Part IIIb

Electronic Data Systems

11

Overview of Electronic Databases in Pharmacoepidemiology

Brian L. Strom

Rutgers Biomedical and Health Sciences, Newark, NJ, USA

Once hypotheses about drug-induced adverse effects are generated, usually from spontaneous reporting systems (see Chapter 10), techniques are needed to test these hypotheses. Usually between 500 and 3000 patients are exposed to the drug during Phase III testing, even if drug efficacy can be demonstrated with much smaller numbers of patients. Studies of this size would be expected to observe a single case of outcomes with an incidence of 1 per 1000 to 6 per 1000 (see Chapter 4). Given this context, postmarketing studies of drug effects must then generally include at least 10 000 exposed persons in a cohort study, or enroll diseased patients from a population of equivalent size for a case–control study. Given a study of this size, the upper 95% confidence limit for the incidence any event that is not identified would be 3 per 10 000 (see Chapter 4). However, prospective studies this large are expensive and difficult to perform. Yet such studies often need to be conducted quickly, to address acute and serious regulatory, commercial, and/or public health crises. For all of these reasons, the past decades have seen a growing use of electronic databases containing healthcare data, sometimes called "automated databases," as potential data sources for pharmacoepidemiologic studies.

Large electronic databases can often meet the need for a cost-effective and efficient means of conducting postmarketing surveillance studies. To meet the needs of pharmacoepidemiology, the ideal database would include records from inpatient and outpatient care, emergency care, mental health care, all laboratory and radiological tests (including pharmacogenomic tests that may not have been performed as part of clinical care), functional assessments, and all prescribed and over-the-counter medications, as well as alternative therapies. The population covered by the database would be large enough to permit discovery of rare events for the drug(s) in question, and the population would be stable over its lifetime. Although it is generally preferable for the population included in the database to be representative of the general population from which it is drawn, it may sometimes be advantageous to emphasize the more disadvantaged groups that may have been absent from premarketing testing. The drug(s) under investigation must of course be present in the formulary and must be prescribed in sufficient quantity to provide adequate power for analyses.

Other requirements of an ideal database are that all parts are easily linked by means of a

Pharmacoepidemiology, Sixth Edition. Edited by Brian L. Strom, Stephen E. Kimmel and Sean Hennessy.

patient's unique identifier, that the records are updated on a regular basis, and that the records are verifiable and reliable. The ability to conduct medical chart review to confirm outcomes is also a necessity for most studies (unless validated algorithms for the study outcome already exist), as diagnoses entered into an electronic database may include rule-out diagnoses or interim diagnoses and recurrent/chronic, as opposed to acute, events. Information on potential confounders, such as smoking and alcohol consumption, may only be available through chart review or, more consistently, through patient interviews. With appropriate permissions and confidentiality safeguards in place, access to patients is sometimes possible and useful for assessing compliance with the medication regimen as well as for obtaining biosamples or information on other factors that may relate to drug effects. Information on drugs taken intermittently for symptom relief, over-the-counter drugs, and drugs not on the formulary must also be obtained directly from the patient.

These automated databases are the focus of this section of the book. Of course, no single database is ideal for all questions. In the current chapter, we will introduce these resources, presenting some of the general principles that apply to them all. In Chapters 12–14 of this book, we will present more detailed descriptions of those databases that have been used in a substantial amount of published research, along with the strengths and weaknesses of each.

Description

So-called automated databases have been used for pharmacoepidemiologic research in North America since 1980, and are primarily administrative in origin, generated by the request for payments, or claims, for clinical services and therapies. In contrast, electronic health record databases were developed for use by researchers in Europe, and similar databases have been developed in the US more recently.

Claims and other Administrative Databases

Claims data arise from billable interactions between patients and the healthcare system. When a patient goes to a pharmacy and gets a drug dispensed, the pharmacy bills the insurance carrier for the cost of that drug, and has to identify which medication was dispensed, the milligrams per tablet, number of tablets, etc. Analogously, if a patient goes to a hospital or to a physician for medical care, the providers of care bill the insurance carrier for the cost of the medical care, and have to justify the bill with a diagnosis. If there is a common patient identification number for both the pharmacy and the medical care claims, these elements could be linked and analyzed as a longitudinal medical record.

Since drug identity and the amount of drug dispensed affect reimbursement, and because the filing of an incorrect claim about drugs dispensed is fraud, claims are often closely audited, for example by Medicaid. Indeed, there have been numerous validity checks on the drug data in claims files that showed that the drug data are of extremely high quality, such as confirming that the patient was dispensed exactly what the claim showed was dispensed, according to the pharmacy record. In fact, claims data of this type provide some of the best data on drug exposure in pharmacoepidemiology (see Chapter 37).

The quality of disease data in these databases is somewhat less perfect. If a patient is admitted to a US hospital, the hospital charges for the care and justifies that charge by assigning diagnosis codes (until recently International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] codes) and a Diagnosis-Related Group (DRG). Hospital diagnosis codes are reasonably accurate diagnoses that are used for clinical purposes, based primarily on the discharge diagnoses assigned by the patient's attending physician. (Of course, this does not guarantee that the physician's diagnosis is correct.) The amount paid by the insurer to the hospital is based on the DRG, so there is no financial incentive to provide incorrect diagnosis codes. In fact, most hospitals have mapped each set of diagnosis codes into the DRG code that generates the largest payment.

In contrast, however, outpatient diagnoses are assigned by the practitioners themselves, or by their office staff. Once again, reimbursement in the US does not usually depend on the actual diagnosis, but rather on the visit intensity during the outpatient medical encounter, and the resulting procedure codes indicate the intensity of the services provided. Thus, there is no incentive for the practitioner to provide incorrect diagnosis codes, but there is also no incentive for them to be particularly careful or complete about the diagnoses provided. For these reasons, the outpatient diagnoses are the weakest link in claims databases.

Some other databases are not made up of actual claims but derive from other administrative processes, such as data from US health maintenance organizations or other data sources. The characteristics of these data are similar in many ways to those of claims data, and they are discussed together as encounter-based databases in Chapter 12.

Electronic Health Record Databases

In contrast, electronic health record databases are a more recent development, arising out of the increasing use of computerization in medical care. Initially, computers were used in medicine primarily as a tool for literature searches. Then, they were used for billing. Now, however, there is increasing use of computers to record medical information at the point of care. In most instances, this is replacing the paper record as the primary medical record. As medical practices increasingly become electronic, this opens up a unique opportunity for pharmacoepidemiology, as larger and larger numbers of patients are available in such systems. The best-known and most widely used example of this approach is the UK Clinical Practice Research Datalink® (CPRD®), along with the newer database, The Health Improvement Network® (THIN®), both described in Chapter 13. As general practice databases, these contain primarily outpatient data. In addition, recently inpatient electronic health record databases are becoming available (Chapter 14).

Electronic health record databases have unique advantages. Important among them is that the validity of the diagnosis data in these databases is probably better than that in claims databases, as these data are being used to document medical care rather than just for billing purposes. When performing a pharmacoepidemiologic study using these databases, there is no purpose in validating the data against the actual medical record, since one is analyzing the data from the actual medical record. However, there are also unique issues one needs to be concerned about, especially the uncertain completeness of the data from other physicians and sites of care. Any given practitioner provides only a piece of the care a patient receives, and inpatient and outpatient care are unlikely to be recorded in a common medical record.

Strengths

Computerized databases have several important advantages, including their potential for providing a very large sample size. This is especially important in the field of pharmacoepidemiology, where achieving an adequate sample size is uniquely problematic. In addition, these databases are relatively inexpensive to use, especially given the available sample size, as they are by-products of existing administrative systems. Studies using these data systems do not need to incur the considerable cost of data collection, other than for those subsets of the populations for whom medical records are abstracted and/or interviews are conducted. The data can be complete; for example, for claims databases, information is available on all medical care provided for covered services, regardless of who the provider was. As indicated above, this can be a problem for electronic health record databases, especially in the US, where primary care providers often do not serve as gatekeepers to specialty care. In addition, these databases can be population based, they can include outpatient drugs and diseases, and there is no opportunity for recall and interviewer bias, as they do not rely on patient recall or interviewers to obtain their data. Another advantage is that these databases can potentially be linked to other external electronic databases (e.g., death records, maternal-child records, police accident records), to expand the capabilities and scope of research. This requires the use of common identification elements (e.g., name and date of birth) and standardized semantics to allow communication across databases.

Weaknesses

The major weakness of such data systems is the uncertain validity of diagnosis data. This is especially true for claims databases, and for outpatient data. For these databases, access to medical record data for validation purposes is usually needed. This issue is less problematic for electronic health record databases; however, the validity of medication data from electronic health record databases in the United States is less certain than pharmacy dispensing data from claims databases. The addition of laboratory results data to these resources can assist in diagnosis validity, as well.

In addition, such databases can lack information on some potential confounding variables. For example, in claims databases there are no data on date of menopause, and diagnosis-based algorithms to identify smoking and alcohol abuse may have poor sensitivity, all of which can be of great importance to selected research questions. This argues that one either needs access to patients or physician records if these contain the data in question, or one needs to be selective about the research questions that one seeks to answer through these databases, avoiding questions that require data on variables which may be important potential confounders that must be controlled for.

Another major disadvantage of administrative data is the instability of the population due to job changes, employers' changes of health plans, and changes in coverage for specific employees and their family members. The opportunity for longitudinal analyses is thereby hindered by the continual enrollment and disenrollment of plan members. Another source of population instability is when patients transfer out of the system due to death or relocation. The effect of this is an inflated list with patients no longer seeking medical care. This will invalidate calculations of patient-time in studies of disease incidence, for example, because the denominator is inflated. The challenge for the investigator is to be creative in devising strategies to guard or correct for this incomplete information in the database (e.g., by performing sensitivity analysis censoring follow-up one or two years after the patient's last recorded entry in the database). Alternatively, strategies can be adopted for selecting stable populations within a particular database and, for example, by examining patterns of prescription refills for chronically used medications and restricting the study population to include only continuously enrolled patients. Of course, the largest such data system, US Medicare, suffers much less from this problem since it covers the elderly, so people never lose eligibility. Even there, however, patients can switch between fee-for-service plans and managed care plans, and the latter may not record all healthcare which is provided (see Chapter 12).

Further, by definition, such databases only include illnesses severe enough to come to medical attention. In general, this is not a problem, since illnesses that are not serious enough to come to medical attention and yet are uncommon enough for one to seek to study them in such databases are generally of lower importance.

Some results from studies that utilize these databases may not be generalizable, for example on healthcare utilization. This is especially relevant for databases created by data from a population that is atypical in some way, such as US Medicaid data.

Finally, as noted briefly above, as an increasing number of electronic health record databases emerge in the US, to date all are problematic in that they do not include complete data on a defined population. In the US health system, unlike other countries, patients can, and often do, seek medical care from a variety of different healthcare providers at unaffiliated institutions with a nonlinked electronic health record systems. Thus, providers' electronic health records are inherently incomplete, and need to be linked to administrative data in order to be useful for quality research. This is different from the situation in, for example, the UK, where electronic health record databases are much more likely to be complete given the general practitioner gatekeeper paradigm and unique patient identifier for all healthcare services.

Particular Applications

Based on these characteristics, one can identify particular situations when these databases are uniquely useful or uniquely problematic for pharmacoepidemiologic research. These databases are useful in situations: (1) when looking for uncommon outcomes because of a large sample size; (2) when a denominator is needed to calculate incidence rates; (3) when one is studying short-term drug effects (especially when the effects require specific drug or surgical therapy that can be used as validation of the diagnosis); (4) when one is studying objective, laboratory-driven diagnoses; (5) when recall or interviewer bias could influence the association; (6) when time is limited; and (7) when the budget is limited.

Uniquely problematic situations include: (1) illnesses that do not reliably come to medical attention; (2) inpatient drug exposures that are not included in some of these databases; (3) outcomes that are poorly captured by the coding system, such as Stevens–Johnson syndrome; (4) descriptive studies, if the population studied is nonrepresentative; (5) delayed drug effects, wherein patients can lose eligibility in the interim; and (6) important confounders about which information cannot be obtained without accessing the patients, such as cigarette smoking, occupation, menarche, menopause, etc.

The Future

Given the frequent use of these data resources for pharmacoepidemiologic research in the recent past, we have already learned much about their appropriate role. As it appears that these uses will be increasing, we are likely to continue to gain more insight in the coming years, especially with the access in the US to Medicare data, and the advent in the US of the FDA's Sentinel system, exceeding 170 million individuals (see Chapter 25). However, care must be taken to ensure that all potential confounding factors of interest are available in the system or addressed in some other way, that diagnoses under study are chosen carefully, and that medical records can be obtained when needed to validate the diagnoses.

In this section of the book, Chapters 12–14, we will review the details of a number of these databases. The databases selected for review have been chosen because they have been the most widely used for published research. They are also good examples of the different types of data that are available. There are multiple others like each of them and undoubtedly many more will emerge over the ensuing years. Each has its advantages and disadvantages, but each has proven it can be useful in pharmacoepidemiologic studies.