

Epidemiologic Tools for Today and Tomorrow

PATRICIA HARTGE

*Division of Cancer Epidemiology and Genetics,
National Cancer Institute, Bethesda, Maryland, USA*

ABSTRACT: As the premier population sciences, epidemiology and demography face common challenges as the U.S. population ages, as the genomic revolution unfolds, and as computing power changes the scale of analysis by several orders of magnitude. Each discipline does need to develop new tools to address the changing research questions, and the best strategy for success includes increased collaboration between the disciplines. The paradigms of each discipline still offer important insights into the problems of both disciplines, so cross-training would be a simple step to begin enlarging the tool box for population science.

KEYWORDS: epidemiology; demography; methods

INTRODUCTION

Epidemiology and demography share their focus on the population, and they are the premier "population sciences." Demography aims to depict the size and shape of the populace and to predict how its contours will change. Epidemiology measures the pattern of disease in the populace so as to understand biology and, thus, to improve public health. To some extent, the disciplines use common statistical tools, and, to a great extent, common data. More precisely, epidemiologists are delighted to use the population data demographers use, but with a distinct purpose and focus.

With different purposes, the disciplines emphasize different paradigms. Epidemiology features the 2×2 table (see TABLE 1). With elegant simplicity, it reveals to us how subdividing the population, the denominator, subdivides the risk of disease. In the case-control context, the actual denominators can be unknown, provided that the control group represents the population from which the cases arose. That is, case-control design may work with an unknown sampling fraction provided the *ratio* of exposed to unexposed people in the control sample equals that in the population. Of course, epidemiologists prefer to conduct a study using the demographer's count of the entire population, because that permits measurement of other important parameters. The ratio or the difference in the rates of disease occurrence in the exposed and the unexposed subsets of the populace nonetheless can be estimated accurately from appropriately chosen groups for comparison.

The 2×2 table extends to a $2 \times N$ table (see TABLE 2) or to continuous measures of exposure, allowing examination of a dose-response effect. By separating the population further, one can remove the confounding influences of third-party variables

Address for correspondence: Patricia Hartge, Sc.D., Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd., EPS 8090, MSC 7246, Bethesda, MD 20892-7246, USA. Voice: 301-496-7887; fax: 301-402-2623.
hartgep@mail.nih.gov

TABLE 1. 2 × 2 table

Stage of Pregnancy	Cases	Controls
Early pregnancy	242	337
Late pregnancy	75	64

TABLE 2. Odds ratios for breast cancer by age at first full-term pregnancy among women with parity at least 2, multistate Collaborative Breast Cancer Studies

Variable	Cases	Controls	Odds Ratio
Age (years) at first full-term pregnancy			
<20	242	337	1.00
20–24	1,080	1,393	0.97
25–39	819	1,024	0.96
30–34	307	281	1.35
≥35	75	64	1.56

SOURCE: Wachter and Finch.¹

or examine how the exposure of interest interacts with other factors. The goal is to find clues to etiology, such as the increasing level of breast cancer risk with each year of postponing first birth.

A classic tool for demographers that epidemiologists and others also employ is the lifetable. This simple approach yields an abundance of parameters that simulate a lifespan—life expectancy, fertility rate, cumulative risk of cancer, and so forth. These and other tools of survival analysis are, and will remain, at the heart of both disciplines. The aging of the population challenges epidemiologists and demographers to apply these tools and to devise new ones. We need to consider how the big demographic wave will push the disciplines closer to each other. Two other enormous changes will radically alter epidemiology and demography—the revolution in computing and the revolution in biology. This is a good time to take inventory of what we do and how we do it, although both will surely change.

POPULATION SURVEYS

Epidemiologists rely on population surveys as the basic source of data in epidemiology, although they rarely survey the entire population of a county, state, or nation. These surveys provide the critical data needed for disease surveillance and for computing true rates of disease occurrence. When the Census Bureau makes a change in the classification of race, for example, it profoundly affects our ability to produce accurate descriptive epidemiology.

To address study questions pertinent to the whole population, we frequently draw a sample of that population and approach the individuals selected for interview. For cancers and for some other diseases, population registries include virtually all cases

occurring in the population. For example, in many states, cancer registrars record virtually all cancers diagnosed in the whole population. Thus, the proper comparison for the exposure histories of cases applies to the population, and the proper control group is a random sample of that general population (stratified on age, sex, and race for statistical efficiency in estimation). To draw a control group, investigators might use random digit dialing, to select people from households with telephones and the Health Care Financing Administration rosters to select for those over age 65. This frame does not produce a perfect sample, but it does have extremely high coverage. In the U.S., only a few alternatives exist, among them driving licenses (much lower coverage) and town lists (in a few states).

Drawing the population sample has been made easier by increased computing capacity and ready access to sophisticated and specialized software. On the other hand, persuading people drawn into the sample to participate in an epidemiologic study has become much harder. People are busier, more burdened by telemarketers, and more skeptical. Increasingly, studies must offer financial incentives, often substantial, or compensation for time spent on the study. Interviewers also spend more time persuading people to participate.

In a similar push-pull situation, routine computerized storage has made it much easier to get critically valuable data from medical and other personal databases, but legal, ethical, and cultural impediments to access have arisen in parallel. Just when the answers to many important epidemiologic questions are within reach, the data may be off limits. In other nations, the particulars of social cooperation, privacy concerns, and population record-keeping differ greatly, but nowhere is truly population-based epidemiologic survey research simple.

QUESTIONNAIRES: MAINSTAY OF EPIDEMIOLOGIC STUDIES

Although the challenges of drawing samples from the population for surveys loom large, the questionnaires are getting better all the time, for demography, epidemiology, and other disciplines. The various computer-assisted methods (CAPI, CASI, CATI, and ACASI) provide more options for administering interviews, for example. Postal and other self-administered formats have gotten better with experience and research.

On questionnaire content issues, epidemiologists and demographers still could learn more from each other and from sociologists. We know we need to work together on race and ethnicity questions. We also need to collaborate on measuring social class. In many epidemiologic studies, we have trouble measuring this concept, and, perhaps, demographers do, too. In part, social class has been slightly to the side for both disciplines, but it moves to the realm of central concerns as we both sharpen our understanding of ethnicity and the effects of ethnicity on health disparities.

BIOSPECIMENS: THE BODY TELLS ITS OWN STORY

The revolution in biology has given us more than insight. It offers a whole new world of "exposures" to measure, the world of biomarkers and intermediates.

Whether from blood products, cheek cells, or the paraffin-embedded normal tissue surrounding a tumor, DNA is now collected in many studies. Even when no specific genes are implicated at the outset, we have great expectations that some genes certainly will play a role; DNA banking is becoming the norm.

The technical and logistic issues in DNA collection vary from setting to setting. In some studies, it is feasible to collect blood by finger-prick. Finger sticks actually hurt more than venipuncture, but most people prefer the finger stick, perhaps because it feels less invasive. Of course, privacy considerations constrain how, when, and why we can collect DNA. IRBs are struggling with these issues now, and the ground rules could change radically during the next ten years. Blood is the major biomarker reservoir for many studies. More than a source of DNA, it provides a measure of current levels of hormones, vitamins, and organic compounds. These current levels sometimes serve as an index of past exposure integrated over time. On the other hand, blood collection has its share of problems. Blood drawing in the general population very frequently attains a response rate of less than 50%. This poses potential threats to validity from selection biases. Confounding also does not vanish merely because the exposure is serum level DHEAS, and so forth. Furthermore, levels in the target tissue (e.g., breast) may be most relevant, and they can be substantially above or below serum or plasma levels. Interpretation of blood assays depends on measures of validity and reproducibility, which are often lacking.

Nonetheless, blood collection is here to stay. Often, the challenges of biospecimen collection lead to a choice of hospital-based designs. Indeed, this is the optimal design for many questions, for example, the recent study of brain cancer that exonerated cell phones. At first, the biologic revolution may appear to push demography and epidemiology further apart. On closer inspection, it seems likely that biospecimen collection will become a routine feature of population research.

MEASURING THE ENVIRONMENT

The measurement of exposure covers a wide range, including truly ecologic variables like air pollution, somewhat aggregated measures like radon, household indicators like pesticides from lawn care, and truly individual measures. In this area, there are several matters of common concern for population sciences. We need good training in existing study designs and analytic approaches that accommodate various effects of aggregation of exposure. We also need development of new approaches.

From an epidemiologist's point of view, there are reasons to lean toward the disaggregated measurement. First, we still need to worry about the correlates of exposure that may confound, typically varying from person to person. Perhaps one county has a uniform and high level of an air pollutant, and a second has a uniform low. Suppose the proportion of cigarette smokers is higher in the first county, and so also is the lung cancer mortality rate. Both aggregated and disaggregated approaches are feasible, but epidemiologists generally want individual smoking history to measure the effects of the air pollutant on lung cancer risk. The challenge is to be sure we understand the menu of design choices and consider the cost and value of each in the mixed variable setting.

Timing of exposure poses another challenge for environmental measures that will be compared to risk of disease, especially chronic disease. Sometimes a weighted summary of 30 years of environmental exposure is the relevant influence on disease risk. Thus, current measured levels in the environment may not measure the relevant index. This concern may dictate individual measurement of exposures that are actually populationwide at the time they occur but that become individual as people move from place to place. For such variables, historical data are invaluable, for example, from the records maintained by public water suppliers.

COMMON GROUND: DESCRIPTIVE EPIDEMIOLOGY AND GIS

Descriptive epidemiology forms the keystone of epidemiology and deals with cancer maps, and race-, age-, sex-specific rates of disease. Within epidemiology, it is the subspecialty closest to demography. It is also the oldest part of epidemiology and not one that typically attracts a lot of attention. Descriptive epidemiologists routinely complain that most epidemiologists do not understand or appreciate the subtleties of descriptive epidemiology.

This classic endeavor is getting a fresh new look because of the web and fresh ideas from the emergence of Geographic Information Systems (GIS). Epidemiologists are trying out large-scale field use of Global Positioning System (GPS) devices. The readings can then link the individual's location (or locations, for historical residential data) to these very large databases of various characteristics ("exposures" to epidemiologists). GIS/GPS seems likely to attract interest from various disciplines and could be a major stimulus to new approaches to ecologic and individual analysis.

The new look of descriptive epidemiology provided by the web may also infuse this area with new ideas. People untutored in statistics or population science have unprecedented access to wonderful data sets, powerful and flexible tools for analysis, and instant communication of their ideas about the patterns they find. We will have to devote some time to simple error surveillance and to educating the public about population science methods.

CROSS-TRAINING: A SIMPLE PRESCRIPTION

Apart from conducting research and developing new methods in population science, we can improve training of demographers and epidemiologists.

First, there are the classic tools of wide application, especially the lifetable and all of its progeny. It is worthwhile for epidemiology students to examine some real lifetables and then to devise their own related tables for other kinds of events that vary greatly with aging. This may be part of a bigger problem in epidemiology and biostatistics in how we teach standardization measures. Many well-trained epidemiology students do not have good intuition for the basic idea. Similarly, the connections among the many survival methods are well worth emphasizing.

Second, I would like epidemiology students to know, at least roughly, the current facts about the size and shape of the U.S. and the world populations. Not to the nearest thousand, but to the nearest 10% would be very helpful. This would increase their

tendency to be practical and concrete in designing studies and realistic in their laudable efforts to pull together results from one or more epidemiologic studies with population survey data.

Third, I hope the training of demography students would include detailed consideration of validity and bias as these issues arise in epidemiologic studies. They need practice in discerning the likely magnitude and direction of that bias. Students in both disciplines need exposure to the techniques of population attributable risk and to the pitfalls, as seen from the two disciplinary perspectives.

It would be quite feasible to accomplish these modest training goals. A demographer and an epidemiologist might co-teach a class designed for students in either or both specialties. Together, they could present common problems that would be approached somewhat differently by the two disciplines and explore the similarities and the differences in approach.

TOMORROW

From the point of view of an epidemiologist, things are about to change very dramatically. Three big waves are crashing over us. The population is aging and diversifying—one reason for this conference. The revolution in biology gives us more information about one person than we can imagine. The revolution in computing means we can push the data around in ways we can hardly conceive. Some false leads are in store in the realm of gene–environment interactions, and we should expect them. We know a lot about how to measure the effect of an exposure or a characteristic on the risk of disease, especially about how easily one can get the wrong answer. We have no choice but to make epidemiology faster and smarter. If epidemiologists do not measure how and why disease sits upon the population, other people will, and they will make avoidable mistakes. If we succeed in riding these waves, we will make a great leap forward in understanding biology of the species, risks to the person, and health of the public. In order to succeed, we will need new statistical approaches, very big studies, and lots of collaboration with demography and our other sister disciplines in population science.

REFERENCE

1. CHI, W.-C., C.-C. HSIEH, P.A. NEWCOMB, *et al.* 2000. Age at any full-term pregnancy and breast cancer risk. *Am. J. Epidemiol.* **151**(7): 715–722.

COMMENTARY AND DISCUSSION OF THE PAPER

Patricia Hartge: I have to tell Jack Caldwell that I am a very deeply skeptical person, first of all. I am an epidemiologist and I am skeptical of my data. I want to tell Doug Ewbank that I am skeptical of a great deal more than the diagnoses. However, I think that you are on to something important in the theory and the gestalt that is a little different between our two disciplines. The commonality is obviously, the *deme*, the epidemiology and the demography, and that we are really, let us pat ourselves on the back, the premier population sciences. As I have tried to think about the tools that we use in epidemiology today, I have tried to think where we have some common problems that we can work on together, and I will try to highlight those. I am also going to try to see if I can touch on some of the genesis of your remark about how epidemiologists use our data. I think probably the thing that you said that I need to address first is something, actually, that Ezra Susser mentioned in his talk.

Why are some individuals at higher risk? This is the question that we ask most, or that I do in my work, and I think I am pretty typical of most epidemiologists. I am engaged in a study of non-Hodgkin's lymphoma. I am asking, Are measurable levels of environmental exposure to pesticides related to the risk of developing non-Hodgkin's lymphoma? This study uses a population-based design; I define the geographic areas that people must reside in to be in my study, but I do not study those populations in the same way that demographers describe the structure of a population; what is fertility, what is mortality? Yet both approaches are population-based research.

My chief question is biologic. Therefore, the attention that I pay to the errors in my data is driven by that need. Validity is my chief concern. Errors in selection, errors in the data that I get and how I get them, and mistakes that I may make in the analysis are my chief concern. The sort of paradigm of what I do, of what epidemiologists do, is the two-by-two table. I think demographers wince when they think of about one column in the two-by-two table, namely the column that is meant to represent the control group, or the referent, or the underlying population. Because, from the epidemiologist's point of view, the job of that column, of those two cells, is to represent the population whence cometh the cases. It is not intrinsically interesting, it is interesting only that it should speak to the same underlying population from which the disease came. I think it is a very powerful tool—that is why I have been doing the work for twenty years. It can easily be taught to demography students along with the difference in perspective. I would also say that a lot of what demographers do and their central paradigms, the life table especially, should be stressed in epidemiology. It would be a good and easy thing for us to make sure that students in epidemiology and demography have a strong sense of the tools and the perspectives of each discipline.

I am going to return to the theme of practical suggestions for training. First, I will do the thing that you asked me to do and talk a little bit about the tools that we are using today. I am going to draw heavily from three studies that I am currently involved in. One is the study I just described, and this study has an interview component; the cases are drawn from cancer registries. Cancer epidemiologists view demographers as measuring two really hard outcomes, birth and death, but almost all other epidemiologists view cancer epidemiologists as measuring a relatively hard

outcome that is, in many parts of the country, very thoroughly detected and documented in cancer registries.

I am also working on a study, a survey that we conducted, now three years ago, in the Jewish Ashkenazi population of Washington, D.C. It was a one-time survey involving a very brief questionnaire and a finger prick. I am going to mention it because I think at the outset Richard said he posed the question of another conference: Should biologic measures be included in social science research? Well, if you change social science research to population research, and that surely makes it epidemiology, the answer is, *absolutely*. However, it is very tricky, and I want to talk about some of the problems when you do it, but that is a survey in which we were able to very quickly, in six weeks, persuade volunteers, more than 5000 people, to give us a little finger prick of blood, and a very short survey that gave us, if I may say, in all modesty, incredibly important data on what the real risk is of developing breast or ovarian cancer if you should carry a mutation in your BRCA1 or BRCA2 genes. It was not population based, but it was, I would argue, *the way* to answer an important question that clearly has implications for the population. The other study that I am currently involved in is of the cohort design. This is actually rolling off of an experiment. The trial assigns people at random to either be screened for prostate, lung, colon, or ovarian cancer, or to have their routine medical care. In the course of the screening trial we collect a lot of biospecimens, we ask a lot of questions, and then we can follow the subjects over time. So I will be drawing my examples from my personal experience with those ongoing studies.

I am not only skeptical, I worry a lot too. Let me tell you what I worry about. I worry about it principally from the perspective of epidemiology, but as I am listening to what everyone else is saying, I am thinking demographers also ought to be worried, and it sounded to me from Richard Suzman's remarks that he has begun to worry about some of the same things. I feel like we demographers are a little fleet of fishing boats, and epidemiologists have a little fleet of fishing boats and they fish in some of the same waters that we do, and there are three giant waves coming. One is the aging of the population, and that is a kind of slow swell. There is a huge wave coming because of the revolution in computing. The amount of data is immense—I started surfing around on the web looking for data and I was impressed at how much you can do in a half an hour with the databases that are out there for the world to use. It has really transformed how you can do demography and epidemiology, and also who can do it. Anybody whatsoever can quickly tool around and get data that used to be, frankly, what we had to have, the keys to the kingdom, to be able to do. I think coming at me, and I think at demography too, is this tsunami of genetic and biologic information, so that from those tiny little finger pricks on the cards that I described there are going to be more data than I could ever have imagined in my wildest dreams. I think that is going to result in the fleet of fishing boats that is now demography and the fleet of fishing boats that is epidemiology bouncing all around, and I am not sure when those three waves have finished kind of crashing over us if we will all be fishing further apart, or in the same waters, or what, but I think our common challenge is to do the things that we know how to do well as these waves are kind of coming at us. Because what I see is an awful lot of freelance interpretation, freelance epidemiology that makes not subtle errors but basic errors, that no one would make who had ever taken Epi 101. Geneticists now say, "It's people, it's genes, it's data, I

have a computer." I think we will find ourselves in a very different situation, within, perhaps, the next ten years, and I am hoping that by working together we actually will be able to not just survive these waves coming over, but that we will say, boy, this was the golden age when we really understood how these population sciences were speaking to the same problem.

In population surveys we have and will continue to rely on demographers to get the denominators, basically. We would much prefer that demographers and government agencies would count who is in a population and we are happy to take on the extra work of counting all the cancer cases that arise in that population. The example of the registries that I mentioned is a classic one. This is a premier system of registries in the United States, the S.E.E.R. registries; they are also registries funded by the CDC. It is, technically, a lot easier to draw these samples of the population for the kinds of studies that we do in epidemiology, but getting the response rates once the samples have been drawn is *really* a lot harder than it was. In the study of non-Hodgkin's lymphoma, in drawing the sample under the age of 65 random-digit telephone dialing was used, and above the age of 65 the Healthcare Financing Administration rosters were used, which are said to be 98% complete. This is the same technology that we actually pioneered in 1978. It has become a whole lot easier to do the drawing but persuading people to give us an hour or two of their time for the good of science is immeasurably, no it is *measurably*, harder than it was. This means that to the skeptic and the worried person, the possibility of response bias *really* is a problem. I can give you small technical solutions that we have tried—mostly, spend more money, train people better, but I do not think it is going to have a real easy answer. Tempting is the fact that sometimes you can answer the problem with banked data. These have also become just so enticingly good and there are, certainly in Scandinavian countries, questions you can answer with banked data that, frankly, cannot be answered anyplace else. At the same time, I feel that we can address some questions; we finally have the data to answer some important questions, but the access is much more restricted. On confidentiality issues, the IRBs are struggling, and every year when I go back to an IRB and ask permission to do anything remotely related to this, I see the level of general social concern about access to those data, which would be extremely helpful for demography or for epidemiology, is increasing. This is one area where we have to make common cause, and other disciplines in population science have to join us.

In questionnaires, I am using a combination of a computer-assisted telephone interview and self-administered questionnaires in the non-Hodgkin's study, for example. The computer-assisted methods, CAPI and CASI, are also great advances. I think that training people and understanding how to train people to do survey research is vastly better than it was twenty years ago. However, I think that the content areas that are weak for epidemiology are actually areas where we could learn a lot by working with demographers and probably also with sociologists. When the census changed its classification of races, this had enormous implications for routine surveillance in epidemiology, but I think that was a tractable issue compared with social class. In epidemiology, we mostly do a very bad job measuring social class, and I think we could probably learn a lot from demographers and sociologists.

In the National Cancer Institute we have a focus on understanding and reducing health disparities. I think that kind of work is going to turn on better understanding

of what the constructs of social class are, and I think it is one area where, clearly, epidemiologists and demographers are going to be working together. Biospecimens have become the norm in our very large, expensive studies. If it is going to cost a lot to find the population of interest, and often it does, DNA banking is the norm. If you need to have a high response rate, and you would certainly like to when you are banking DNA, you have to use some of the less invasive measures, such as the collection of cheek cells, which can be done with a swab or a swish in a bottle of mouthwash. This works very well for a modest amount of DNA, so it is fine for a one-shot problem, or even ten genes or twenty genes. It is not okay if you want to then address the question of a thousand interesting genes. For that you still need blood collection. Even though venipuncture actually hurts less, most people do not like it; they find it invasive. In my studies it is typical that half of people, even when you say, "I can learn a whole lot more about your exposure to pesticides and your medical history if you give me blood," will say, "No thanks," or nearly half (forty percent) will say, "No thanks, but I will spit in a bottle for you." Thus, this is a rate-limiting factor when you are thinking about really doing population research and getting biological specimens. You need the blood for many things other than DNA; you need it for dietary and hormonal issues, for environmental pollutants. Even if you're lucky enough to get blood, you may wonder what is that level of the compound of interest? For example, you measure the circulating androgen, but you want to measure the androgen that the prostate is seeing. That kind of question pushes you away from population designs. If you are starting to think about taking biospecimens from individuals to get tissue-levels of things, of course you will still be pushed to a two-by-two table, but now the focus is on a population that I can imagine, but not one that I can enumerate. Thus, the population I imagine is the population of people who will come to the George Washington Hospital if they develop prostate cancer and will become my cases. Now the controls I have to find are those people that do not have prostate cancer who are in the underlying population at risk.

Demographers seldom want or need these kinds of research designs, but you can see that there is no alternative, and you can also see that good collaboration between a person with a strong population perspective and a person doing this very biologically driven work can yield, in the end, a combined sense of the biology of the problem and how it sits in the population.

The environmental measures that I am taking in the non-Hodgkin's study are pretty typical of the kind of work that's going on now in epidemiology. What the interviewers ask for is the used vacuum cleaner bag because it turns out, after some research, to be not a bad dosimeter of pesticides that are tracked in from the lawn. It is not perfect, but it has the great advantage of being objective. It is something that can really add nicely to the question, "when you lived on Jennifer Street, did you treat for cockroaches?" which is arguably influenced by the fact that six months ago, you were diagnosed with non-Hodgkin's lymphoma, whereas what is in your vacuum cleaner bag cannot have been influenced by your knowledge that you recently were diagnosed with cancer. Some of the pesticides are not stored in the dust, some are available in the blood, so for the people who are also willing to give me blood, I can also use that. It means that there is a tremendously challenging analysis when you try to combine measures that came from the water, the dust, the blood, and the personal questionnaire; and that is why I cannot get studies out in—I cannot get an

important analysis out in—a period measured in weeks. It takes longer than weeks to do it.

This kind of work too, I think, is an area where we can learn a lot from each other. Those measures, at a moment in time, are intrinsically ecologic, but especially for chronic disease, or any disease where it is probably the accumulation of many years of exposure that matters; when people move around, you lose the ability to just say “I think I will just take the value of air pollution in the greater Washington area, and that will constitute my main measurement of the important exposure.” The analytic possibilities for combining things that are *truly* ecologic do truly cover the whole population and the things that are individual. This, I think, is a developing area, it is not one where most practicing epidemiologists are yet exploring in a concrete way, but I see that coming, and I see that being another area where all of the population sciences will work together sensibly to combine data that are ecologic.

What I think will push us there fastest is GIS. I think the Geographic Information Systems are amazing. You can—with location—link to lots and lots of databases. In our lymphoma study, we have actually found that it is better to take a little GPS and at each of the 2,400 homes that we go to we take a little reading, which turns out to be rather more accurate than just writing down that I live at 4646 Langdon Lane. This very inexpensively links us to volumes and volumes of environmental data.

I think that perhaps the part of epidemiology that could be demography, in fact a lot of people who practice in this area are probably trained in demography, is what we call *descriptive epidemiology*, meaning the study of person, place, or time. We are staying at a more aggregate level and depending on our questions.

Within the field of epidemiology, the descriptive epidemiologists may feel underappreciated. They share the demographer’s skepticism about data collected on less than 100 percent of the people. They feel that the core of analytic training should stress standardization, survival, and life-table methods. The subspecialty of descriptive epidemiology has become small because it has already told us a lot of what there is to tell. Either it will change radically, perhaps with the use of geographical information systems (GIS), or by interacting more with the other social sciences, or it will be a minor part of training on the way to doing analytic epidemiology.

I make a very modest proposal for a little cross-training. It would be easy to inject a little bit more epidemiology into the curriculum of demographers and vice versa, and I think it would be fun and very simple for there to be courses that are co-taught by demographers and epidemiologists. The instructors would pick very key problems and go at them from the two perspectives. So that is my modest little proposal for something to do. At the larger level, I do not have a prescription. I think I will say to the demographers who may feel that they are not getting the attention they might need right now from epidemiology, that it is partly because a lot of epidemiologists do see this biologic revolution as our biggest challenge. Yes, we are worried about having a better sense of population context. Ezra is completely right, that we are moving into a new era for epidemiology, but we do not know exactly what it will be. We have three big problems; and the problem that is crashing over us first is really biology. The pressure many epidemiologists feel is to take a quick refresher course on genetics rather than demography. Indeed we go back to the question, “Shall we incorporate these biological measures in our studies?” Absolutely; no question in my mind. These are the kinds of studies we are doing that incorporate a much stronger

case for making the biologic inferences and epidemiologic inferences we want to make, and ultimately they will be stronger for characterizing what is going on in the population. I would appreciate some questions from the group here.

Jim Koopman: That is a beautiful image about the three waves coming and the boats sitting around. However, the image that you give of the wave of biologic data somehow moving the epidemiologists away from the population because you have to get such a narrow set of population ... there is a way that biologic data can move us toward the population better as well, and that is with phylogenetic relationships. Both for the human phylogenetic relationships and also for the infectious agent, phylogenetic relationships tell us a lot about how people are connected in different ways. It is a huge amount of data and it is hard to think that for a couple of dollars you are going to be able to get a hundred thousand base pairs of information that tell you a lot about how people are connected in the population.

Patricia Hartge: That is a wonderful comment. It was instructive to me to have to learn a little bit about founder populations. The Ashkenazi study that I described was done not because of the intrinsic interest in the BRCA1 effect in that population, it was because of the happenstance that the population, therefore, had only three mutations of any prevalence; thus, the test could be done for a dollar instead of eight hundred. That made me look a little bit at which populations *are* relatively isolated and they are few. There are little pockets—an Icelandic population here.... The phylogenetic tree is vastly illuminating, and that will pull us together. The way that I worry that the biologic revolution may drive us apart is, I suppose, two-fold. Within the current culture of the discipline, there is a tendency to downplay the core theory of epidemiologic design. I am a little bit afraid that as this vast quantity of biologic data comes at epidemiologists, we will be tempted to abandon what we know is right. We will be tempted to make inferences from studies that have unacceptable response rates. We will be tempted to say, if it has a very long name and it is in the serum, we do not have to worry about confounding. Confounding is confounding; selection bias is selection bias. I think, on balance, we will remember the core principles. However, the worry part is that we may just say, it just looks so cool, let us go off and learn some biology and we will forget those fundamental principles of design and inference that have brought us to where we are today.

I think that the other problem is a statistical one: our ways of handling vast, orders-of-magnitude-larger datasets; where we know that genetic pathways intertwine but we do not know how they do; we know that there might be five genes that probably interact in how you metabolize estrogen. We do not yet have really good analytic tools that will let us embrace that elaborate amount of data, and I think we will see many wrong studies of gene-environment interaction. I am *sure* we will. That is another big challenge that I think may distract us from the question of how the disease sits in the population.

Jim Koopman: Just to carry one more comment. The third wave of statistical—increased data—analytical power. In the past we have not integrated our analytical models, models of response bias and models of the other biases, into one big model. I think in the evolution of MCMC methods for handling missing data, if we can integrate models of response bias into our analysis as well, it is not an easy and complete answer, but that wave can also help overcome that problem.

Richard Suzman: I think you have put your finger on one of the main driving forces of what is going to induce change, and that is the changing nature of datasets,

where one can now imagine analyzing datasets of ten to the eleventh, twelfth, or thirteenth powers. What we begin to see are integrated databases, in which a demographic surveillance site and longitudinal survey embed clinical studies. Within the clinical studies are blood and DNA samples. Moreover, you have experience-based sampling coming out now, where a little beeper or palm-pilot will beep randomly and say where you are, what is your context, and will give us a saliva sample to measure cortisol or well-being. Thus, you are getting these datasets with much more time variation on individuals, let alone the FMRI datasets. I think to some extent a lot of the science seems to be driven in some fields by the availability or non-availability of data, with the economists being the traditional lamppost researchers of looking, accepting what data there are, and devising methods to look for the keys under the lamppost rather than where you lost them.

The question I have is that there seems to be a very big difference between epidemiologists and demographers on sharing data. Somewhere along the line, I do not know if the gene pools divided and the sharing gene went somewhere and the hoarding protecting gene went somewhere else, but is it the shame gene that is most important because you know where some of the problems are in your data and you do not want to have that displayed publicly. However, can you explain, other than by the fact that epidemiological data are usually known by place names—Framingham, Westinghouse, Charleston, and so forth—what it is about the sharing of data that somehow has been bypassed?

Patricia Hartge: I have to make two disclaimers before I answer you substantively. First of all I came from economics through demography to epidemiology, so on a sample of one, I do not accept the genetic explanation for the way I handle data. The second is, because I work for the National Cancer Institute, I work for you—my data are in the public domain, what is mine is yours. Now, when you ask me for my data, then I will say it is very difficult for epidemiologists to completely describe everything they did. After I have written my paper, then I feel sure that it will be reviewed and reanalyzed and I have had that experience many times.

The real problem is, rather, what you were saying earlier. That it would really be a bad idea to assume that what I get in this study, where my response rates *will not* be one hundred percent or even eighty percent, that it would be simple to take a geographic database that shows pounds of pesticide sold in different markets, a census database that shows what our apartment buildings are—I am being concrete, but just so you get the general sense here—it would not be a good idea to have a simple layer cake marriage that takes U.S. data, sales data, and then my *very* expensive, *very* detailed examination of the people who are willing to give me blood, dust, and so forth. Now, even within the number of people who are willing to talk to me, some of them have not lived with their carpets long enough to give me a carpet sample that actually speaks to their own pesticide exposure. I think scientifically it is everybody's feeling that this work, if done well, is brilliant, and this work, if done badly, is *disastrous*. I think that is the impulse.

Richard Suzman: Imagine how demography would be different if Framingham were a public dataset.

Bob Hauser: I just want to follow up on this discussion of data sharing and your comments earlier about where IRBs are or are not, and the wave of data and computer availability. I think *that* wave and the one that Richard Suzman is attempting to

encourage are both under really serious threat right now. The National Bioethics Advisory Commission issued a report about six weeks ago—with the comment period ending on February 17th—which included in its recommendations, among other things, the following three elements: first, that data are identifiable if any identifying link exists anywhere; second, that there are group as well as individual or family interests in disclosure; and third, that there should be no waiver of informed consent unless it is possible to contact the individuals whose information is being used in a study at some future date. Put those three things together and you have an end to public data and to much of scientific work as we understand it, I believe, in both of our disciplines.

Patricia Hartge: It is an important comment and I have a feeling that this is something we are all going to have to spend a lot of time kind of educating people about and reminding them that the great threats to their privacy are not from us, they are not from research.

Doug Ewbank: I want to address the issue of population base, and I think demographers do get frustrated with some epidemiologic studies and that sometimes it is the way they are designed, sometimes it is the way they are published. However, I do not think we have been very good ourselves at thinking through exactly what it is and why it is. I think the perception, the simple perception, is that the issue is one of representativeness. Ashkenazis are not representative in some sense. I do not fundamentally think that is the main problem or the main difference. I think it goes back to this: you started off by saying that as epidemiologists you are interested in biological mechanisms, which are an individual level of analysis, and that is quite appropriate. However, the population level, from a demographic perspective, involves understanding a phenomenon that *only* occurs at the population level, or may be very different at the population level than at an individual level. For example, the incidence rates changing with age may change very differently in a population than in individuals. You can get some handle on that, for example, looking at Ashkenazi Jews and asking how the importance of a BRCA mutation changes with age since epidemiologists are generally interested in demonstrating that this is important, that it *is* a risk factor, that there is probably a biologic mechanism. Often they do not publish things in a way that allows us to go beyond that and ask, at the population level, what might the population dynamics be?

Patricia Hartge: I would argue, actually, it is when we *do* try to pop it up to the population level that we usually make worse mistakes. If you have seen the number of population-attributable risk calculations, and they are commonly what goes in the media report of a study, they will say, “thirty-seven percent of bladder cancer appears to be attributable to...” and the epidemiologists will wince because the factor might be coffee-drinking, which has been extensively studied, and shown not to be related, and all that has happened is that in one study someone has done that arithmetic.

I think the question about representativeness is key and I agree with you that this is probably not why demographers wince, but I should say it. The Ashkenazi study is a good example. Since that time, many other studies of other designs have confirmed that our penetrance rates were correct; that is, the biology turns out to be from the gene, not just the Ashkenazi population. We were lucky to be able to study this population because it made the work feasible, but it turns out that it is biologically

representative. It was not essential to say, "it is true of these people whom I can enumerate in this area." It essentially had much wider representativeness.

The individual level and the fact that there are some variables that are intrinsically not individual—yes, that is really hard. The work that is beginning to distinguish between how wealthy is the *person* and how wealthy is the *neighborhood*. Do you know that sort of work? That begins to get there, and social epidemiologists are willing to go there, but there is not a lot of it; I think it is an area for a lot of development.

The other thing I wanted to say is that the reason we know that we are not doing everything completely wrong, is cross validation. I mentioned purposely the three studies that I am working on because one is a case control study, and I select according to whether or not you have the disease; one is a cohort study: you come in and I follow you over time and I see what happens; and one is a cross-sectional survey. Now one of the ways that I know that the practice of epidemiology must have some merit is that we get confirmation from different kinds of studies in very different populations of the same underlying biologic phenomena. That is, why, yes, representativeness is valuable, and certainly, if you are going to try to take your epidemiology into public health, which is what Doug Weed wants us to do, it is what Ezra Susser wants us to do, then, yes, you had better be more attuned to it. However, the bread and butter is the biologic representativeness issue.

Doug Ewbank: I am suggesting that we are asking different kinds of questions often and that what you are doing actually can help us to answer the kinds of questions we are interested in if it is presented properly.

Patricia Hartge: That is so true, that there is a certain room here for simply communicating, translating, saying, okay, now he just spoke—I do this frequently in the office with both geneticists and statisticians. There is a tremendous amount of simultaneous translation that now happens in cancer epidemiology because we are also multidisciplinary. Thus, if you speak a little statistics, you do not have to *be* one, but then you can say to the other person on the team "What he just said was ...". I think we need a little bit more of that. I could not agree more.

Doug Ewbank: Let me say that I have been working with ApoE, which is unusual because there is such a huge, phenomenal amount that has been published on it.

Patricia Hartge: Does everyone know this is the gene related to Alzheimer's and heart disease?

Doug Ewbank: ...and is therefore related to overall mortality in a big way. However, there are so many genes that you have to investigate, so many new enzymes, and proteins that a large part of what epidemiology is doing now is just trying to sort through all of these and figure out which ones seem to be interesting or seem to be important. The amount of research on ApoE has gotten to the point where I have seen eighty articles that say four are associated with Alzheimer's, and another one is not going to tell me anything and they are at a stage where they can address different kinds of questions, and have either failed or chosen not to move on to ask some very different population dynamic kinds of questions, which are what the demographers are interested in.

Patricia Hartge: It will probably be problem-driven and the intrinsic nature of the problem is that people from different disciplines will cohere around it. So maybe the best thing we do today is have coffee together.

Ezra Susser: I think that the genomic revolution, whatever we call it, is going to take us in some unanticipated directions. One of them is genetic lineage, and looking

at the global picture in that way, historically, as well as in the present; but the other is in the area of infectious disease. A lot of the research on infectious disease is now sometimes hard to distinguish from molecular biology or genomics—detection of new pathogens and so on. In terms of the comment that epidemiologists are going to start with the refresher course on genetics, I think they are, but mostly so that they can be multilingual, so they can talk with the geneticists. Because as soon as you start interacting with geneticists or people in infectious disease who are geneticists looking for new pathogens—if you would like to think of it like that—the questions that they start asking you generally have to do with population dynamics. They know how to do the genetics but they want to ask you, “how do we think about these new pathogens” in the context of populations and where we should find them. Thus, I find, at least in my own experience, that the interaction with the molecular biologists is leading me back to demographic questions and demography because that is what they are calling on me to do. So, what do you think?

Patricia Hartge: I think that is true. I think that at our investigator’s retreat, I now go and visit the posters on genetics and phylogenetics that I did not before. I guess I will say about my little fishing fleets, that I am not sure exactly when all three waves crash, but we may be actually much closer together than we think now.