Published in final edited form as:

J Clin Epidemiol. 2007 July ; 60(7): 648–650. doi:10.1016/j.jclinepi.2006.10.019.

Understanding secondary databases: a commentary on “Sources of bias for health state characteristics in secondary databases”

Sebastian Schneeweiss

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, USA

Imagining an ideal cohort study with prospectively collected study-specific data items is always a useful exercise when evaluating the appropriateness of secondary health care utilization databases to answer a specific study question. It will reveal the fact that many factors are not assessed at all [1], measured factors may be misclassified or missing [2], and some patients are less likely to show up in the databases than expected. We will further realize that the cost of implementing a database study is substantially less and that it will be completed before we reach tenure (which is a good proxy for reaching retirement age at my institution).

Terris et al. [3] provided a comprehensive framework for understanding the factors influencing the creation of databases that is in part built on work by Andersen [4]. For any use of secondary data, it is critical to fully understand the process that has generated a database. This includes not only official documentation and regulations but also the mundane realities of how health care encounters translate into standardized codes. Many of my colleagues and I work with only one or very few databases, because it takes some time and several studies to fully understand their potential and limitations. This involves meeting with service providers (physicians, nurses, technicians), coders and office assistants, health plan programmers, and administrators. Professionals who have worked in the system for a longer time period and can provide historical explanations for the reasons behind methods of coding and processing are particularly valuable resources. During this process of evaluating a database we are likely to encounter several of the issues described by Terris et al. that may have important consequences.

One key consequence is that coded information needs to be understood and analyzed as a set of proxies that indirectly describe the health status of patients through the lenses of health care providers and coders operating under the constraints of a specific health care system. Often, several levels of proxies are involved; for example, the health state of a patient can be assessed through the dispensing of a drug that was prescribed by a physician who had made a diagnosis in a patient who visited her practice and complained about symptoms. This chain of proxies is influenced by issues of access to care, severity of the condition, diagnostic ability of the physician, her preference for one drug over another [5], the patient's ability to pay the medication copayment [6], and the accurate recording of the dispensed medication. In this scenario, the chain of proxies leads to a reasonable interpretation that the patient indeed had a condition that was severe enough to be treated by a physician and troubled the patient enough to see the physician in the first place and eventually pay a copayment for the medication. Obviously, such interpretations are not always possible. In fact, in most cases we do not need a specific interpretation, but it is sufficient to know that on average an increasing number of medications used by a patient is just as predictive for worse health as more complex scores and algorithms [7].

The issues raised by Terris et al. are known to have fundamental implications not only for the internal validity of studies conducted with secondary data but also for their generalizability to specific patient subgroups, health care systems or jurisdictions. Depending on our area of research, we are concerned about different attributes of databases. As drug safety researchers or when studying the comparative effectiveness of treatment strategies in routine care, we are mostly concerned with the internal validity of study results. Increasingly, newer study designs and analytic techniques that help reduce residual confounding are used in database studies, including cross-over designs [8], instrumental variable methods [9], two-stage sampling designs using detailed clinical information from medical records in a subsample [10], or propensity score calibration [11]. Some of the points raised by Terris et al. may not affect the internal validity in a meaningful way, although it is difficult to make general statements. Patients who have less access to the health care system are less likely to be included in a study, which reduces external but not internal validity. Random non-differential misclassification of study outcomes (mis-, under-, or overreporting independent of the study exposure) will lead to minimally biased relative risk estimates in most situations if specificity of the coding is close to 100% [12].

Time trend analyses of longitudinal health care utilization data are very robust techniques frequently used in health services research to evaluate the effectiveness of new programs or policies. By establishing a stable baseline trend of the study outcome rate, any sudden changes in that rate in close temporal relation to the program initiation are likely attributable to the program in the absence of any co-interventions. This approach does not require detailed characterization of patients' health states, because the baseline trend is an aggregate characterization sufficient for valid inference in such designs [13]. As health services researchers, we need to not only pay attention to internal validity but also achieve an exact understanding of which population is characterized in the specific study and to which other populations the findings may be generalized. The lack of access to care by underprivileged populations critically limits the generalizability of databases based on health insurance data. Changes in coding patterns or differences in codes themselves (ICD-9 vs. ICD-10 diagnostic code, CPT vs. ICD procedure codes, ATC vs. NDC drug codes) often make it difficult to compare health services use and health outcomes over time and between jurisdictions and health plans.

Most secondary data sources, including electronic medical records, health insurance claims, or worker compensation files are longitudinal databases containing strings of information for each individual over many years. Time is one of the few highly reliable items in such databases. Miscoding of dates of medical interventions, tests, or drug dispensing is unlikely because of their clinical and financial relevance. The performance of clinical procedures involves scheduling of physicians and staff and because of their clinical importance such procedures are very likely to be recorded in the medical chart on the day they were performed. Procedures that were recorded in medical records are again very likely to be coded to insure the corresponding charges will be claimed. These charges are likely to be coded with a correct procedure date because claims are frequently rejected by insurances if dates are either missing or implausible.

Over time, coding patterns may change and complicate studies that stretch over long periods. However, time windows can also be purposefully expanded to more fully describe patients' health state. Instead of assessing a patient's status at a single office visit, during which only one or two diagnoses may be coded (resulting in an insufficient assessment of the health state), one can expand the time window to 6 months, during which there may have been several office visits, drug dispensing, and medical procedures possibly including a hospitalization, which together will provide a more complete and detailed description of the health state. Of course, measuring the instantaneous health status remains a challenge in databases, and therefore rapid

changes in disease status related to both prescribing of the study exposure (e.g., a drug or procedure) and the study outcome often lead to confounding that is difficult to adjust.

Although it is important to fully understand the limitations of databases, this is no reason for diving into an episode of acute depression. Many important research questions can be answered, though we need the wisdom to recognize which cannot. With more detailed clinical data becoming available in many secondary databases, including lab test results, imaging results, and other diagnostic information, the uses of databases in clinical epidemiology will continue to expand.

Acknowledgments

Funded by the National Institute on Aging (RO1-AG21950, RO1-AG023178) and the Agency for Healthcare Research and Quality (2 RO1-HS010881). Dr. Schneeweiss received funding as investigator of the DEcIDE Network funded by the Agency for Healthcare Research and Quality.

References

1. Schneeweiss S, Wang P. Association between SSRI use and hip fractures and the effect of residual confounding bias in claims database studies. *J Clin Psychopharmacol* 2004;24:632–8. [PubMed: 15538126]
2. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *J Clin Epidemiol* 2004;57:131–41. [PubMed: 15125622]
3. Terris DD, Litaker DG, Koroukian SM. Sources of bias for health state characteristics in secondary databases. *J Clin Epidemiol*. 2006 [in this issue].
4. Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? *J Health Soc Behav* 1995;36:1–10. [PubMed: 7738325]
5. Schneeweiss S, Glynn RJ, Avorn J, Solomon DH. A Medicare database review found that physician preferences increasingly outweighed patient characteristics as determinants of first-time prescriptions for COX-2 inhibitors. *J Clin Epidemiol* 2005;58:98–102. [PubMed: 15649677]
6. Roblin DW, Platt R, Goodman MJ, Hsu J, Nelson WW, Smith DH, et al. Effect of increased cost-sharing on oral hypoglycemic use in five managed care organizations: how much is too much? *Med Care* 2005;43:951–9. [PubMed: 16166864]
7. Schneeweiss S, Seeger J, Maclure M, Wang P, Avorn J, Glynn RJ. Performance of comorbidity scores to control for confounding in epidemiologic studies using claims data. *Am J Epidemiol* 2001;154:854–64. [PubMed: 11682368]
8. Corrao G, Zambon A, Faini S, Bagnardi V, Leoni O, Suissa S. Short-acting inhaled beta-2-agonists increased the mortality from chronic obstructive pulmonary disease in observational designs. *J Clin Epidemiol* 2005;58:92–7. [PubMed: 15649676]
9. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects in claims databases using physician-specific prescribing preferences as an instrumental variable. *Epidemiology* 2006;17:268–75. [PubMed: 16617275]
10. Collet JP, Schaubel D, Hanley J, Sharpe C, Boivin JF. Controlling confounding when studying large pharmacoepidemiologic databases: a case study of the two-stage sampling design. *Epidemiology* 1998;9:309–15. [PubMed: 9583424]
11. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Correcting effect estimates for unmeasured confounding in cohort studies with validation studies using propensity score calibration. *Am J Epidemiol* 2005;162:279–89. [PubMed: 15987725]
12. Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in observational epidemiology*. 2. New York: Oxford University Press; 1996. p. 341–390.
13. Schneeweiss S, Maclure M, Soumerai SB, Walker AM, Glynn RJ. Quasi-experimental longitudinal designs to evaluate drug benefit policy changes with low policy compliance. *J Clin Epidemiol* 2002;55:833–41. [PubMed: 12384199]