Epidemiologic Research Using Administrative Databases

Garbage In, Garbage Out

David A. Grimes, MD

Administrative databases stem from claims made for services by health care providers and institutions.[1] Simply put, they are billing systems. These databases were created for reasons other than epidemiologic research–a key limitation. Data fields commonly include only basic demographic information, drug dispensing, provider visits, and hospitalization. Examples of administrative databases often used by researchers include Medicare, Medicaid, and those of health maintenance organizations such as Kaiser Permanente.

Vital records, such as birth certificates, represent another administrative database commonly used for epidemiologic research.[2,3] Again, these data are collected for civil and legal purposes, not for research.

Research using administrative databases has important strengths and weaknesses. Sample sizes are often large, which provide power to find differences. Those enrolled may be representative of the community of interest. Recording of drug prescriptions occurs contemporaneously, which avoids the problems of imperfect recall of drug exposures and nonresponse. No informed consent is needed from patients, and using an existing database is fast and inexpensive compared with generating a new one.

The drawbacks, however, are substantial–and often insurmountable. Huge sample sizes can lead to spurious statistical associations ("mass significance"). Persons enrolled (eg, those in Medicare and Medicaid, the elderly and poor) may not be representative of the population of interest. Comorbidities (preexisting conditions) such as diabetes and hypertension are poorly recorded, which can introduce bias. Paradoxically, diabetes can be shown to be protective. In complex and lengthy hospitalizations, particularly if the patient dies, diabetes tends not to be coded; patients who do have diabetes coded thus appear to be more likely to survive their hospitalizations than those who do not.[4] The temporal sequence of events may be unclear: for example, did pneumonia lead to hospitalization, or was it a complication of hospitalization? In billing databases, recording of events is related to the probability of reimbursement.

Epidemiologic analyses with birth certificate data are popular but treacherous. This stems from the uneven quality of the data.[3] Some information, such as mother's age, parity, and insurance status, is accurate. However, other critical information, such as smoking, parental work, comorbidities, complications, procedures, and birth defects is missing or poorly coded.[3] For example, a birth certificate study of maternal smoking and birth defects[5] was futile because neither smoking nor birth defects were recorded accurately.[3]

The Danish National Patient Registry has been used extensively for epidemiologic research, including obstetric and gynecologic studies. In

Dr. Grimes is from Family Health International, Research Triangle Park, North Carolina and the Department of Obstetrics and Gynecology, UNC School of Medicine, Chapel Hill, North Carolina; e-mail: dgrimes@fhi.org.

Financial Disclosure
Dr. Grimes serves as a consultant (DSMB member) for Bayer.

ISSN: 0029-7844/10

Denmark, all persons get a unique identifier at birth and health care is provided by the government; hence, linkage studies are easy. A highly publicized study of venous thromboembolism concluded that "third-generation" oral contraceptives were more dangerous than "second-generation" pills.[6] The report acknowledged "inclusion of about 10% uncertain diagnoses."

In contrast, an independent validation of 1,100 venous thromboembolism diagnoses in this registry found gross misclassification. Only 59% of diagnoses could be confirmed by chart review.[7] Often, physicians entered a code for confirmed venous thromboembolism instead of "observation for venous thromboembolism." This misclassification likely was related to the exposure of interest ("generation" of oral contraceptive), which would bias the results. Other conditions, including rupture of the uterus, hypertension, and rheumatoid arthritis, are coded poorly in this database as well.[8]

Research using vital records should be limited to simple descriptive reports with caveats about data accuracy. Using birth certificate information for epidemiologic analyses is inappropriate because of well-documented deficiencies in information quality.[3] Similarly, epidemiologic research using adminstrative databases, such as the Danish National Patient Registry, must at a minimum validate each reported outcome by chart review[9] or by patient interview.

In recent decades, the computer science concept of "GIGO" ("garbage in, garbage out") has somehow come to mean "garbage in, gospel out."[10] When computer software tackles a large database, many accept the "computerized" output as trustworthy, regardless of the quality of the input. Sadly, no fancy statistical machinations can compensate for poor-quality data. Publications relying on unconfirmed database reports of venous thromboembolism should be ignored.[11] Trying to study obstetric and neonatal outcomes from data on birth certificates is analogous to trying to study the cause of motor vehicle accidents from data on drivers' licenses (eg, sex, height, eye color, hair color). The information available is simply inadequate. When using administrative databases for epidemiologic research, if garbage goes in, garbage (not gospel) comes out.

REFERENCES

1. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects–advantages and disadvantages. Nat Clin Pract Rheumatol 2007;3:725–32.
2. Schoendorf KC, Branum AM. The use of United States vital statistics in perinatal and obstetric research. Am J Obstet Gynecol 2006;194:911–5.
3. Northam S, Knapp TR. The reliability and validity of birth certificates. J Obstet Gynecol Neonatal Nurs 2006;35:3–12.
4. Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SL. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. Circulation 2007;115:1518–27.
5. Honein MA, Paulozzi LJ, Watkins ML. Maternal smoking and birth defects: validity of birth certificate data for effect estimation. Public Health Rep 2001;116:327–35.
6. Lidegaard O, Lokkegaard E, Svendsen AL, Agger C. Hormonal contraception and risk of venous thromboembolism: national follow-up study. BMJ 2009;339:b2890.
7. Severinsen MT, Kristensen SR, Overvad K, Dethlefsen C, Tjonneland A, Johnsen SP. Venous thromboembolism discharge diagnoses in the Danish National Patient Registry should be used with caution. J Clin Epidemiol 2010;63:223–8.
8. Pedersen M, Klarlund M, Jacobsen S, Svendsen AJ, Frisch M. Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. Eur J Epidemiol 2004;19:1097–103.
9. Schulz KF, Cates W Jr, Grimes DA, Selik RM, Tyler CW Jr. Reducing classification errors in cohort studies: the approach and a practical application. Stat Med 1983;2:25–31.
10. Ault MR. Combating the garbage-in, gospel-out syndrome. Radiation Protection Management 2004;20:26–30.
11. Shapiro S, Dinger J. Risk of venous thromboembolism among users of oral contraceptives: a review of two recently published studies. J Fam Plann Reprod Health Care 2010;36:33–8.