energy intake versus reported intake and, with the addition of data on resting metabolic rate, energy expenditure of physical activity. This approach substantially reduced the known underreporting biases of energy intake, especially in the intervention arms of dietary trials. In the past, many epidemiologic studies have reported that total energy intake was less in obese persons than it was in nonobese persons, whereas the total energy biomarker was strongly positively related to body mass index.

The availability of specific blood and urinary markers of protein intake (such as urinary nitrogen) and of vegetable intake (such as carotenoids) coupled with newer techniques,

including metabolomics, amino acid analyses, and others, will vastly improve nutrition studies. Such improvements are critical based on a growing interest in improving our understanding of the role of nutrition in health and disease. An ongoing feasibility study in 150 Women’s Health Initiative participants in Seattle, Washington, is focusing on the development of objective measures of nutrients and foods.

HOW TO UTILIZE NEW TECHNOLOGY IN EPIDEMIOLOGIC STUDIES

The success of an epidemiologic study begins with the selection of populations at risk and appropriate study designs. Dr. Bracken focused on the utilization of new approaches for epidemiologic study design. A major current challenge in epidemiology remains the enthusiasm for big cohort studies that is driven largely by the need for large numbers of people to detect genetically mediated small changes in risk. Four strategies exemplify successful large scale research: 1) large, simple randomized trials to improve the precision of estimating rare outcomes; 2) large genomic collaborative studies that often use case-control designs; 3) studies that are driven primarily to more accurately estimate the number of events or incidence rates, particularly in less technically developed countries and especially for diseases with low incidence rates; and 4) meta-analysis of large administrative data sets to quantify disease outcome and to estimate the quantity, quality, and costs of health care. Despite the success of these strategies, Dr. Bracken challenged the concept that big cohort studies represent the future of epidemiology. Most big cohort studies, including mega cohort studies, claim to be prepared to study numerous current and future hypotheses. In essence, however, they are studying no specific hypothesis, and for most post hoc hypotheses, these studies are substantially under- or overpowered. Measures of exposures, for example, diet, exercise, and social and physical environmental variables, are of low quality because difficulties with excessive respondent burden preclude detailed evaluation. Moreover, outcome assessments often depend on health care delivery systems, which may both be biased and have incomplete or inaccurate ascertainment of endpoints. The detailed phenotyping mentioned earlier in this article is frequently lacking. Low response percentages may select against specific populations that are of the greatest interest, especially when studying gene-environmental interactions.

The absence of any specific hypotheses will limit our ability to take full advantage of new technologies. The populations most likely to provide useful information, that is, those at highest risk, may not be included even with large sample sizes. It would be preferable to focus on specific well-defined hypotheses using the very best new technologies and to test these hypotheses in populations likely to provide answers.

For example, if we are interested in studying the effects of exposure to shale gas drilling on health outcomes, then it is likely that even a very large national cohort study might not include enough individuals with the exposures of interest. Similarly, exposures to environmental teratogens are most likely to cause birth defects when they occur very early in pregnancies. Risk factors before conception may also be very important in pregnancy outcomes, for example obesity, smoking, elevated blood pressure, diabetes, and vitamin intake. Large cohort studies

such as the US National Children's Study (NCS) that select populations late in pregnancy or even at birth will miss important risk variables.

SUMMARY

Epidemiology research has recently been criticized for lack of relevance and cost, and it is being driven more and more toward large sample sizes that utilize new information technology, such as electronic medical records. Epidemiologic studies have been very successful not only for investigating epidemics with a short incubation period, for example, food-borne outbreaks, but also for identifying and quantifying risk factors for longer-term diseases, such as human immunodeficiency virus/acquired immunodeficiency syndrome, coronary heart disease, diabetes, osteoporosis, and site-specific cancers.

The future of epidemiology depends on innovation in generating interesting and important testable hypotheses that are relevant to population health. These new strategies will depend on new technology, both in measurement of agents and environment and in pathophysiology and outcomes, such as cellular epidemiology and molecular pathology. The populations to be studied, sample sizes, and study designs will depend on the hypotheses to be tested and include case-control, cohort, and clinical trials. Developing large national cohorts without attention to specific hypotheses is inefficient, will fail to address many associations with high-quality data, and may well produce spurious results.

Advances in genomics have provided key measures of host susceptibility, genetic attributes, and possible biological effects of agents on the host and include studies not only of primary genome sequences and genotypes but also of epigenetics (e.g., CpG methylation), microRNA, somatic mutations, etc. Epidemiologic studies focusing on the host alone, genomics, or crude measurements of environmental agents or poor definition of the disease will have limited abilities to enhance the field of epidemiology.

Finally, we believe that the etiology of a specific disease often appears at first glance to deal with a very complex combination of variables that all have small effects. However, such epidemiology studies also often turn out to include imprecise and inaccurate measures of exposure to an agent (environment), of host susceptibility (genomics), and of outcome, especially with respect to specificity and definition of outcome. Improved identification of the specific agent(s), better measures of host susceptibility, and more precise definitions of outcomes often clarify the "puzzle," such as identification of viral etiology of cervical cancer and human immunodeficiency virus, dietary intake of saturated and polyunsaturated fat and blood cholesterol levels and coronary heart disease, aspirin and risk of Reye's syndrome, and high exogenous and endogenous estrogens and breast cancer. Epidemiology remains the study of epidemics in time, place, and person. The new technologies provide the opportunities to probe host, agent, and environment with greater precision and detail and to help test hypotheses regarding disease prevention and treatment.

RELEVANT QUESTIONS FOR FURTHER DISCUSSION

1. Epidemiology focuses on population or group differences. Science is moving to individualized, that is, personalized

medicine (individualized therapies, etc.), genomics, proteomics, etc. How should epidemiology respond? Will new molecular biology and technology approaches (i.e., DNA methylation, epigenetics, microRNA, etc.) be able to replace traditional measures of environmental exposures?

2. Many of the new molecular biology technologies are attempting to quantify very low levels of markers. How do we evaluate the within- and between-individual variability, especially among studies that use different technologies and populations, that is, populations with or without diseases, etc.? Should studies be published without reproducibility data?
3. Many new molecular measurements are probably very highly correlated with other measurements (e.g., acute phase responses, fat metabolism, liver disease). How should epidemiologic studies deal with the complex interactions of these variables; are we just finding ways to measure the "head of a pin?" Are any of the measures of "energy metabolism" any better than a scale and exercise questionnaire?
4. Epidemiologic studies often succeed by studying populations with very different rates of disease, risk characteristics, etc. Is it feasible to obtain specimens or use new technologies in populations outside of major research centers? What are the limits to transporting and storing specimens? Many chronic diseases have long incubation periods. Can stored samples from the past be utilized with new molecular biology and technologies? Will future epidemiology studies be restricted to major urban areas?
5. Many molecular biology characteristics are tissue- or cell-specific, for example, microRNAs and somatic mutations of cancer cells, yet we often use urine, blood, plasma, blood cells, energy metabolism, etc. What are the risks? How do we evaluate whether our nontissue- or cell-specific measurements are useful (e.g., telomere length in white blood cells, microRNA in blood, cytokines in blood)?
6. How should future epidemiologists be trained and current epidemiologists be retrained to be able to utilize and understand new molecular biology and technology in their research? Should the next generation of epidemiologists come from scientists with training in molecular biology, genetics, or computer sciences?
7. Is it possible to integrate the new methods of molecular biology, cellular epidemiology, genomics, and metabolomics into proposed mega cohorts? Is there a place for such large cohorts for epidemiologic research?
8. Are there any new molecular biology and technological approaches for preventing rabbits from eating Dr. Kuller's vegetable plants?

ACKNOWLEDGEMENTS

Author affiliations: Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania (Lewis H. Kuller); Center for Perinatal, Pediatric and Environmental Epidemiology, School of Medicine, School of Public Health, Yale University, New Haven, Connecticut (Michael B. Bracken); Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Shuji Ogino); Department of Pathology,

Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Shuji Ogino); Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts (Shuji Ogino); Public Health Services, Fred Hutchinson Can-

cer Research Center, Seattle, Washington (Ross L. Prentice); and Department of Pathology, University of Vermont, Colchester, Vermont (Russell P. Tracy).

Conflict of interest: none declared.