Kenneth J. Rothman, DrPH
Distinguished Fellow, Research Triangle Institute
160th Cutter Lecture
The Growing Rift Between Epidemiologists and Their Data
December 3, 2014
Cutter Lecture on Preventative Medicine

Since 1912, the Cutter Lecture on Preventive Medicine has been one of the most respected presentations, especially in the field of epidemiology. The lectures are administered by the Department of Epidemiology at the Harvard T.H. Chan School of Public Health according to the bequest from John Clarence Cutter, MD (1851 - 1909), a graduate of the Harvard Medical School. He specified that the lectures be delivered in Boston, free of charge to medical professionals and the press. Covering a range of public health topics, the lectures remain dedicated to enhancing the physical and social welfare of the world’s population.

Dr. Rothman is a Distinguished Fellow and Vice President for Epidemiology Research at the Research Triangle Institute, an independent nonprofit dedicated to improving the human condition. He is also a Professor in the Department of Epidemiology at the Boston University School of Public Health. His research interests in epidemiology span a wide range of health problems, including cancer, cardiovascular disease, neurologic disease, birth defects, injuries, environmental exposures, and drug epidemiology. He is the author of two textbooks: Modern Epidemiology (1986), now in its third edition, and Epidemiology—An Introduction (2002), now in its second edition. Rothman is the founding editor of the journal Epidemiology, Assistant Editor of the American Journal of Public Health, Editor of the American Journal of Epidemiology, and a member of the Editorial Board of the New England Journal of Medicine and the International Advisory Board of The Lancet. Among several awards, he recently received the Abraham Lilienfeld Award for Excellence in Teaching, Research, and Leadership from the American College of Epidemiology.

The Growing Rift Between Epidemiologists and Their Data

Many of today’s leading epidemiologists have abandoned the detailed data that once populated published research papers documenting important discoveries about the incidence, distribution, and control of human disease. In the 160th Cutter Lecture on Preventive Medicine at the Harvard T.H. Chan School of Public Health on December 3, Dr. Kenneth Rothman purposefully illustrated this growing problem and encouraged colleagues to restore intimacy with their data. He began by taking the audience on a historical journey spanning five centuries of epidemiological publications.

Despite being an early adopter and enthusiastic user of the leading-edge programmable calculator in the seventies—the subject of his first book on epidemiology—Rothman feared that this walk through the ages might lead some listeners to view him as a “leadite”—another aging professor who thinks pioneering epidemiologists who worked with pencil and paper rather than computers did things better in the old days. “The reason I’m worried about that,” he warned, “is because I’m about to tell you that we did do things better in the old days.”

Excursion Begins in the Seventeenth Century

The earliest publication Rothman cited and noted as “still a joy to read”—John Graunt’s Natural and Political Observations on the Bills of Mortality—dates back to 1662, when the venerable Harvard University was already nearly three decades old. Graunt, a haberdasher by day, moonlighted as an epidemiologist. Likely instigated by the sharply rising number of deaths from the plague in the late 1500s and early 1600s, his remarkably detailed weekly and annual accounts of births and deaths (and other milestones) by location included one of the first known life tables, first-time trends for and descriptions of diseases, and the first credible population estimate for the city of London. He also proffered evidence refuting the theory that the plague was spread by contagion and the notion that plague epidemics were coincident with the reign of a new king.

Two centuries later, John Snow published a study featuring similarly detailed tables of raw data depicting cholera deaths by district and water supply, the focus of his research. Using only the data presented in this nineteenth-century paper, Rothman was able to graph the onset of fatal cases of cholera in a particular section of London during a thirty-day period in 1854. Jumping ahead to the 1950s, the data in just one of many
Check out the full album [here](#).
Check out the full album [here](#).
tables embedded in a lung cancer study published by Ernest Wynder and Evarts Graham in the Journal of the American Medical Association allowed Rothman to calculate standardized risk rates for various exposure categories.

In the 1970s, new technologies began to facilitate the separation that Rothman proposes has become significant between epidemiologists and their data. However, in the midst of this troubling transition, the prize for detailed data goes to the highly respected colleague and friend of many epidemiologists—Dimitrios Trichopoulos—on the minds of all who knew him, given his passing only a few days prior to this fall’s Cutter Lecture. His 1976 paper on induced abortion and secondary infertility included exposure data for every person in the control study.

These epidemiological forebears and their contemporaries believed it was an inherent part of their responsibility to present data reflecting a level of detail that granted their readers easy access to all the information they needed to conduct their own analyses, evaluate findings, and draw conclusions using only the original study as a source.

**Building the Case for Data Intimacy**

Perhaps Rothman’s historical tour convinced most lecture attendees of the merits of publishing papers with complete and accurate data, if only because this approach allows others to verify or disprove research findings. That may be reason enough for full data disclosure. By the time Rothman shifted the focus to articles appearing in twenty-first century academic journals, many in the audience likely were ready to support his quest to restate the commitment to hard data that had been slowly sacrificed over the last several decades.

A 2000 New England Journal of Medicine report on asthma and inhaled corticosteroids features a suspiciously smooth and consistent curve graph that Rothman says was clearly generated by a logistical model, demonstrating the growing abyss between raw data and published findings. Referring to a more recent retrospective cohort study on drugs and mortality, Rothman labeled the identical benign entries in the “P value” column of one table as “beneath comment.” He went on to explain that many of today’s published papers present minimal information, often restricted to a comparison of baseline values, exposure distributions, and impact estimates derived from regression models rather than actual data.

**Sparse Data Yield Faulty Conclusions**

Next, Rothman took a deep dive into a 1992 National Institute of Occupational Safety and Health (NIOSH) retrospective cohort mortality study examining lung cancer cases in workers exposed to beryllium at seven US processing plants. He and others have been compelled to scrutinize the historical data in this paper since no new data is forthcoming regarding this known human carcinogen. The facilities have been cleaned up and relevant exposure eliminated.

The NIOSH study investigators had plenty of data—more than five hundred cases of lung cancer stratified into sixteen age categories, fourteen time (on the job) categories, two race categories, and four exposure categories. But the problem with having so many categories (1,792 here) is that they yield very few cases in any given strata. When small numbers are averaged across numerous categories, the resulting rate ratios used to measure effects tend to be inflated.

Rothman and colleagues conducted simulations using the NIOSH cohort, then reanalyzed the NIOSH study results using a regression model, and proved that the original investigators’ key finding regarding an increased risk of lung cancer among workers in the highest exposure category was inaccurate due to sparse-data bias.

Despite their in-depth stratified analysis, the NIOSH investigators lost touch with their data and were unaware of the underlying problem associated with their approach and key findings. While Rothman maintains that stratified analysis remains a critical element of epidemiological research, he says his critical examination of the NIOSH study highlights the importance of first acquiring deep knowledge of the data and subsequently applying regression modeling techniques. Stratification alone can lead to sparse-data bias and erroneous conclusions.

We live in an increasingly digital world in which many of us may often feel more connected to our devices than each other. In addition to being concerned about this human disconnect, Rothman suggests that epidemiologists consider whether their reliance on emerging technologies and new methodologies contributes to a lack of critical intimacy with their data. Reconnecting to data will “keep us honest,” says Rothman, and allow publication readers to validate or refute results. And that, after all, is what propels the dynamic cycle of discovery.
How to Subscribe

Subscriptions to Epicenter, the newsletter of Harvard’s T.H. Chan School of Public Health Department of Epidemiology, are available to you at no charge in print or PDF form. To subscribe, send your email or mailing address to:

Eric DiGiovanni - edigiova@hsph.harvard.edu

(Ideas for submission or comments are always welcomed)

Contact The Department of Epidemiology

Harvard T.H. Chan School of Public Health
Department of Epidemiology
677 Huntington Ave, Kresge 902
Boston, MA 02115
hsph.harvard.edu/epidemiology

Tel: 617-432-1050
Editorial Board: Eric DiGiovanni & David Vaughn
Writer: Stacy Moran