Big Data Equals Big Challenges for Prostate Cancer

The term "big data" is a favorite mantra of health care today. Big data sets are starting to drive much of what is done in medicine including directing research, drug development, clinical pathways, insurance coverage and public opinion. The official definition of "big data" in health care is subject to interpretation by different sources. One dictionary defines big data as "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges." According to another definition, health care data are considered "big" if it represents many more subjects than a typical randomized clinical trial (tens of thousands or more) and/or includes a broad range (hundreds or more) of clinically relevant patient and provider characteristics.[1]

But big data sets while large and typically considered robust, are not infallible. Two recent high profile prostate cancer data sets raise concerns on how erroneous data may harm progress in this disease.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute contains over 9 million entries spanning from 1973 to 2013, the most recent data available. While some might argue this may not fit the purist's modern definition of big health care data, any mere mortal would likely consider 9 million entries "big" data. While not all SEER cases are prostate cancer, a significant portion of this information relates to prostate cancer directly or as a comparator to other cancers. One concept everyone can agree on is that any dataset relies on the accuracy of the data entered to make sound conclusions.

In late 2014 auditors identified a potential problem with PSA data entered in several years of the SEER database. The error related to the simple misalignment of a decimal point in how the PSA data were entered. Initially almost 18% of the PSA data were felt to have been miscoded. The audit is ongoing and while the error rate may prove not to be as high as first thought, the damage is done as PSA data have been pulled from the 2014 SEER update. More importantly, confidence in this well regarded large and mature dataset has been eroded.

The mere suggestion of inaccuracies in the revered SEER database sent shock waves across the academic and research communities. Many well known researchers expressed concerns that previously reported studies, particularly in the controversial area of routine screening for prostate cancer, may have been flawed if they relied on this erroneous SEER PSA data.

Speaking of the prostate cancer screening controversy, much of the recent debate can be traced to the findings of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. The U.S. Preventive Services Task Force (USPSTF) relied heavily on data from this large trial to issue their 2012 recommendation not to screen for prostate cancer. With approximately 76,000 men, PLCO is the largest US prostate cancer screening trial. Men were randomly assigned to annual PSA screening or standard care. The results showed no difference in prostate cancer mortality whether screened or not. However, a recent NEJM re-analysis indicated that up to 90% of men in the control arm had at least 1 PSA test before or during the trial resulting in contamination of the control group.[2] With the lack of a true control group this "big" PLCO trial seems to harbor erroneous prostate cancer data as well.

Modern medicine is poised to continue collecting massive amounts of information on all aspects of health care. Mining big data for research and health care policy determinations have been with us for some time and will assume an even greater role in the future. Bigger usually means better but as these two prostate cancer experiences demonstrate, that might not always be true.

Leonard G. Gomella, MD
Thomas Jefferson University
Philadelphia, PA

1. Frakt AB, Pizer SD. The promise and perils of big data in healthcare. *Am J Manag Care* 2016;22(2)98-99.
2. Shoag JE, Mittal S, Hu JC. Reevaluating PSA testing rates in the PLCO trial. *N Engl J Med* 2016; 374(18):1795-1796.