Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
22(1) 7–13

DOI: 10.1177/0962280211403598
smm.sagepub.com
SAGE

Sensible use of observational clinical data

J Marc Overhage[1,2] and Lauren M Overhage[1]

Abstract

Observational data sets offer many potential advantages for medical research including their low cost, large size and generalisability. Because they are collected for clinical care and health care operations purposes, observational data sets have some limitations that must be considered in order to perform useful analyses. Sensible use of observational data sets can yield valuable insights, particularly when clinical trials are impractical.

Keywords

bias (epidemiology), computerised, data Interpretation, medical records systems, statistical

1 Why use observational clinical data?

Randomised controlled clinical trials (RCTs) are the established method used to support inference. Because of their strict design, RCTs ensure effective control of internal biases and high internal validity, allowing inferences to be drawn about cause and effect. However, this very restrictive design also means that RCTs provide limited insight into real-world clinical practice, where there are wider patient profiles and variable health care practices, prescribing habits and patient adherence to the regimen. Non-randomised studies using large observational databases have increasingly been used to help evaluate the effectiveness of interventions in producing the required health outcomes. The nature of such observational studies allows the inclusion of very large numbers of subjects and can be an important source of both effectiveness and safety for medicines and allow comparison between interventions in a fashion that is unlikely to occur in the RCT setting. In addition, observational studies are typically less expensive and allow interventions and conditions to be studied that would not be economically feasible to evaluate using RCTs.

Health care researchers have begun to take advantage of large administrative claim databases that have been developed over the last two decades for clinical effectiveness research, pharmacoepidemiology, health services research and public health research among

[1]Regenstrief Institute, Inc, Suite 2000, 410 West 10th Street, Indianapolis, IN 46202 USA
[2]Indiana University School of Medicine, 340 West 10th Street, Suite 6200, Indianapolis, IN 46202-3082, USA

Corresponding author:
J Marc Overhage, Regenstrief Institute, Inc, Suite 2000, 410 West 10th Street, Indianapolis, IN 46202 USA
Email: joverhag@iupui.edu

others. Studies conducted using these databases are observational in nature and have a number of methodological limitations. However, if appropriate techniques are applied, the information obtained can be reliable and useful.[1]

There are two major sources of observational data and clinical systems. To date, most of the research studies that have used observational data has relied on claims data rather than clinical data. While claims data offer many advantages including their size (data sets representing 160 million individuals are available), relative completeness (they include information about care across different health care providers across broad geographies) and structure (almost all data are coded), they are crude representations of the patient. Data from clinical systems including electronic medical records provide a more nuanced representation of the patient including severity of the condition, richer covariates and potentially better completeness but have challenges of their own.

Observational data sets are particularly useful for effectiveness studies in which outcomes across a wide range of patients and health care providers.[2] In contrast, traditional randomized controlled trials (RCTs) are described as efficacy studies which provide data about outcomes in carefully selected patients and providers. Relative to effectiveness studies performed in observational data sets, RCTs have the inherent strength of balancing measured and unmeasured patient characteristics between different treatment groups controlling for a variety of bias that can arise from differences in patient characteristics rather than 'intervention' effects. RCTs are not without their own biases. The 'intent-to-treat' analytic approach, in which people are considered to have used the assigned treatment regardless of actual compliance attempts to account for the *bias towards the null*. Understanding the role and value of both approaches, the June, 2009 IOM report, Initial National Priorities for Comparative Effectiveness Research, discusses the need for both traditional randomised trials complemented by analyses of observational data. The report calls for 'large-scale clinical and administrative data networks that enable observational studies of patient care while protecting patient privacy and data security' recognising that 'new methods for linking patient-level data from multiple health care organisations will promote inclusion of populations frequently omitted from clinical trials'.[3]

2 What are the issues encountered in using observational clinical data?

2.1 Data are not patients

Observational data, whether claims or from clinical systems, are a cartoon of the patient they are derived from. Claims data provide broad brush strokes and clinical data begin to fill in details but they are never complete and do not accurately nor completely represent the patient – so do not assume that they do. Claims data not only lack detail but are known to be incomplete and may contain intentional and unintentional inaccuracies. Intentional inaccuracies may be introduced through efforts to maximise reimbursement.[4] The sometimes byzantine reimbursement rules and claims which are all about reimbursement, may provide a higher payment for care attributed to a specific set of diagnostic codes than another leading providers to invest in staff and tools to 'optimise' reimbursement. These diagnostic codes are not wrong *per se* but are inconsistent across providers and over time making it difficult to properly categorise patients. Another example of intentional inaccuracies is diagnoses that are present when a patient is admitted to a hospital. Claims data will commonly include discharge diagnoses and it is often impossible to distinguish a diagnosis for a condition acquired during a hospital admission from one that was present on admission. This problem was so prevalent that

payors have worked with providers to begin to add a 'present on admission' flags to discharge diagnoses. While this flag will help distinguish whether the condition developed before or during the hospital admission, the analysis has to properly utilise this flag. Yet another reason that diagnoses may be intentionally omitted is when recording the diagnosis could have serious consequences for the patient. A diagnosis of seizure disorder, for example, could prevent a patient from driving a commercial vehicle and keep them from earning a living so a provider may choose not to record the diagnosis.

Examples of unintentional inaccuracies results in a number of ways. First, there can be simple errors in recording the data, and ICD-9 code 240.0 (simple goiter) is recorded instead of 250.0 (diabetes without complication). Second, reimbursement rules may require a diagnostic code be provided as a rational for a test or procedure which may results in diagnostic codes being recorded that the patient does not have. An order for a routine screening mammogram, for example, may result in a diagnostic code for breast cancer being recorded in an observational database because that is the problem the clinician is trying to detect or 'rule out'.[4–6]

2.2 Data are Swiss cheese

A great strength of claims data is that they are likely to include some representation of almost all care that a patient receives. There are important exceptions, however, including over the counter drugs (those that are available without a prescription). Important medications such as aspirin, anti-histamine and H2-blockers used to treat heartburn and many others are all available over the counter. The advent of low-price generic medications has further eroded the completeness of claims data. Even some prescription drugs may not be submitted to the payor for reimbursement because they are so low cost.

Clinical data are more problematic when it comes to completeness. You simply do not know what you do not know which has led some researchers to make the statement that 'Absence of evidence is not evidence of absence'.[7] While a few patients receive almost all their care within an integrated delivery system where nearly all their clinical data are recorded in a single medical record, this is the exception. Almost all patients receive care from a variety of providers including their primary care physician, specialists, hospitals, diagnostic centers, visiting nurse service, hospice, extended care facility, pharmacy and many others. These providers all typically maintain their own medical record and thus the data for a single patient are spread in separate silos. Even a primary care provider that is trying to serve as a focal point and coordinate much of a patient's care is unlikely to maintain comprehensive data. In order to overcome this challenge and in order to use the data for clinical care, public health, quality, research and other purposes, groups have begun to develop methods to standardise and aggregate data from across multiple providers. This is a difficult task requiring issues of privacy and security be addressed, along with business issues and a variety of technological issues including how to match data for the same patient when each provider organisation uses its own identifier for the patient and how to standardise how the data are labelled or described so that analyses can recognise the same test result across providers.[8] Depending on the goal of the analysis, providers may need to be matched as well. Health care organisations like hospitals, payors and other providers all identify providers in their own way. A clinician may have dozens of identifiers in a single community as well as multiple identifiers assigned at a national level and these identifiers may need to be linked and various attributes such as specialty rationalised to perform certain analyses.

Another way in which data are Swiss cheese is that diagnostic tests and other assessments are performed as needed rather than on any regular schedule. Even in patients with chronic diseases in

which these tests and assessments are recommended by professional societies and through clinical guidelines, they are only performed slightly more than 50% of the time.[9]

2.3 Data hide their meaning

The meaning of observational data may be hidden in a variety of ways. First, the data may not reliably contain the data needed for a particular analysis such as income level. Fortunately, there are often ways to infer the necessary data. The mean family income for residents of the census block in which the patient lives may be an adequate proxy for level of income for example. Second, you may have to rely on a combination of data to understand the meaning. Identifying all patients with hypertension, for example, might rely on a combination of vital signs data (systolic and diastolic blood pressures), medications and diagnoses recorded on the problem list. A third way that data may hide their meanings is by making assumptions when you find specific data. An observational data set, for example, might include occult blood measurements performed on stool and an analyst could assume that these represent all the screenings performed for colorectal cancer when, in fact, a large proportion of patients are being screened but the data from these screenings are not being recorded in the observational database.

While in data captured specifically for research, explicit definitions of conditions may be incorporated for the condition under study, conditions recorded in observational data may be based on quite variable definitions. For example, a RCT may specify that, for a patient to be considered to have systolic congestive heart failure (CHF), they must have had a cardiac echocardiogram with estimated ejection fraction of less than 45%. A clinician caring for a patient who has an ejection fraction less than 45% but who does not exhibit signs and symptoms of heart failure may not record CHF as a diagnosis.

Observational data consist primarily of transactions that occur during the care of a patient which means that there is rarely data about when a patient stopped taking a medication or when a condition was resolved. While you may be able to assume that chronic conditions like diabetes will continue in most patients, others like low back pain or pneumonia may be more difficult to track. For medications, there may be clues about when medication use may have ended – the days supply. Medication dispensing events will often provide an estimate, based on the quantity of medication dispensed and the instructions to the patient, and of the number of days the dispensed supply will last if the patient takes the medication as prescribed. These data will allow an analysis to create estimated periods of continuous exposure. There is no equivalent data for diagnoses.

2.4 Data are dynamic over time

One of the most challenging aspects of working with observational data sets, and particularly those from clinical sources, is that their contents are dynamic over time. Data that previously were not available electronically may become available and data that were previously being captured electronically may stop being captured. In claims data, medication claims for a group of patients may not have been available because they did not have a medication benefit, but then they become eligible and medication claims data become available. They are eligible for medical benefits this entire time and procedure and diagnosis data are continuously available but medication data are suddenly present. In clinical data, the nature of a test may change over time.

Figure 1 illustrates the change in measures of cardiac contractility measured by echocardiography from a single source over a 20-year period. An analysis that spanned the years April 2002 through

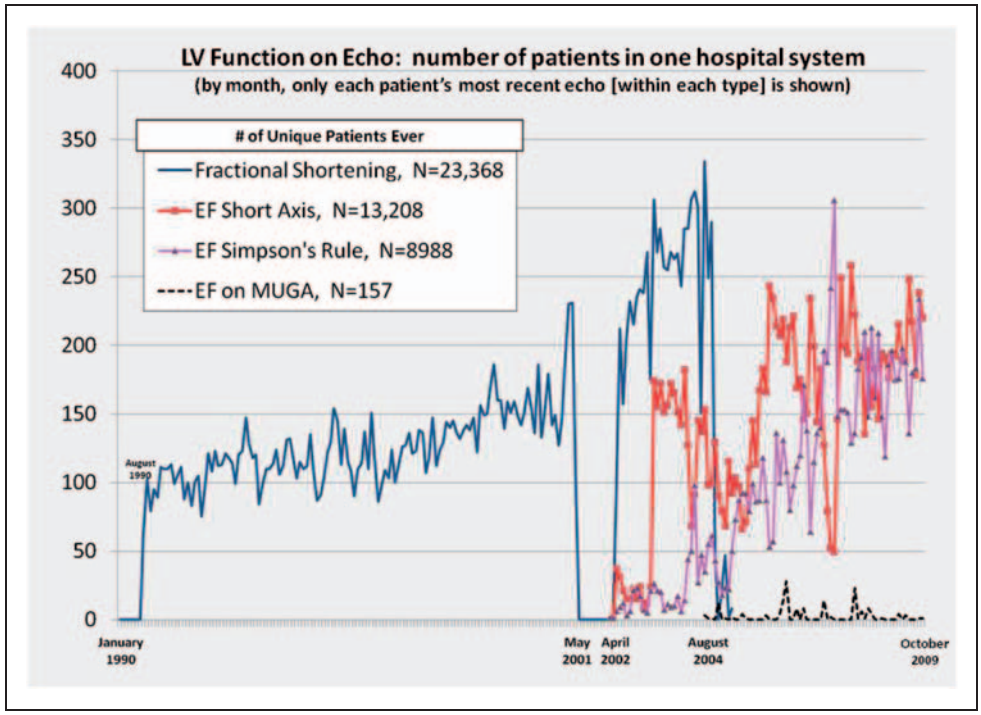

Figure 1. Data from clinical sources varies dynamically over time. The figure illustrates the change over time in the character of cardiology data over a 20-year period in one clinical data set.

August 2004 would require careful attention to which measure of contractility was used. In addition, for some reason, data are missing for May 2001 through April 2002. It is unlikely, given the trend over 20 years that no echocardiograms were performed over this period but, rather, that some technological change or error disrupted the data availability in that interval.

2.5 Data may be truncated temporally

Claims data begin and end with the enrollment or coverage period. In the USA, approximately 15% of health plan members 'turn over' each year resulting in a relatively short period of observation. If all the events of interest occur during that observation period, this does not represent a serious problem – 30-day re-admission rates or in hospital mortality, for example, are likely to be captured – but longer term outcomes may not. In addition, these relatively short observation periods may limit the availability of data on co-morbidities or prior treatments. A colonoscopy which is recommended once over 10 years may well have been performed, while the patient was part of a different health plan and not visible during a particular observation period with a different health plan. Similarly, censoring at the end of the observation period requires special attention. Data from clinical systems are even more problematic. With claims data, there is usually an 'eligibility' or 'enrollment' date when someone starts charging for coverage and a disenrollment date when coverage stops which means the beginning and end of the observation period are clearly documented. In clinical systems, there is not usually a demarcated beginning and end. It is easy to identify the first and last dates on

which a datum about the patient is recorded but the observation period may extend beyond those dates. A patient may be well and not seek care for a year after the last datum is recorded or they may have died minutes after they left the providers office and you do not know which. In addition, just as a patient may change health plans, they may change providers, perhaps because they changed health plans or because they moved, resulting in truncated observation period.

2.6 Data are not data

The populations included in observational data sets vary widely. A dataset derived from the VA administration's patient care will contain few females and be older than the general population. A data set derived from care delivered in a health maintenance organisation may represent a younger, healthier population. Medicaid populations will include more younger women and children than the overall population and may include a sub-population of severely ill patients that are also eligible for Medicare. There are several implications of these variations including how generaliseable the finding may be and how consistent results may be across different observational data sets.

2.7 Data are biased

The great strength of observational data is that they represent the care of patients as it is delivered but, as is often the case, this is also observational data's greatest weakness. Tests and treatments are not chosen at random. Only selected patients coming to the emergency department will be evaluated with a computed tomogram or CT of the head. The patients who are evaluated with a head CT are almost certain to be different from patients who are not; at least in terms of the severity of their symptoms if not their risk factors. Similarly, the choice of treatment is almost always confounded with severity, co-morbidities and other factors. Specific types of treatment such as the choice of specific medication within a class may be less subject to these biases if providers do not perceive significant clinical and cost differences between them.[10]

2.8 Data are never as abundant as they appear

One of the attractive aspects of observational data is the large numbers of observations available. Unfortunately, the number of observations that are available for a particular analysis may be dramatically smaller depending on the purpose of the analysis. One commercial claims data source, for example, has data on 160 million unique individuals. For a pharmacoepidemiology analysis that examines outcomes that are likely to arise during hospital stays, for example, we estimated that only one million of these individuals would be likely to have the data necessary for the analysis. Even once these types of issues are taken into account, the number of patients from a data set relevant to a specific analysis can be orders of magnitude smaller than the data set. Consider, for example, a study of a procedure performed for a particular condition like prostatectomy (removal of the prostate) for prostate cancer. A study might compare a traditional 'open' procedure with a minimally invasive procedure performed using robot-assisted surgery. Prostate cancer is quite common with an incidence of approximately 150 out of 100 000 men. In an observational data set with one million patients, only about half will be male, so we would expect 750 new cases of prostate cancer for each year and approximately 10% of those will undergo prostatectomy, so over a year, only 75 cases are available for study. Fortunately, observational data sets may have multiple years of data and may have data for more than one million individuals but

the example highlights that, even for common conditions, the number of relevant patients may be limited.

2.9 Not all data comes from patients

Studies performed with observational data will often require additional data that characterises the health care system. One example is drug formularies, which are lists of drugs that a health plan will reimburse for, change over time and may directly impact which medications a prescriber chooses for a patient. Formularies change over time and are typically difficult to obtain retrospectively. Another example is availability of specific procedures or resources in a market. Observational data sets may span a variety of markets, and local resources can determine the choice of diagnostic approach or treatment in some cases. Whether or not cardiac catheterisation facilities are available within specific time frame may change how a patient with unstable angina or myocardial infarction is managed. Data about the availability of catheterisation facilities including their locations would be important in addition to the patient data.

3 Conclusion

If observational data are analysed thoughtful with careful attention to their potential limitations, valuable insights can be gained. In order to use observational data sensibly often requires careful attention to the detailed characteristics of the data needed for the analysis in the specific data set being analysed. In addition, several of the assumptions that can be safely made in RCTs must be validated in order to draw strong conclusions.

Acknowledgement

Dr. Marc Rosenman provided several of the examples used in this article.

References

1. Strom BL and Melmon KL. The use of pharmacoepidemiology to study beneficial drug effects. In: Strom BL (ed.) *Pharmacoepidemiology*, 3rd ed. Chichester, UK: John Wiley and Sons, Ltd, 2000, pp.553–572.
2. Ayanian JZ. Using administrative data to assess health care outcomes. *Eur Heart J* 1999; **20**: 1689–1691.
3. IOM (Institute of Medicine). *Initial national priorities for comparative effectiveness research*. Washington, DC: The National Academies Press, 2009.
4. Tang PC, Ralston M, Arrigotti MF, Qureshi L and Graham J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J Am Med Inform Assoc* 2007; **14**: 10–15.
5. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH and Mark DB. Discordance of databases designed for claims payment versus clinical information systems. *Implications for outcomes research*. Ann Intern Med *1993; **119**(8): 844–850*.
6. Ceratti A, Roger France FH and Beguin C. Health data quality improvement by comparing administrative medical data and billing data. *Int J Med Inform* 2008; **77**(8): 527–533.
7. Mark Weiner MD. Presentation at the Midwest Biostatistical Society Meeting, Muncie, Indiana, May 2009.
8. Grannis SJ, Banger A and Harris D. 'Perspectives on Patient Matching: Approaches, Findings, and Challenges [White paper]', Office of the National Coordinator for Health Information Technology, http://healthit.hhs.gov/portal/server.pt?open=512& objID=1240& mode=2 on (2009, accessed 4 May 2010).
9. McGlynn EA, Asch SM, Adams J, et al.. The quality of health care delivered to adults in the United States. *N Engl J Med* 2003; **348**(26): 2635–2645.
10. Hunter D. First, gather the data. *N Engl J Med* 2006; **354**: 329–331.