Part IIId

Choosing a Data Source

17

Choosing among the Available Data Sources for Pharmacoepidemiology Research

Brian L. Strom

Rutgers Biomedical and Health Sciences, Newark, NJ, USA

As discussed in previous chapters, pharmacoepidemiologic studies apply the techniques of epidemiology to the content area of clinical pharmacology. Between 500 and 3000 individuals are usually studied prior to drug marketing. Most postmarketing pharmacoepidemiologic studies need to include at least 10 000 subjects, or draw from an equivalent population for a case–control study, in order to contribute sufficient new information to be worth their cost and effort (see Chapter 4). This large sample size raises logistical challenges. Chapters 10–16 presented many of the different data collection approaches and data resources that have been developed to perform pharmacoepidemiologic studies efficiently, meeting the need for these very large sample sizes. This chapter is intended to synthesize this material, to assist the reader in choosing among the available approaches.

Choosing among the Available Approaches to Pharmacoepidemiologic Studies

Once one has decided to perform a pharmacoepidemiologic study, one needs to decide which of the data collection approaches or data resources described in the earlier chapters of this book should be used. Although to some degree the choice may too often be based upon a researcher's familiarity with given data resources and/or the investigators who have been using them, it is very important to tailor the choice of pharmacoepidemiologic resource to the question to be addressed. One frequently may want to use more than one data collection strategy or resource, in parallel or in combination. If no single resource is optimal for addressing a question, it can be useful to use a number of approaches that complement each other. Indeed, this is probably the preferable approach for addressing important questions. Regardless, investigators are often left with a difficult and complex choice.

In order to explain how to choose among the available pharmacoepidemiologic data resources, it is useful to synthesize the information from the previous chapters on the relative strengths and weaknesses of each of the available pharmacoepidemiologic approaches, examining the comparative characteristics of each (see Table 17.1). One can then examine the characteristics of the research question at hand, in order to choose the pharmacoepidemiologic

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

Table 17.1 Comparative characteristics of pharmacoepidemiologic data resources.

Pharmacoepidemiologic approach	Relative size	Relative cost	Relative speed	Representativeness	Population based	Cohort studies possible	Case–control studies possible
Spontaneous reporting	++++	+	++++	++	—	—	+ (with external controls)
Health maintenance organizations/health plans	++	+++	+++	+++	++	++++	++++
Commercial insurance databases	++	+++	+++	+++	++	++++	++++
US government claims databases	+++	++	++	variable	++++	++++	++++
UK medical record databases	++	++	+++	+++	+++	++++	++++
In-hospital databases	+	++	+++	++	—	++	++
Canadian provincial databases	++	++	+++	++++	++++	++++	++++
Pharmacy-based medical record linkage systems	++	++	+++	++++	++++	++++	++++
Ad hoc studies							
Case–control surveillance	variable	+++	+	variable	—	—	++++
Prescription Event Monitoring	+++	+++	+	+++	++	++++	+ (nested)
Registries	variable	+++	+	variable	variable	+++	+++
Field studies							
Ad hoc case–control studies	as feasible	+++	+	as desired	as desired	—	++++
Ad hoc cohort studies	as feasible	++++	—	as desired	as desired	++++	++ (nested)
Randomized trials	as feasible	++++	—	—	—	++++	++ (nested)

Spontaneous reporting	+++	++	—	+++	+++	N/A
Health plans	++++	+++	++	—	++	3–15%/year
Commercial insurance databases	++++	+++	++	—	++	about 25%/year
US government claims databases	++++	+++	++	—	++	variable
UK medical record databases	+++	++++	++	—	++	nil
In-hospital databases	++++	+++	++	++++	—	nil
Canadian provincial databases	++++	+++	++	—	++	nil
Pharmacy-based medical record linkage systems	++++	+	+	—	—	nil
Ad hoc studies						
Case–control surveillance	++	++++	+++	—	+	N/A
Prescription Event Monitoring	+++	+++	++	—	+++	variable
Registries	+++	+++	++	+	variable	N/A
Field studies						
Ad hoc case–control studies	++	++++	+++	++	+	N/A
Ad hoc cohort studies	+++	+++	+++	++	++++	variable
Randomized trials	++++	+++	++++	++	++++	N/A

Note: See the chapter text for descriptions of the column headings, and previous chapters for descriptions of the data resources. N/A, not applicable.

approach best suited to addressing that question (see Table 17.2). The assessment and weights provided in this discussion and in the accompanying tables are arbitrary. They are not being represented as a consensus of the pharmacoepidemiologic community, but represent the judgment of this author alone, based on the material presented in earlier chapters of this book. Nevertheless, I think that most would agree with the general principles described, and even many of the relative ratings. My hope is that this synthesis of information, despite some of the arbitrary ratings inherent in it, will make it easier for the reader to synthesize the large amount of information presented in prior chapters.

Note that there are a number of other data sources not discussed here, some of which have been, or in the future may be, of importance to pharmacoepidemiologic research. Examples include the old Boston Collaborative Drug Surveillance data [1], MEMO [2], Pharmetrics® [3], Aetna [4], Humana [5], and many others, many reviewed in prior editions of this book. Given the wonderful proliferation of pharmacoepidemiologic data resources, we are making no attempt to include them all. Instead, we will discuss them in categories of types of data, as we did in the chapters themselves.

Comparative Characteristics of Pharmacoepidemiologic Data Resources

Table 17.1 lists each of the different pharmacoepidemiologic data resources that were described in earlier chapters, along with some of their characteristics.

The *relative size* of the database refers to the population it covers. Only spontaneous reporting systems, US Medicare, some of the pharmacy-based medical record linkage systems, and Prescription Event Monitoring in the UK cover entire countries or large fractions thereof. Of course, population databases differ considerably in size, based on the size of their underlying populations. Aggregations of Medicaid databases are the next largest, with the commercial databases approaching that. The UK electronic health record databases would be next in size, as would the health maintenance organizations (HMOs), depending on how many are included. The Canadian provincial databases again could be equivalently large, depending in part on how many are included in a study. The other data resources are generally smaller. Case–control surveillance, as formerly conducted by the Slone Epidemiology Unit, can cover a variable population, depending on the number of hospitals and metropolitan areas included in the network for a given study. The population base of registry-based case–control studies depends on the registries used for case finding. *Ad hoc* studies can be whatever size the researcher desires and can marshal resources for.

As to *relative cost*, studies that collect new data are most expensive, especially randomized trials and cohort studies, for which sample sizes generally need to be large and follow-up may need to be prolonged. In the case of randomized trials, there are additional logistical complexities. Studies that use existing data are least expensive, although their cost increases when they gather primary medical records for validation. Studies that use existing data resources to identify subjects but then collect new data about those subjects are intermediate in cost.

With regard to the *relative speed* of study completion, studies that collect new data take longer, especially randomized trials and cohort studies. Studies that use existing data are able to answer a question most quickly, although considerable additional time may be needed to obtain primary medical records for validation. Studies that use existing data resources to identify subjects but then collect new data about those subjects are intermediate in speed.

Representativeness refers to how well the subjects in the data resource represent the population at large or a more specific population of

Table 17.2 Characteristics of research questions and their impact on the choice of pharmacoepidemiologic data resources.

Pharmacoepidemiologic approach	Hypothesis generating[a]	Hypothesis strengthening[b]	Hypothesis testing[c]	Study of benefits (versus risk)	Incidence rates desired	Low incidence outcome	Low prevalence exposure
Spontaneous reporting	++++	+	—	—	—	++++	++++
Health plans	++	++++	+++	++	+++	+++	+++
Commercial insurance databases	++	++++	+++	++	+++	+++	+++
US government claims databases	++	++++	+++	++	+++	++++	++++
UK medical record databases	++	++++	+++	++	++++	+++	+++
In-hospital databases	+	++++	+++	++	+++	+	+
Canadian provincial databases	++	++++	+++	++	+++	+++	+++
Pharmacy-based medical record linkage systems	+	++	++	++	+++	+++	+++
Ad hoc studies							
Case–control surveillance	+++	+++	+++	+++	—	++++	+
Prescription Event Monitoring	++	++	+++	+++	+++	+++	+++
Registries	+	+++	+++	+++	+++	+++	+++
Field studies							
Ad hoc case–control studies	+	++	+++	+++	+	++++	+
Ad hoc cohort studies	+	++	+++	+++	++++	++	+++
Randomized trials	+	+	++++	++++	++++	+	++++

(*Continued*)

Table 17.2 (Continued)

Pharmacoepidemiologic approach	Important confounders	Drug use inpatient (versus outpatient)	Outcome does not result in hospitalization	Outcome does not result in medical attention	Outcome a delayed effect	Exposure to a new drug	Urgent question
Spontaneous reporting	—	+++	++++	+	+	++++	++++
Health plans	+++	—	+++	—	+	++	+++
Commercial insurance databases	++	—	+++	—	+	+++	+++
US government claims databases	++	—	+++	—	+ to +++	++	++
UK medical record databases	+++	—	+++	—	+++	+++	+++
In-hospital databases	++	++++	—	—	—	+++	+++
Canadian provincial databases	++	—	+++	—	+++	++	+++
Pharmacy-based medical record linkage systems	+	—	—	—	++	+++	+++
Ad hoc studies							
Case–control surveillance	+++	+	—	—	++	+	+
Prescription Event Monitoring	++	+	++++	+	+	++++	+
Registries	++	++	+	++	++	+++	+
Field studies							
Ad hoc case–control studies	+++	++++	++	—	++	+	+
Ad hoc cohort studies	+++	+++	++++	+++	+	++++	+
Randomized trials	++++	+++	++++	++++	+	++++	+

Notes: See the text of the chapter for descriptions of the column headings, and previous chapters for descriptions of the data resources.

[a] Hypothesis-generating studies are designed to raise new questions about possible unexpected drug effects, whether adverse or beneficial.

[b] Hypothesis-strengthening studies are designed to provide support for, although not definitive evidence for, existing hypotheses.

[c] Hypothesis-testing studies are designed to evaluate in detail hypotheses raised elsewhere.

interest. US Medicare, Prescription Event Monitoring in the UK, the provincial health databases in Canada, and the pharmacy-based medical record linkage systems include entire countries, provinces, or states, and so are typical populations. Spontaneous reporting systems are drawn from entire populations, but of course the selective nature of their reporting could lead to less certain representativeness. Medicaid programs are limited to the disadvantaged, and so include a population that is least representative of a general population. Analogously, randomized trials include populations limited by the various selection criteria plus their willingness to volunteer for the study. The Clinical Practice Research Datalink® (CPRD®) and The Health Improvement Network® (THIN®) use a nonrandom large subset of the total UK population, and so may be representative of the overall UK population. Health plans and commercial databases are closer to representative populations than a Medicaid population would be, although they include a largely working population and, so, include few patients of low socioeconomic status and fewer than normal elderly. Some of the remaining data collection approaches or resources are characterized in Table 17.1 as "variable," meaning their representativeness depends on which hospitals are recruited into the study. *Ad hoc* studies are listed in Table 17.1 "as desired," because they can be designed to be representative or not, as the investigator wishes.

Whether a database is *population based* refers to whether there is an identifiable population (which is not necessarily based on geography), all of whose medical care would be included in that database, regardless of the provider. This allows one to measure incidence rates of diseases, as well as being more certain that one knows of all the medical care that any given patient receives. As an example, assuming little or no out-of-plan care, the Kaiser programs are population based. One can use Kaiser data, therefore, to study medical care received in and out of the hospital, as well as diseases that may result in repeat hospitalizations. For example, one could study the impact of the treatment initially received for venous thromboembolism on the risk of subsequent disease recurrence. In contrast, hospital-based case–control studies conducted outside a closed network like Kaiser are not population based: they include only the specific hospitals that belong to the system and do not capture all healthcare services a patient may receive. Thus, a patient diagnosed with and treated for venous thromboembolism in a participating hospital could be readmitted to a different, nonparticipating hospital if the disease recurred. This recurrence would not be detected in a study using such a system. The data resources that are population based are those that use data from organized healthcare delivery or payment systems. Registry-based and *ad hoc* case–control studies can occasionally be conducted as population-based studies, if all cases in a defined geographic area are recruited into the study [6], but this is unusual (see also Chapters 3 and 16).

Whether cohort studies are possible within a particular data resource would depend upon whether individuals can be identified by whether or not they were exposed to a drug of interest. This would be true in any of the population-based systems, as well as any of the systems designed to perform cohort studies.

Whether case–control studies are possible within a given data resource depends upon whether patients can be identified by whether or not they suffered from a disease of interest. This would be true in any of the population-based systems. Data from spontaneous reporting systems can be used for case finding for case–control studies, although this has been done infrequently [7].

The *validity of the exposure data* is most certain in hospital-based settings, where one can be reasonably certain of both the identity of a drug and that the patient actually ingested it. Exposure data in spontaneous reporting systems come mostly from healthcare providers

and so are probably valid. However, one cannot be certain of patient adherence in spontaneous reporting data. Exposure data from claims data and from pharmacy-based medical record linkage systems are unbiased data recorded by pharmacies, often for billing purposes, a process that is closely audited as it impacts reimbursement. These data are likely to be accurate with regard to medication possession, although, again, one cannot assure adherence. Refill adherence, though, has been found to correlate closely with adherence measured using microchips embedded in medication bottles (see Chapter 38). However, there are drugs that may fall beneath a patient's deductibles or co-payments, or not be on formularies, so dispensed by the pharmacy but paid for in cash. In claims databases, these scenarios may result in misclassification of true medication exposure, as the patient would falsely appear unexposed. Also, since drug benefits vary depending upon the plan, pharmacy files may not capture all prescribed drugs if beneficiaries reach the drug benefit limit or pay for the prescription out of pocket. In the UK medical record systems, drugs prescribed by physicians other than the general practitioner could be missed, although continued prescribing by the general practitioner would be detected. *Ad hoc* case–control studies generally rely on patient histories for exposure data. These may be very inaccurate, as patients often do not recall correctly the medications they are taking [8]. However, this would be expected to vary, depending upon the condition studied, type of drug taken, questioning technique used, and so on [8–16] (see Chapter 37).

The *validity of the outcome data* is also most certain in hospital-based settings, in which the patient is subjected to intensive medical surveillance (see Chapter 14). It is least certain in outpatient data from organized systems of medical care. There are, however, methods of improving the accuracy of these data, such as using drugs, laboratory data, and procedures as markers of the disease and obtaining primary medical records (see Chapter 37). The outcome data from automated databases are listed as variable, therefore, depending upon exactly which data are being used and how. The UK medical record systems analyze the actual medical record, rather than claims, and can access additional questionnaire data from the general practitioner as well. Thus, their outcome data may be more accurate.

Control of confounding refers to the ability to control for confounding variables. As discussed in Chapter 3, randomization is the most convincing way of controlling for unknown, unmeasured, or unmeasurable confounding variables. Approaches that collect sufficient information to control for known and measurable variables are next most effective. These include health plans, the UK medical record systems, case–control surveillance, *ad hoc* case–control studies, and *ad hoc* cohort studies. Users of health databases in Canada, commercial databases, and Medicaid (sometimes) can obtain primary medical records, but not all information necessary is always available in those records. They generally are unable to contact patients directly to obtain supplementary information that might not be in a medical record. Finally, spontaneous reporting systems do not provide enough systematically collected information for control of confounding.

Relatively few of the data systems have data on *inpatient drug use*. The exceptions include spontaneous reporting systems, in-hospital databases (see Chapter 14), and some *ad hoc* studies if designed to collect such.

Only a few of the data resources have sufficient *data on outpatient diagnoses* available without special effort to be able to study them as outcome variables. *Ad hoc* studies can be designed to be able to collect such information. In the case of *ad hoc* randomized clinical trials, this data collection effort could even include tailored laboratory and physical examination measurements. In some of the resources, the

outpatient outcome data are collected observationally, but directly via the physician, and so are more likely to be accurate. Included are spontaneous reporting systems, the UK medical record systems, HMOs, Prescription Event Monitoring, and some *ad hoc* cohort studies. Other outpatient data come via physician claims for medical care, including Medicaid databases, commercial databases, and the provincial health databases in Canada. Finally, other data resources can access outpatient diagnoses only via the patient, so they are less likely to be complete; although the diagnosis can often be validated using medical records, it generally needs to be identified by the patient. These include most *ad hoc* case–control studies.

The degree of *loss to follow-up* differs substantially among the different resources. They are specified in Table 17.1.

Characteristics of Research Questions and Their Impact on the Choice of Pharmacoepidemiologic Data Resources

Once one is familiar with the characteristics of the pharmacoepidemiologic resources available, one must then examine more closely the research question, to determine which resources can best be used to answer it (see Table 17.2).

Pharmacoepidemiologic studies can be undertaken to generate hypotheses about drug effects, to strengthen hypotheses, and/or to test *a priori* hypotheses about drug effects. *Hypothesis-generating studies* are studies designed to raise new questions about possible unexpected drug effects, whether adverse or beneficial. Virtually all studies can and do raise such questions, through incidental findings in studies performed for other reasons. In addition, virtually any case–control study could be used, in principle, to screen for possible drug causes of a disease under study, and virtually any cohort study could be used to screen for unexpected outcomes from a drug exposure under study. In practice, however, the only settings in which this has been attempted systematically have been health plans, case–control surveillance, Prescription Event Monitoring, and Medicaid databases. To date, the most productive source of new hypotheses about drug effects has been spontaneous reporting. However, this is the goal of Sentinel, a Congressionally mandated data system of over 100 million US lives, initially built primarily for hypothesis strengthening as "Mini-Sentinel," although now being used for hypothesis generation as well, in addition to the traditional approach of using such data for hypothesis testing (see Chapter 25). In the future, new approaches using the internet (e.g., health websites with consumer posting boards and other social media) could potentially be used for hypothesis generation of events, including those not coming to medical attention.

Hypothesis-strengthening studies are designed to provide support for, although not definitive evidence for, existing hypotheses. The objective of these studies is to provide sufficient support for, or evidence against, a hypothesis to permit a decision about whether a subsequent, more definitive study should be undertaken. As such, hypothesis-strengthening studies need to be conducted rapidly and inexpensively. They can include crude analyses conducted using almost any dataset, evaluating a hypothesis which arose elsewhere. Because not all potentially confounding variables would be controlled, the findings could not be considered definitive. Examples would be the modular studies conducted within Sentinel (see Chapter 25). Alternatively, hypothesis-strengthening studies can be more detailed, controlling for confounding, conducted using the same data resource that raised the hypothesis. In this case, because the study is not specifically undertaken to test an *a priori* hypothesis, the hypothesis-testing type of study can only serve to strengthen, not test, the hypothesis. Spontaneous reporting systems are useful for raising hypotheses, but are

not very useful for providing additional support for those hypotheses. Conversely, randomized trials can certainly strengthen hypotheses, but are generally too costly and logistically too complex to be used for this purpose. (*Post-hoc* analyses of randomized trials can obviously be reanalyzed, for the purposes of generating or strengthening hypotheses, but then they are really being analyzed as cohort studies.) Of the remaining approaches, those that can quickly access, in computerized form, both exposure data and outcome data are most useful. Those that can rapidly access only one of these data types, only exposure or only outcome data, are next most useful, while those that need to gather both data types are least useful, because of the time and expense that would be entailed.

Hypothesis-testing studies are designed to evaluate in detail hypotheses raised elsewhere. Such studies must be able to have simultaneous comparison groups and must be able to control for most known potential confounding variables. For these reasons, spontaneous reporting systems cannot be used for this purpose, as they cannot be used to conduct studies with simultaneous controls (with rare exceptions, see [2]). The most powerful approach, of course, is a randomized clinical trial, as it is the only way to control for unknown or unmeasurable confounding variables. Instrumental variable analyses can approximate a randomized clinical trial, but only in the circumstances, to date limited, that all the underlying assumptions are met. (On the other hand, studies of dose response, duration response, drug–drug interactions, determinants of response, etc. are more readily done in nonrandomized than randomized studies; see Chapter 3.) Techniques which allow access to patients and their medical records are the next most powerful, as one can gather information on potential confounders that might only be reliably obtained from one of those sources or the other. Techniques which allow access to primary records but not the patient are next most useful.

The research implications of questions about the *beneficial effects* of drugs are different, depending upon whether the beneficial effects of interest are expected or unexpected. Studies of *unexpected beneficial effects* are exactly analogous to studies of unexpected adverse effects, in terms of their implications for one's choice of approach; in both situations one is studying side effects. Studies of *expected beneficial effects*, or drug efficacy, raise the special methodologic problem of confounding by the indication: patients who receive a drug are different from those who do not in a way which usually is related to the outcome under investigation in the study. This issue is discussed in detail in Chapter 33. As described there, it *is* sometimes possible to address these questions using nonexperimental study designs. Generally, however, the randomized clinical trial is far preferable, when feasible.

In order to address questions about the *incidence of a disease* in those exposed to a drug, one must be able to quantify how many people received the drug. This information can be obtained using any resource that can perform a cohort study. Techniques that need to gather the outcome data *de novo* may miss some of the outcomes if there is incomplete participation and/or reporting of outcomes, such as with Prescription Event Monitoring, *ad hoc* cohort studies, and outpatient pharmacy-based cohort studies. On the other hand, *ad hoc* data collection is the only way of systematically collecting information about outcomes that need not come to medical attention (see below). The only approaches that are free from either of these problems are hospital-based approaches. Registry-based case–control studies and *ad hoc* case–control studies can occasionally be used to estimate incidence rates, if one obtains a complete collection of cases from a defined geographic area. The other approaches listed cannot be used to calculate incidence rates.

To address a question about a *low incidence outcome*, one needs to study a large population

(see Chapter 4). This can best be done using spontaneous reporting, US Medicare, Prescription Event Monitoring, or the pharmacy-based medical record linkage systems, which can or do cover entire countries. Alternatively, one could use commercial databases, health plans, or aggregates of Medicaid databases, which cover a large proportion of the US, or the medical record systems in the UK. Canadian provincial databases can also be fairly large, and one can perform a study in multiple such databases. *Ad hoc* cohort studies could potentially be expanded to cover equivalent populations. Case–control studies, either *ad hoc* studies, studies using registries, or studies using case–control surveillance, can also be expanded to cover large populations, although not as large as the previously mentioned approaches. Because case–control studies recruit study subjects on the basis of the patients suffering from a disease, they are more efficient than attempting to perform such studies using analogous cohort studies. Finally, randomized trials could, in principle, be expanded to achieve very large sample sizes, especially large simple trials (see Chapter 32), but this can be extremely difficult and costly.

To address a question about a *low prevalence exposure*, one also needs to study a large population (see Chapter 4). Again, this can best be done using spontaneous reporting, US Medicare, the pharmacy-based medical record linkage systems, or Prescription Event Monitoring, which cover entire countries. Alternatively, one could use commercial databases, large health plans, or aggregates of Medicaid databases, which cover a large proportion of the US, or the medical record databases in the UK. *Ad hoc* cohort studies could also be used to recruit exposed patients from a large population. Analogously, randomized trials, which specify exposure, could assure an adequate number of exposed individuals. Case–control studies, either *ad hoc* studies, studies using registries, or studies using case–control surveillance, could theoretically be expanded to cover a large enough population, but this would be difficult and expensive.

When there are *important confounders* that need to be taken into account in order to answer the question at hand, then one needs to be certain that sufficient and accurate information is available on those confounders. Spontaneous reporting systems cannot be used for this purpose. The most powerful approach is a randomized trial, as it is the most convincing way to control for unknown or unmeasurable confounding variables. Techniques which allow access to patients and their medical records are the next most powerful, as one can gather information on potential confounders that might only be reliably obtained from one of those sources or the other. Techniques which allow access to primary records but not the patient are the next most useful.

If the research question involves *inpatient* drug use, then the data resource must obviously be capable of collecting data on inpatient drug exposures. The number of approaches that have this capability are limited, and include spontaneous reporting systems and inpatient database systems. *Ad hoc* studies could also, of course, be designed to collect such information in the hospital.

When the *outcome under study does not result in hospitalization, but does result in medical attention*, the best approaches are randomized trials and *ad hoc* studies, which can be specifically designed to be sure this information can be collected. Prescription Event Monitoring and the UK medical record systems, which collect their data from general practitioners, are excellent sources of data for this type of question. Reports of such outcomes are likely to come to spontaneous reporting systems as well. Medicaid databases and commercial databases can also be used, as they include outpatient data, although one must be cautious about the validity of the diagnosis information in outpatient claims. Canadian provincial databases are

similar, as are health plans. Finally, registry-based case–control studies could theoretically be performed, if they included outpatient cases of the disease under study.

When the *outcome under study does not result in medical attention at all*, the approaches available are much more limited. Only randomized trials and prospective cohort studies can be specifically designed to be certain this information is collected. Finally, occasionally one could collect information on such an outcome in a spontaneous reporting system, if the report came from a patient or from a healthcare provider who became aware of the problem while the patient was visiting for medical care for some other problem. In the future, as already noted, new approaches using the internet (e.g., health websites with consumer posting boards) could potentially be used for hypothesis generation of events not coming to medical attention.

When the *outcome under study is a delayed drug effect*, then one obviously needs approaches capable of tracking individuals over a long period of time. The best approach for this are some of the provincial health databases in Canada. Drug data are available in some of these for more than 25 years, and there is little turnover in the population covered. Thus, this is an ideal system within which to perform such long-term studies. Some health plans have even longer follow-up time available. However, as health plans they suffer from substantial turnover, albeit more modest after the first few years of enrollment. Commercial databases are similar. Any of the methods of conducting case–control studies can address such questions, although one would have to be especially careful about the validity of exposure information collected many years after the exposure. Medicaid databases have been available since 1973. However, the large turnover in Medicaid programs, due to changes in eligibility with changes in family and employment status, makes studies of long-term drug effects problematic. Similarly, one could conceivably perform studies of long-term drug effects using Prescription Event Monitoring, the pharmacy-based medical record linkage systems, *ad hoc* cohort studies, or randomized clinical trials, but these approaches are not as well suited to this type of question as the previously discussed techniques. Theoretically, one also could identify long-term drug effects in a spontaneous reporting system. This is improbably, however, as a physician is unlikely to link a current medical event with a drug exposure long ago.

When *the exposure under study is a new drug*, then one is, of course, limited to data sources that collect data on recent exposures, and preferably those that can collect a significant number of such exposures quickly. *Ad hoc* cohort studies or a randomized clinical trial are ideal for this, as they recruit patients into the study on the basis of their exposure. Spontaneous reporting is similarly a good approach, as new drugs are automatically and immediately covered, and in fact reports are much more common in the first three years after a drug is marketed. The major databases are next most useful, especially the commercial ones, as their large population base will allow one to accumulate a sufficient number of exposed individuals rapidly, so one can perform a study sooner. In some cases, there is a delay until the drug is available on the program's formulary; however, that especially can be an issue with HMOs. The US government claims databases (Medicare and Medicaid) have a delay in availability of their data, which makes them less useful for the newest drugs. *Ad hoc* case–control studies, by whatever approach, must wait until sufficient drug exposure has occurred that it can affect the outcome variable being studied.

Finally, if *one needs an answer to a question urgently*, potentially the fastest approach, if the needed data are included, is a spontaneous reporting system; drugs are included in these systems immediately, and an extremely large population base is covered. Of course, one cannot rely on any adverse reaction being detected

in a spontaneous reporting system. The computerized databases are also useful for these purposes, depending on the speed with which the exposures accumulate in them; of course, if the drug is not on the formulary in question, it cannot be studied. Modular analyses in Sentinel were designed for exactly this purpose (see Chapter 25). The remaining approaches are of limited use, as they take too long to address a question. One exception to this is Prescription Event Monitoring, if the drug in question happens to have been a subject of one of its studies. The other, and more likely, exception is case–control surveillance, if the disease under study is available in adequate numbers in its database, either because it was the topic of a prior study or because there was a sufficient number of individuals with the disease collected to be included in control groups for prior studies.

Examples

As an example, one might want to explore whether nonsteroidal anti-inflammatory drugs (NSAIDs) cause upper gastrointestinal bleeding and, if so, how often. One could examine the manufacturer's premarketing data from clinical trials, but the number of patients included is not likely to be large enough to study clinical bleeding, and the setting is very artificial. Alternatively, one could examine premarketing studies using more sensitive outcome measures, such as endoscopy. However, these are even more artificial. Instead, one could use any of the databases to address the question quickly, as they have data on drug exposures that preceded the hospital admission. Some databases could only be used to investigate gastrointestinal bleeding resulting in hospitalization (e.g., Kaiser Permanente, except via chart review). Others could be used to explore inpatient or outpatient bleeding (e.g., Medicare, Medicaid, Canadian provincial databases). Because of confounding by cigarette smoking, alcohol, and so on, which would not be well measured in these databases, one also might want to address this question using case–control or cohort studies, whether conducted *ad hoc* or using any of the special approaches available, for example case–control surveillance or Prescription Event Monitoring. If one wanted to be able to calculate incidence rates, one would need to restrict these studies to cohort studies, rather than case–control studies. One would be unlikely to be able to use registries, as there are no registries, known to this author at least, which record patients with upper gastrointestinal bleeding. One would not be able to perform analyses of secular trends, as upper gastrointestinal bleeding would not appear in vital statistics data, except as a cause of hospitalization or death. Studying death from upper gastrointestinal bleeding is problematic, as it is a disease from which patients usually do not die. Rather than studying determinants of upper gastrointestinal bleeding, one would really be studying determinants of complications from upper gastrointestinal bleeding, diseases for which upper gastrointestinal bleeding is a complication, or determinants of physicians' decisions to withhold supportive transfusion therapy from patients with upper gastrointestinal bleeding, for example age, terminal illnesses, and so on.

Alternatively, one might want to address a comparable question about nausea and vomiting caused by NSAIDs. Although this question is very similar, one's options in addressing it would be much more limited, as nausea and vomiting often do not come to medical attention. Other than a randomized clinical trial, for a drug that is largely used on an outpatient basis one is limited to systems which request information from patients, or *ad hoc* cohort studies.

As another example, one might want to follow up on a signal generated by the spontaneous reporting system, designing a study to investigate whether a drug which has been on the market for, say, five years is a cause of a relatively rare condition, such as allergic hypersensitivity

reactions. Because of the infrequency of the disease, one would need to draw on a very large population. The best alternatives would be Medicare or Medicaid databases, health plans, commercial databases, case–control studies, or Prescription Event Monitoring. To expedite this hypothesis-testing study and limit costs, it would be desirable if it could be performed using existing data. Prescription Event Monitoring and case–control surveillance would be excellent ways of addressing this, but only if the drug or disease in question, respectively, had been the subject of a prior study. Other methods of conducting case–control studies require gathering exposure data *de novo*.

As a last example, one might want to follow up on a signal generated by a spontaneous reporting system, designing a study to investigate whether a drug which has been on the market for, say, three years is a cause of an extremely rare but serious illness, such as aplastic anemia. One's considerations would be similar to those just described, but even Medicare or Medicaid databases would not be sufficiently large to include enough cases, given the delay in the availability of their data. One would have to gather data *de novo*. Assuming the drug in question is used mostly by outpatients, one could consider using Prescription Event Monitoring or a case–control study.

Conclusion

Once one has decided to perform a pharmacoepidemiologic study, one needs to decide which of the resources described in the earlier chapters of this book should be used. By considering the characteristics of the pharmacoepidemiologic resources available as well as the characteristics of the question to be addressed, one should be able to choose those resources that are best suited to addressing the question at hand.

References

1 Lawson DH, Beard K. *Intensive hospital-based cohort studies*. In: Strom BL, ed. Pharmacoepidemiology, 2nd edn. Chichester: John Wiley & Sons, 1994, pp. 157–70.

2 Wei L, Parkinson J, MacDonald TM. The Tayside Medicines Monitoring Unit (MEMO). In: Strom BL, ed. *Pharmacoepidemiology*, 4th edn. Chichester: John Wiley & Sons, 2005, pp. 323–36.

3 Prescott JD, Factor S, Pill M, Levi GW. Descriptive analysis of the direct medical costs of multiple sclerosis in 2004 using administrative claims in a large nationwide database. *J Manag Care Pharm* 2007; **13**: 44–52. http://www.amcp.org/data/jmcp/44-52.pdf

4 Maddison P, Kiely P, Kirkham B, *et al*. Leflunomide in rheumatoid arthritis: recommendations through a process of consensus. *Rheumatology (Oxford)* 2005; **44**: 280–6.

5 Shafazand S, Yang Y, Amore E, O'Neal W, Brixner D. A retrospective, observational cohort analysis of a nationwide database to compare heart failure prescriptions and related health care utilization before and after publication of updated treatment guidelines in the United States. *Clin Ther* 2010; **32**: 1642–50.

6 Risks of agranulocytosis and aplastic anemia. A first report of their relation to drug use with special reference to analgesics. The International Agranulocytosis and Aplastic Anemia Study. *JAMA* 1986; **256**: 1749–57.

7 Strom BL, West SL, Sim E, Carson JL. Epidemiology of the acute flank pain syndrome from suprofen. *Clin Pharmacol Ther* 1989; **46**: 693–9.

8 Klemetti A, Saxen L. Prospective versus retrospective approach in the search for environmental causes of malformations. *Am J Public Health* 1967; **57**: 2071–5.

9 Glass R, Johnson B, Vessey M. Accuracy of recall of histories of oral contraceptive use. *Br J Prev Soc Med* 1974; **28**: 273–5.

10 Stolley PD, Tonascia JA, Sartwell PE, *et al*. Agreement rates between oral contraceptive users and prescribers in relation to drug use histories. *Am J Epidemiol* 1978; **107**: 226–35.

11 Paganini-Hill A, Ross RK. Reliability of recall of drug usage and other health-related information. *Am J Epidemiol* 1982; **116**: 114–22.

12 Rosenberg MJ, Layde PM, Ory HW, Strauss LT, Rooks JB, Rubin GL. Agreement between women's histories of oral contraceptive use and physician records. *Int J Epidemiol* 1983; **12**: 84–7.

13 Schwarz A, Faber U, Borner K, Keller F, Offermann G, Molzahn M. Reliability of drug history in analgesic users. *Lancet* 1984; **2**: 1163–4.

14 Coulter A, Vessey M, McPherson K. The ability of women to recall their oral contraceptive histories. *Contraception* 1986; **33**: 127–39.

15 Mitchell AA, Cottler LB, Shapiro S. Effect of questionnaire design on recall of drug exposure in pregnancy. *Am J Epidemiol* 1986; **123**: 670–6.

16 Persson I, Bergkvist L, Adami HO. Reliability of women's histories of climacteric oestrogen treatment assessed by prescription forms. *Int J Epidemiol* 1987; **16**: 222–8.