Annals of Surgical Oncology **15**(2):395–396
DOI: 10.1245/s10434-007-9657-4

Secondary Analyses of Large Population-Based Data Sets: Issues of Quality, Standards, and Understanding

Clifford Y. Ko, MD, Janak Parikh, MD, and David Zingmond, MD

Center of Surgical Outcomes and Quality, University of California, Los Angeles, Los Angeles, CA, USA

Nathan and Pawlik have written an important commentary regarding secondary analyses of large population-based data sets[1] for this issue of the *Annals of Surgical Oncology*. The commentary highlights not only some of the important issues that *investigators* need to address when performing such studies, but also some of the issues that the *reder* needs to understand when evaluating these types of studies. Each of the issues raised by the authors is important and could be the focus of a detailed dissertation; we would like to build on their commentary and discuss five items/ideas as the field moves forward.

1. Investigators: As a starting point, the important bottom-line message of the commentary is that investigators should be aware of the inherent limitations of large population-based data set analyses. While it is true that many of these data sets are relatively simple to obtain and inexpensive—coupled with the fact that performing analyses is becoming "easier" with the availability of menu-driven statistical software packages, it remains essential that rigorous, methodologically sound studies are performed. Similar to surgery, a little knowledge can be a dangerous thing—and as a start for performing these analyses, more than a rudimentary knowledge of statistics and epidemiology is needed. Working with investigators/collaborators experienced and trained in these areas is prudent.
2. Publishing Standards: It may be useful to have guidelines or standards for these types of studies when publishing. This would include, at the very least, a thorough discussion of the limitations and how they might affect the findings and implications of the study. Going further, it may be an aim for the future to have a set of criteria developed for the publication of large population-based data sets, similar to the CONSORT criteria used for reporting randomized controlled trials.[2]
3. Peer Review: Along the same lines as providing criteria for authors who publish such studies, help for the reviewers performing peer assessments of submitted manuscripts may also be beneficial. More specifically, with the increasing number of manuscripts using secondary data analyses being submitted to journals such as the *Annals of Surgical Oncology*, standards, criteria, and information about the common data sets may help to improve consistent and rigorous peer review. At the American College of Surgeons, the Committee on Trauma has its registry data set, the National Trauma Data Bank. Those in charge of the database have developed a "guideline/description" that not only explains the data set and its content, but also highlights the known and potential limitations to help peer review studies for journals using this database. The same type of paper is currently being developed for the National Cancer Data Base. Similar papers that are specific to the common data sets may be developed and made available to reviewers. Additionally, a general "secondary data analysis" guideline could be

Received September 15, 2007; accepted September 17, 2007; published online: December 11, 2007.
Address correspondence and reprint requests to: Clifford Y. Ko, MD; E-mail: cko@mednet.ucla.edu

developed that highlights many of the points addressed in the commentary by Nathan and Pawlik. Again, as these submissions are likely to increase, some type of standard will probably help to improve peer review and the consequent quality of published studies.

4. Quality and Appropriateness of Care: Currently in our health-care system, we are being increasingly regulated by performance measurement. In this regard, the population-based data sets are potentially very helpful and robust tools for examining and identifying problematic areas of quality of care. They allow us to obtain, for example, rates of procedures, concordance to various performance measures, and rates of outcomes, among other things. Concomitantly, these population-based data sets also allow us to describe and study some of the factors associated with variation of these metrics/rates, such as underuse of surgery, chemotherapy, etc. While the statistics of such reports may be simple, given the studies may be solely descriptive, it is essential that the investigators understand the quality and validity of the data that are being used. For example, it is important to recognize that underuse of a particular therapy can be actual underuse, or it may simply be undercoding of the therapy.
5. Randomized Controlled Trials vs Observational Studies: Observational treatment studies using large population-based data have often been touted as being too biased to contribute substantially to our knowledge base—but the jury may still be out in this regard. Certainly the randomized controlled trial is considered the best study design for studying treatment effects; however, these trials do have their own limitations such as generalizability, they are often difficult to perform, and they are expensive. With regard to the performance of other study designs lower in the hierarchy, it is interesting to recognize that well-performed observational studies may be extremely useful in the literature. In fact, studies have found that when observational trials and randomized controlled trials on the same topic are compared, there is little evidence that the results of observational trials are different from those of randomized controlled trials. In other words, they have been found to identify the same qualitative result as well as virtually the same magnitude of treatment effect.[3,4] Further study to examine these same issues in surgical oncology is warranted.

In conclusion, data are key to investigation. Incumbent on moving forward is high-quality data and high-quality analysis while recognizing the limitations of both. Nathan and Pawlik highlight a number of important issues one needs to address while working with such data sets, and similar studies will only likely increase. Given the advances in technology and growing data availability, data will be increasingly merged—this includes potentially combining cancer registries, claims data, pharmacy files, laboratory tests, inpatient and outpatient data, physician information, etc. Even adding in some primary clinical data to these data sets is occurring. As a result, a potentially better and more complete picture may be obtained when performing such studies. As all of this happens, an appropriate understanding of the data and techniques to analyze them is paramount. As highlighted in the commentary, the understanding of issues relevant to study design, statistics, and epidemiology, among other topics, is required to responsibly perform such work.

REFERENCES

1. Nathan H, Pawlik TM. Limitations of claims and registry data in surgical oncology research. *Ann Surg Oncol* 2007. doi:10.1245/s10434-007-9658-3.
2. Altman D, Schulz KF, Moher D for the CONSORT Group et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; 134(8):663–94.
3. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000; 342(25):1887–92.
4. Benson J, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000; 342(25):1878–86.