37

Validity of Drug and Diagnosis Data in Pharmacoepidemiology

Mary Elizabeth Ritchey[1], *Suzanne L. West*[2,3], *and George Maldonado*[4]

[1] *RTI Health Solutions, RTI International, Research Triangle Park, NC, USA*
[2] *RTI International, Research Triangle Park, NC, USA*
[3] *Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA*
[4] *Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN, USA*

To provide evidence-based care, clinicians need to know the benefits and risks of the medications they are prescribing (see Chapter 35), and this information needs to come from robust research. For example, evidence for medication efficacy typically comes from randomized controlled trials, whereas establishing the magnitude of a drug safety concern often comes from observational studies using self-reported data or from electronic data such as administrative claims data or electronic health records (EHRs). Previous editions of this chapter focused primarily on self-reported data or administrative claims, but with the growing availability of EHR data, pharmacoepidemiologists are increasingly using these data for research because they contain more granular information such as the reason for medication prescription, laboratory test results, and patient vitals (e.g., blood pressure and weight).

Clinical Problems to be Addressed by Pharmacoepidemiologic Research

Of particular concern to the subject of this book is the validity of data on drug exposure and disease occurrence because the typical focus of pharmacoepidemiologic research is often the association between a medication and an adverse drug event. Further, many potential confounders of importance in pharmacoepidemiologic research (although certainly not all) are either drugs or diseases. Clinicians recognize that patients very often do not know the names of the drugs they are taking currently. Thus, it is a given that patients have difficulty recalling past drug use accurately, at least in the absence of aids to enhance recall. Superficially at least, patients cannot be considered reliable sources of diagnosis information either; in some instances, they may not have even been told

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

the correct diagnosis, let alone recall it. Yet, these data elements are crucial to pharmacoepidemiologic studies that ascertain data using questionnaires. Special approaches have been developed by pharmacoepidemiologists to obtain such data more accurately when using self-report for data collection, but the success of these approaches needs to be considered in detail.

Besides self-reported data, pharmacoepidemiologists have been using administrative claims data for more than 30 years to evaluate drug safety. We discuss validity issues with using these data for research. However, the changing landscape of healthcare requires reassessing the validity of the data pharmacoepidemiologists are now using for their research and how these data impact clinical practice.

More and more, pharmacoepidemiologists are turning to EHR data for their research.

Whereas the increased granularity of EHR data is a benefit for their use in pharmacoepidemiology, important limitations of these data include their potential incompleteness and lack of interoperability across health systems. Unless EHR data arise from "closed" healthcare systems where patients receive all their outpatient and inpatient care, then the EHR data may represent only a portion of the patients' health problems and care received. If EHR data from multiple health systems are used, even if the health systems use the same EHR vendor, the data may need to be restructured so that they are consistent across all data arising from all health systems. The clinician reviewing evidence for patient care that arises from studies using EHR data trusts that these data have been curated sufficiently to produce robust and valid study findings.

Methodologic Problems to be Solved by Pharmacoepidemiologic Research

There are five major methodologic problems associated with validity of data for pharmacoepidemiologic research: indices of measurement error, quantitative measurement of validity, quantitative measurement of reliability, measurement error in pharmacoepidemiologic research, and adjusting measures of association for measurement error.

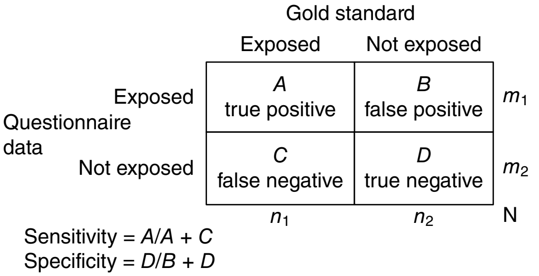
Indices of Measurement Error

Two main comparisons may be drawn between two (or more) methods of data collection or sources of information on exposure or outcome: validity and reliability. Many different terms have been used to describe each, resulting in some confusion. Although the literature uses the term *validation study* or *verification* to describe the agreement between two sources of information, *concordance* or *agreement* might be a more appropriate term to describe the comparison between data sources because validation requires a "gold standard." In the following discussion, we define and differentiate between validity and reliability. Validity is assessed using sensitivity and specificity, while reliability is typically measured using percent agreement and kappa [1].

Quantitative Measurement of Validity

Only when one of the methods or sources is clearly superior to the other can the comparison be said to measure validity. The superior method or source is often called a "gold standard." In recognition that a method or source can be superior to another method or source without being perfect, the term *alloyed (or tarnished) gold standard* has been used [2].

For a binary exposure or outcome measure, such as "ever" versus "never" use of a particular drug, two measures are used to assess validity. Sensitivity measures the degree to which the potentially inferior source or method correctly identifies individuals who, according to the superior method or source, possess the characteristic of interest (i.e., ever used the drug). Specificity measures the degree to which the inferior source



		Gold standard		
		Exposed	Not exposed	
Questionnaire data	Exposed	*A* true positive	*B* false positive	m_1
	Not exposed	*C* false negative	*D* true negative	m_2
		n_1	n_2	N

Sensitivity = *A*/*A* + *C*
Specificity = *D*/*B* + *D*

Figure 37.1 Formulas for calculating sensitivity and specificity.

or method correctly identifies individuals who, according to the superior method or source, lack the characteristic of interest (i.e., never used the drug). Figure 37.1 illustrates the calculation of sensitivity and specificity.

Sensitivity and specificity are the two sides of the validity coin for a dichotomous exposure or outcome variable. In general, sources or methods with higher sensitivity tend to have lower specificity, and methods with higher specificity tend to have lower sensitivity. In these very common situations, neither of the two sources or methods compared can be said to have superior overall validity. Depending on particulars of the study setting in which the research question is addressed, either sensitivity or specificity may be the more important validity measure. Moreover, absolute values of these measures can be deceiving. For instance, if the true prevalence of ever use of a drug is 5%, then an exposure classification method or information source with 95% specificity (and perfect sensitivity) will incorrectly double the measured prevalence to 10%. The ultimate criterion of importance of a given combination of sensitivity and specificity is the degree of bias exerted on a measure of effect such as an estimated relative risk due to measurement error.

As measures of validity, sensitivity and specificity have “truth” (i.e., the classification according to a gold standard or an alloyed gold standard) in their denominators. Investigators should take care not to confuse these measures with the predictive values of positive and negative classifications, which include the inferior measure in their denominators. We distinguish here between the persons who *actually* do or do not have an exposure or outcome with those who are *classified* as having it or not having it (using the potentially inferior or alternative data source). The proportion of persons classified as having the exposure or outcome who truly do have the exposure or outcome is the positive predictive value. The proportion of persons correctly classified as lacking the exposure or outcome is the negative predictive value.

Assessment of the positive predictive value (as is performed in many validation studies in administrative claims and EHR data) of an outcome does not directly measure the validity of the data source. Predictive values are measures of performance of a classification method or information source, not measures of validity. Predictive values depend not only on the sensitivity and specificity (i.e., on validity) but also on the prevalence of the exposure or outcome. Thus, if a method or information source for classifying persons with respect to outcome or exposure has the same sensitivity and specificity in two populations but those populations differ in their outcome or exposure prevalence, the source or method will have different predictive values in the two populations. Nonetheless, all measures are useful and the most important one will depend on the question being answered. Ideally, one would design a validation study to calculate sensitivity and specificity as well as positive and negative predictive values.

In some validation studies, one method or source may be used as a gold standard or as an alloyed gold standard to assess another method or source with respect to only one side of the validity coin. Studies that focus on the completeness of one source, such as studies in which interview responses are compared with prescription dispensing records to identify drug exposures that were forgotten or otherwise not reported by the respondents, may measure (more or less accurately) the sensitivity of the interview data. However, such studies are silent on the specificity unless one acknowledges strong assumptions (e.g., that the respondent could not have obtained the drug in a way that would not be recorded in the prescription dispensing records). Similarly, in administrative claims data, prescriptions that are filled outside the insurance plan may not be captured in the database, especially for generic drugs that are less costly to purchase outright rather than using a co-pay.

For a drug exposure, a true gold standard would be a list of all drugs the study participant has taken (i.e., ingested), including dose, duration, and dates of exposure. This drug list might be a diary of prescriptions the study participants kept or, perhaps more readily available, a computerized database of filled prescriptions, although neither of these data sources is a genuine gold standard. Prescription diaries cannot be assumed to be kept in perfect accuracy. For instance, participants may tend to record that drug use was more regular and complete than it actually was or that use adhered to the prescribed regimen. Similarly, substantial gaps may exist between the point at which a prescription is filled and when it is ingested, if it is ingested at all. See Chapter 38 for further discussion of adherence.

Two methods are used to quantify the validity of continuously distributed variables, such as duration of drug usage. The mean and standard error of the differences between the data in question and the valid reference measurement are typically used when the measurement error is constant across the range of true values (i.e., when measurement error is independent of where an individual's true exposure falls on the exposure distribution in the study population) [3]. With the caveat that it is generalizable only to populations with similar exposure distributions, the product–moment correlation coefficient may also be used.

High correlation between two measures does not necessarily mean high agreement. For instance, the correlation coefficient could be very high (i.e., close to 1), even though one of the variables systematically overestimates or underestimates values of the other variable. The high correlation means that the over- or underestimation is systematic and very consistent. When the two measures being compared are plotted against each other and they have the same scale, full agreement occurs only when the points fall on the line of equality, which is 45° from either axis [4]. However, perfect correlation occurs when the points lie along any straight line parallel to the line of equality. It is difficult to tell from the value of a correlation coefficient how much bias will be produced by using an inaccurate measure of disease or exposure.

Quantitative Measurement of Reliability

When the same data collection method or source of information is used more than once for the same information on the same individual, comparisons of the results measure the reliability of the method or information source. An example of a reliability study is a comparison of responses in repeat interviews using the same interview instrument. Reliability is not validity, although the term is sometimes used, inaccurately, as such. In general, studies that measure mere agreement are all too commonly interpreted as though they measured validity or accuracy. The term *reliability* tends to be used far too broadly to refer variously not only to reliability itself, but to agreement or validity as well. Researchers and others should take greater care with the way they use such terms.

When different data collection methods or different sources of information are compared (e.g., comparison of prescription dispensing records with interview responses), and neither of them can be considered distinctly superior to the other, the comparisons measure mere agreement. Agreement between two sources or methods does not imply that either is valid.

To evaluate reliability or agreement for categorical variables, the percentage agreement between two or more sources and related (kappa) coefficient are used. They are used only when two imperfect classification schemes are being compared, not when one classification method may be considered *a priori* superior to the other [3,5]. The kappa statistic is the percentage agreement corrected for chance [3]. Agreement is conventionally considered poor for a kappa statistic less than zero, slight for a kappa between zero and 0.20, fair for a kappa of 0.21–0.40, moderate for a kappa of 0.41–0.60, substantial for a kappa of 0.61–0.80, and almost perfect for a kappa of 0.81–1.00 [1]. Figure 37.2 illustrates the percentage agreement and kappa calculations for a reliability assessment between questionnaire data and medical record information.

The intraclass correlation coefficient is used to evaluate the reliability of continuous variables [5]. It reflects both the average differences in mean values as well as the correlation between measurements. The intraclass correlation coefficient indicates the degree to which the total measurement variation is due to the differences between the subjects being evaluated and to differences in measurement for one individual. When the data from two sets of measurements are identical, the intraclass correlation coefficient equals 1.0. Under certain conditions, the intraclass correlation coefficient is exactly equivalent to Cohen's weighted kappa [3]. It is impossible to translate values of measures of agreement, such as kappa, into expected degrees of bias in exposure or disease associations.

		Medical record		
		Exposed	Not exposed	
Questionnaire data	Exposed	*A*	*B*	m_1
	Not exposed	*C*	*D*	m_2
		n_1	n_2	N

Accuracy = $A + D/N$

Chance agreement (expected) = $((n_1 \times m_1) + (n_2 \times m_2)/N^2$

$$\kappa = \frac{\text{accuracy–chance agreement}}{\text{1–chance agreement}}$$

Figure 37.2 Formulas for calculating the percent agreement and K.

Measurement Error in Pharmacoepidemiologic Research

Epidemiologic assessments of the effects of a drug on disease incidence depend on an accurate assessment of the study exposure, disease occurrence, and variables to be adjusted in the statistical analysis. Measurement error for any of these factors may incorrectly identify a risk factor in the study that does not exist in the population or, conversely, may fail to detect a risk factor when one truly exists.

In an epidemiologic study, the measure of association is often based on the number of subjects categorized by the cross-classification of presence or absence of disease and exposure. For example, when questionnaire data are used to study the association between drug A and disease B, study participants who forget their past exposure to drug A would be incorrectly classified as nonexposed. Similarly, if a provider uses a diagnosis code to document the process of testing and ruling out a disease and then a researcher uses the diagnosis code as a study outcome, then the person would be incorrectly classified as having the outcome. This misclassification is a measurement error. Although the measurement process often involves some error, if this measurement error is of sufficient magnitude, the validity of the study's findings is diminished.

Surprisingly, measurement error is often ignored in epidemiologic studies. Jurek *et al.* [6] reported the results of a random survey of studies published in three major epidemiology journals; they concluded the following for exposure-measurement error (EME): "Overall, the potential impact of EME on error in epidemiologic study results appears to be ignored frequently in practice" (p. 871).

Measurement error is a potentially serious cause for concern in epidemiologic studies, and therefore, for several reasons, this should not be ignored when analyzing and interpreting pharmacoepidemiologic study results. First, small amounts of measurement error can cause large amounts of error in study results. For example, consider a pharmacoepidemiologic study of nonsteroidal antiinflammatory drug (NSAID) A versus NSAID B on gastrointestinal (GI) bleed (Figure 37.3). In a study with a total number of study subjects equal to more than 22 000, if only 10 subjects are misclassified with respect to their exposure or disease (five who actually took NSAID B are incorrectly classified as having taken NSAID A, and five users of NSAID A without GI bleed are incorrectly classified as having GI bleed), the observed odds ratio (OR) would be 2.1 when the correct OR is in fact 1.0.

Second, measurement error can cause study results to overestimate or underestimate true effect sizes, and there is no simple rule for predicting the direction of the error in real-life situations. We now understand that these old and often-cited heuristics are not necessarily true, except under special conditions that are not likely to occur in practice: (1) nondifferential misclassification always produces bias toward the null, and (2) bias toward the null always produces an observed relative risk that is an underestimate of

Case-Control Study of NSAIDs and Myocardial Infarction (MI)

No exposure misclassification

MI cases recall exposure just as well as those without MI

	NSAID use	No NSAIDs	
MI	200	200	OR = 2.5
No MI	240	600	

MI cases recall exposure better than those without an MI

MI	Sensitivity=	0.95
	Specificity=	0.9
No MI	Sensitivity=	0.8
	Specificity=	0.7

	NSAID use	No NSAIDs	
MI	210	190	OR = 1.4
No MI	372	468	

MI cases do not recall exposure as well as those who did not have an MI

MI	Sensitivity=	0.8
	Specificity=	0.9
No MI	Sensitivity=	0.9
	Specificity=	0.7

	NSAID use	No NSAIDs	
MI	180	220	OR = 0.9
No MI	396	444	

Figure 37.3 Example of differential misclassification of exposure.

the true relative risk. These heuristics are unlikely to be true in practice for the following reasons.

- Conditions beyond nondifferentiality are required to guarantee bias is toward the null [7–12] (e.g., when the degree of exposure measurement error systematically differs across levels of a polychotomous or continuous exposure variable, or when errors in measuring the exposure and outcome are not independent).
- Even when the above conditions beyond nondifferentiality are met, exact nondifferentiality is required to guarantee bias is toward the null [13,14].
- Also required to guarantee bias is toward the null is either (1) the absence of other study biases (e.g., absence of confounding, absence of bias due to nonrandom subject selection/participation) or (2) the combined effect of all other biases is also toward the null [13].
- Bias is a statistical term that is defined as the difference between the true value and the expected value of an estimator (i.e., the average of study results over hypothetical repetitions of the study). Bias is not the difference between the observed estimate for one repetition of the study and the true value. This important distinction was not appreciated in earlier writings on this topic, and even today we epidemiologists are not careful in our use of the term bias. Therefore, when bias is toward the null, the expected value of the estimator is shifted toward the null, but an observed estimate can be an overestimate of the true relative risk due to the influence of random error [13]. (Similarly, when there is no bias of any kind, one observed estimate can be an overestimate or an underestimate of the true relative risk simply due to random error.)

Third, error in measuring variables to be adjusted in the analysis can result in only partial adjustment for the mismeasured variables [15].

Adjusting Measures of Association for Measurement Error

One can use sensitivity analysis methods (also known as uncertainty analysis and bias analysis) [16–24] to adjust measures of association for measurement error as well as for other study biases. (As used in this context, the meaning of the term *sensitivity* differs from its other epidemiologic meaning as the counterpart to specificity as a measure of classification validity.) Sensitivity analysis is the last line of defense against biases after every effort has been made to eliminate, reduce, or control them in study design, data collection, and data analysis. In a sensitivity analysis, one alters key assumptions or methods reasonably to see how sensitive the results of a study are to those variations. (See Chapter 38 for discussion of sensitivity analyses in pharmacoeconomic studies.)

One key assumption, usually implicit, in any study that does not quantitatively account for the possibility of error in measuring the study exposure or study outcome is that the exposure and the outcome in a study have been measured accurately. With estimates of sensitivity and specificity from validation studies (from previous research or from a subsample within the study analyzed) or “guesstimates” from expert experience and judgment, one can modify this assumption and use sensitivity analysis methods to “back calculate” what the results might have looked like if more accurate methods had been used to classify participants with respect to outcome, exposure, or both [17,25].

For many years, a qualitative and informal version of this kind of assessment has been conducted. However, the net result is controversy, with investigators judging the bias small and critics judging it large. Further, in the absence of a formal bias analysis, intuitive judgments, even those of the most highly trained and widely experienced investigators, can be poorly calibrated in such matters. Formal sensitivity analysis makes the assessment of residual bias transparent and quantitative and forces the investigator (and other critics) to defend criticisms that in earlier times would have remained qualitative and unsubstantiated. An important and well-known historical example is the bias from nondifferential misclassification of disease proposed by Horwitz and Feinstein [26] to explain associations between early exogenous estrogen preparations and endometrial cancer. When proper sensitivity analyses were conducted to assess this bias, only a negligible proportion of those associations were explained by bias [26–28].

Epidemiologic applications of quantitative methods with a long history in the decision sciences have become accessible for quantifying uncertainties about multiple sources of systematic error in a probabilistic manner [24,29–31]. These methods permit the incorporation of available validation data as well as expert judgment about measurement error, uncontrolled confounding, and selection bias along with conventional sampling error, and prior probability distributions for effect measures themselves, to form uncertainty distributions. These approaches have been used practically in pharmacoepidemiology studies such as in assessing selection bias in a study of topical coal tar therapy and skin cancer among severe psoriasis patients [30]; exposure misclassification and selection bias in a study of phenylpropanolamine use and stroke [24]; and selection bias, confounder misclassification, and unmeasured confounding in a study of less than standard therapy and breast cancer mortality [29], as well as in other clinical and nonclinical applications [18,31–39].

Sometimes biases can be shown to be of more concern and sometimes of less concern than intuition or simple sensitivity analysis might suggest. Almost always, the probabilistic uncertainty about these sources of systematic error dwarfs the uncertainty reflected by conventional confidence intervals (CIs). By the use of these methods, the assessment of systematic error can move from a qualitative discussion of "study limitations," beyond sensitivity analyses of one scenario at a time for one source of error at a time, to a comprehensive analysis of all sources of error simultaneously. The resulting uncertainty distributions can not only supplement but also supplant conventional likelihood and *P* value functions, which reflect only random sampling error. As a result, much more realistic, probabilistic assessments of total uncertainty attending to effect measure estimates are in the offing [19].

Currently Available Solutions

Conducting Validation Studies to Assess Self-Reported Data

In 1979, Leon Gordis commented that epidemiologists have become so enamored with analyzing their data that they have paid too little attention to the validity of the raw data being analyzed with these sophisticated techniques [40]. Gordis' comment reflects a time when pharmacoepidemiologic research was typically conducted by using questionnaires to gather data. The field was only just beginning to use data that arose from the provision of healthcare, including health insurer data such as Medicaid claims.

This section of the chapter focuses on the collection and validation of self-reported data for pharmacoepidemiologic research. We begin this section with a brief discussion of how individuals store and retrieve information from memory, tasks that are required when responding to a questionnaire. We use an example of how a person might recall a depression episode to illustrate retrieval of specific information from memory. Recognizing the challenges of information retrieval, we discuss best practices for designing questions to elicit specific drug and diagnosis information. Separately for drugs and diagnoses, we discuss the influence of comparator selection when validating self-reported data, the accuracy of recall, and the factors influencing recall and provide examples for illustration.

Autobiographical Memory and the Response Process

Pharmacoepidemiologic research that relies on self-reported data requires asking study respondents to recall events or exposures that occurred at some time in the past, with recall intervals spanning from days to years. The types of temporal questions study respondents are often asked and that require the memory processes are as follows [41]:

- Time of occurrence, which requires respondents to provide a date when an event occurred, such as when they were diagnosed with a particular condition.
- Duration questions such as, "How long did you take drug A?"
- Elapsed time, which asks how long it has been since an event occurred, including questions such as, "How many months has it been since you last took drug A?"
- Temporal frequency questions that ask respondents to report the number of events that occurred over a specific time period, such as "How many visits did you make to your primary care practitioner in the past six months?"

To appreciate the accuracy of data derived by respondent recall for addressing these types of questions, it is important to understand how we process, organize, and recall autobiographical information, which is key to the response process. Creating and retrieving information from autobiographical memories is a three-step process. Information that comes in via sensory or emotional input (e.g., visual, hearing, semantic) is *encoded* into a construct that can be stored within the brain. The next step is *storage*, which refers to how the brain retains the information, typically in either short- or long-term memory. *Retrieval or recall* of memories requires reaccessing information that was previously encoded and stored. Recall effectively returns a memory from long-term storage to short-term memory, where it can be accessed for retrieval purposes [42,43]. Current thinking is that retrieval of information from autobiographical memory is goal oriented, where the retrieval process requires bringing together spatial, temporal, and social information with information derived from the emotions and senses [42].

The recall of encoded or catalogued information from memory is thought to be facilitated by using important personal milestones [41]. Thus, when respondents are asked to recall a visit to a doctor that may have occurred at a particular point in time, researchers believe that the respondents use scripts (a generic mental representation of the event) to help retrieval. For example, the respondent first contemplates a doctor visit in general and then supplements this script with details relevant to the particular visit that require contemplation for specific criteria (e.g., diagnosis) and timing (e.g., a particular year). In general, underreporting of medical conditions and health visits is more widespread as the interval since the event increases [44–46].

Recent evidence suggests that age affects memory details, with older individuals recalling slightly more details than younger individuals. Using an instrument focused on words used in everyday spoken and written language to measure autobiographical memory, Gardner and colleagues noted that recall of content and details for events and objects was slightly greater in adults 46–78 years old compared to those 26–45 years of age for both recent and remote memories [47]. There was little difference between the two age groups for recalling individuals and temporal details of events.

Applying what we know about how autobiographical memory is organized and the recall process in general helps us to understand survey response. A respondent undergoes four key tasks when asked to answer a questionnaire: (1) question comprehension and interpretation, (2) search for and retrieval of information to construct an answer to the question depending on whether appropriate cues are given, (3) judgment to discern the completeness and relevance of memory for formulating a response, and (4) development of the response based on retrieved memories [41,48–50]. If survey instrument developers pay too little attention to the first two key tasks, their questions can be too vague or complex for respondents to marshal retrieval processes appropriately.

The following example best illustrates the response process [41] for recalling the date on which a respondent's depression was diagnosed (January 2015). The recall process begins with the respondent being uncertain whether the depression was diagnosed in 2014 or 2015. To work towards identifying the correct year, the respondent recalls that the depression occurred after he lost his job. The job loss was particularly traumatic because he and his wife just purchased their first home a few months previously, and now, with the loss of his income, they were at risk of losing the house. The home purchase was a landmark event for this respondent, and he remembers that it occurred in mid-2014, just as their children finished the school year. So, in 2014 he lost his job, near the end of the year because the holiday season was particularly grim. He remembers that his depression was diagnosed after the holidays, but was it January or February of 2015? It was January 2015 because he was already taking antidepressants by Valentine's Day, when he went out to dinner with his wife and he could not drink wine with his meal. This chronology is illustrated in Figure 37.4. We describe below how to use the response process to design questions to elicit the self-reported information requested.

As illustrated in Figure 37.4, landmark events probably serve as the primary organizational units of autobiographical memory and, as such, anchor information retrieval [51]. In particular, the example shows how the respondent used landmark and other notable events, relationships among datable events, and general knowledge (holiday period and children finishing the school year) to reconstruct when his major depression

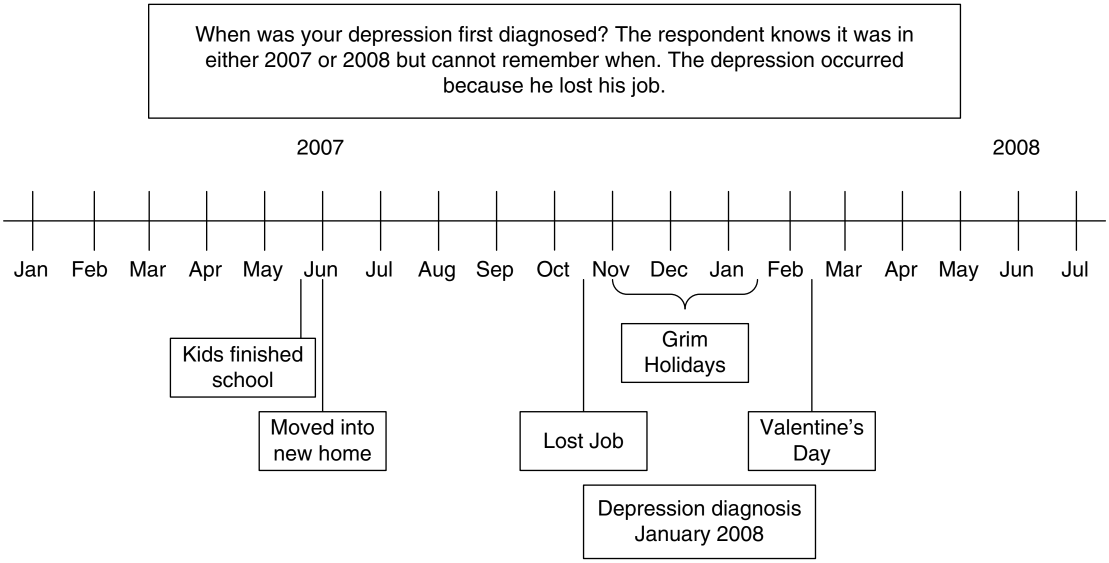

When was your depression first diagnosed? The respondent knows it was in either 2007 or 2008 but cannot remember when. The depression occurred because he lost his job.
2007
2008
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Kids finished school
Moved into new home
Grim Holidays
Lost Job
Valentine's Day
Depression diagnosis January 2008

Figure 37.4 Recall schematic for showing how date of depression diagnosis was determined.

was first diagnosed. An important caveat is that this respondent was willing to expend considerable effort to search his memory to determine when his depression was diagnosed – this may not be the situation for all respondents.

The next section takes what we know about autobiographical memory and the response process to develop questionnaires for pharmacoepidemiologic research.

Best Practices for Questionnaire Design

Designing a questionnaire for collecting pharmacoepidemiologic data requires consideration of the challenges and limitations of autobiographical memory as described above and careful planning and pretesting [52] before fielding the study. Survey researchers encourage use of several general techniques to assist respondents in recalling information accurately, including use of reference periods (e.g., "in the past 12 months, that is, since December 1, 2017, how many times did you…"), event histories and calendars like the one in Figure 37.4, diaries, and photos of medications. We provide a more in-depth discussion of questionnaire design for collecting medication and diagnosis data later. We recommend that, after collecting the self-reported data using techniques to maximize their accuracy, and prior to the analysis, researchers assess their accuracy for addressing the study hypothesis by comparing the data to another data source such as health insurer claims or EHRs [53].

We suggest the following steps be considered during the design and initial analysis stages of a study requiring data collection via questionnaire.

- Use validated instruments or validated questions whenever possible.
- Consider question banks if new questions are required, such as World Bank's Living Standards Measurement Study [54] and Q-Bank [55,56].
- Use question assessment tools to determine the likelihood of response error. These tools include the Question Appraisal System [57], the Survey Quality Predictor (SQP) [58], and the Question Understanding Aid (QUAID) [59,60].
- Strive for a fifth-grade literacy level if you must develop new survey questions to be used for a general population [61].
- Pretest the questions using cognitive testing [62–64] to assess respondent comprehension of new questions.

The process of *satisficing* occurs when respondents expend the least psychological and emotional effort possible to provide an acceptable answer to a survey question rather than an optimal answer [65,66]. To minimize satisficing, questionnaire developers should consider the length of the instrument and the number of response categories. When faced with a long list of choices and depending on the mode of questionnaire administration (i.e., telephone versus self-administered), respondents may choose answers from either the top or the bottom of the list to minimize effort. For this reason, it is often recommended to randomize response options. Respondents with lower cognitive skills and less education, when challenged with discerning the best possible response, are more apt to settle for a satisfactory rather than an optimal response. Because accuracy of response is critical for pharmacoepidemiologic research, questionnaire developers must consider methods to minimize response burden leading to satisficing.

With the increasing availability of broadband and the population's access to the internet, more surveys are moving away from face-to-face and telephone interviewer administration to web-based surveys. This modality requires the same considerations for question design as described earlier, but because no interviewer is available, usability should be tested as well. Usability evaluates the survey–respondent interaction: essentially, how efficiently and effectively respondents can answer the web-based questions [67,68]. For example, usability evaluates screen size, button placement, and formatting issues specific to web applications, especially for questionnaires using mobile technologies. Usability assessments can be combined with other pretesting modalities [52], including cognitive interviews, by embedding probes that allow respondents to explain why they provided their answers [69], a parallel to face-to-face cognitive testing without requiring an interviewer [52].

The earlier discussion focused on measurement error related to survey design and to respondent motivation. Measurement error can also be attributed to improper training of interviewers and poor data entry quality. The degree to which one understands the measurement error associated with key variables critical to the analysis can be assessed by using several different modelling approaches, which Biemer discusses in more detail [70].

Assessing the Accuracy of Self-Reported Data

Despite researchers using the best methods for designing questionnaires to elicit specific information on medications used previously and past diagnoses, self-reported data still require evaluation for accuracy to ensure valid findings. Ideally, researchers will have access to a truly accurate comparison source (i.e., gold standard) so that sensitivity and specificity can be calculated for use in bias analyses. For example, we can use pill counts, chemical markers inserted into the pills, electronic monitoring caps, or pharmacy dispensing databases to assess self-reported medication use. As discussed earlier, depending on the comparison data source, it may only be possible to calculate either sensitivity or specificity

Methodologic studies that use alternative data sources, such as prospectively collected drug data (e.g., from diaries), or databases of dispensed drugs can measure both sensitivity and specificity if one assumes that these databases are true gold standards. In pharmacoepidemiology, lower sensitivity is often more of a concern than lower specificity. Questionnaires that underreport diseases or miss drug exposures because the medication was filled without using the pharmacy plan (e.g., when the co-pay is higher than the cost of the medication) – that is, data sources with low sensitivity – cannot be used to rigorously evaluate drug–disease associations. Alternatively, low specificity is often less of a problem in pharmacoepidemiology unless the characteristic with low specificity also has very low prevalence in the population being studied. For example, because the incidence of Stevens–Johnson syndrome is low, a small degree of misclassification when using administrative claims data in which the case definition uses the *International Classification of Diseases, Ninth Edition, Clinical Modification* (ICD-9-CM) code 695.1 will include several skin problems other than Stevens–Johnson (i.e., the false-positive rate would be high) [71].

Besides the need for completeness on the individual level, the comparator database must have information for all persons whose information is to be assessed for accuracy. Systematic omissions of specific population groups, such as certain ethnic or racial groups, diminish the quality of the database.

In the next section of the chapter, we discuss issues in using the medical record as a comparator data source to evaluate the accuracy and completeness of survey data on medication and diagnoses ascertained via self-report. We discuss use of automated databases as a comparator data source for assessing validity and reliability of self-reported information in a later section.

Influence of Comparator Selection for Assessing Self-Report Accuracy

The early work on evaluating the completeness of self-reported diagnosis and medication data typically used paper medical records for comparison [72–74]. In summary, several studies from the late 1980s through 2000 indicated that inpatient medical records were often missing outpatient medications [75–77]. Similarly, outpatient medical records were also often incomplete, and completeness varied by the number and type of medication the patients were taking [78–82]. Diagnoses or other relevant inpatient information were often omitted from patient records as well [83–86]. These studies indicate that the paper medical record may not be that useful for validation of diagnosis and medication data. With the greater availability of EHR software and policy levers incentivizing their use, providers and hospitals in the United States have been moving to EHRs, making paper medical records obsolete.

Nonetheless, regardless of whether the medical record is paper or electronic, one needs to understand its availability, completeness, and accuracy to determine whether it is adequate for evaluating the accuracy of self-reported information. Retrieval of medical records depends not only on a person's ability to remember and report who prescribed the drug or diagnosed the condition in question, but on whether the healthcare provider recorded the information (and recorded it accurately) and on the availability of the medical record for review. If the medical record cannot be retrieved because the healthcare provider could not be identified, the provider had retired, or the record was destroyed or lost, the events cannot be verified.

While paper medical records are often incomplete, how complete are EHR data for assessing the accuracy of self-reported diagnosis and medication data? This question requires reframing to consider EHR completeness at both the individual patient and the institution level. In the US, healthcare is fragmented. Patients see multiple providers, are treated in several different health settings (e.g., chiropractors, podiatrists), and may become inpatients at several different hospitals [87,88]. Thus, accessing patients' outpatient and inpatient medical records does not guarantee that a researcher will have all medical care provided and drugs prescribed to the patient. For example, if a researcher is able to access only the patient's primary care records, it is possible that the results of cardiology tests to confirm a diagnosis or medications for that diagnosis will not be available. However, when patients are seen by integrated delivery systems that include primary care, multiple specialties, and inpatient care, there is a greater likelihood that the EHR will contain most of the care provided and medications prescribed to the patient.

In addition, the EHR data themselves may be no more accurate than paper records if the EHR data simply substitute for paper records. For example, exposure information about medications or important confounders (e.g., smoking) may be incomplete if clinicians do not ascertain this information and correctly enter it into the EHR. Another problem introduced by EHRs is the potential for errors inherent to electronic data entry, such as copying and pasting of incorrect data from other parts of the record, of expired or irrelevant clinical information, or of incorrect and/or unverified medication lists [89].

Self-Reported Drug Data From De Novo Questionnaire Studies

This section summarizes what is known on how well respondents recall prescription and over-the-counter medication use, factors that influence recall, such as the type of medications being queried, as well as questionnaire design features suggested to improve recall accuracy.

Accuracy and Recall

Several studies have evaluated self-reported recall accuracy for current or past medication use compared with prospectively collected cohort data or pharmacy, hospital, and outpatient medical record documentation. Overall, published studies indicate that people accurately remember ever using a medication and when they first began using some medications, although they do not remember brand names and duration of use as well [90–98]. Current use of chronically used medications, such as statins, beta-blockers, and calcium channel blockers, was recalled with ≥95% sensitivity and specificity when a mailed medication inventory was compared to pharmacy records [99]. In general, greater inaccuracies have been noted as more time elapsed between occurrence of exposure and its subsequent reporting [91,95,97]; this tendency was especially true for over-the-counter NSAID use in contrast with prescription NSAID use for recall over a two-month period [100].

Accuracy of self-reporting medication use varies by several factors. For example, chronically used medications (especially those with more refills) are recalled more often than acute exposures, as are the first and most recent brands in a class; a person recalls multiple medications in one class more frequently than single medication exposure; and salient exposures (those that prompted study initiation) are more accurately recalled than common and less disconcerting exposures [90,91,96,101–105]. For prescription drugs, recall between self-reported use and medical records was moderately accurate, but for over-the-counter medications and vitamin supplements, accurate recall was poor [106]. Discrepancies are due to both underreporting (e.g., respondent forgot medication was taken) and underdocumenting (e.g., physician was unaware of medication use or did not record patient's use in chart) [79–81 92,101,103,105–108] and differed by therapeutic class [106,107,109–116]. When self-reported data were compared to multiple sources (e.g., medical records and pharmacy dispensing), verification for self-reported use was higher than that for a single source [117].

Influences on Accuracy

Influence of Questionnaire Design

As reported in a systematic review, several factors affect the accuracy of medication exposure reported via questionnaire [118]. Researchers can facilitate recall and reporting of medication use by indication-specific questions, memory prompts (such as drug photo), a list of drug names, or a calendar to record life events [70–74,93,119]. Medication-specific or indication-specific questions can identify most medications respondents are currently using, rather than a general medication question such as, "Have you taken any other medications?" [105]. Similarly, open-ended questions such as, "Have you ever used any medications?" yielded less than half of the affirmative responses for use of three different medications [120]. Using the filter question "Did you use any medications in the three months before or during your pregnancy?", van Gelder and colleagues noted that many women failed to report medications that they had been dispensed for pain or infections. These findings could be attributed to poor recall, but they may also be due to women having chosen not to take the dispensed medications [121]. If researchers choose to use open-ended medication questions, adding indication-specific questions that facilitate recall of medication exposures may be useful. Finally, 20–35% of respondents reported drug exposure only when asked medication (name)-specific questions [120].

Response order may affect recall, as noted with malaria medications when respondents had more than one episode of malaria [122]. Regardless of how frequently the medication is used for treating malaria in general, medications listed earlier in the response set tended to be selected more frequently than those listed later – a finding that may be related to satisficing, as discussed earlier [65].

A comparison of self-report of current and recent medication use (within the past two years) to pharmacy records of dispensed prescriptions for multiple drug classes found that the number of drug dispensings recalled was highest for cardiovascular medications (66%) and poorest for alimentary tract medications (48%) [123]. Recall was influenced by the number of chronically used medications: 71% for one drug, 64% for two drugs, and 59% for three or more drugs, although duration of use was not related to recall. However, the questionnaire did not allow sufficient space to record all medications used in the time period of this study. Thus, if respondents were unable to record all medications due to space limitations, a misleading validation might have occurred: it appeared that respondents were unable to recall all the medications dispensed according to the database.

Another methodologic study evaluated whether question structure influences the recall of currently used medications in 372 subjects with hypertension who had at least 90 days of dispensings in the PHARMO database [105]. The questionnaire had indication-specific questions first (e.g., medications used for hypertension, diabetes), followed by an open-ended question that asked if the subjects used any other medications not already mentioned. For hypertension, the sensitivity was 91% for indication-specific questions and 16.7% for open-ended questions. About 20% of subjects listed medications on the questionnaire that were not in the database, and a similar proportion failed to list medications on the questionnaire that were in use according to the pharmacy database. Based on these recall sensitivity results, indication-specific questions appear to invoke better recall accuracy. However, to adequately address the issue of question structure, a questionnaire could be designed to ask open-ended questions first, followed by indication-specific questions. This sequencing would allow a comparison of the number of medications recalled by each question structure.

Influence of Patient Population

Few studies have evaluated whether demographic and behavioral characteristics influence the recall of past medication use, but results to date suggest that recall does vary by these factors as well as by therapeutic class and study design. For example, research suggests that education attainment [104,108,124] and race/ethnicity [91,95] may affect recall accuracy. Studies are inconsistent for age [77,91,95–97,101,103,107,116], socioeconomic status [64,101,103,107,124], and smoking [95,97] as predictors of recall accuracy, and no study found that recall accuracy varies by gender [97,99,123]. The inconsistencies in the effect of age on recall accuracy might arise from differing study designs. The two studies that reported an age effect were methodologic studies evaluating recall accuracy [97,123], whereas the two that reported no age effects [91,95] were etiologic studies that reported verification of drug use as a measure of exposure misclassification for the association under study. Because of the paucity of information on predictors of recall, further research in this area is warranted.

Example

As indicated previously, accuracy of *de novo* questionnaire studies has been determined via comparison with pharmacy, general practitioner, and hospital records. To find an example of available study types, we conducted a literature scan of published studies, specifically searching for validation of NSAID use in questionnaire studies, and summarized our findings in Table 37.1.

Comparing use recalled during telephone interviews to a pharmacy database, West and colleagues found that 57% (95% CI 50–65%) of "any" NSAID use during the previous 12 years was accurately reported [97]. While a single dispensing was reported only 41% (95% CI 32–50%) of the time, repeated use was reported 85% (95% CI 76–94%) of the time, using the pharmacy records as the gold standard. Thirty percent of interviewees reported NSAID name and 15% reported both name and dose. Report was poorer with a shorter duration of use or over a longer recall period.

In summary, the methodologic literature on recall accuracy discussed above indicates that study participants have difficulty remembering drug use from the distant past, which contributes to misclassification of exposure in *de novo* studies. Researchers are using best practices in questionnaire design, including medication-specific and indication-specific questions, along with recall enhancements, which have been shown to produce better data. Calendars and photos of drugs augment recall to a greater degree than listing only the brand names of the

Table 37.1 Validation of NSAID exposure in studies using questionnaires.

Author	Recall period	Questionnaire and sample size	Study question	Memory aids	Comparison data source	Findings
West 1995 [97]	2–3 years 7–11 years	Telephone interviews n = 319	Nonsteroidal antiinflammatory drugs (NSAIDs)	Pictures of NSAIDs	Pharmacy database	Recall percentage for any NSAID use: 57 (95% CI 50–64) Single NSAID dispensed in 12 year period: 41 (95% CI 32–50) Repeated NSAID use: 85 (95% CI 76–94)
			For those with repeated NSAID use, a single NSAID was selected as the target drug for assessing name, dose, and dates of use			NSAID name: 30 (95% CI 24–36) NSAID name and dose: 15 (95% CI 10–20) Agreement: (a) ±6 months, (b) ±1 year, (c) ±2 years (a) (b) (c) First use 20 28 51 Last use 17 24 42 Duration 67 71 80
Smith 1999 [115]	Current use	Personal interview and medication inventory n = 55 users	Aspirin	None	Serum levels	0.16 (0.0–0.32)

drugs in question. These techniques – namely, photos, calendars, and the two different types of drug questions – have become the state of the art for collecting self-reported drug data by personal or telephone interview.

The literature to date suggests that recall accuracy of self-reported medication exposures is sometimes, but not always, influenced by type of medication, drug use patterns, design of the data collection materials, and respondent characteristics. Given the current state of the literature, epidemiologists who plan to use questionnaire data to investigate drug–disease associations will need to consider which factors may influence recall accuracy in the design of their research protocols.

Self-Reported Diagnosis and Hospitalizations from De Novo Studies

Accuracy and Recall

Just as recall accuracy of past medication use varies by the type of drug, the ability of respondents to remember disease conditions varies by disease, particularly when it is chronic, like hypertension, or is viewed as threatening, such as sexually transmitted infections. The best reporting has been noted with conditions that are specific and familiar, such as diabetes mellitus [113,125–131], hypertension [113,126,128,129,132], asthma [125,127,128], and cancers such as breast, lung, large bowel, and prostate [129,132–134]. However, assessing reporting accuracy is likely more difficult for common, recurring symptom-based conditions, such as sinusitis, arthritis, low back pain, and migraine headaches, which many people may have, or believe they have, without having been diagnosed by a clinician. For recall of acute conditions such as fractures, there is typically good agreement between self-report and the comparison data source, although the one methodologic study of fracture incidence indicated a slight tendency for overreporting of hand, finger, rib, or facial fractures [135], which might be attributed to confusing a fracture with other similar orthopedic problems like sprains and strains. Recall of acute conditions is likely to depend on the length of the recall period: mild traumatic brain injury that occurred prior to age 10 years was poorly recalled 15 years later [136].

Three studies assessed the recall accuracy for self-reported mental illnesses, comparing respondent information to clinical evaluation [127,128,137]. The results indicated poor agreement between the two data sources, with underreporting as the primary reason for poor agreement. It is unclear from these studies whether the reason for underreporting was the respondent's unwillingness to admit to mental illness or whether the conditions were actually underdiagnosed.

Both underreporting and overreporting of diagnoses have been noted in studies comparing self-reported diagnoses to clinical records [127,128], with overreporting occurring for conditions in which the diagnostic criteria are less explicit [138]. For common ailments, underreporting was often the major cause of disagreement [113,125,129,131]. Both overreporting and underreporting were noted for cardiovascular conditions, depending on the data source used for comparison [113,126,128,129,131,132,134,139–141]. In most instances of recall error, many respondents who had incorrectly reported myocardial infarctions (MIs) and stroke had other conditions that they may have mistakenly understood as coronary heart disease, MI, or stroke, based upon communication with their physician during their diagnostic visits [134,139–141].

Influences on Accuracy

Influence of Questionnaire Design

Questionnaire design also influences validity of disease and hospitalization data obtained by self-report. Simpler questions yield better responses than more complex questions, presumably because complex questions require the respondent to first comprehend what is being asked and then provide an answer. Inherent redundancy in longer questions and allowing more time to develop an answer to the question may increase recall [142]. However, longer questions could increase the cost of the research and could needlessly tire the respondents, leading to satisficing. Facilitating recall by providing respondents with a checklist of reasons for visiting the doctor improves recall of all medical visits [143].

Although specific guidance on best practices for improving the ascertainment of diagnoses and hospitalizations is lacking, there are several general approaches to questionnaire design that are useful (see, for example, Sudman and Bradburn [144] and McColl and colleagues [145] for further details). Briefly, researchers developing questionnaires should be mindful of question wording and sequencing and response formats. With regard to question wording, to increase response accuracy, questionnaire designers should attend to the cognitive processes involved in developing a response, especially those related to saliency for the respondent. Whether a respondent recalls having been diagnosed with a particular condition previously is likely to depend on the seriousness of the condition. Use of a filter question such as, "Have you had any side effects from use of drug X in the past year?" must be done with caution because respondents who avoid the filter are not asked subsequent questions that may be important for the study. As noted for recall of medications, open-ended questions are not recommended, particularly if the questionnaire is self-administered. That said, all potential response categories must be listed when using closed-ended questions or when an "other" category is provided. Questionnaire design experts suggest that demographic questions be placed at the end because they may be regarded as threatening.

The typical rule of thumb for question sequencing is to ask general questions before delving into specific topics and to group questions according to topic. When laying out the questions in a questionnaire, researchers should consider whether ordering effects are possible: for example, ask about heart disease in general before asking about a heart attack. Ordering might influence response rates to particular questions and may vary with the topic and make-up of the respondent population. With regard to response formats, the response categories should be unambiguous, nonoverlapping, and exhaustive. When there is a possibility of biased response due to response ordering, it is best to randomize the response options to minimize the bias. Finally, satisficing is also possible when respondents are asked to identify the diagnoses they have been given previously.

Influence of Patient Population

Factors influencing accuracy of past diagnoses and hospitalizations include the number of physician services for that condition and the recency of services [44–46,146–148]. For reporting of diagnoses, the longer the interval between the date of the last medical visit for the condition and the date of interview, the poorer the recall was for that condition [44–46]. These differences in recall may be explained in part by recall interval, patient age, a cohort (generational) effect, or some intertwining of all three factors. Diagnoses considered sensitive by one generation may not be considered as such by subsequent generations. Further, terminology changes over time, with prior generations using different nomenclature compared with recent generations.

Conditions with substantial impact on a person's life are more accurately reported than those with little or no impact. More patients with current restrictions on food or beverages due to medical problems reported chronic conditions that were confirmed in medical records than did those without these restrictions [44]. Similarly, those who had restrictions on work or housework reported their chronic conditions more often than those who did not have these restrictions [44]. The major determinant of recall for spontaneous abortions was the length of the pregnancy at the time the event occurred: nearly all respondents who experienced spontaneous abortions occurring more than 13 weeks into the pregnancy remembered them compared with just over half of respondents who experience such abortions occurring in the first six weeks of pregnancy.

Perhaps as a result of the emotional stress, lifestyle changes, and potential financial strain, hospitalizations tend to be reported accurately [147]. Further, underreporting of hospitalizations occurred in only 9% of patients who received surgery compared to 16% of patients without a surgical procedure. Underreporting in those with only a one-day hospital stay was 28% compared with 11% for 2–4-day stays and approximately 6% for stays lasting five or more days.

Surgical procedures are also more likely to be accurately recalled. General practitioner records confirmed 90% of the surgeries reported during one study interview. For the remaining 10%, the medical record may have lacked the needed information [149]. Recall of surgery date (±1 year) was correct for 87.5% of patients interviewed. Researchers also agree that respondents remember the type of surgery accurately [116,148–150]. Recall accuracy was very good for hysterectomy and appendectomy [110,125,129], most likely because these surgeries are both salient and familiar to respondents. Cholecystectomy [129] and oophorectomy [110] were not as well recalled and were subject to some overreporting. However, overreporting may have been due to the potential incompleteness of the medical records used for comparison [110]. For induced abortions, marginal agreement occurred, as noted by records from a managed care organization: 19% of women underreported their abortion history, 35% overreported abortions, and 46% reported accurately according to their medical record [151].

The influence of demographic characteristics on reporting of chronic illnesses has been evaluated in many studies, although the results are conflicting. The most consistent finding is that overall recall accuracy decreases with age [113,116,131,133,152], although this may be confounded by recall interval or cohort (generational) effects. Whether gender influences recall accuracy is uncertain. Men have been reported to recall more accurately than women, independent of age [125], whereas conflicting evidence found that women reported more accurately than men [127], especially in older age groups [44]. Further studies indicate that gender and age differences depended upon the disease under investigation [127], with women overreporting malignancies and men overreporting stroke [131]. No differences were found for reporting of hospitalizations by age or gender [147].

Reporting of illnesses, procedures, and hospitalizations tends to differ by race/ethnicity, but most studies had much larger proportions of whites than nonwhites [44,116,125,127,147,151]. Reporting by education level was equivocal; one study showed no difference [46] while another study indicated better recall for those with less education [44], and others suggested more accurate responses for those with a college education [131,133,135,151]. Those with a poor or fair current health status reported conditions more completely than those with good to excellent health status [44].

Although menarche and menopause are not medical conditions *per se*, the age at which they occur is often of interest in pharmacoepidemiologic studies. In the Menstrual and Reproductive Health Study, which had recall periods ranging from 17 to 53 years (mean 33.9 years), the exact age of menarche was recalled by 59%, and age within one year was recalled by 90% [153]. Similarly, for menopause, 45% of women were able to report their exact age at natural menopause and 75.5% reported age within one year. The percentage agreements for surgical menopause were 55.6% and 83.4%, respectively, for exact age and age within one year. The lower percentage agreement for age at which natural menopause occurred compared to that for surgical menopause may be attributed to the gradual occurrence of natural menopause compared to the definitive nature of hysterectomy [154].

Example

We conducted a literature scan of published studies searching for outcomes of MI and GI bleeding associated with use of NSAIDs to provide specific examples of validation and reliability studies for diagnoses (Table 37.2). Many of those identified were methodologic studies conducted specifically to determine the accuracy of the questionnaire; however, some of the accuracy assessments were embedded in empirical studies. Fourrier-Reglat and colleagues compared reported medical data from patient and prescriber self-administered questionnaires [155]. Myocardial infarction showed substantial agreement (kappa = 0.75; 95% CI 0.71–0.80), while upper GI bleeding had only slight agreement (kappa = 0.16; 95% CI 0.11–0.22) between the two reporting groups. When the prescriber data were used as the gold standard, patient reports of MI provided moderately complete data (sensitivity 77.7%; specificity 99.6%; positive predictive value [PPV] 77.1%; negative predictive value [NPV] 99.6%), and reports of upper GI bleeding by patients were not typically confirmed by the prescriber reports (sensitivity 44.6%; PPV 10.4%).

Jarernsiripornkul and colleagues also used a multistage process to develop a questionnaire to

Table 37.2 Validation of myocardial infarction (MI) or gastrointestinal (GI) outcomes in patients with nonsteroidal antiinflammatory drugs (NSAIDs) in questionnaire data.

Author	Questionnaire and sample size	Study question	Comparison data source	Conditions	Findings
Ambegaonkar 2004 [192]	Gastrointestinal Toxicity Survey (NSAID Induced) (GITS [NI]) – 11 questions, n = 400 patients	To test a new questionnaire designed to identify patients at high risk for NSAID-associated GI events	Stanford Calculator of Risk for Events (SCORE) – 6 questions	56.0% rheumatoid arthritis	The overall correlation between results for GITS (NI) responses and the total score for the SCORE questionnaire was 0.96 (P<0.001) Comparison: ordinary least square $R^2 = 0.91$ feasible generalized least squares (FGLS) $R^2 = 0.93$ Use of the FGLS regression analysis and comparison of the risk levels predicted by the SCORE questionnaire and the GITS (NI) questionnaire demonstrated a 79.8% agreement for all four risk categories and an 88.8% agreement when the two highest risk categories were collapsed into a single category The multinomial logistic regression (MNL) analysis showed agreement of 75.8% for four risk categories and an agreement of 86.8% for three risk categories For both methods, disagreement was equally distributed among overprediction and underprediction of risk levels by the GITS (NI) questionnaire relative to the SCORE questionnaire. In the case of four risk categories, disagreement by two risk levels was limited to 0.6% and 1.5% for the FGLS and MNL regression methods, respectively

(Continued)

Table 37.2 (Continued)

Author	Questionnaire and sample size	Study question	Comparison data source	Conditions	Findings
Fourrier-Reglat 2010 [155]	CADEUS cohort (French national cohort study of traditional NSAIDs and COX-2 users conducted between September 2003 and August 2004 that employed self-administered questionnaires to obtain medical data from patients and their prescribers) n = 18 530 pairs of patients and prescribers	To compare patients and prescribers reported medical data	Prescribers report as gold standard		Previous medical history: MI: kappa = 0.75 (95% CI 0.71–0.80) Sensitivity: 77.7% Specificity: 99.6% PPV: 74.1% NPV: 99.6% Upper digestive hemorrhage: kappa = 0.16 (95% CI 0.11–0.22) Sensitivity: 44.6% Specificity: 98.5% PPV: 10.4% NPV: 99.8% NSAID indication: For index NSAID indication, the proportion of agreement ranged from 84.3% to 99.4% and concordance was almost perfect (kappa = 0.81–1.00) for inflammatory rheumatism, flu-like symptoms, dysmenorrhea and dental pain; substantial for arthritis, back pain and headache; moderate for osteoarticular pain.
Singh 1996 [193]	Stanford Health Assessment Questionnaire (HAQ)	To evaluate the event rates for all NSAID-induced GI complications in patients with rheumatoid arthritis, describe the time course of these events, and evaluate the role of prophylactic therapy with antacids and H2 receptor antagonists	Face validity and hospital records (2.4% hospitalized)		Face validity has been studied by surveying patients to ensure their understanding of the symptoms that are listed in lay language on the questionnaire; appropriate modification of the confusing symptoms has been made. Patient recall and accuracy in reporting side effects have been evaluated by repeat questionnaire administration, interview, and review of physician records. To minimize underreporting by patients, those events that are severe enough to require hospitalization are also ascertained by record review of all hospitalizations.

CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.

have patients self-report potential adverse reactions to NSAIDs [156]. The questionnaire was designed to elicit adverse effects that would be reported to health professionals, to determine how well patient report compared with health professionals reporting to the Adverse Product Reaction Monitoring (APRM) Centre of the Thai FDA. The questionnaire was cognitively tested to finalize the version sent to the test sample along with pictures to facilitate recall. Of the 694 (42%) of questionnaires returned, 60% reported ≥1 symptom deemed as a possible or probably adverse drug reaction by a pharmacist. By comparison, only 5% of the self-reported symptoms indicative of adverse events from the questionnaires were recorded in the outpatient medical records.

These examples demonstrate the variation in methods used to collect and determine accuracy of questionnaire data. Although many methods are available for use, researchers should remember the principles discussed earlier in the chapter when they validate questionnaire data: not all validation is equivalent. Full disclosure of the process is important when reporting findings of any study.

In summary, the decision as to whether a person reports an illness during an interview appears to be related to age and type of illness, when it occurred, and its saliency, but is less likely to be mediated by demographic characteristics such as gender, race, and education. Illnesses that are considered embarrassing and that do not substantially alter the person's lifestyle are not reported completely, and these types of illnesses may change with each generation. Likewise, reporting accuracy depends on the consistency of documentation and the terminology utilized – from the questionnaire to the medical records – and finally, what has been communicated to the individual. Although difficult to measure, respondent motivation appears to influence the completeness of reporting as well [44,127,147].

Conducting Validation Studies to Assess Data Collected During Provision of Healthcare

In addition to conducting *de novo* studies to evaluate drug–disease associations, a variety of computerized, administrative claims, and EHR databases are available for pharmacoepidemiologic research, the structure, strengths, and limitations of which are reviewed in Chapters 14–14. One major advantage of using such databases for pharmacoepidemiologic research is the comparative validity of the drug data in lieu of questionnaire data, where recall bias is always a concern, as previously described.

In general, the administrative claims and EHR differ widely on many factors, such as size (e.g., from several hundred thousand to several million covered lives), number of health insurance plans and health systems included, the type of health services provided and therefore available for analysis (e.g., prescriptions, mental health benefits, general practice versus specialist visit data), inclusion of out-of-plan claims in the main database versus other databases, and the timeliness of the data (e.g., the lag for cleaning and obtaining data from a data vendor may be six or more months). The databases also differ on the number of available demographic variables: all have age (some may have date of birth) and sex, EHRs may have race (but administrative claims typically do not), or a measure of health status [157]. Because the administrative claims data were developed primarily for reimbursement, they all have relatively complete data on health service use and charges that are covered by the plan (and relatively incomplete data for services not covered by the plan). EHRs provide in-depth, granular data for a specific office or hospital visit but may not provide all health information for an individual in a longitudinal fashion, especially if the patient sees more than one healthcare provider. Linkage of EHRs and administrative claims can be resource intensive but may elucidate whether the data sufficiently capture the patient experience.

The drawbacks and limitations of these data systems are important to keep in mind. Their most critical limitation for pharmacoepidemiologic research is the manner in which health insurance is currently covered in the US, typically through the place of employment. If the employer changes plans, which may occur on an annual basis, or the employee changes among the plans offered by the employer, or the employee changes jobs, the plan no longer covers that employee or his or her family. In addition, the healthcare delivery system coverage for an employer may change over time. Thus, the continual enrolment and disenrolment of plan members hinder the opportunity for extended longitudinal analyses in both administrative claims and EHRs.

Best Practices for Validation Studies in Administrative Claims or EHR Databases

For the data in administrative claims or EHRs to be considered valid, people who appear in the computerized files as having a drug exposure or disease should truly have that attribute and those without the exposure or disease should truly not have the attribute. Validity and completeness are determined by comparing the database information with other data sources, such as comparison of paper medical records or EHRs, administrative claims, pharmacy dispensings, or procedure logs. Choice of an appropriate comparator varies by study question, variables used for the research study, the comparator, and availability of other data sources.

The study investigator must be aware of the limitations of both the administrative claims database and the chosen comparison dataset. The chosen comparator should provide sufficient data to validate both the exposure and outcome used for the study. A variable that provides linkage between the files in a data source, such as a medical record number, should be available so that accuracy can be evaluated within a subset of known study patients. For example, if a single claim contains six diagnosis codes and six months of claims were used to determine outcomes in patients, then all six diagnosis codes for all claims across the six-month study should be available in a comparison dataset to establish the validity of the outcome. As described earlier in the chapter, a validation assessment should include evaluation of patients with and without the exposure or outcome. Positive predictive value, negative predictive value, sensitivity, and specificity combined provide a complete picture of the agreement between the two data sources.

The following is a broad overview of how to conduct a validation study in administrative claims or EHR data. First, choose a meaningful number of patients for validation. This sample size should be statistically grounded; however, considerations of data availability, cost, and labor are understandable. Next, extract the variables needed to determine cohort selection, exposure, outcome, and other variables for validation. Calculate measures of agreement and error rates (e.g., standard deviations) between the two datasets. Finally, consider strengths and limitations of the two datasets to ascertain validity and completeness of the data source to answer the study question.

Influence of Data System

Completeness and validity of data are the most critical elements in the selection of a database for research. Completeness is defined as the proportion of all exposures, events of interest or both that occurred in the population covered by the database that appear in the computerized data. Missing subjects, exposures, or events could introduce bias in the study results [158]. For example, completeness of the drug data might vary by income level if persons with higher incomes and drug co-payments choose to obtain their medications at pharmacies not participating in a prescription plan, which is how pharmacy data are collected. Similarly, a bias may be introduced in the association between a drug and a serious adverse drug reaction if hospitalizations for that adverse reaction are missing, for example if the researcher only has access to the outpatient clinic EHR database.

Influence of Clinical Coding Systems

Diagnoses, procedures, medications, and other therapeutics are included in administrative claims and EHR data through structured coding systems. Each coding system has its own ontology and is separated into specific codes, based on an established hierarchy. Further, the coding systems are updated periodically to reflect changes in the practice of healthcare as well as to incorporate new therapies and processes. Both codes and the general structure and hierarchy differ between coding systems. In many cases, a single code from a coding system is insufficient to define a variable and an *algorithm* needs to be developed. The algorithm may contain multiple codes, a required timing for codes, and/or a sequential process for determining the level(s) of the variable. It is likely that an algorithm developed in one coding system will require translation to be comparable to another coding system. Algorithms for each clinical concept should be developed and validated separately

The International Classification of Diseases (ICD) is the standard for classification of diseases for clinical and research purposes [159] and is used in many administrative claims systems, such as for billing purposes. The ICD is updated periodically, and adoption is asynchronous by country. For example, through the fall of 2015, most US administrative claims systems were using the clinical modification version of ICD-9, while many European administrative claims systems began using ICD-10 in the 1990s, and ICD-11 was released in 2018. As with the transition to ICD-10, we anticipate a staggered approach to implementation, with countries in Europe and Canada adopting the ICD-11 system (long) before the US. The ontology differs between the ICD systems, and codes have been mapped between ICD-9 and ICD-10 [160]. However, there is not a one-to-one correlation between codes; there are approximately 14 000 diagnosis codes in ICD-9 compared with approximately 70 000 in ICD-10 [161]. Mapping between codes can be used as a starting point to develop algorithms [162], but various techniques in mapping may yield different results [163]. As with other coding systems, any validation should be conducted separately between ICD systems.

Influence of Structured and Unstructured Healthcare Data in Computerized Databases Containing Administrative Claims or EHR data

In addition to the structured data in administrative claims and EHR, many components of healthcare are captured within clinician notes, images and descriptions of procedure results, and other unstructured data. The performance of an algorithm can be enhanced through use of this unstructured information in addition to the structured data from coding systems. Unstructured data can be converted into structured information (e.g., manually) for a specific project, or the algorithm can be modified to improve algorithm performance as cases are identified over time [164,165]. Liao and colleagues compared the performance of algorithms including unstructured data to detect coronary artery disease to algorithms using only structured elements in the same data source to assess validity in three chronic disease patient cohorts [166]. They found that inclusion of unstructured data increased sensitivity in all three cohorts, with the most improvement seen in the cohort where coronary artery disease prevalence was lowest. Note that while previous algorithms are sometimes used for comparison [165], patient charts (electronic or paper) are still often used as the reference standard for assessing validation [167].

Influence of Distributed Data Systems

Multiple health data sources may be included within a single study or for ongoing surveillance.

Simultaneous assessment of multiple data sources allows for better understanding of a larger population while also observing a diverse set of patients [168]. These multidatabase studies or distributed data systems may have differences in information collected, coding systems, language (e.g., across different countries), and even the underlying practice of medicine and overarching system of healthcare. Thus, even in the situation where distributed data systems use a common data model, careful consideration is warranted regarding how to assess validity of drugs, other therapeutics, diagnoses, procedures, and health-related events within each administrative claims data source contributing to the distributed data system. Whenever EHR data are utilized, differences across sites warrant assessment of validity within each health system to improve overall accuracy [167].

Validity of Drug and Other Medical Intervention Data in Administrative Claims or EHR Databases

Accuracy

Drug data in administrative claims databases are often not validated. Administrative claims data contain billing of a prescription that is dispensed (i.e., "filled") but do not contain information on the provider writing the prescription (or on the underlying condition the prescription is intended to treat). While prescriptions that are dispensed but unclaimed by patients should be removed from billing, they may remain within the administrative claims data. Furthermore, dispensing data cannot address drug ingestion or adherence, and over-the-counter medications are not typically included in the database at all. Thus, despite the widespread use of claims data to assess drug use, the data may not be accurate and validity should be tested, particularly when using a new drug exposure or database (e.g., some data sources may not contain a drug because it is not "on formulary" and thus is unavailable within the health system or allowed by the health insurer). Similarly, sensitivity analyses should be performed to determine the susceptibility of the results to possible misclassification, even within known data sources.

Unclaimed prescriptions, estimated to occur for approximately 2% of all prescriptions, present an adherence issue in administrative claims data [169]. For every 1000 new prescriptions, an average of 16.5 are unclaimed [170]. Antiinflammatory and antiinfective drugs tend to be the therapeutic class most often unclaimed [170–173]. Two-thirds of unclaimed prescriptions were for new prescriptions [171], and a similar proportion tended to be for nonessential medications [174]. Many unclaimed prescriptions were telephoned into the pharmacy [169,171], and the most frequent reason patients cited for not picking up a prescription was that they determined that they did not need the medication or they forgot to pick it up [169,172]. However, cost and having a similar medication at home were also often cited [169,172,174].

Drug data in EHRs represent the actual prescribing practice. EHR data account for the written prescription and may have sufficient detail to ascribe the prescription to the underlying condition it is intended to treat. Prescriptions or other documentation may also be available in the EHR for over-the-counter medications. However, they may not present a valid picture of the patient experience with the medication. The dispensing, ingestion, adherence, and pattern of use are typically not included as structured fields. Some of this information may be available in unstructured text such as the clinician visit notes.

Influences on Accuracy

Population and Representation in Data Source

One might ask how unclaimed prescriptions might affect the validity and completeness of pharmacy data. Many individuals have some type of pharmacy benefits plan in which reimbursement for medication costs is processed through a third-party payer. Entry into the reimbursement software is predicated on dispensing of the drug. However, a drug that is dispensed but not claimed should be returned to stock and the appropriate adjustment be made to the patient's pharmacy benefits plan – failure to do this would be insurance fraud.

Unfortunately, when conducting research with pharmacy data, we do not know whether all such insurance adjustments have been made. So while we believe a substantial number of prescriptions were dispensed, they may not have been used at all. To the extent that dispensings in the database were not picked up, there is no chance that the individual had the drug exposure and our study would suffer from exposure misclassification. Exposure misclassification can occur even when dispensings were picked up but not actually used by patients. For these reasons, some researchers require a minimum of two dispensings for assessing patient exposure to chronic medications. This rule of thumb is thought to improve to the likelihood that the drug was taken by the patient.

Example

A handful of studies to date have assessed dispensing associated with prescriptions via linked administrative claims and EHR data. These studies indicated that 70–77% of initial prescriptions are dispensed [175,176]. Prescribed analgesics (i.e., pain medications, including NSAIDs) and lifestyle drugs (e.g., phosphodiesterase type 5 inhibitors) are least likely to be dispensed, while antimicrobials are most likely to be dispensed for an initial prescription. Substantial variation in dispensing was seen across medications within a class. In addition, results from Rowan suggest that <20% of patients taking analgesics and NSAIDs possessed adequate medication to be adherent throughout a 12-month period [176]; this finding may be consistent with intermittent or "as needed" utilization.

In summary, drug and medical intervention data are often considered to be correct when using administrative claims data and EHRs for research. Although this is generally the case, researchers should be aware of whether and how prescribing, dispensing, and administration of drugs are captured within each database they are contemplating using. We will likely see greater emphasis on data linkage and incorporation of more unstructured data from clinical notes into pharmacoepidemiologic research, which may lead to increased need for validation of drug and medical intervention exposures in the future.

Validity of Diagnosis, Procedure, and Hospitalizations in Administrative Claims and EHR Databases

Accuracy

Unlike the drug data, where many researchers are comfortable with data accuracy and completeness, inpatient and outpatient diagnoses in these databases raise considerable concern for investigators. The accuracy of outpatient diagnoses is more uncertain than inpatient diagnoses for several reasons. Hospitals employ experienced people to code diagnoses for reimbursement, which may not occur in individual physicians' offices where outpatient diagnoses are determined. Also, hospital personnel scrutinize inpatient diagnoses for errors [177], monitoring that does not typically occur in the outpatient setting.

Systematic errors as a result of diagnostic coding may influence the validity of both inpatient and outpatient diagnostic data. For example, diseases listed in administrative claims databases are often coded using the ICD coding system. Poorly defined diseases are difficult to code using the ICD system, and no way exists to indicate that an ICD code is coded for "rule-out" purposes. How healthcare plans deal with "rule-out" diagnoses is unclear; for example, should they be included or excluded from the diagnoses in the physician claims files? In a study of transdermal scopolamine and seizure occurrence, many patients with ICD codes indicating seizures had this diagnosis as a "rule-out" code when medical records were reviewed to confirm the diagnosis, indicating that “rule-out” codes do become part of administrative claims data [178]. In addition, reimbursement standards and patient insurance coverage limitations may influence the selection of ICD codes for billing purposes [179]. The potential for abuse of diagnostic codes, especially outpatient codes, may occur when physicians apply to either an insurance carrier or the government for reimbursement and may be less likely in staff or group model health maintenance organizations (HMOs) such as Kaiser Permanente.

Influences on Accuracy

Validation Study Design

Abstraction of electronic data for validation studies is not subject to the issues of questionnaire design that are present with self-reported *de novo* studies; however, manual abstraction is subject to human error. Algorithms that are complex require substantial understanding of the healthcare environment in which the data were collected and necessitate review of lengthy portions of the patient chart, which may increase risk of error during record abstraction. Understanding of each specific healthcare system may be warranted to understand nuances of documentation practices. In one study, medical record documentation within a single multispecialty medical group showed that documentation varied across measures (e.g., medications documented 92% of the time, smoking history documented 38% of the time, and drug allergies documented in 62% of encounters) [180]. While no systematic patterns were noted across clinician and patient characteristics, differences in documentation were found between internists and pediatricians as well as between male and female providers.

Population and Representation in Data Source

At an institutional level, informaticists in the US have been concerned about the completeness of EHR data for research use [181,182]. Patient information in an EHR may be considered complete if it has sufficient detail regarding clinical encounters, if ongoing encounters are included over calendar time, if multiple types of data (e.g., labs, medications, and diagnoses) are available, and/or if sufficient information is available across a patient record to predict the condition of interest. In 2013, Weiskopf *et al.* reviewed all four of these definitions within a single healthcare system data warehouse in the United States and found that 26.9% of patients had complete records according to any one of the four definitions (8.4–18.5% of patient records were complete for each measure), and only 0.6% of patients had complete records according to all of these definitions of completeness [182].

Example

Continuing with the NSAID example, we conducted a literature scan of published studies validating MI or GI bleeding outcomes with use of NSAIDs in administrative claims databases; these studies are summarized in Table 37.3. Administrative claims data are often compared with medical records in a validation study. Most of these studies provide only a PPV that indicates whether the coding scheme is accurately classifying observed measures compared with another source. Validation measures such as sensitivity and specificity are not often calculated in these comparative studies.

In claims data, MI, denoted as ICD-9-CM code 410.xx, has been assessed in computerized health databases of Quebec [183], Saskatchewan Health [184], and the HealthCore® Integrated Research Database [185]. In all of these databases, this ICD-9-CM code had substantial or nearly perfect ability to validate the diagnosis of MI, with the PPV ranging from 88.4% to 96% across studies. Other ICD-9-CM codes used for possible detection of MI have shown poor ability to classify MI.

Both the overall PPV for ICD-9-CM 410.xx to measure MI and the PPV for MI among patients taking NSAIDs were evaluated in the HealthCore® Integrated Research Database. Among all the patients with a code for MI, the PPV was 88.4% (95% CI 83.2–92.5%). Among patients taking

Table 37.3 Validation of myocardial infarction (MI) and gastrointestinal (GI) bleeding events in studies using administrative claims data to evaluate harms of nonsteroidal antiinflammatory drug (NSAID) exposure.

Author	Dataset	Study aim and sample size	Comparison data source	Conditions	Findings
Abraham 2006 [188]	VA	To validate Veterans Affairs (VA) administrative claims data for the diagnosis of NSAID-related upper gastrointestinal events (UGIE) and to develop a diagnostic algorithm n = 906 ICD-9-CM codes and CPT procedure codes in patient treatment and outpatient care databases indicating upper gastrointestinal events (n = 606) Controls (n = 300)	Medical records	Case definition for UGIE was any of the following: Gastric ulcer 531.0, 531.1, 531.2, 531.3, 531.4, 531.5, 531.6, 531.7, 531.9 Duodenal ulcer 532.0. 532.1, 532.2, 532.3, 532.4, 532.5, 532.6, 532.7, 532.9 Peptic ulcer 533.0, 533.1, 533.2, 533.3, 533.4, 533.5, 533.6, 533.7, 533.9 Gastrojejunal ulcer with perforation 534.0, 534.1, 534.2, 534.3, 534.4, 534.5, 534.6, 534.7, 534.9 Gastrointestinal hemorrhage 578.0, 578.1, 578.9	Only ICD-9 codes for UGIE: Sens: 100% Spec: 96% PPV: 27% NPV: 100% ICD-9 and CPT for UGIE: Sens: 82% Spec: 100% PPV: 51% NPV: 99% ICD-9 and CPT algorithm for UGIE: Sens: 66% Spec: 88% PPV: 67% NPV: 88% Algorithm validated in additional 44 patients, PPV among NSAID users: 80%
Brophy 2007 [183]	Computerized health databases of Quebec, Canada	To determine whether a history of MI modified the risk of acute MI associated with the use of various NSAIDs n = 234 MI survivors	Previous validation of MI claims [194]; no validation of NSAID use	MI: hospitalization with ICD-9 code 410, considered fatal if person died within 30 days of admission	PPV = 0.96 (95% CI 0.94–0.98)

(*Continued*)

Table 37.3 (Continued)

Author	Dataset	Study aim and sample size	Comparison data source	Conditions	Findings
Castellsague 2009 [195]	Saskatchewan Health	To estimate the risk of upper gastrointestinal complications associated with use of cyclooxygenase-2 (COX-2) selective (celecoxib and rofecoxib) and individual nonselective NSAIDs compared with nonuse of these drugs Specific codes: n = 38 (10% sample) Nonspecific codes: n = 742 (all potential cases)	Medical records	Upper gastrointestinal complications: ICD-9 codes 531.0–531.2, 531.4–531.6, 532.0–532.2, 532.4–532.6, 533.0-533.2, 533.4–533.6, 534.0–534.2, 534.4–534.6, 569.3, 569.4, 569.8, 578	Previous research: PPV for site- and lesion-specific peptic ulcer disease codes in Saskatchewan = 91% PPV for nonspecific codes = 68% This study: Specific PPV = 92% Nonspecific code PPV (ranged across codes) = 60% for unspecified hemorrhage, 4% for hemorrhage of rectum/anus
Curtis 2008 [196]	Medicare	To evaluate the feasibility of adapting data mining methods using the empirical Bayes Multi-item Gamma Poisson Shrinkage (MGPS) algorithm to longitudinal administrative claims data Number not specified	Public use data files supplemented with specific medication data from CMS for greater precision in defining current NSAID exposure	Linked survey information, medical claims, and medication use data from the Medicare Current Beneficiary Survey (MCBS) for the years 1999–2003	"Identified current NSAID exposure using the MCBS medication data and all medical events using the linked Medicare claims"
van Staa 2009 [197]	GPRD	To evaluate the external validity of published cost-effectiveness studies by comparing the data used in these studies to observational data from actual clinical practice and whether these studies should have been used to inform prescribing policies. Selective COX-2 inhibitors (coxibs) and upper GI events were used as an example n = 96	Medical records	Upper GI events: ICD-10 codes K25–K29 NSAIDS: any prescription in GPRD	PPV = (95/96) = 99.0%

Varas-Lorenzo 2009 [184]	Saskatchewan Health	To evaluate risk of fatal and nonfatal acute MI with NSAID use n = 200	Medical records	ICD-9 code 410–414, 427.5, 798 ICD-10 code I20–I22, I23.3, I24–I25, I46, R96.0, R96.1, R98 Abstraction items included available information on cardiac symptoms; copies of available electrocardiograms recorded during the first 72 hours after hospital admission and the last one before hospital discharge; serum biomarkers levels: troponins, CPK-MB, or CPK measured within first 72 hours and compared with later measures; necropsy and other cardiac diagnostic test findings. Based on abstracted information, two cardiologists classified events as definite or probable/possible (either fatal or nonfatal) according to adapted standardized criteria recently adopted by American Heart Association/European Society of Cardiology. Classification of exposure to NSAIDs was based on the days between the index date and the end of supply of the most recent dispensing before the index date	PPV for ICD-9 code 410 = 0.95 (95% CI 0.91–0.98) PPV for ICD-9 code 411 for intermediate coronary syndrome = 0.73 (95% CI 0.70–0.77) PPV for ICD-9 code 411 for AMI = 0.09 (95% CI 0.07–0.11)

(*Continued*)

Table 37.3 (Continued)

Author	Dataset	Study aim and sample size	Comparison data source	Conditions	Findings
Wahl 2010 [185]	HealthCore® Integrated Research Database	To validate administrative claims codes with medical chart review for MI, ischemic stroke, and severe upper gastrointestinal (UGI) bleed events in a large, commercially insured US population n = 200 charts per outcome	Medical charts	MI: ICD-9 code 410.xx excluding 410.x2 and a length of stay (LOS) between 3 and 180 days, or death if LOS is <3 days Severe UGI bleed events were defined as a hospitalization for either UGI hemorrhage or peptic ulcer disease, including perforation. In the claims data, this was defined as ICD-9 codes 531.x, 532.x, 533.x, 534.x, 578.0, 578.1, 578.9, or a physician service code for GI hemorrhage (CPT code 43255 or ICD-9 procedure code 44.4x)	Overall: PPV for MI = 88.4% (177/200; 95% CI 83.2–92.5%) PPV for ischemic stroke = 87.4% (175/200; 95% CI 82.0–91.7%) PPV for severe UGI bleed = 56.5% (109/193; 95% CI 49.2–63.6%) Among those taking NSAIDs: PPV for MI = 92.3% (97/105; 95% CI 85.4–96.6%) PPV for ischemic stroke = 78.9% (57/72; 95% CI 67.6–87.7%) PPV for severe UGI bleed = 57.9% (70/121; 95% CI 48.5–66.8%)

NSAIDs, the PPV for MI was 92.3% (95% CI 85.4–96.6%). The difference between the overall PPV and PPV among patients taking NSAIDs could highlight the potential for differential coding by patient status. Further study of differences in diagnosis coding by medication or disease status is needed to know whether validating the drug and disease pair is warranted or whether validation of the exposure and outcome separately is sufficient to imply veracity of the results.

A substantial proportion of cases identified by algorithms for probable or definite MI within all databases are confirmed as probable or definite MI in medical records, with PPV ranging from 55% to 97%. Validity for MI has been measured in the Group Health Cooperative (now Kaiser Permanente Washington) [186] (sensitivity 86.5%; specificity 85.4%) and in the General Practice Research Database [187] (sensitivity 89.3%), with substantial agreement between the administrative claims and medical records.

Measurement of GI bleeding is more varied across databases, and several algorithms using different combinations of ICD-9-CM and CPT (Current Procedural Terminology) codes have been used to determine event occurrence. The PPV for the studies range from 60% to 100%. In general, the PPV has been higher when both ICD-9-CM and CPT codes are used. However, in the US Department of Veterans Affairs (VA) administrative claims data, where sensitivity and specificity were also assessed, the higher PPVs with use of both coding systems resulted in a lower sensitivity and specificity [188]. Limiting further to only those patients using NSAIDs, the PPV increased to 80%. Both the overall PPV for severe GI bleeding and the PPV for GI bleeding among patients taking NSAIDs were determined in the HealthCore® Integrated Research Database [185]. Among all patients with an ICD-9-CM or CPT code indicative of GI bleeding, the PPV was 56.5% (95% CI 49.2–63.6%). Among patients taking NSAIDs, the PPV for GI bleeding was 57.9% (95% CI 67.6–87.7%).

The variation seen in comparisons of GI bleeding in administrative claims and EHR may be due to the differences in algorithms used to determine GI bleeding. The variation may be due also to differences in GI bleeding in the underlying populations captured in each database. Validation, including measures of sensitivity and specificity, of the same algorithm in multiple databases will aid in determining whether GI bleeding can be adequately assessed in administrative claims and/or EHRs.

In summary, validating the case definition developed for observational studies using administrative claims databases with original documents such as inpatient or outpatient medical records is an important step to enhance the quality and credibility of the research. Although many studies in the past few years have reviewed original documents to validate the diagnoses under study or have referenced those validation studies, a need still exists for validation of drug exposures and disease diagnoses in databases in which no previous validation has been performed. As medical practice changes over time, further validation of previously validated claims is also warranted. Evaluating the completeness of the databases is much more difficult, as it requires an external data source that is known to be complete [143,187,189,190]. Although administrative claims and EHR databases have greatly expanded our ability to undertake pharmacoepidemiologic research, we need to ensure that our tools, including the databases used for our analyses, are complete and of the highest quality.

The Future

Methods for conducting pharmacoepidemiologic studies have shifted over the past several decades from reliance on studies requiring *de novo* data collection from individuals, to extensive use of electronic data from either administrative claims or EHRs, to linked data sources and distributed data networks. Yet *de novo* data collection will continue to be required to ascertain information on quality of life, patient-reported outcomes (see Chapter 42), and medications either not included in pharmacy dispensing files or not reliably entered into EHRs, such as herbal and over-the-counter medications. In fact, with the advent of wearables and the Internet of Things, we anticipate that *de novo* collection of health data may increase in the coming years.

The improved computer technology that resulted in faster processor speeds and increased storage capacity facilitated storage of healthcare data in an electronic format, such as EHRs, and allowed development of distributed data networks using data from multiple health plans. The availability of these data for research has improved our ability to conduct studies [168] and the increasing uptake of EHRs is leading to increased availability of more granular clinical data for pharmacoepidemiologic research (e.g., lab results, clinical notes). Initial evaluation of EHR data suggests great promise, but increased data quality and standardization of terminology and codes will be required to make these data, collected for clinical care, useful for research purposes [191]. Similar processes will be warranted for use of data from wearables and prior to integration of new data from biobanks, mobile apps, social media, or other sources into a rigorous research framework.

As part of the standardization process, data holders will have to document that their data are valid for conducting research and surveillance activities. This will require investigators to apply their knowledge and practices from use of administrative claims and EHR data to linked data and to these novel data sources. Both medication exposure and outcome diagnosis data from these novel data sources do not carry the same level of comfort regarding validity as claims data and EHR data. As these data are considered for research, we hope and expect to see studies validating their use.

References

1 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**(1): 159–74.

2 Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. *Am J Epidemiol* 1993; **137**(11): 1251–8.

3 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; **126**(2): 161–9.

4 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**(8476): 307–10.

5 Kelsey JL, Thompson WD, Evans AS. Methods in observational epidemiology. In: *Monographs in Epidemiology and Biostatistics*, vol. **10**. New York: Oxford University Press, 1986.

6 Jurek AM, Maldonado G, Greenland S, Church TR. Exposure–measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Epidemiol* 2006; **21**: 871–6.

7 Chavance M, Dellatolas G, Lellouch J. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *Int J Epidemiol* 1992; **21**(3): 537–46.

8 Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol* 1990; **132**(4): 746–8.

9 Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol* 1991; **134**: 1233–44.

10 Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 1992; **3**(3): 210–15.

11 Sorahan T, Gilthorpe M. Nondifferential misclassification of exposure always leads to an underestimate of risk: an incorrect conclusion. *Occup Environ Med* 1994; **51**: 839–40.

12 Wacholder S, Dosemeci M, Lubin JH. Blind assignment of exposure does not always prevent differential misclassification. *Am J Epidemiol* 1991; **134**: 433–7.

13 Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol* 2005; **34**(3): 680–7.

14 Maldonado G, Greenland S, Phillips C. Approximately nondifferential exposure misclassification does not ensure bias toward the null. *Am J Epidemiol* 2000; **151**: S39.

15 Rothman KJ, Greenland S, Lash TL. Validity in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 2008, pp. 144–5.

16 Greenland S. Multiple bias modelling for analysis of observational data [with discussion]. *J Roy Stat Soc Ser A* 2005; **168** (2): 267e306.

17 Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 2008.

18 Jurek AM, Maldonado G, Greenland S, Church TR. Uncertainty analysis: an example of its application to estimating a survey proportion. *J Epidemiol Commun Health* 2007; **61**: 650–4.

19 Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer, 2009.

20 Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014; **43**: 1969–85.

21 Leamer EE. Sensitivity analyses would help. *Am Econ Rev* 1985; **75**: 308–13.

22 Maldonado G. Adjusting a relative-risk estimate for study imperfections. *J Epidemiol Commun Health* 2008; **62**(7): 655e63.

23 Phillips C, Maldonado G. Using Monte Carlo methods to quantify the multiple sources of error in studies. *Am J Epidemiol* 1999; **149**: S17.

24 Phillips CV. Quantifying and reporting uncertainty from systematic errors. *Epidemiology* 2003; **14**(4): 459–66.

25 Jurek AM, Maldonado G, Greenland S. Adjusting for outcome misclassification: the importance of accounting for case–control sampling and other forms of outcome-related selection. *Ann Epidemiol* 2013; **23**: 129–35.

26 Horwitz RI, Feinstein AR. Alternative analytic methods for case–control studies of estrogens and endometrial cancer. *N Engl J Med* 1978; **299**(20): 1089–94.

27 Hutchison GB, Rothman KJ. Correcting a bias? *N Engl J Med* 1978; **299**(20): 1129–30.

28 Greenland S. A mathematic analysis of the "epidemiologic necropsy". *Ann Epidemiol* 1991; **1**(6): 551–8.

29 Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* 2003; **14**(4): 451–8.

30 Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal* 2001; **21**(4): 579–83.

31 Phillips CV, LaPole LM. Quantifying errors without random sampling. *BMC Med Res Methodol* 2003; **3**: 9.

32 Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. *J Am Stat Assoc* 2003; **98**: 47–54.

33 Steenland K, Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *Am J Epidemiol* 2004; **160**(4): 384–92.

34 Graff JJ, Sathiakumar N, Macaluso M, Maldonado G, Matthews R, Delzell E. The effect of uncertainty in exposure estimation on the exposure–response relation between 1,3–butadiene and leukemia. *Int J Environ Res Public Health* 2009; **6**: 2436–55.

35 Jurek AM, Lash TL, Maldonado G. Specifying exposure classification parameters for sensitivity analysis: family breast cancer history. *Clin Epidemiol* 2009; **1**: 109–17.

36 Jurek AM. Maldonado G. Quantitative bias analysis in an asthma study of rescue–recovery workers and volunteers from the 9/11 World Trade Center attacks. *Ann Epidemiol* 2016; **26**: 794–801.

37 Jurek AM, Maldonado G, Spector LG, Ross JA. Periconceptional maternal vitamin supplementation and childhood leukaemia: an uncertainty analysis. *J Epidemiol Commun Health* 2009; **63**: 168–72.

38 Maldonado G, Delzell E, Tyl RW, Sever LE. Occupational exposure to glycol ethers and human congenital malformations. *Int Arch Occup Environ Health* 2003; **76**: 405–23.

39 Scott LLF, Maldonado G. Quantifying and adjusting for disease misclassification due to loss to follow-up in historical cohort mortality studies. *Int J Environ Res Public Health* 2015; **12**: 12834–46.

40 Gordis L. Assuring the quality of questionnaire data in epidemiologic research. *Am J Epidemiol* 1979; **109**(1): 21–4.

41 Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. Cambridge, MA: Cambridge University Press, 2000.

42 Berntsen D, Rubin DC. *Understanding Autobiographical Memory: Theories and Approaches*. Cambridge, UK: Cambridge University Press, 2012.

43 Kasabova A. *On Autobiographical Memory*. Newcastle upon Tyne: Cambridge Scholars, 2009.

44 Madow WG. Interview data on chronic conditions compared with information derived from medical records. *Vital Health Stat* 1967; **2**(23): 1–84.

45 Naleway AL, Belongia EA, Greenlee RT, Kieke BA Jr, Chen RT, Shay DK. Eczematous skin disease and recall of past diagnoses: implications for smallpox vaccination. *Ann Intern Med* 2003; **139**(1): 1–7.

46 Watson DL. Health interview responses compared with medical records. *Vital Health Stat* 1965; **1**(46): 1–74.

47 Gardner RS, Mainetti M, Ascoli GA. Older adults report moderately more detailed autobiographical memories. *Front Psychol* 2015; **6**: 631.

48 Warnecke RB, Sudman S, Johnson TP, O'Rourke D, Davis AM, Jobe JB. Cognitive aspects of recalling and reporting health-related events: Papanicolaou smears, clinical breast examinations, and mammograms. *Am J Epidemiol* 1997; **146**(11): 982–92.

49 Tourangeau R. Cognitive sciences and survey methods. In: Janine TB, ed. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC: National Academies Press, 1984, pp. 73–100.

50 Willis GB, Royston P, Bercini D. The use of verbal report methods in the development and testing of survey questionnaires. *Appl Cognit Psychol* 1991; **5**: 251–67.

51 Belli RF. The structure of autobiographical memory and the event history calendar: potential improvements in the quality of retrospective reports in surveys. *Memory* 1998; **6**(4): 383–406.

52 Geisen E, Murphy J. Compendium of web and mobile survey pretesting methods. International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, 2017.

53 Presser S, Couper MP, Lessler JT, *et al*. Methods for testing and evaluating survey questions. In: Presser S, ed. *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken: John Wiley & Sons, 2004.

54 World Bank. Living Standards Measurement Survey – Questionnaires. file:///C:/Users/Owner/Downloads/bihmane.pdf (accessed May 5, 2019).

55 Smith TW. Optimizing Questionnaire Design in Cross-National and Cross-Cultural Surveys. International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, 2016.

56 Q-Bank. https://wwwn.cdc.gov/qbank/ (accessed May 5, 2019).

57 Dean E, Caspar R, McAvinchey G, Reed L, Quiroz R. Developing a low-cost technique for parallel cross-cultural instrument development: the Question Appraisal System (QAS-04). *Int J Soc Res Methodol* 2007; **10**(3): 227–41.

58 Survey Quality Predictor (SQP) 2.1. http://sqp.upf.edu/ (accessed May 5, 2019).

59 Graesser AC, Wiemer-Hastings K, Kreuz R, Wiemer-Hastings P, Marquis K. QUAID: a questionnaire evaluation aid for survey methodologists. *Behav Res Methods Instrum Comput* 2000; **32**: 254–62.

60 University of Memphis. QUAID: Question Understanding Aid Tool. quaid.cohmetrix.com/ (accessed May 5, 2019).

61 Biemer PP, Lyberg L. *Introduction to Survey Quality*.Hoboken: Wiley Interscience, 2003, p. xiv.

62 Miller K. *Cognitive Interviewing Methodology*. Hoboken: Wiley, 2014.

63 Willis G. *Cognitive Interviewing*. Thousand Oaks: Sage Publications, 2005.

64 Willis GB. *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford: Oxford University Press, 2015.

65 Krosnick JA. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. *Survey Methods Newsletter* 2000; **20**: 4–8.

66 Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opin Quart* 1987; **51**: 201–19.

67 Geisen E, Bergstrom JR. Respondent–survey interaction. In: *Usability Testing for Survey Research*. Cambridge, MA: Morgan Kaufmann, 2017.

68 Nichols E, Anderson Riemer AE, Holland T, Olmsted Hawala EL. Best Practices of Usability Testing Online Questionnaires at the Census Bureau: How Rigorous and Repeatable Testing can Improve Online Questionnaire Design. International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, 2016.

69 Fowler S, Willis G. The Practice of Cognitive Interviewing Through Web Probing. International Conference on Questionnaire Design, Development, Evaluation, and Testing (QDET2), Miami, FL, 2016.

70 Biemer P. Measurement errors in sample surveys. In: Pfeffermann D, Rao CR, eds. *Handbook of Statistics – Sample Surveys: Design, Methods and Applications*. Amsterdam: North-Holland, pp. 281–316.

71 Strom BL, Carson JL, Halpern AC, *et al.* A population-based study of Stevens–Johnson syndrome. Incidence and antecedent drug exposures. *Arch Dermatol* 1991; **127**(6): 831–8.

72 West SL, Ritchey ME, Poole C. Validity of Pharmacoepidemiologic drug and diagnosis data. In: Strom BL, Kimmel SE, Hennessy S, eds. *Pharmacoepidemiology*, 5th edn. Chichester: John Wiley & Sons, 2012, pp. 757–94.

73 West SL, Strom BL, Poole C. Validity of pharmacoepidemiology drug and diagnosis data. In: Strom BL, Kimmel SE, Hennessy S, eds. *Pharmacoepidemiology*, 3rd edn. Chichester: John Wiley & Sons, 2000, pp. 661–706.

74 West SL, Strom BL, Poole C. Validity of pharmacoepidemiology drug and diagnosis data. In: Strom BL, Kimmel SE, Hennessy S, eds. *Pharmacoepidemiology*, 4th edn.

Chichester: John Wiley & Sons, 2000, pp. 709–766.

75 Lau HS, Florax C, Porsius AJ, de Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. *Br J Clin Pharmacol* 2000; **49**(6): 597–603.

76 Strom BL, Carson JL, Halpern AC, *et al*. Using a claims database to investigate drug-induced Stevens–Johnson syndrome. *Stat Med* 1991; **10**(4): 565–76.

77 Guess HA, West R, Strand LM, *et al*. Fatal upper gastrointestinal hemorrhage or perforation among users and nonusers of nonsteroidal anti-inflammatory drugs in Saskatchewan, Canada 1983. *J Clin Epidemiol* 1988; **41**(1): 35–45.

78 Kirking DM, Ammann MA, Harrington CA. Comparison of medical records and prescription claims files in documenting prescription medication therapy. *J Pharmacoepidemiol* 1996; **5**: 3–15.

79 Monson RA, Bond CA. The accuracy of the medical record as an index of outpatient drug therapy. *JAMA* 1978; **240**(20): 2182–4.

80 Christensen DB, Williams B, Goldberg HI, Martin DP, Engelberg R, LoGerfo JP. Comparison of prescription and medical records in reflecting patient antihypertensive drug therapy. *Ann Pharmacother* 1994; **28**(1): 99–104.

81 West SL, Strom BL, Freundlich B, Normand E, Koch G, Savitz DA. Completeness of prescription recording in outpatient medical records from a health maintenance organization. *J Clin Epidemiol* 1994; **47**(2): 165–71.

82 Buchsbaum DG, Boling P, Groh M. Residents' underdocumentation in elderly patients' records of prescriptions for benzodiazepine. *J Med Educ* 1987; **62**(5): 438–40.

83 Lain SJ, Roberts CL, Hadfield RM, Bell JC, Morris JM. How accurate is the reporting of obstetric haemorrhage in hospital discharge data? A validation study. *Aust N Z J Obstet Gynaecol* 2008; **48**(5): 481–4.

84 Crofts JF, Bartlett C, Ellis D, Fox R, Draycott TJ. Documentation of simulated shoulder dystocia: accurate and complete? *Br J Obstet Gynaecol* 2008; **115**(10): 1303–8.

85 Lefter LP, Walker SR, Dewhurst F, Turner RW. An audit of operative notes: facts and ways to improve. *ANZ J Surg* 2008; **78**(9): 800–2.

86 Haberman S, Rotas M, Perlman K, Feldman JG. Variations in compliance with documentation using computerized obstetric records. *Obstet Gynecol* 2007; **110**(1): 141–5.

87 Stange KC. The problem of fragmentation and the need for integrative solutions. *Ann Fam Med* 2009; **7**(2): 100–3.

88 West SL, Blake C, Zhiwen L, McKoy JN, Oertel MD, Carey TS. Reflections on the use of electronic health record data for clinical research. *Health Informat* 2009; **15**(2): 108–21.

89 Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. *Appl Clin Inform* 2017; **8**(1): 12–34.

90 Glass R,Johnson B, Vessey M. Accuracy of recall of histories of oral contraceptive use. *Br J Prev Soc Med* 1974; **28**(4): 273–5.

91 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**(3): 226–35.

92 Adam SA, Sheaves JK, Wright NH, Mosser G, Harris RW, Vessey MP. A case–control study of the possible association between oral contraceptives and malignant melanoma. *Br J Cancer* 1981; **44**(1): 45–50.

93 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**(1): 84–7.

94 Persson I, Bergkvist L, Adami HO. Reliability of women's histories of climacteric oestrogen

treatment assessed by prescription forms. *Int J Epidemiol* 1987; **16**(2): 222–8.

95 Goodman MT, Nomura AM, Wilkens LR, Kolonel LN. Agreement between interview information and physician records on history of menopausal estrogen use. *Am J Epidemiol* 1990; **131**(5): 815–25.

96 Nischan P, Ebeling K, Thomas DB, Hirsch U. Comparison of recalled and validated oral contraceptive histories. *Am J Epidemiol* 1993; **138**(9): 697–703.

97 West SL, Savitz DA, Koch G, Strom BL, Guess HA, Hartzema A. Recall accuracy for prescription medications: self-report compared with database information. *Am J Epidemiol* 1995; **142**(10): 1103–12.

98 Strom BL, Schinnar R. An interview strategy was critical for obtaining valid information on the use of hormone replacement therapy. *J Clin Epidemiol* 2004; **57**(11): 1210–13.

99 Drieling RL, LaCroix AZ, Beresford SAA, Boudreau DM, Kooperberg C, Heckbert SR. Validity of self-reported medication use compared with pharmacy records in a cohort of older women: findings from the Women's Health Initiative. *Am J Epidemiol* 2016; **184**(3): 233–8.

100 Lewis JD, Brensinger C, Bilker WB, Strom BL. Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. *Pharmacoepidemiol Drug Saf* 2002; **11**(3): 211–18.

101 Klemetti A, Saxen L. Prospective versus retrospective approach in the search for environmental causes of malformations. *Am J Public Health Nations Health* 1967; **57**(12): 2071–5.

102 Coulter A, Vessey M, McPherson K, Crossley B. The ability of women to recall their oral contraceptive histories. *Contraception* 1986; **33**(2): 127–37.

103 Feldman Y, Koren G, Mattice K, Shear H, Pellegrini E, Macleod SM. Determinants of recall and recall bias in studying drug and chemical exposure in pregnancy. *Teratology* 1989; **40**(1): 37–45.

104 West SL, Savitz DA, Koch G, *et al.* Demographics, health behaviors, and past drug use as predictors of recall accuracy for previous prescription medication use. *J Clin Epidemiol* 1997; **50**(8): 975–80.

105 Klungel OH, de Boer A, Paes AH, Herings RM, Seidell JC, Bakker A. Influence of question structure on the recall of self-reported drug use. *J Clin Epidemiol* 2000; **53**(3): 273–7.

106 Bryant HE, Visser N, Love EJ. Records, recall loss, and recall bias in pregnancy: a comparison of interview and medical records data of pregnant and postnatal women. *Am J Public Health* 1989; **79**(1): 78–80.

107 Mackenzie SG, Lippman A. An investigation of report bias in a case–control study of pregnancy outcome. *Am J Epidemiol* 1989; **129**(1): 65–75.

108 de Jong PCMP, Berns MPH, van Duynhoven YTHP, Nijdam WS, Eskes TKAB, Zielhuis GA. Recall of medication during pregnancy: validity and accuracy of an adjusted questionnaire. *Pharmacoepidemiol Drug Saf* 1995; **4**: 23–30.

109 Hulka BS, Kupper LL, Cassel JC, Efird RL, Burdette JA. Medication use and misuse: physician–patient discrepancies. *J Chronic Dis* 1975; **28**(1): 7–21.

110 Paganini-Hill A, Ross RK. Reliability of recall of drug usage and other health-related information. *Am J Epidemiol* 1982; **116**(1): 114–22.

111 Landry JA, Smyer MA, Tubman JG, Lago DJ, Roberts J, Simonson W. Validation of two methods of data collection of self-reported medicine use among the elderly. *Gerontologist* 1988; **28**(5): 672–6.

112 Johnson RE, Vollmer WM. Comparing sources of drug data about the elderly. *J Am Geriatr Soc* 1991; **39**(11): 1079–84.

113 Kehoe R, Wu SY, Leske MC, Chylack LT Jr. Comparing self-reported and

physician-reported medical history. *Am J Epidemiol* 1994; **139**(8): 813–18.

114 Law MG, Hurley SF, Carlin JB, Chondros P, Gardiner S, Kaldor JM. A comparison of patient interview data with pharmacy and medical records for patients with acquired immunodeficiency syndrome or human immunodeficiency virus infection. *J Clin Epidemiol* 1996; **49**(9): 997–1002.

115 Smith NL, Psaty BM, Heckbert SR, Tracy RP, Cornell ES. The reliability of medication inventory methods compared to serum levels of cardiovascular drugs in the elderly. *J Clin Epidemiol* 1999; **52**(2): 143–6.

116 Clegg LX, Potosky AL, Harlan LC, *et al.* Comparison of self-reported initial treatment with medical records: results from the prostate cancer outcomes study. *Am J Epidemiol* 2001; **154**(6): 582–7.

117 de Jong PCM, Berns M, van Duynhoven Y, *et al.* Accessibility and validity of data on medical drug use during pregnancy collected from various sources. *J Pharmacoepidemiol* 1991; **2**: 45–57.

118 Gama H, Correia S, Lunet N. Questionnaire design and the recall of pharmacological treatments: a systematic review. *Pharmacoepidemiol Drug Saf* 2009; **18**(3): 175–87.

119 Ademi Z, Turunen JH, Kauhanen J, Enlund H. A comparison of three questionnaire-based measures of analgesic use over 11 years in adult males: a retrospective analysis of data from a prospective, population-based cohort study. *Clin Ther* 2007; **29**(3): 529–34.

120 Mitchell AA, Cottler LB, Shapiro S. Effect of questionnaire design on recall of drug exposure in pregnancy. *Am J Epidemiol* 1986; **123**(4): 670–6.

121 van Gelder MM, van Rooij IA, de Walle HE, Roeleveld N, Bakker MK. Maternal recall of prescription medication use during pregnancy using a paper-based questionnaire: a validation study in the Netherlands. *Drug Saf* 2013; **36**(1): 43–54.

122 Lunet N, Bastos J, Cumaio F, Silva P, Dias E, Barros H. Recall of drug utilization depends on subtle structural questionnaire characteristics. *Pharm World Sci* 2008; **30**(2): 175–81.

123 Van den Brandt PA, Petri H, Dorant E, Goldbohm RA, van de Crommert S. Comparison of questionnaire information and pharmacy data on drug use. *Pharm Weekbl* 1991; **13**(2): 91–6.

124 Cotterchio M, Kreiger N, Darlington G, Steingart A. Comparison of self-reported and physician-reported antidepressant medication use. *Ann Epidemiol* 1999; **9**(5): 283–9.

125 Linet MS, Harlow SD, McLaughlin JK, McCaffrey LD. A comparison of interview data and medical records for previous medical conditions and surgery. *J Clin Epidemiol* 1989; **42**(12): 1207–13.

126 Bush TL, Miller SR, Golden AL, Hale WE. Self-report and medical record report agreement of selected medical conditions in the elderly. *Am J Public Health* 1989; **79**(11): 1554–6.

127 Commission on Chronic Illness. Chronic illness in a large city: the Baltimore Study. In: *Chronic Illness in the United States*, vol. **IV**. Cambridge, MA: Harvard University Press, 1957, pp. 297–328.

128 Heliovaara M, Aromaa A, Klaukka T, Knekt P, Joukamaa M, Impivaara O. Reliability and validity of interview data on chronic diseases. The Mini-Finland Health Survey. *J Clin Epidemiol* 1993; **46**(2): 181–91.

129 Spitz MR, Fueger JJ, Newell GR. The development of a comprehensive, institution-based patient risk evaluation program: II. Validity and reliability of questionnaire data. *Am J Prev Med* 1988; **4**(4): 188–93.

130 Midthjell K, Holmen J, Bjørndal A, Lund-Larsen G. Is questionnaire information valid in the study of a chronic disease such as diabetes? The Nord-Trondelag diabetes

study. *J Epidemiol Commun Health* 1992; **46**(5): 537–42.

131 Kriegsman DM, Penninx BW, van Eijk JT, Boeke AJ, Deeg DJ. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy. *J Clin Epidemiol* 1996; **49**(12): 1407–17.

132 Colditz GA, Martin P, Stampfer MJ, *et al.* Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Am J Epidemiol* 1986; **123**(5): 894–900.

133 Bergmann MM, Calle EE, Mervis CA, Miracle-McMahil HL, Thun MJ, Heath CW. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol* 1998; **147**(6): 556–62.

134 Paganini-Hill A, Chao A. Accuracy of recall of hip fracture, heart attack, and cancer: a comparison of postal survey data and medical records. *Am J Epidemiol* 1993; **138**(2): 101–6.

135 Nevitt MC, Cummings SR, Browner WS, *et al.* The accuracy of self-report of fractures in elderly women: evidence from a prospective study. *Am J Epidemiol* 1992; **135**(5): 490–9.

136 McKinlay A, Horwood LJ. The accuracy of adult recall for early mild traumatic brain injury. *Disabil Rehabil* 2017; **39**(13): 1296–9.

137 Takayanagi Y, Spira AP, Roth KB, Gallo JJ, Eaton WW, Mojtabai R. Accuracy of reports of lifetime mental and physical disorders: results from the Baltimore Epidemiological Catchment Area Study. *JAMA Psychiatry* 2014; **71**(3): 273–80.

138 Zhu K, McKnight B, Stergachis A, Daling JR, Levine RS. Comparison of self-report data and medical records data: results from a case–control study on prostate cancer. *Int J Epidemiol* 1999; **28**(3): 409–17.

139 Tretli S, Lund-Larsen PG, Foss OP. Reliability of questionnaire information on cardiovascular disease and diabetes: cardiovascular disease study in Finnmark county. *J Epidemiol Commun Health* 1982; **36**(4): 269–73.

140 Rosamond WD, Sprafka JM, McGovern PG, Nelson M, Luepker RV. Validation of self-reported history of acute myocardial infarction: experience of the Minnesota Heart Survey Registry. *Epidemiology* 1995; **6**(1): 67–9.

141 Walker MK, Whincup PH, Shaper AG, Lennon LT, Thomson AG. Validation of patient recall of doctor-diagnosed heart attack and stroke: a postal questionnaire and record review comparison. *Am J Epidemiol* 1998; **148**(4): 355–61.

142 Cannell CF, Marquis KH, Laurent A. A summary of studies of interviewing methodoloty. *Vital Health Stat* 1977; **1**(series 2): i–viii, 1–78.

143 Lessler JT, Harris BSH. *Medicaid Data as a Source for postmarketing Surveillance Information*. Research Triangle Park: Research Triangle Institute, 1984.

144 Sudman S, Bradburn N. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass, 2004.

145 McColl E, Jacoby A, Thomas L, *et al.* Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess* 2001; **5**(31).

146 Wilcox AJ, Horney LF. Accuracy of spontaneous abortion recall. *Am J Epidemiol* 1984; **120**(5): 727–33.

147 Biering-Sorensen U. Reporting of hospitalization in the Health Interview Survey. *Vital Health Stat* 1965; **1**(61): 1–71.

148 Corwin RG, Krober M, Roth HP. Patients' accuracy in reporting their past medical history: a study of 90 patients with peptic ulcer. *J Chronic Dis* 1971; **23**(12): 875–9.

149 Coulter A, McPherson K, Elliott S, Whiting B. Accuracy of recall of surgical histories: a

comparison of postal survey data and general practice records. *Community Med* 1985; **7**(3): 186–9.

150 Wingo PA, Ory HW, Layde PM, Lee NC. The evaluation of the data collection process for a multicenter, population-based, case–control design. *Am J Epidemiol* 1988; **128**(1): 206–17.

151 Udry JR, Gaughan M, Schwingl PJ, van den Berg BJ. A medical record linkage analysis of abortion underreporting. *Fam Plann Perspect* 1996; **28**(5): 228–31.

152 Linton KL, Klein BE, Klein R. The validity of self-reported and surrogate-reported cataract and age-related macular degeneration in the Beaver Dam Eye Study. *Am J Epidemiol* 1991; **134**(12): 1438–46.

153 Bean JA, Leeper JD, Wallace RB, Sherman BM, Jagger H. Variations in the reporting of menstrual histories. *Am J Epidemiol* 1979; **109**(2): 181–5.

154 Colditz GA, Colditz GA, Stampfer MJ, *et al*. Reproducibility and validity of self-reported menopausal status in a prospective cohort study. *Am J Epidemiol* 1987; **126**(2): 319–25.

155 Fourrier-Reglat A, Cuong HM, Lassalle R, *et al*. Concordance between prescriber- and patient-reported previous medical history and NSAID indication in the CADEUS cohort. *Pharmacoepidemiol Drug Saf* 2010; **19**(5): 474–81.

156 Jarernsiripornkul N, Chaisrisawadsuk S, Chaiyakum A, Krska J. Patient self-reporting of potential adverse drug reactions to non-steroidal anti-inflammatory drugs in Thailand. *Pharm World Sci* 2009; **31**(5): 559–64.

157 Bonito AJ, Farrelly M, Han J, *et al*. *Assessment of the Feasibility of Creating a Managed Care Encounter-Level Database*. RTI Project Report 6703–003. Prepared for the Center for Organization and Delivery Studies. Rockville: Agency for Health Care Policy and Research, 1997.

158 Stergachis AS. Record linkage studies for postmarketing drug surveillance: data quality and validity considerations. *Drug Intell Clin Pharm* 1988; **22**(2): 157–61.

159 World Health Organization. Classification of Diseases (ICD 11). www.who.int/classifications/icd/en/ (accessed May 5, 2019).

160 Centers for Medicare and Medicaid Services. 2016 ICD-10-CM and GEMs. www.cms.gov/Medicare/Coding/ICD10/2016–ICD–10–CM–and–GEMs.html (accessed May 5, 2019).

161 National Center for Health Statistics. International Classification of Diseases, (ICD-10-CM/PCS) Transition – Background. www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm (accessed May 5, 2019).

162 Sun JW, Rogers JR, Her Q, *et al*. Adaptation and validation of the combined comorbidity score for ICD-10-CM. *Med Care* 2017; **55**: 1046–51.

163 Fung KW, Richesson R, Smerek M, *et al*. Preparing for the ICD-10-CM transition: automated methods for translating ICD codes in clinical phenotype definitions. *eGEMs* 2016; **4**(1): 4.

164 Liao KP, Cai T, Savova GK, *et al*. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 2015; **350**: h1885.

165 Upadhyaya SG, Murphree DH Jr, Ngufor CG, *et al*. Automated diabetes case identification using electronic health record data at a tertiary care facility. *Mayo Clin Proc Innov Qual Outcomes* 2017; **1**(1): 100–10.

166 Liao KP, Ananthakrishnan AN, Kumar V, *et al*. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS One* 2015; **10**(8): e0136651.

167 Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013; **20**: e147–e154.

168 Trifiro G, Coloma PM, Rijnbeek PR, *et al*. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? *J Intern Med* 2014; **275**: 551–61.

169 Schering Laboratories. *The Phantom Patient and Community Pharmacy Practice*. Schering Report SVII. Kenilworth: Schering, 1996.

170 Craghead RM, Wartski DM. An evaluative study of unclaimed prescriptions. *Hosp Pharm* 1991; **26**(7): 616–17, 632.

171 Farmer KC, Gumbhir AK. Unclaimed prescriptions: an overlooked opportunity. *Am Pharm* 1992; **NS32**(10): 55–9.

172 Hamilton WR, Hopkins UK. Survey on unclaimed prescriptions in a community pharmacy. *J Am Pharm Assoc* 1997; **NS37**(3): 341–5.

173 Kirking MH, Kirking DM. Evaluation of unclaimed prescriptions in an ambulatory care pharmacy. *Hosp Pharm* 1993; **28**(2): 90–1, 94, 102.

174 Kinnaird D, Cox T, Wilson JP. Unclaimed prescriptions in a clinic with computerized prescriber order entry. *Am J Health Syst Pharm* 2003; **60**(14): 1468–70.

175 Fischer MA, Stedman MR, Lii J, *et al*. Primary medication non-adherence: analysis of 195,930 electronic prescriptions. *J Gen Intern Med* 2010; **25**(4): 284–90.

176 Rowan CG, Flory J, Gerhard T, *et al*. Agreement and validity of electronic health record prescribing data relative to pharmacy claims data: a validation study from a US electronic health record database. *Pharmacoepidemiol Drug Saf* 2017; **26**(8): 963–72.

177 Bright RA, Avorn J, Everitt DE. Medicaid data as a resource for epidemiologic studies: strengths and limitations. *J Clin Epidemiol* 1989; **42**(10): 937–45.

178 Strom BL, Carson JL, Schinnar R, Snyden ES, Shaw M, Waiter SL. No causal relationship between transdermal scopolamine and seizures: methodologic lessons for pharmacoepidemiology. *Clin Pharmacol Ther* 1991; **50**(1): 107–13.

179 Wynia MK, Cummins DS, VanGeest JB, Wilson IB. Physician manipulation of reimbursement rules for patients: between a rock and a hard place. *JAMA* 2000; **283**(14): 1858–65.

180 Soto CM, Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. *BMC Health Serv Res* 2002; **2**(1): 22.

181 Kahn MG, Callahan TJ, Barnard J, *et al*. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs* 2016; **4**(1): 18.

182 Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Informat* 2013; **46**: 830–6.

183 Brophy JM, Levesque LE, Zhang B. The coronary risk of cyclo-oxygenase-2 inhibitors in patients with a previous myocardial infarction. *Heart* 2007; **93**(2): 189–94.

184 Varas-Lorenzo C, Castellsague J, Stang MR, Perez-Gutthann S, Aquado J, Rodriguez LA. The use of selective cyclooxygenase-2 inhibitors and the risk of acute myocardial infarction in Saskatchewan, Canada. *Pharmacoepidemiol Drug Saf* 2009; **18**(11): 1016–25.

185 Wahl PM, Rodgers K, Schneeweiss S, *et al*. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. *Pharmacoepidemiol Drug Saf* 2010; **19**(6): 596–603.

186 Newton KM, Wagner EH, Ramsey SD, *et al*. The use of automated data to identify complications and comorbidities of diabetes: a validation study. *J Clin Epidemiol* 1999; **52**(3): 199–207.

187 Van Staa TP, Abenhaim L. The quality of information recorded on a UK database of primary care records: a study of hospitalizations due to hypoglycemia and other conditions. *Pharmacoepidemiol Drug Saf* 1994; **3**: 15–21.

188 Abraham NS, Cohen DC, Rivers B, Richardson P. Validation of administrative data used for the diagnosis of upper gastrointestinal events following nonsteroidal anti-inflammatory drug prescription. *Aliment Pharmacol Ther* 2006; **24**(2): 299–306.

189 Stergachis AS. Evaluating the quality of linked automated databases for use in pharmacoepidemiology. In: Hartzema AG, Porta MS, Tilson HH, eds. *Pharmacopeidemiology: An Introduction*. Cincinnati: Harvey Whitney, 1991, pp. 222–34.

190 Harris BL, Stergachis A, Ried LD. The effect of drug co-payments on utilization and cost of pharmaceuticals in a health maintenance organization. *Med Care* 1990; **28**(10): 907–17.

191 Kudyakov R, Bowen J, Ewen E, *et al.* Use of an electronic health record to classify patients with newly diagnosed versus pre-existing type 2 diabetes: building methodology infrastructure for comparative effectiveness research. *Population Health Manage* 2012; **15**(1): 3–11.

192 Ambegaonkar A, Livengood K, Craig T, Day D. Predicting the risk for gastrointestinal toxicity in patients taking NSAIDs: the Gastrointestinal Toxicity Survey. *Adv Ther* 2004; **21**(5): 288–300.

193 Singh G, Ramey DR, Morfeld D, Shi H, Hatoum HT, Fries JF. Gastrointestinal tract complications of nonsteroidal anti-inflammatory drug treatment in rheumatoid arthritis. A prospective observational cohort study. *Arch Intern Med* 1996; **156**(14): 1530–6.

194 Levy AR, Tamblyn RM, Fitchett D, McLeod PJ, Hanley JA. Coding accuracy of hospital discharge data for elderly survivors of myocardial infarction. *Can J Cardiol* 1999; **15**(11): 1277–82.

195 Castellsague J, Holick CN, Hoffman CC, Gimeno V, Stang MR, Perez-Gutthann S. Risk of upper gastrointestinal complications associated with cyclooxygenase-2 selective and nonselective nonsteroidal antiinflammatory drugs. *Pharmacotherapy* 2009; **29**(12): 1397–407.

196 Curtis JR, Cheng H, Delzell E, *et al.* Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example. *Med Care* 2008; **46**(9): 969–75.

197 van Staa TP, Leufkens HG, Zhang B, Smeeth L. A comparison of cost effectiveness using data from randomized trials or actual clinical practice: selective cox-2 inhibitors as an example. *PLoS Med* 2009; **6**(12): e1000194.