

Bigger, Better, Sooner—Scaling Up for Success

Michael J. Thun¹, Robert N. Hoover², and David J. Hunter^{3,4}

Abstract

Over the last twenty years, the field of epidemiology has seen a rapidly increasing interest in, and need for, addressing low-level risks, interactions as well as main effects, and simultaneous assessment of vast numbers of biomarkers. Multiple examples over this time have shown the necessity for very large, high-quality individual studies (e.g., biobanks) or consortia of studies for these efforts to be successful. The need for this will continue to increase in the foreseeable future. It will also be important to analyze and publish aggregated data much earlier in the discovery process than typical for past efforts. *Cancer Epidemiol Biomarkers Prev*; 21(4): 571–5. ©2012 AACR.

A year ago, the Editor wrote a commentary commemorating the 20th year of publication of CEBP. He proposed a series of invited commentaries from various disciplines covered by the journal to consider major advances over the last 2 decades, and, more importantly, "what lies ahead." What followed was an impressive series from a distinguished and erudite cadre of leaders covering a broad range of issues in cancer research. In this Commentary, we chose not to address a specific research area, but rather to consider a sea-change in how epidemiologists increasingly do business—specifically, the evolving recognition over the last 20 years of the value of very large-scale studies. These include large individual studies, but more often, highly collaborative and frequently interdisciplinary consortial efforts involving multiple studies. Although such efforts have been launched sporadically over the history of cancer epidemiology, they have become an integral part of the discipline over the last 2 decades and have been critical to achieving some of the more important discoveries during this time. This development is a likely consequence of the maturation of our field to include increasing interest in identifying important, but low-level risks, in discovering interactions as well as main effects, and in taking advantage of scientific and technological advances in molecular biology allowing targeted or agnostic assessment of the role of thousands to millions

of potential biomarkers. We have chosen 3 recent examples to illustrate this trend.

Menopausal Hormone Therapy

Treatment with unopposed estrogen (ET) for the menopause was introduced in the mid-1940s; use escalated rapidly in the United States in the 1960s and 1970s in concert with an aggressive pharmaceutical marketing campaign. Despite extensive laboratory and clinical evidence linking circulating estrogens to increased breast cancer risk, before 1976, there were few attempts to assess systematically the impact of ET on risk in human populations. Indeed, the limited evidence on ET at that time, based primarily on clinical follow-up studies, was that this exposure was profoundly protective against breast, as well as all other cancers (1). Beginning in 1976, a series of epidemiologic investigations emerged that generally showed weakly positive associations of breast cancer risk with ever use of ET, with some evidence of higher risks with higher dose (2). However, not all studies showed this; even when positive, the risks with ever use were frequently not significant; significant associations of risk with increased dose sometimes involved differing measures of dose across studies; extent of control for confounding varied widely; and often overall associations were driven by subgroups for which the associations were not consistent across studies.

In 1991, a meta-analysis of 16 studies sought to address these apparent inconsistencies by increasing statistical power (3). The meta-analysis confirmed that duration of use was positively associated with breast cancer risk among ever-users (most early studies did not distinguish current from former use). Women with < 5 years of use had no increase in risk compared with never users, whereas for those with ≥15 years of use the relative risk (RR) was 1.3 (95% CI: 1.2–1.6). Although the meta-analysis was important in establishing the general relationship, it had limited ability to assess a broad range of exposure metrics, covariates, and their joint effects.

Authors' Affiliations: ¹Epidemiology Research Program, American Cancer Society, Atlanta, Georgia; ²Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland; ³Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health; and ⁴Department of Medicine, Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts

Corresponding Author: Robert N. Hoover, 6120 Executive Blvd, EPS/8094, Bethesda, MD 20892. Phone: 301-496-3004; Fax: 301-402-2623; E-mail: hoover@mail.nih.gov

doi: 10.1158/1055-9965.EPI-12-0191

©2012 American Association for Cancer Research.

In 1997, the Collaborative Group on Hormonal Factors in Breast Cancer compiled individual-level data from 51 studies covering >50,000 women with breast cancer—and >100,000 without—for a pooled analysis of many of these issues (4). By this time, the use of a combination therapy consisting of an estrogen and a progestin (EPT) in women with an intact uterus was becoming increasingly popular. However, in the studies included in this pooled analysis, only 12% of the women had used this regimen, so the pooled analyses essentially reflected the impact of ET. The large sample size of this analysis provided the first opportunity to assess the separate contributions of correlated patterns of use. Specifically, the central importance of current or recent use became clear, with the highest excess risks occurring in current users. The association weakened rapidly after cessation of use, such that no excess risk was seen among former users who had stopped 4 to 5 years earlier. Among current and recent users, no increase in risk was seen with <5 years of use. The RR estimate for 5 to 9 years of use was 1.31, rising to 1.56 for ≥ 15 years of use. Further analyses that stratified on 14 breast cancer risk factors, found no indication of effect modification except for the 2 highly correlated factors of weight and body mass index (BMI). Whereas no significant excess risks were seen among overweight and obese women, for normal weight and lean women, the risks for current or recent users of ≥ 5 years was 1.5.

As treatment with combination hormones became increasingly common, first in Europe, and later in the United States, the results of early studies of EPT and breast cancer were also inconsistent. A report in 1989 suggested, based on small numbers, that the risk of breast cancer might be substantially greater for use of EPT versus ET (5). Over the ensuing decade, a series of small-scale studies produced mostly conflicting results. Beginning in 1999 (6), a number of studies with larger sample sizes of EPT users reported reasonably consistent associations between EPT use and breast cancer. These associations were discernable in individual studies because they were stronger than those with ET use and because the hypothesis had been focused by the 1997 pooled analysis. Specifically, the subgroup of greatest interest was defined *a-priori* as lean or normal weight women with current or recent use. Here, the association with EPT was consistent with, but stronger than, that with ET. Subsequently, the Women's Health Initiative (WHI) confirmed the findings from observational studies in a large randomized clinical trial (7). Initially, the WHI did not seem to replicate the association with ET. However, once duration of use, adiposity of study subjects, and, in particular, the interval between menopause and beginning ET were taken into account, much of the initially apparent differences were resolved (8–10).

Thus, sixty years since the introduction of HT, and >25 years and 100 studies since the first epidemiologic evidence suggesting that it might be a cause of breast cancer, we had a reasonably coherent understanding of this

association. It was also obvious that if initial concerns had been followed with adequately powered studies, we would have known this decades earlier.

Excess body fat

Evidence that increased body fat and/or excessive caloric intake might increase the risk of certain cancers first emerged in the early 20th century, decades before the development of chronic disease epidemiology. Experimental studies by Rous (11), Tannenbaum (12), and others showed that caloric restriction of rodents inhibits the spontaneous development of certain cancers and the growth of transplantable tumors. Also in the early 1900s, large actuarial studies conducted by life insurance companies (13–15) observed that policy holders in the upper categories of weight adjusted for height had higher death rates from cancer compared with men and women at average body weight. The insurance company results were relatively inaccessible to researchers in other fields, however, since they were published in monographs for the American Actuarial Society rather than in the general scientific literature. Neither the animal studies nor the actuarial data could determine whether the mechanism involved body fat, energy balance, or intake of some specific source of energy such as fat.

Several case series and one hospital-based case-control study (16) published from the 1940s through the mid-1960s implied strong associations between obesity and endometrial cancer, whereas case-control studies of breast and colorectal cancer produced conflicting results. No large study systematically examined the relationship between increased body weight and multiple cancer sites until the publication in 1979 by Lew and Garfinkel (17). This analysis, based on follow-up of 750,000 participants in the ACS Cancer Prevention Study-I from 1959 to 1972, confirmed that overall cancer mortality was increased in persons $\geq 40\%$ above average body weight. The excess risk chiefly involved cancers of the colon and rectum in men and endometrium, biliary tract, breast, cervix, and ovary in women.

The pace of epidemiologic research on excess adiposity in relation to cancer accelerated during the 1980s and early 1990s, yet much of the evidence remained inconclusive. Sample size limitations required investigators to combine what WHO now defines as underweight (BMI < 18 kg/m²) with "normal weight" (BMI 18–24.9 kg/m²) and to group "overweight" (BMI: 25–29.9 kg/m²) with "obese" (BMI ≥ 30 kg/m²). This diluted the contrast between "obese" and "normal" weight. The resultant uncertainty is reflected in both the 1982 and 1996 editions of the Schottenfeld and Fraumeni text on Cancer Epidemiology and Prevention, which found scientific consensus emerging only for the associations of body weight with endometrial (18, 19) and biliary tract cancer (20).

The dramatic increase in the prevalence of obesity that began in the United States in the 1970s made it progressively easier to study the relationship of BMI to specific cancer sites. Larger studies were able to show that some of

the inconsistent results observed in earlier studies were actually reproducible characteristics of the relationship. The fact that obesity increased breast cancer risk in postmenopausal women but decreased risk before menopause, and that it was more strongly associated with colon cancer in men than women were biological realities rather than flaws in the evidence for causality.

The first true scientific consensus review of the topic by the International Agency for Research on Cancer (IARC) in 2002 (21) expanded the list of cancer sites for which the evidence was considered sufficient (specifically that the avoidance of weight gain would reduce risk) to include cancers of the colon, breast (postmenopausal), kidney (renal cell), and esophagus (adenocarcinoma), as well as endometrial cancer. The evidence linking "body fatness" to gall bladder cancer was described as "probable," as was the evidence for abdominal fatness and pancreatic cancer.

Publication of the landmark study by Calle and colleagues (22) in 2003 further underscored the magnitude of the problem. With 900,000 study subjects, 16 years of follow-up and high prevalence of overweight and obesity, it could examine a broader array of cancer sites than previous studies. In addition to the sites for which the evidence was designated as "sufficient" by IARC, the study found statistically significant associations with mortality from cancers of the pancreas, gallbladder, liver, cervix, and stomach in at least one sex. This article, together with the IARC report focused increasing scientific attention and research funding on the issue.

Comprehensive reviews and pooled studies published since 2005 illustrate progress in understanding the relationships between cancer and excess body fat/fat distribution. The 3rd edition of the Schottenfeld and Fraumeni text, (2006; ref. 23), and the 2007 edition of Food, Nutrition, and Physical Activity for the Prevention of Cancer (24) both devote entire chapters to this issue. In addition, a recent pooled analysis of 5 prospective studies provides reasonably strong evidence that obesity is an independent risk factor for thyroid cancer (25). Among the factors contributing to the overall progress are larger studies, more mature cohort studies, meta- and pooled analyses, and populations that represent a wide range of excess body fat. These allow assessment of lower level risks, a wider range of exposure at high levels of adiposity, more precise categorization of the amount, timing, and anatomic distribution of excess body fat, variations in the association for biologically meaningful tumor subgroups (defined by histology, location, and time of onset) and effect modification by covariates such as menopausal hormone therapy.

Genetic Susceptibility Research

Perhaps no other field of Cancer Epidemiology has undergone more change than genetic epidemiology. Twenty years ago, the field was dominated by family-based studies, sprinkled with some highly specific designs such as twin studies. Genetic epidemiologists

made inference largely from patterns of cancer aggregation in families. The arrival of the technical capacity to carry out genome-wide linkage analyses in the 1980's changed much of the focus to establishing the chromosomal localization of single-gene Mendelian disorders, an activity usually led by geneticists. The ability to genotype specific genetic variants using techniques such as RFLP analysis attracted epidemiologists using conventional population-based (case-control, nested case-control) studies to compare the prevalence of specific variants in case series of specific cancers and controls—the "candidate gene" approach.

In many ways, the late 1990's and early 21st century was a "lost decade" as we pursued candidate hypotheses, usually to extinction. Only a tiny fraction of genetic associations published (many of them in CEBP) were ever replicated (26). An even smaller proportion of claims for gene-environment interaction have been reproduced. In search for reasons for this dismal lack of success, "population stratification" (confounding by ancestry) in studies of unrelated individuals was proposed to be the culprit, and family-based designs were proposed as the panacea (27). However, subsequent developments suggest that population stratification is rarely a major source of confounding in studies in which some effort is made to match or control for self-reported ancestry, and the real problem was that we were very poor at picking candidate genes, and quite ignorant of the types of genetic variants that can influence gene function.

A timely coincidence of progress in establishing a database of human genetic variation and the development of "SNP chips" able to genotype a single human DNA sample at hundreds of thousands of variant loci, transformed this picture over the last 5 years (28). Suddenly, it was possible to hunt for cancer-associated variants by agnostically scanning the whole genome, and letting statistical association drive discovery of true positive loci (29). These genome-wide association studies (GWAS) have led to the discovery of >100 new robust genetic associations for >2 dozen specific cancers (30), all in the space of a few years. A glass half empty view of this is that most of these associations have low RRs, that for most loci we still do not know the causal variant, and that for most cancers the variants discovered only account for a small fraction of the estimated heritability. The glass half full view is that more new "risk factors" for cancer have been discovered using this approach in the last 5 years than in the previous several decades, that SNP-based risk prediction scores are at least as good for some cancers as those based on nongenetic predictors (31), and that newer studies are increasing the proportion of heritability explained.

What have these studies taught us about Cancer Epidemiology? From a mechanistic perspective, these studies show how little we still understand about the complex biology of most human cancers. Most of the genetic variants discovered are not in any gene previously hypothesized to be associated with cancer (another reason

for the failure of the candidate gene approach). Many of the variants are in intergenic regions that have not yet been attached to any gene function, and presumably are involved in a much more complex process of gene regulation than previously imagined. From a disciplinary perspective, we have learned the great virtue of establishing multistudy Consortia to maximize sample size and power to detect true positive associations, and to narrow the CIs around large numbers of probably true negatives (32). Although Consortia such as the Diet and Cancer Pooling Project (33) and the Collaborative Group on Hormonal Factors in Breast Cancer (4) are long established in Epidemiology, previous Consortia have largely relied on amalgamating results from previously published individual studies, while the GWAS have proceeded direct to publication of pooled results without spending years on interim publication of underpowered individual studies.

Conclusions

Each of the above examples provides its own lessons to be learned. The common theme, however, is that many of the important contemporary questions in biology and public health can only be addressed by aggregating large amounts of high quality epidemiologic data. It is beyond the scope of this commentary to specify details of how this is done, but several general principles are noteworthy. These "big science" efforts clearly present formidable challenges to traditional practices with respect to study organization, leadership, funding, communication, recognition (including authorship), opportunities for junior investigators, and many other issues (34). To the credit of our discipline, based on scientific imperatives, we did not wait to solve these issues before launching such efforts, confident that we could develop solutions as we gained more experience. This seems to be happening, as multiple viable ways of addressing these concerns are in use or

under development (35). With respect to study design, an important distinction has emerged between genetic research, for which data from case-control (retrospective) and cohort (prospective studies) studies can be combined, and studies of gene-environment interactions or environmental exposures alone, for which prospectively collected data from cohort studies are frequently optimal. Also important are advance efforts to ensure that data collected prospectively on critical variables use a format that can be readily harmonized across studies. Finally, a trend is emerging toward direct publication of pooled results without the delay of awaiting separate publication from underpowered individual studies. This is likely to become more common in the future across a broader range of cancer epidemiology, offering as it does, the ability to establish new and robust associations in a timely and reliable fashion, rather than letting an unreliable process of literature based replication over many years to winnow out many false negative results and eventually confirm a smaller number of true positive results. True to the underlying principle of epidemiology, established by John Graunt in 1662, of "the uniformity and predictability of many important biological phenomena taken in the mass" (36), we need to continue to calibrate the size of the "mass" to the nature of the question.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: M. Thun, R.N. Hoover, D.J. Hunter.
Analysis and interpretation of data: M. Thun, R.N. Hoover, D.J. Hunter.
Writing, review, and/or revision of the manuscript: M. Thun, R.N. Hoover, D.J. Hunter.
Administrative, technical, or material support: R.N. Hoover.
Study supervision: R.N. Hoover.

Received February 14, 2012; accepted February 15, 2012; published OnlineFirst February 28, 2012.

References

- Defares JG. Is cancer preventable? *Lancet* 1971;1:135-6.
- Hormone replacement therapy and cancer: is there cause for concern? *Lancet* 1989;2:368.
- Steinberg KK, Thacker SB, Smith J, Stroup DF, Zack MM, Flanders WD, et al. A meta-analysis of the effect of estrogen replacement therapy on the risk of breast cancer. *JAMA* 1991;265:1985-90.
- Collaborative group on hormonal factors in breast cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* 1997;350:1047-59.
- Bergkvist L, Adami HO, Persson I, Hoover R, Schairer C. The risk of breast cancer after estrogen and estrogen-progestin replacement. *N Engl J Med* 1989;321:293-7.
- Persson I, Weiderpass E, Bergkvist L, Bergström R, Schairer C. Risks of breast and endometrial cancer after estrogen and estrogen-progestin replacement therapy. *Cancer Causes Control* 1999; 10:253-60.
- Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321-33.
- Prentice RL, Chlebowski RT, Stefanick ML, Manson JE, Pettinger M, Hendrix SLA, et al. Estrogen plus progestin therapy and breast cancer in recently postmenopausal women. *Am J Epidemiol* 2008;167: 1207-16.
- Beral V, Reeves G, Bull DG, Green J for the Million Women Study Collaborators. Breast cancer risk in relation to the interval between menopause and starting hormone therapy. *J Natl Cancer Inst* 2011; 103:296-305.
- Chlebowski RT, Anderson GL. The influence of time from menopause and mammography on hormone therapy-related breast cancer risk assessment. *J Natl Cancer Inst* 2011;103:284-5.
- Rous P. The influence of diet on transplanted and spontaneous tumors. *J Exp Med* 1914;20:433-51.
- Tannenbaum A. Relationship of body weight to cancer incidence. *Arch Pathol* 1940;30:509-17.
- Directors of the Actuarial Society of America and Association of Life Insurance. *Medico-actuarial mortality investigation*. New York: Spectator Company; 1913.

14. Dublin L. Some aspects of cancer mortality. *Proc Am Life Insur Med Directors of America* 1929;15:402.
15. Society of Actuaries. *Build and blood pressure study*. Chicago; 1959.
16. Wynder EL, Escher GC, Mantel N. An epidemiological investigation of cancer of the endometrium. *Cancer* 1966;19:489–520.
17. Lew EA, Garfinkel L. Variations in mortality by weight among 750,000 men and women. *J Chronic Dis* 1979;32:563–76.
18. Armstrong BK, McMichael AJ, MacLennan R. Diet. In: Schottenfeld D, Fraumeni JP, editors. *Cancer epidemiology and prevention*, 1st ed. Philadelphia, PA: W.B. Saunders Company; 1982.
19. Willett W. Diet and nutrition. In: Schottenfeld D, Fraumeni JF, editors. *Cancer epidemiology and prevention*, 2nd ed. New York, NY: Oxford University Press Inc.; 1996.
20. Fraumeni JF, DeVesa S, McLaughlin JK, Stanford JL. Biliary tract cancer. In: Schottenfeld D, Fraumeni J, editors. *Cancer epidemiology and prevention*, 1st ed. Philadelphia, PA: W.B. Saunders Company; 1982.
21. IARC. *IARC handbooks of cancer prevention. Weight control and physical activity*. Lyon: International Agency for Research on Cancer; 2002.
22. Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med* 2003;348:1625–38.
23. Ballard-Barbash R, Friedenreich C, Slattery M, Thune I. Obesity and body composition. In: Schottenfeld D, Fraumeni J, editors. *Cancer epidemiology and prevention*, 3rd ed. New York, New York, Oxford University Press Inc.; 2006.
24. World Cancer Research Fund & American Institute for Cancer Research. *Body composition, growth, and development. In: food, nutrition, physical activity, and the prevention of cancer: a global perspective*. Washington, DC: AICR; 2007.
25. Kitahara CM, Platz EA, Beane-Freeman LE, Hsing AW, Linet MS, Park Y. Obesity and thyroid cancer risk among U.S. men and women: a pooled analysis of five prospective studies. *Cancer Epidemiol Biomarkers Prev* 2011;20:464–72.
26. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.
27. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037–48.
28. Hunter DJ, Thomas G, Hoover RN, Chanock SJ. Scanning the horizon: what is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention? *Cancer Causes Control* 2007;18:479–84.
29. Hunter DJ, Chanock SJ. Genome-wide association studies and "the art of the soluble". *J Natl Cancer Inst* 2010;102:836–7.
30. Chung CC, Chanock SJ. Current status of genome-wide association studies in cancer. *Hum Genet* 2011;130:59–78.
31. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* 2010;362:986–93.
32. Beckmann L, Hüsing A, Setiawan VW, Amiano P, Clavel-Chapelon F, Chanock SJ, et al. Comprehensive analysis of hormone and genetic variation in 36 genes related to steroid hormone metabolism in pre- and postmenopausal women from the breast and prostate cancer cohort consortium (BPC3). *J Clin Endocrinol Metab* 2011;96:e360–7.
33. Smith-Warner SA, Spiegelman D, Ritz J, Albanes D, Beeson WL, Bernstein L, et al. Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol* 2006;163:1053–64.
34. Hoover RN. The evolution of epidemiologic research from cottage industry to "big" science. *Epidemiology* 2007;18:13–7.
35. Ness RB. "Big" science and the little guy. *Epidemiology* 2007;18:9–12.
36. Graunt J. *Natural and political observations made upon the bills of mortality*. Wilcox WF editor. Baltimore, MD. The John Hopkins Press, 1939.