Analytic Challenges in Studies That Use Administrative Databases or Medical Registries

Draft White Paper for Third Edition of

"Registries for Evaluating Patient Outcomes: A User's Guide"

Introduction

The use of administrative databases and medical registries to provide data for epidemiologic research has blossomed in the last decade,[1] fulfilling prophecies that date to the mid-1970s.[2] Studies that use data collected for a primary purpose other than research have contributed substantial information to understanding the incidence, prevalence, outcomes, and other descriptive characteristics of many diseases. For simplicity, this paper will refer to all such studies as "registry-based research" and to their datasets as "registries." Registry-based research has also contributed information to understanding disease etiology, patterns of treatment and disparities in care, adverse effects, and late events associated with disease treatments, and the comparative effectiveness of some therapies. Despite these achievements, registry-based research sometimes receives criticism because of the potential for it to yield invalid results.[1,3] Weiss,[3] for example, points out the potential for registry-based research to ascertain exposures, outcomes, and potential confounding variables with poor accuracy, or to provide an invalid reference group (unexposed in a cohort design or controls in a case-control design). Ray provides a table of potential pitfalls in "automated database studies," which includes a similar warning about inaccurate measurement of exposure, outcomes, and covariates, a warning about the potential for unmeasured confounding and missing data, and a warning about the potential to include immortal person-time.[4]

While these examples and lists of pitfalls provide valuable guidance, none is unique to registry-based research. Nonrandomized studies of all designs are susceptible to systematic errors arising from mis-measurement of analytic variables,[5] unmeasured confounding,[6] and poor choice of reference group.[7,8] Immortal person-time bias is not limited to registry-based research; it has even plagued a secondary analysis of data gathered in a randomized trial.[9,10] Taking a different approach, this paper begins with a review of the fundamentals of sound study design and analysis. These fundamentals apply to epidemiologic research nested in any study population, but the paper will focus on and illustrate the topics with examples that use registry-based research. In the subsequent section, important considerations in

registry-based research will be discussed, with the recognition that studies nested in other study populations may have the same considerations, but perhaps less often or to a lesser degree than registry-based research.

Fundamentals of Design and Analysis in Registry-Based Research

Statement of Objective

Most productive epidemiologic research begins with a clear statement of objective. This objective might be descriptive, such as to measure the incidence of a particular disease in some population, to characterize the patterns or costs of treatment for a particular disease in some population, or to measure the occurrence of outcomes among patients with a particular disease. The objective might also involve a comparison, such as to compare the incidence of a particular disease in two or more subgroups defined by common characteristics (e.g., etiologic research), to compare the cost or quality of care for a particular disease in two or more subgroups (e.g., health services research or disparities research), or to compare the rate of outcomes among two or more subgroups of patients (often defined by different types or levels of treatment) with a particular disease (e.g., clinical research). In all cases, the overarching objective is to obtain an accurate (valid and precise) and generalizable estimate of the frequency of an outcome's occurrence, or its relative frequency compared across groups.[11] A valid estimate is one that can be interpreted with little influence by systematic error. A precise estimate is one that can be interpreted with little influence by random error. A generalizable estimate is one that provides information pertinent to the target population, the population for which the study's information provides a basis for potential action, such as a public health or medical intervention. Often times the objective will be accompanied by a specific hypothesis, although that is less important than a statement of objective

Selection of a Study Population

Once the study's objectives have been stated, the next step in the research plan is to select a study population. Selection of a study population requires identifying potential participants in time and place, including inclusion/exclusion (admissibility) criteria related to the study's objectives and feasibility. Admissibility criteria related to the study's objectives include focusing on a clinically relevant study population of persons in whom sufficient events will occur to provide adequate precision for the estimates of disease frequency, and in whom the exposure categories will occur at sufficient frequency to provide adequate precision for the estimates of association. These criteria are also used to exclude persons with characteristics that can introduce significant bias into the estimates of disease frequency or estimates of association, and that cannot be controlled easily or adequately in the analysis. Precision and validity criteria for admissibility pertain to all studies, regardless of whether they are nested in a health registry.

Admissibility criteria related to feasibility center on access to the data. Many ongoing cohort studies have established procedures for data sharing. Similarly, most publicly funded registries have established procedures for data access. Investigators must ordinarily provide a statement of the study's objective, a protocol for data collection from the registry and for data analysis, a list of persons who will have access to the data, and a study timeline. Some registries charge a fee for data access, although many do not.

An advantage of registry-based research is the potential to study associations between rare exposures and rare outcomes in a population large enough to provide sufficient precision, with nearly complete follow-up, and with few exclusion criteria pertaining to age, comorbidity, or other factors that sometimes limit participation in clinical trials.[12,13] For example, surveillance databases that monitor adverse events potentially associated with pharmaceuticals identified signals suggesting that use of HMG Co-reductase inhibitors (statins) might increase the risk of amyotrophic lateral sclerosis (ALS).[14,15] The only available epidemiologic evidence came from pooling 41 randomized trials, in which ten ALS cases occurred among 56,352 persons assigned to placebo and nine ALS cases occurred among 64,602 persons assigned to the statins arm.[15] Using Danish medical registries, a case-control study identified 556 cases of ALS or other motor neuron syndromes and 5560 population controls.[16] The odds ratio associating disease occurrence with statins use was 0.96 (95% CI=0.73, 1.28), thereby rapidly and cost-efficiently providing evidence to counter the drug-monitoring studies and with far greater precision than provided by the pooled clinical trials.

Selection of a study population inevitably involves balancing accuracy and generalizability concerns, as well as cost and feasibility considerations. For example, restriction is one of the most effective strategies for control of confounding through study design.[17] If one is concerned about confounding by sex, a simple and effective strategy to control that confounding is to restrict the study population to a single sex. However, such restriction reduces the study's precision by decreasing the sample size, and may also reduce the generalizability of the results (only applicable to half of the target population). An alternative would be to include both sexes and to stratify the analysis by sex. While this approach would improve the generalizability of the results, and allow an evaluation of confounding, the precision of the estimated association would be reduced, and perhaps substantially reduced, if the estimate of effect in men was substantially different from the estimate of effect in women. In this circumstance, the study becomes effectively two studies.

Definition of Analytic Variables

The protocol for an epidemiology study should provide a clear, unambiguous definition of the study's outcome, a description of how it will be measured, and a discussion of the accuracy of that measurement.

When sensitivity of a dichotomous disease classification is non-differential and independent of any errors in classification of exposure categories, and there are expected to be few false positives (near perfect specificity), there will usually be little bias of a ratio measure of association.[5] This exception to the rule that "non-differential misclassification biases towards the null" has important design implications. It suggests that registry-based research studies should be designed to optimize specificity; in fact, to ideally make the specificity perfect so there will be no false positives. Such a design might require more stringent criteria applied to the outcome definition than are ordinarily applied in a clinical setting, and therefore more stringent than might be found in a disease registry. For example, the estimated prevalence of dementia in a cohort of men and women 65 years old or older varied by a factor of ten depending on the diagnostic criteria that were applied.[18] Strategies to reduce inclusion of false-positive cases can include requiring evidence in the patient record of medical procedures (e.g., cholecystectomy for gallstone disease or podiatry examination for diabetes) or interventions (e.g., insulin or glucose lowering medications for diabetes) that provide greater confidence in the validity of the case-finding definition.[19] Such an approach often results in fewer included cases and reduced precision, but improved validity.[20]

If the study objective is to compare the frequency of outcome across subgroups, then the protocol should provide a definition of the exposure contrast(s). It is critical that both the index condition (e.g., the "exposed" or "treated" group) and the reference condition (e.g., the "unexposed" or "untreated/placebo" group) are well defined.[7,21] One frequent shortcoming of epidemiologic research is to compare the occurrence of disease in an index group with the occurrence of disease in all others who do not satisfy the index group definition. Studies of this design are easily constructed with registry-based research, because of the abundance of participants who do not meet the index group definition. This "all others" reference group is therefore usually a poorly defined mixture of persons.[22] For example, if one uses a pharmaceutical registry to compare the incidence of a disease in statins users with the incidence of disease in those who do not use statins, the reference group of non-users will contain persons with indications for statin use but who have not been prescribed statins, as well as persons without indications for statins use. Non-users also differ from users in the frequency of contact with medical providers, which raises the potential for differential accuracy of ascertainment of health outcomes. It is therefore preferable to first ensure that the reference group of non-users contains persons with indications for use of the treatment,[19] and if possible, are receiving alternative therapies for the same indication.[23] If one has a biologic basis to separate statins into categories, such as hydrophilic and hydrophobic statins, then a comparison of users of hydrophilic statins with users of hydrophobic statins would often be more valid. With these definitions, only persons with indications for statins, and treated with statins, are included in the analysis, thereby reducing the potential for confounding by indication and differential follow-up.[24]

Last, considerable attention should be given to identifying and accurately measuring potential confounders and effect-modifiers.[5,25] The opportunity to examine important etiologic questions with considerable precision has expanded significantly with the availability of large registries, but systematic error due to confounding by unmeasured or poorly measured confounders remains a central concern. Fortunately, registries generally capture inpatient and outpatient clinical events and medication use that can characterize co-morbidities and healthcare resource utilization, which can aid in the control of confounding. As discussed further below, information on behavioral and lifestyle factors (e.g., cigarette smoking, alcohol use, diet) is infrequently captured or is poorly measured in registries. Thus, registry-based researchers should carefully consider the available information on confounders before initiating studies. When data on critical confounders cannot be obtained in a registry, and cannot be obtained by linking to another data source, an alternative dataset might be better suited to accomplish the study's objectives. Alternatively, in the presence of unmeasured confounding, researchers can use bias analysis[6,26] to assess the potential impact of residual confounding on their observed findings.[27,28]

Validation Substudies

The goal of quality study design and analysis is to reduce the amount of error in an estimate of association. With that goal in mind, investigators have an obligation to quantify how far they are from this goal, and bias analysis is one method by which to achieve this goal.[6,26] Bias analysis methods require data to inform the bias model, and these data are obtained from internal or external validation substudies. Registry-based research is often amenable to collection of internal validation data, for example by medical record review. In addition, many registries have internal protocols that constantly validate at least some aspects of the data in the registry. The validation data generated by these protocols can provide an initial indication of the data quality. To facilitate data collection for study-specific internal validation studies, investigators should consider the important threats to the validity of their research while designing their study and allocate project resources accordingly. This consideration should immediately suggest the corresponding bias analyses, which then inform the data collection that will be required to complete the bias modeling.

For example, in the study of statin use related to ALS and neurodegenerative diseases described above,[16] the ICD-10 code used to identify cases (G12.2) corresponded to diagnoses of ALS or other motor neuron syndromes. The investigators therefore selected a random sample of 25 persons from among all those who satisfied the case definition and a clinician investigator reviewed their discharge summaries. The proportion of these twenty-five who did not have ALS (32%) was used to inform a bias analysis to model the impact of these false-positive ALS diagnoses. Assuming a valid bias model, the bias analysis results

showed that the null association was unlikely to result from the non-differential misclassification of other diseases as ALS.

In this example, there was no effort to validate that non-cases of ALS were truly free of the disease. Non-cases are seldom validated, because false-negative cases of rare diseases, especially, occur very rarely. Furthermore, validating the absence of disease often requires a study-supported medical examination of the non-case patients, an expensive, time-consuming, and invasive procedure. Prevalent diseases with a lengthy preclinical period and relatively simple diagnostic tests, such as diabetes, are more amenable to validation of non-cases. The ALS example also illustrates that an internal validation study requires protocol planning and allocation of study resources to collect the validation data. A protocol should be written that specifies how participants in the validation sample will be selected from the study population. Participation in the validation substudy might require informed consent to allow medical record review, whereas the registry data itself might be available without individual informed consent. These aspects should be resolved in the planning stage, and the analytic plan should include a section devoted to bias modeling and analysis.[6]

Important Considerations

Once an investigator decides to pursue a research objective using registry-based research, there are a number of important considerations to evaluate before undertaking the study. These considerations mostly pertain to the quality and completeness of the registry's data,[29,30] and especially to the potential for systematic errors in the database to affect the validity of the study's result.

Structural Framework for Data Collection

Health registries collect data for various primary purposes[31] and can be categorized as follows: 1) data collected for the purpose of reimbursing health care providers; 2) data collected for the purpose of monitoring care provided to beneficiaries of an integrated healthcare system; 3) data collected for the purpose of surveillance regarding a particular disease or disease category; 4) data collected for the purpose of surveillance for individuals with a specific exposure; and 5) data collected on persons with a single admission-defining disease or medical procedure. Each has strengths and limitations (Table 1) to consider when evaluating for use in studies.

Registries that collect information for reimbursement (e.g., Medicare, Medicaid, or Ingenix), which are sometimes called "claims" or "administrative" databases, are quite useful for understanding healthcare costs and can provide important surveillance information on clinical practices and outcomes. However,

they may be susceptible to systematic errors if data entries are manipulated by the data generators to affect (likely increase) their reimbursement. For instance, certain clinical conditions with high reimbursement rates may be preferentially reported on claims in patients with those conditions, but who present in the hospital or outpatient setting with other clinical issues, particularly if the presenting conditions are reimbursed at lower rates. The accuracy of some claims datasets have been questioned for diagnoses and procedures including dialysis,[32] weight management,[33] neutropenia,[34] heart failure,[35] diabetes,[36] and functional outcomes after prostatectomy,[37] as examples. On the other hand, the accuracy of registered diagnoses can be quite good.[38] The accuracy of the claims data for its intended objective should therefore be considered, and preferably estimated quantitatively by an internal validation substudy.[39,40] Alternatively, estimates of the accuracy may be available from an analogous study population from the same or a similar claims dataset; an example is an external validation study. Claims data often lack important information on laboratory parameters, diagnostic test results, and behavioral and lifestyle characteristics, which may limit their utility for research in some topic areas.

The second type of registry collects information on the healthcare provided to beneficiaries within an integrated healthcare system. This system can be a health insurer (e.g., Kaiser Permanente), a benefits program provided to selected individuals (e.g., Veteran's Health Administration), or a national healthcare system (e.g., General Practice Research Database). These registries typically use an integrated electronic medical records (EMR) system to capture healthcare information directly from physicians' offices, hospitals, pharmacies and other sites where care is provided (e.g., infusion centers, surgical centers). The granularity and quality of data captured in these registries is quite good and includes demographic and clinical characteristics, medication use, major clinical events including death, and importantly, results of diagnostic tests and laboratory assays. As with many epidemiology studies, some registries are limited in their geographic coverage and in the demographic characteristics of their patient populations. This lack of representativeness may affect the generalizability of results from studies nested in them.

The intended purpose of a third set of registries is surveillance of the incidence and outcomes related to a particular disease or disease category. These surveillance registries often pertain to infectious diseases, cancer, and end-stage renal disease (ESRD). Surveillance for infectious diseases sometimes recognizes that only a proportion of cases will be reported, but assumes that the sensitivity and specificity of reporting remain constant over time, so that changes in the relative frequency of reported incidence provides a signal regarding the true incidence in the population. Thus, although the data quality is high, the completeness may be low. In contrast, both the data quality and completeness in most cancer registries are quite high, and the motivation for manipulation to influence reimbursement does not exist

because the registry data are not used for that purpose. For example, the US Cancer Surveillance, Epidemiology, and End Results (SEER) registry has a history of quality control and improvement dating to its inception in 1973 (http://seer.cancer.gov/about/uses.html) and has been linked to the Medicare administrative database to provide data on cancer treatments and outcomes. In the US and some other countries, patients with ESRD (patients receiving chronic dialysis or who are transplant recipients) are guaranteed coverage of all dialysis services including medications, procedures, and hospitalizations. These benefits extend throughout the patient's life and require significant resources. As such, countries have established surveillance programs, such as the United States Renal Data System (USRDS, http://www.usrds.org/) to monitor the healthcare provided to these patients and the costs associated with their healthcare.

The fourth type of registry collects data on patients with a common exposure, and is commonly used as part of a post-marketing pharmacovigilance program related to a biologic or pharmaceutical product or a medical device. This type of registry is typically designed to monitor the incidence of adverse events related to the exposure.

A last type of registry is a clinical registry of persons with a single admission-defining disease or medical procedure. In fact, the first known health-related registry was the Leprosy Registry in Norway, initiated in 1856. In keeping with this history, many of the current clinical registries are found in Scandinavia. For example, the Danish government supports clinical databases used for quality assurance and research (e.g., breast cancer, colorectal cancer, hip arthoplasty, and rheumatologic diseases), as well as disease registries (e.g., the multiple sclerosis registry) used for monitoring and research.[41] In fact, a central objective of disease-specific registries may be to provide an infrastructure for clinical trials pertaining to treatments for the disease. The main advantage of these registries and databases is the quality of data on disease characteristics, received treatments, and outcomes related to the disease. The main disadvantage is that they are difficult to use for studies of etiology of the disease that initiates membership in the registry, since the registry includes only persons with the disease.

Changes in Coding Conventions over Time

A common problem with registry-based research is the impact of changes in coding conventions over the lifetime of the registry. These can take the form of diagnostic drift,[42] changes in discharge coding schemes, changes in the definition of grading of disease severity, or even variations in the medications on formulary in one region but not others at different points in time. The Danish National Registry of Patients (DNRP) is a registry of patient contacts at Danish hospitals. From 1977 to 1993, discharge diagnoses were coded according to ICD-8, and from 1994 forward discharge diagnoses were coded

according to ICD-10. ICD-10 included a specific code for Chronic Obstructive Pulmonary Disease (J44), whereas ICD-8 did not [ICD-8 496 (COPD not otherwise specified) did not appear in the DNRP]. In addition, from 1977 to 1994 the DNRP registered discharge diagnoses for only in-patient admissions, but from 1995 forward discharge diagnoses from outpatient admissions and emergency room contacts were also registered. COPD patients seen in outpatient settings before 1995 were therefore not registered, and that excluded patients who likely had less severe COPD on average. The change in ICD coding convention in 1994 and the exclusion of outpatient admissions before 1995 presented a barrier to estimating the time trend for incidence of all admissions for COPD in any period that overlapped these two changes to the DNRP.[43]

The General Practice Research Database (GPRD) is a medical records database capturing information on approximately 5% of patients in the United Kingdom.[44] Information is directly entered into the database by general practitioners who were trained in standardized data entry. In 1987 when the GPRD was initiated, diagnoses were recorded using Oxford Medical Information Systems (OXMIS) codes, which were similar to ICD-9 codes. In 1995, the GPRD adopted the Read coding system, a more detailed and comprehensive system that groups and defines illnesses using a hierarchical system. Without knowledge of this shift in coding and how to align codes for specific conditions across the different coding schemes, studies using multiple years of data could produce spurious findings.

Other Data Quality Considerations

Selection of Registered Population

An important advantage of some registry-based research is that it is population-based, and therefore provides good representativeness for the target population. However, not all registry-based research provides this advantage. For example, the US Veterans Health Administration medical registries provide an important resource for registry-based research. A recent analysis of persons receiving Veterans Health Administration services in fiscal years 2004 and 2005 reported a mortality rate due to accidental poisoning of about 20 per 100,000 person-years.[45] However, this registry includes only US military veterans, a selected subpopulation of the US population, with a higher proportion of men than the overall population, and probably an unrepresentative proportion of other characteristics as well. The rate of accidental poisonings was thus almost twice that of the US general population, after adjusting for differences in the age and sex distributions. Similarly, the US Medicare administrative registry provides an important resource for registry-based research, including its links with the US SEER cancer registry mentioned above. However, the registry includes only Medicare recipients, almost all of whom are sixty-five years old or older, and excludes members of this population who participate in managed health care

plans. Whether the lack of representativeness in these two examples, and others like them, affects inference made to the target population depends on the particular topic.

Probability of Registration in Relation to Disease Severity

A second type of incomplete data arises at the level of registered persons, rather than afflicting the whole registry. In an earlier example, cases of COPD were registered in the Danish National Registry of Patients in reference to ICD-8 before 1994 and in reference to ICD-10 thereafter. Only inpatient diagnoses of COPD were registered in the DNRP before 1995, and inpatient, outpatient, and emergency department contacts were registered thereafter. At no time has the DNRP registered COPD cases diagnosed and treated only by a Danish General Practitioner. The least severe cases of this progressive disease are, therefore, missing from the DNRP throughout its history,[46] and patients treated as outpatients are missing from the DNRP before 1995. Similar problems occur with hospital-based registries of other progressive diseases, such as diabetes, Alzheimer's disease, or Parkinson's disease. Patients treated by their general practitioner will often appear in the hospital registry with the proper discharge diagnosis eventually, since these progressive diseases become more severe over time. The less severe cases do not appear in hospital discharge registries, which presents a barrier to studies of population-based incidence or prevalence, as well as presenting a barrier to accurately identifying whether exposure to a potential etiologic agent preceded the disease diagnosis,[47] since neither the date of first diagnosis by the general practitioner, nor the date of symptom onset, is recorded.

Registry data often lack accurate measurements of lifestyle and behavioral factors, such as tobacco use, alcohol drinking, exercise habits, and diet. Some registries can provide proxy measurements of these behavioral factors. For example, poor lung function or diagnosis of COPD is a proxy marker for tobacco smoking history, alcohol related diseases such as cirrhosis or prescriptions for disulfiram can be used as proxy markers for alcohol abuse, and medically diagnosed obesity may be a proxy marker for poor diet and lack of exercise. None of these proxies provides a reliable measure of the actual concept, however. For diseases that can be identified by use of specific medications, one could compare the incidence of that medication use with the incidence in the hospitalization registry to estimate the proportion of total cases that are registered. Comparison of the date of onset of the medication use with the date of first outpatient or inpatient diagnosis of the disease would provide an estimate of the typical delay between diagnosis by a general practitioner and progression of the disease to a severity level treated in the outpatient or inpatient setting.

Missing Data

Item non-response and missing data at the level of an individual record are often less of a problem for registry-based research than for comparable cohort studies. Cohort studies that rely on participation by study subjects are subject to attrition and non-response. Attrition occurs when participants early in the cohort's follow-up stop replying to regularly mailed surveys, telephone interviews, or emailed data collection instruments. These losses to follow-up are sometimes related to exposure characteristics and health outcomes, which introduces a form of selection bias,[48] even if subjects rejoin the study at a later time.[49] Item non-response occurs when a participant answers a survey or interview, but does not provide a response for one or more of the data fields. Item non-response can also occur when data on an exposure or outcome are collected by other methods, such as when a biospecimen is unavailable to provide tissue for an assay of a genetic or protein biomarker. This missing data may also be related to exposure and disease characteristics, and can introduce a bias, although reliable methods have been developed to resolve bias from item non-response (missing data) in many circumstances.[50] Likewise, inverse probability weighting can sometimes be used to address selection bias and loss-to-follow-up,[51] although it has seldom been implemented to date.

Registry-based research ordinarily uses data collected for a primary purpose other than research. Item non-response (missing data) is also often less of a concern, since the registries often have inherent quality control methods to assure high data completeness. Missing data can, however, plague registry-based research in other ways. For example, left-truncation is sometimes an important problem in registry-based research, and is basically a missing data problem (although it can also be conceptualized as an information bias).[52] Left truncation occurs when information required to characterize prevalent exposures, covariates, or diseases precedes the establishment of the registry. With left truncation, unexposed persons (e.g., non-users of a medication) may have been users before the registry was established and apparently incident cases of a disease may have been diagnosed before the registry was established, which would make them prevalent cases. Furthermore, covariate information collected at the inception of the registry might have been affected by the medical history before the registry was established. For example, blood pressure measured soon after a registry began might be affected by blood pressure medications prescribed before the registry began. Characterizing this initial measurement as baseline (i.e., preceding the first recorded prescription for blood pressure medications) would fail to account for the effect of the prevalent prescription for blood pressure medications, which was prescribed during the left truncation period.

As a second example, in a study of the association between metformin use and the occurrence of breast cancer, the prescription registry used to ascertain use of metformin among diabetic patients was not established until after the medication came to market.[53] Data on use of metformin were therefore left-

truncated, which can be conceptualized as a missing data problem for time-varying characterization of metformin use in the years preceding the registry (see Ibrahim et al.[54] for a review of methods to model time-varying data), or can be conceptualized as the more general problem of having poor sensitivity of ever/never classification of metformin use.

Left-truncation is a common problem whenever prevalent conditions may have preceded the establishment of a registry. For example, many etiologic epidemiology and clinical epidemiology studies exclude prevalent cases of the outcome at the inception of follow-up. However, some cases of disease may have occurred before follow-up began and even before the registry's inception, and these prevalent cases would be impossible to identify unless they also appeared in the registry after its inception but before the follow-up time began. For many prevalent diseases with good survival, contact with the medical system is frequent, so most prevalent cases should be identifiable after the registry is five to ten years old. However, the potential for left truncation to mask some prevalent cases of the disease under study should be considered as a question specific to the research topic.

Right-censoring can also occur in registry-based research. For example, studies that use birth-registries to ascertain congenital defects usually fail to detect defects, such as congenital heart anomalies, that are diagnosed later in life. These defects are usually never recorded in the birth-registry, so must be ascertained by some other method. Without such continued follow-up, the measurement of the outcome is right-censored at the date of last follow-up by the birth registry.

Left-truncation and right censoring are specific examples of the more general problem of data gaps. Data gaps occur when registries pertain to only a particular subgroup of the larger population, and membership in that subgroup is dynamic. Examples include persons covered by Medicaid and members enrolled in managed care plans. In both examples, the registries pertain to participants in a health insurance program, and membership in those programs can change frequently. Data are collected only while the participants are members. If membership is lost and restored again later, there will be a data gap. Importantly, membership in these plans might be related to other characteristics that affect health, such as socioeconomic status or employment.[55] Similar problems can arise when there are gaps in residency and the registry is based on national health care data, or when persons have health insurance from more than one source.

Data gaps in registry-based research can also arise when medications are dispensed in the hospital, since many registries do not capture in-hospital medication use, leading to a form of information bias. In drug safety studies examining mortality risk related to the use of a particular medication, missing in-hospital

medication use can result in spurious estimates of treatment effects.[56] This bias was illustrated in a case-control study examining mortality risk related to inhaled corticosteroid use from the Saskatchewan, Canada database. Analyses that failed to account for missed corticosteroid use during hospitalization events preceding death or the matched date for controls showed a beneficial effect (RR=0.6, 95% CI: 0.5, 0.73). The RR estimates changed markedly once the missing in-hospital corticosteroid use was included (RR=0.93, 95% CI: 0.76, 1.14 and RR=1.35, 95% CI: 1.14, 1.60).[56] This bias has also been observed in studies of injectable medications in dialysis patients where hospitalization events preceding death resulted in spuriously low effect estimates.[57]

Confounding by Indication

Confounding by indication may occur in non-randomized epidemiologic research that compares two treatments (or treatment with no treatment).[58] In the absence of randomization, the indications for selecting one treatment in preference to another (or in preference to no treatment) are often also related to the outcome meant to be achieved or prevented by the treatment.[59] For example, randomized trials in younger breast cancer patients have shown that chemotherapy prevents breast cancer recurrence.[60] However, in a nonrandomized study of older breast cancer patients, those who received chemotherapy had a higher rate of recurrence than those who did not, probably because chemotherapy was offered only to the women with the most aggressive cancers.[61] This example is a classic illustration of confounding by indication. Importantly, this study collected complete detailed data on every prognostic marker of recurrence and all of the other breast cancer treatments, yet adjustment for this detailed suite of variables did not resolve the confounding by indication, even using more advanced methods.[24]

Registry-based research is as susceptible to confounding by indication as any other design, although strategies to reduce the strength of this confounding have been proposed[22] and may be most successful when used in the large study populations often achievable only in registries.[62] Explained here is a special class of confounding by indication, which might arise especially in registry-based research: time-dependent confounding by indication generated by dynamic dosing.[63] Dynamic dosing refers to the clinical situation in which a medication's dose is titrated (increased or decreased) in response to a changing biomarker or clinical measurement on which the medication acts (i.e., a clinical intermediate).[63] Examples include diabetes medications titrated in reaction to hemoglobin A1c (HbA1c) measurements, erythropoiesis stimulating agents (ESAs) titrated in reaction to hemoglobin levels, blood pressure medications titrated in reaction to systolic and diastolic blood pressure values, and anti-retroviral therapy titrated in reaction to CD4 counts. The clinical intermediate is therefore both a consequence of therapy and a predictor of future therapy. Time-dependent confounding arises when the clinical intermediate is

also a prognostic indicator.[64] For example, hemoglobin concentration is a time-dependent confounder of the effect of ESA therapy on survival because it is a risk factor for mortality, it predicts future ESA dose, and past ESA therapy predicts future hemoglobin concentration. Dynamic dosing therefore introduces time-dependent confounding of the treatment's association with outcomes in the presence of this structure of confounding-by-indication.[63]

It is important to recognize that the structure requires the clinical intermediate to be both a causal intermediate and a confounder. If it is only a confounder, such as baseline comorbidity or perhaps time-dependent comorbidity, the confounding can be addressed by conventional analytic methods. However, when the causal structure indicates that the clinical intermediate is both a causal intermediate and a confounder, inverse probability of treatment weighting (IPTW) with marginal structural models (MSM) has been proposed as one method for valid adjustment.[65] Pharmacoepidemiologic studies that have used marginal structural models to address time-dependent confounding have shown significant improvements in confounding control relative to traditional time-dependent analysis.[66-68] In a study of the effect of highly active antiretroviral therapy (HAART) on the time to acquired immunodeficiency syndrome (AIDS), the hazard ratio using standard time-dependent Cox regression to adjust for time-varying covariates such as CD4 count and HIV RNA level was 0.81 (95% CI: 0.61, 1.07). Using a marginal structural model, this effect was strengthened substantially (HR=0.54, 95% CI: 0.38, 0.78), providing stronger evidence of the benefit of HAART therapy.[66] Studies examining the effect of titrated ESA doses on mortality risk in dialysis patients that have used marginal structural models have found hazard ratio estimates at or below the null,[67,68] whereas results from traditional models found substantially elevated hazard ratio estimates.[67]

Precision Considerations When Standard Errors Are Small (Over-Powered)

The large size of the study population that can often be included in a registry-based research study is both a strength and a limitation. The sample size allows adjustment for multiple potential confounders with little potential for over-fitting or sparse data bias[69] and allows design features such as comparisons of different treatments for the same indication (comparative effectiveness research) to reduce the potential for confounding by indication.[22] Nonetheless, systematic errors remain a possibility, and these systematic errors dominate the uncertainty when estimates of association are measured with high precision as a consequence of a large sample size.[70] When confidence intervals are narrow, systematic errors remain, and inference or policy action will potentially result, investigators have been encouraged to employ quantitative bias analysis to more fully characterize the total uncertainty.[26] Bias analysis methods have

been used to address unmeasured confounding,[71] selection bias,[72] and information bias[71,73] in registry-based research.

A second potential problem is the possibility of overweighting results from registry-based research in a quantitative meta-analysis of an entire body of research on a particular topic. In such meta-analyses, weights are in proportion to the inverse of variance, so large studies carry most of the weight. The variance, however, measures only sampling error; it does not measure systematic error. This problem of large studies dominating the weights pertains to any meta-analysis that includes one or two studies much larger than the others. However, given the large sample sizes often achieved by registry-based research, the high-weight studies may often come from studies nested in registries. For example, in a 2004 quantitative meta-analysis of eleven prospective studies of the association between pregnancy termination and incident breast cancer,[74] the two registry-based studies[75,76] accounted for 54% of the weight in the meta-analysis, but only 18% (2 of 11) of the studies. Random effects meta-analyses[77] and other weighting methods[78] provide only a partial solution to this potential overweighting, and only in some circumstances. Meta-analysts should therefore consider the potential for registry-based research to be overweighted in their quantitative summary estimates. A plot of the inverse-normal of rank percentile against the corresponding study's estimate of association and confidence interval provides a visual depiction of the distribution of study results,[79] without undue influence by overpowered studies (see, for example, the aforementioned meta-analysis of the association between pregnancy termination and breast cancer risk[74]).

Special Opportunities

As noted earlier, registry-based research runs the gamut of research topics. There are, however, several research areas to which registry-based research studies are particularly well-suited.

Rapid Response to Emerging Problems, With Prospective Data

Registry-based research is ordinarily secondary to another primary purpose. While the collected data may not be optimized to a particular research topic, it is often possible to use the collected data for rapid response to emerging research problems.[80] The study mentioned above of the association between statins medication and incident ALS is also a suitable example here. Drug surveillance databases had identified a higher than expected prevalence of statins medications associated with reports of ALS. A pooled analysis of trials data revealed no association, but was limited by the small number of ALS cases, short duration of follow-up, and potential for cross-over from the placebo arm to statins treatment after the trial finished.[81] Thus, there was little evidence to evaluate the potential causal association between this highly

effective drug class—which prevents the incidence of cardiovascular morbidity and mortality[82,83]—and the incidence of ALS, a progressive neurodegenerative, terminal, and dread disease.[84]

The precisely measured null association reported in the case-control study[16] provided a rapid and reliable basis to assuage concerns about an etiologic association between statin use and ALS occurrence. Imagine what would have been required for a purposefully designed study to evaluate the association. The pooled trials result had included nearly 120,000 persons observed over more than 400,000 person-years, yet included only 19 cases of this rare disease. Few existing cohort studies would have had sufficient person-time to expect substantially more cases, and a cohort study designed to evaluate the association would have required a substantial investment of time and financial support.

A case-control study might have been feasible, but imagine the resources required to enroll and interview an equivalent number of ALS cases as were included in the registry-based study (~550) and their matched controls. Furthermore, a case-control study of this design would likely have been susceptible to recall bias and selection bias.[5,8] The registry-based research study avoided both of these biases.[16] Recall bias was avoided by ascertaining statins use from a prescription registry. These prescriptions were recorded before the ALS incidence, so could not have been affected by the subsequent disease occurrence. Selection bias was avoided because all ALS cases in the region during the follow-up period were included, and controls were selected from the Civil Registration System. Neither case nor control status, nor use of statins, was likely to be associated with participation. Thus, the registry-based research study on this topic provided a rapid, cost-efficient, and precise result on an important public health topic, which otherwise would have gone unevaluated or would have required a substantial investment of time and finances to achieve an equivalent, or possibly more biased, result. This study provides a good example of the value of registry-based research in such circumstances.

Cost-Efficient Hypotheses-Scanning Analyses

Registry-based research can sometimes evaluate multiple associations with only a marginal increase in cost over the evaluation of a single association. The United States Food and Drug Administration's Sentinel Initiative will use an active surveillance system within electronic data from healthcare information holders to monitor the safety of all FDA-regulated products (http://www.fda.gov/Safety/FDAsSentinelInitiative).[85] Similarly, the EU-ADR project aims to use clinical data from health registries, combined with prescription databases, to detect adverse drug reactions.[86,87] The project uses text mining, epidemiological, and computational techniques to analyze the electronic health records, with the goal of detecting combinations of drugs and adverse events that merit further investigation (http://www.alert-project.org).

As a second example, Latourelle et al. used registry-based research to evaluate the association between estrogen-related diseases, such as osteoporosis or endometriosis, and the occurrence of Parkinson's disease.[88] To be categorized as "exposed" to these diseases, cases or controls had to have them appear as discharge codes in the hospital registry before the first discharge code for Parkinson's disease. For relatively little additional cost, Latourelle et al also evaluated the association between two-hundred other diseases and the subsequent diagnosis of Parkinson's disease as a hypothesis scanning study, with the objective of suggesting new ideas regarding Parkinson's disease etiology.[88] The analysis adjusted for multiple comparisons using empirical Bayes methods designed to reduce the emphasis on potentially false-positive associations.[89] This potential for cost-effective hypotheses scanning studies as an explicit objective of registry-based research should be viewed as a strength of such research, not a limitation, so long as the objective is appropriately labeled as such. Hypotheses suggested by these types of studies are often further investigated using studies designed specifically for the topic.

Hybrid Designs

Registry-based research does not necessarily have to be limited to data collection from secondary data sources. Hybrid designs allow the use of registry-based research for some aspects of data collection, and primary data collection for others. For example, a study of drug-drug and gene-drug interactions that might reduce the effectiveness of tamoxifen therapy began by identifying eligible breast cancer patients using the Danish Breast Cancer Cooperative Group's clinical registry.[90] This clinical registry also provided data on prognostic factors, such as tumor diameter and lymph node evaluation, and on treatments, such as chemotherapy and radiation therapy. Linkage with the Danish Civil Registration System provided data on vital status, linkage with the Danish National Patient Registry provided data on comorbid diseases, and linkage with the Danish National Registry of Medicinal Products provided data on use of prescription medications. Thus, for relatively low cost, a cohort of breast cancer patients with complete medical, prognostic, and breast cancer treatment data was assembled. A case-control study was then nested in this cohort by identifying cases of breast cancer recurrence and matching controls to them by risk-set sampling.[8] Once cases and controls had been identified, their tumor blocks were collected from the Danish National Pathology Registry,[91] and these were used for the necessary bioassays. Thus, registry-based research allowed identification of the source population, selection of cases and controls, and provided all but the bioassay data. This data, which is expensive to collect, was only obtained for about 13% of the members of the total cohort. This hybrid design demonstrates that registry-based research will remain an important contributor, even in the era of personalized medicine.

A second example of a hybrid design linked survey data collected over the internet with registry-based research.[92,93] Participants were recruited by advertisements on web sites likely to be visited by women who intended to become pregnant. They were linked to the study's web site, where they completed an enrollment screening questionnaire, and then an interview regarding socio-demographics, reproductive and medical history, lifestyle, and other factors. Enrolled participants were then contacted bimonthly by e-mail for 12 months or until they reported that conception had occurred. Data obtained from the web-based questionnaires were also linked to nationwide registries, which allowed collection of additional data on confounders and outcomes, as well as an assessment of the validity of some of the self-reported data, such as prescription drug use. This study again demonstrated the use of registry-based research, in combination with primary data collection, to provide a cost-efficient resource for collecting some aspects of the study data. In contrast to the previous example, the cohort in this example was enrolled following more typical cohort study strategies, and not by using the registries to identify a source population.

Hybrid designs have also been used to collect data by medical record review for data fields that are available for a subset of participants in a registry.[94] Thus, the registry data provide a cost-efficient resource for initial data collection, which is then supplemented as necessary by medical record review or another primary data collection method to complete the dataset. Once an investigator is open to the potential for hybrid designs, and there are registry-based resources suitable to the research topic, the opportunities for combining registry data with primary data collection are limited only by the investigator's creativity.

Ample Data Allows for Novel Designs

As mentioned above, the ample data often available from registry-based research can lead to overweighting of such studies in quantitative meta-analyses. While this problem may be disadvantageous, a compensating advantage is the opportunity to use registry-based research to implement novel study designs. For example, confounding by indication and other biases often plague clinical epidemiology,[3,24] even in the era of comparative effectiveness research. However, the ample study size often provided by registry-based research can overcome these threats to validity in some situations. The large sample size might allow a design with carefully restricted exposure groups,[1] such as including only new users of a pharmaceutical,[95] whereas conventionally sized cohort studies would not always have sufficient study size to implement such a design. The new user design in turn facilitates other advanced designs, such as propensity score matching and instrumental variable analyses,[22] which are intended to further counteract these threats to validity. These and other novel designs can be implemented in studies of any size, but are likely most effective when the study size is large.[96]

Data Pooling Methods

Although registry-based research often provides relatively large study size within a research topic area, a study's power may still be insufficient if the study must be restricted to rare exposure subgroups or if the study outcome is rare. In these cases, data pooling across similar registries may allow sufficient sample size to provide adequate power, and provides advantages over conventional meta-analyses because they allow simultaneous and consistent data analyses. Such pooling projects face substantial challenges.

First among these challenges is harmonization of the data elements. To pool an analysis, data collected from different registries must be able to provide analytic variables (exposure, confounders, modifiers, and outcomes) with equivalent categorizations and definitions. Such data harmonization can be quite challenging. Harmonization of data elements categorized differently in two or more registries, or differentially available data elements, may pose an insurmountable barrier to pooling. For example, one registry might include data on behaviors like alcohol and tobacco use, whereas a second registry might not. The pooling project would then face the unenviable decision of controlling for these behaviors for some, but not all, data centers (in which case the analysis becomes comparable to a conventional meta-analysis), or abandoning control for these variables at all centers in order to achieve the data harmonization goal. Differences in the conceptual underpinnings of data elements may be more common. Even a variable as conceptually simple as the Charlson comorbidity index[97] can present surprising challenges when subject to harmonization considerations. The Charlson index includes nineteen comorbid conditions, including for example diabetes. As mentioned above, some registries might be able to ascertain diabetes diagnosed in all medical settings (general practitioner, outpatient setting, and inpatient setting), whereas others might be able to ascertain diabetes diagnosed in only a subset (e.g., only general practitioner or only outpatient specialty clinics). Diabetes is defined differently in the different registries, which are not strictly harmonious, and therefore contribute differently to the Carlson index. While the definition of the Charlson variable may be harmonious across the pooled registries, the underlying conceptualization is different, and this difference could result in differences in the strength of confounding by the comorbidity variable or in the degree to which it modifies the association between an exposure contrast and outcome.

Ethical and legal constraints, which are often placed on data sharing, present a second important challenge to pooling projects. Pooling of de-identified datasets can sometimes be arranged through "data use agreements," but even these arrangements can be quite challenging and time-consuming. Rassen et al. compared four methods of pooling de-identified datasets:[98] (a) full covariate information, which may violate privacy concerns, (b) aggregated data methods, which aggregate patients into mutually stratified

cells with common characteristics, but usually delete cells with low frequency counts that might defeat the privacy protections of large frequency counts, (c) conventional fixed or random effects meta-analysis, which provides only summary estimates of association for pooling, and (d) propensity score-based pooling, for which a propensity score summarizes each individual's covariate information. They reported that the last alternative provided reasonable analytic flexibility and also strong protection of patient privacy, and advocated its use for studies that require pooling of databases, multivariate adjustment, and privacy protection.[98]

More recently, Wolfson et al. proposed a pooling method that requires no transfer of record level data to a central analysis center.[99] Rather, the central analysis center implements statistical computing code over a secure network, accessing record-level data maintained on servers at the individual study centers. Data aggregation occurs through return of anonymous summary-statistics from these harmonized individual-level databases, and even iterative regression modeling can be implemented. The advantage is a reduced burden to comply with ethical and legal requirements to protect privacy, since no record-level data are ever transferred. The disadvantages include requirements for strong data harmonization, secure networks that satisfy regulatory oversight, and assurances that no record-level data are transmitted. It is possible that some summary statistics could violate standards for de-identification, but safeguards can be implemented to prevent transmission of such summary statistics.

These new methods for pooling provide exciting opportunities for pooled projects. At this writing, investigators who choose to undertake them should expect delays required to explain the methods to regulators with oversight of data protection, who are not yet familiar with them. In addition, it is likely that implementing the methods for the first few projects will be challenging. With those caveats in mind, the path should be blazed, because once the methods are familiar and reliable, new research opportunities and efficiencies will inevitably arise. Investigator teams without the time, resources, or patience to implement these new methods can ordinarily rely on conventional meta-analysis methods,[100] which solves the privacy protection concerns, but also has some important disadvantages by comparison.[98,99]

Conclusion

Registry-based research has made important contributions to descriptive epidemiology, public health epidemiology targeted at disease prevention, and clinical epidemiology targeted at improving disease outcomes or estimates of disease prognosis. Investigators who conduct registry-based research should first focus on the fundamentals of epidemiologic design and analysis, with the goal of achieving a valid, precise, and generalizable estimate of disease frequency or association. Beyond the fundamentals,

registry-based research presents special challenges for design and analysis, and special opportunities as well, so researchers should be aware of both in order to optimize the yield from their work.

Table 1: Types of registries used for registry-based research, and their typical advantages and disadvantages

Registry Type	Strengths	Limitations	Examples
Reimbursement purposes ("claims" data)	– Population-based – Captures inpatient and outpatient clinical events – Captures mortality data – Captures oral & injectable medication use – All claims are adjudicated	– Specific patient populations (65+ yrs old/disabled; employed) – Limited information on subject characteristics (e.g., lifestyle factors) – Does not capture laboratory or test results – Missing medication use in the hospital – Reflects regional practices	– Medicare – Ingenix – Marketscan
Monitor health care provided to beneficiaries	– Population-based – Captures inpatient and outpatient clinical events – Captures oral & injectable medications – Captures subject characteristics (e.g., BMI, smoking, blood pressure)	– Limited racial and ethnic diversity – Specific healthcare practices (e.g. selected formulary)	– Nordic Hospital Registries (Denmark, Sweden, Finland) – Kaiser Permanente, Group Health Cooperative, General Practice Research Database (GPRD)
Disease or disease category surveillance	– Population-based – Captures granular disease-specific data (e.g., cancer stage) – Captures outcome events	– Variable amounts of healthcare utilization information – Limited information on subject characteristics (e.g., lifestyle factors)	– SEER, SEER-Medicare – United States Renal Data System (USRDS)

Registry Type	Strengths	Limitations	Examples
Exposure surveillance	– Prospectively designed – Typically collects granular information on relevant covariates – Designed to capture all potential drug-related adverse events	– Limited information on comparator treatments	
Quality assurance or research regarding patients with a single admission-defining disease or procedure	- High quality data on the index disease or procedure - High quality data on the treatment and outcomes associated with the index disease or procedure - Potential to link with other data sources to obtain more complete data	– Absence of an equivalent comparison group without the index disease or procedure – Limited data on health conditions and treatments not related to the index disease or procedure – Limited data on behavioral health (tobacco use, diet, exercise, and alcohol consumption)	– Danish Breast Cancer Cooperative Group – Danish Multiple Sclerosis Registry – Danish Hip Arthoplasty registry

References

(1) Ray WA. Improving automated database studies. *Epidemiology.* 2011;22(3):302-304.

(2) Federspiel CF, Ray WA, Schaffner W. Medicaid records as a valid data source: the Tennessee experience. *Med Care.* 1976;14(2):166-172.

(3) Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? *Epidemiology.* 2011;22(3):292-294.

(4) Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol.* 2008;167(4):492-499.

(5) Rothman KJ, Greenland S, Lash TL. Validity in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 128-47.

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

(7) Rothman KJ, Greenland S, Poole C, Lash TL. Causation and Causal Inference. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 5-31.

(8) Rothman KJ, Greenland S, Lash TL. Case-Control Studies. In: Rothman K, Greenland S, Lash TL, eds. *Modern Epidemiology, 3rd Edition*. Philadelphia: Lippincott, Williams & Wilkins, 2008: 111-27.

(9) Souhami L, Bae K, Pilepich M, Sandler H. Impact of the duration of adjuvant hormonal therapy in patients with locally advanced prostate cancer treated with radiotherapy: a secondary analysis of RTOG 85-31. *J Clin Oncol.* 2009;27(13):2137-2143.

(10) Collette L, Studer UE. Selection bias is not a good reason for advising more than 5 years of adjuvant hormonal therapy for all patients with locally advanced prostate cancer treated with radiotherapy. *J Clin Oncol.* 2009;27(33):e201-e202.

(11) Greenland S, Rothman KJ. Measures of Occurrence. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 32-50.

(12) Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Jama.* 2004;291(22):2720-2726.

(13) Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch Intern Med.* 2002;162(15):1682-1688.

(14) Edwards IR, Star K, Kiuru A. Statins, neuromuscular degenerative disease and an amyotrophic lateral sclerosis-like syndrome: an analysis of individual case safety reports from vigibase. *Drug Saf.* 2007;30(6):515-525.

(15) Colman E, Szarfman A, Wyeth J et al. An evaluation of a data mining signal for amyotrophic lateral sclerosis and statins detected in FDA's spontaneous adverse event reporting system. *Pharmacoepidemiol Drug Saf.* 2008;17(11):1068-1076.

(16) Sorensen HT, Riis AH, Lash TL, Pedersen L. Statin use and risk of amyotrophic lateral sclerosis and other motor neuron disorders. *Circ Cardiovasc Qual Outcomes.* 2010;3(4):413-417.

(17) Rothman KJ, Greenland S, Lash TL. Design strategies to improve study accuracy. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 168-82.

(18) Erkinjuntti T, Ostbye T, Steenhuis R, Hachinski V. The effect of different diagnostic criteria on the prevalence of dementia. *N Engl J Med.* 1997;337(23):1667-1674.

(19) Jick SS, Bradbury BD. Statins and newly diagnosed diabetes. *Br J Clin Pharmacol.* 2004;58(3):303-309.

(20) Brenner H, Savitz DA. The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. *Am J Epidemiol.* 1990;132(1):181-192.

(21) Jick H, Garcia Rodriguez LA, Perez-Gutthann S. Principles of epidemiological research on adverse and beneficial drug effects. *Lancet.* 1998;352(9142):1767-1770.

(22) Sturmer T, Jonsson FM, Poole C, Brookhart MA. Nonexperimental comparative effectiveness research using linked healthcare databases. *Epidemiology.* 2011;22(3):298-301.

(23) Jick SS, Kaye JA, Russmann S, Jick H. Risk of nonfatal venous thromboembolism with oral contraceptives containing norgestimate or desogestrel compared with oral contraceptives containing levonorgestrel. *Contraception.* 2006;73(6):566-570.

(24) Bosco JL, Silliman RA, Thwin SS et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J Clin Epidemiol.* 2010;63(1):64-74.

(25) Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol.* 1980;112(4).

(26) Greenland S, Lash TL. Bias Analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 345-80.

(27) Bradbury BD, Wilk JB, Kaye JA. Obesity and the risk of prostate cancer (United States). *Cancer Causes Control.* 2005;16(6):637-641.

(28) Lash TL, Schmidt M, Jensen AO, Engebjerg MC. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiol Drug Saf.* 2010;19(6):638-644.

(29) Goldberg J, Gelfand HM, Levy PS. Registry evaluation methods: a review and case study. *Epidemiol Rev.* 1980;2:210-220.

(30) Roos LL, Mustard CA, Nicol JP et al. Registries and administrative data: organization and accuracy. *Med Care.* 1993;31(3):201-212.

(31) Sorensen HT, Baron JA. Registries and medical databases. In: Trichopoulos D, Olsen JH, Saracci R, eds. *Teaching Epidemiology*. New York, NY: Oxford University Press, 2010: 455-67.

(32) Clement FM, James MT, Chin R et al. Validation of a case definition to define chronic dialysis using outpatient administrative data. *BMC Med Res Methodol.* 2011;11:25.

(33) Aphramor L. Validity of claims made in weight management research: a narrative review of dietetic articles. *Nutr J.* 2010;9:30.

(34) Kim SY, Solomon DH, Liu J, Chang CL, Daniel GW, Schneeweiss S. Accuracy of identifying neutropenia diagnoses in outpatient claims data. *Pharmacoepidemiol Drug Saf.* 2011.

(35) Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. *Can J Cardiol.* 2010;26(8):306-312.

(36) Chen G, Khan N, Walker R, Quan H. Validating ICD coding algorithms for diabetes mellitus from administrative data. *Diabetes Res Clin Pract.* 2010;89(2):189-195.

(37) Tollefson MK, Gettman MT, Karnes RJ, Frank I. Administrative Data Sets are Inaccurate for Assessing Functional Outcomes After Radical Prostatectomy. *J Urol.* 2011;185(5):1686-1690.

(38) Thygesen SK, Christiansen CF, Christensen S, Lash TL, Sorensen HT. The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. *BMC Med Res Methodol.* 2011;11:83.

(39) Cai S, Mukamel DB, Veazie P, Temkin-Greener H. Validation of the Minimum Data Set in identifying hospitalization events and payment source. *J Am Med Dir Assoc.* 2011;12(1):38-43.

(40) Tirschwell DL, Longstreth WT, Jr. Validating administrative data in stroke research. *Stroke.* 2002;33(10):2465-2470.

(41) Sorensen HT, Christensen T, Schlosser HK, Pedersen L. *Use of Medical Databases in Clinical Epidemiology*. Second ed. Aarhus, Denmark: SUN-TRYK Aarhus Universitet; 2009.

(42) Anderson IB, Sorensen TI, Prener A. Increase in incidence of disease due to diagnostic drift: primary liver cancer in Denmark, 1943-85. *BMJ.* 1991;302(6774):437-440.

(43) Lash TL, Johansen MB, Christensen S et al. Hospitalization rates and survival associated with COPD: a nationwide Danish cohort study. *Lung.* 2011;189(1):27-35.

(44) Rodriguez LA, Perez-Gutthann S, Jick SS. The UK General Practice Research Database. In: Strom BL, ed. *Pharmacopepidemiology*. 3rd ed. Chichester: John Wiley & Sons, 2000: 375-85.

(45) Bohnert AS, Ilgen MA, Galea S, McCarthy JF, Blow FC. Accidental poisoning mortality among patients in the Department of Veterans Affairs Health System. *Med Care.* 2011;49(4):393-396.

(46) Hansen JG, Pedersen L, Overvad K, Omland O, Jensen HK, Sorensen HT. The Prevalence of chronic obstructive pulmonary disease among Danes aged 45-84 years: population-based study. *COPD.* 2008;5(6):347-352.

(47) Alonso A, Jick SS, Jick H, Hernan MA. Antibiotic use and risk of multiple sclerosis. *Am J Epidemiol.* 2006;163(11):997-1002.

(48) Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15(5).

(49) Clough-Gorr KM, Fink AK, Silliman RA. Challenges associated with longitudinal survivorship research: attrition and a novel approach of reenrollment in a 6-year follow-up study of older breast cancer survivors. *J Cancer Surviv.* 2008;2(2):95-103.

(50) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol.* 2006;59(10):1087-1091.

(51) Howe CJ, Cole SR, Chmiel JS, Munoz A. Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *Am J Epidemiol.* 2011;173(5):569-577.

(52) Cain KC, Harlow SD, Little RJ et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *Am J Epidemiol.* 2011;173(9):1078-1084.

(53) Bosco JL, Antonsen S, Sorensen HT, Pedersen L, Lash TL. Metformin and incident breast cancer among diabetic women: a population-based case-control study in Denmark. *Cancer Epidemiol Biomarkers Prev.* 2011;20(1):101-111.

(54) Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol.* 2010;28(16):2796-2801.

(55) Riley GF. Administrative and claims records as sources of health care cost data. *Med Care.* 2009;47(7 Suppl 1):S51-S55.

(56) Suissa S. Immeasurable time bias in observational studies of drug effects on mortality. *Am J Epidemiol.* 2008;168(3):329-335.

(57) Bradbury BD, Wang O, Critchlow CW et al. Exploring relative mortality and epoetin alfa dose among hemodialysis patients. *Am J Kidney Dis.* 2008;51(1):62-70.

(58) Walker AM. Confounding by indication. *Epidemiology.* 1996;7(4):335-336.

(59) Miettinen OS. The need for randomization in the study of intended effects. *Stat Med.* 1983;2(2):267-271.

(60) NCCN Practice Guidelines in Oncology, Breast Cancer - v.2.2011. Invasive breast cancer, systemic adjuvant treatment. [National Comprehensive Cancer Network]. 2010.

(61) Geiger AM, Thwin SS, Lash TL et al. Recurrences and second primary breast cancers in older women with initial early-stage disease. *Cancer.* 2007;109(5):966-974.

(62) Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology.* 2006;17(4):373-374.

(63) Bradbury BD, Brookhart MA, Winkelmayer WC et al. Evolving statistical methods to facilitate evaluation of the causal association between erythropoiesis-stimulating agent dose and mortality in nonexperimental research: strengths and limitations. *Am J Kidney Dis.* 2009;54(3):554-560.

(64) Weiss NS, Dublin S. Accounting for time-dependent covariates whose levels are influenced by exposure status. *Epidemiology.* 1998;9(4):436-440.

(65) Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5):561-570.

(66) Cole SR, Hernan MA, Robins JM et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol.* 2003;158(7):687-694.

(67) Zhang Y, Thamer M, Cotter D, Kaufman J, Hernan MA. Estimated effect of epoetin dosage on survival among elderly hemodialysis patients in the United States. *Clin J Am Soc Nephrol.* 2009;4(3):638-644.

(68) Wang O, Kilpatrick RD, Critchlow CW et al. Relationship between epoetin alfa dose and mortality: findings from a marginal structural model. *Clin J Am Soc Nephrol.* 2010;5(2):182-188.

(69) Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol.* 2000;151(5):531-539.

(70) Greenland S. Randomization, statistics, and causal inference. *Epidemiology.* 1990;1(6):421-429.

(71) Lash TL, Schmidt M, Jensen AO, Engebjerg MC. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiol Drug Saf.* 2010;19(6):638-644.

(72) Fink AK, Lash TL. A null association between smoking during pregnancy and breast cancer using Massachusetts registry data (United States). *Cancer Causes Control.* 2003;14(5):497-503.

(73) Lash TL, Fox MP, Thwin SS et al. Using probabilistic corrections to account for abstractor agreement in medical record reviews. *Am J Epidemiol.* 2007;165(12):1454-1461.

(74) Lash TL, Fink AK. Null association between pregnancy termination and breast cancer in a registry-based study of parous women. *Int J Cancer.* 2004;110(3):443-448.

(75) Melbye M, Wohlfahrt J, Olsen JH et al. Induced abortion and the risk of breast cancer. *N Engl J Med.* 1997;336(2):81-85.

(76) Goldacre MJ, Kurina LM, Seagroatt V, Yeates D. Abortion and breast cancer: a case-control record linkage study. *J Epidemiol Community Health.* 2001;55(5):336-337.

(77) Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol.* 1999;150(5):469-475.

(78) Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Stat Med.* 2010;29(12):1259-1265.

(79) Cunnane C. Unbiased plotting positions — a review. *J Hydrology.* 1978;37:205-222.

(80) Grady D, Hearst H. Utilizing existing databases. In: Hully SB, Cummings SR, Browner WS et al, eds. *Designing Clinical Research*. 3rd ed. Philadelphia, PD: Lippincott Williams & Wilkins, 2007: 207-21.

(81) Sorensen HT, Lash TL. Statins and amyotrophic lateral sclerosis--the level of evidence for an association. *J Intern Med.* 2009;266(6):520-526.

(82) Thavendiranathan P, Bagai A, Brookhart MA, Choudhry NK. Primary prevention of cardiovascular diseases with statin therapy: a meta-analysis of randomized controlled trials. *Arch Intern Med.* 2006;166(21):2307-2313.

(83) Aronow HD, Topol EJ, Roe MT et al. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: an observational study. *Lancet.* 2001;357(9262):1063-1068.

(84) Mitchell JD, Borasio GD. Amyotrophic lateral sclerosis. *Lancet.* 2007;369(9578):2031-2041.

(85) Rosati K. Using electronic health information for pharmacovigilance: the promise and the pitfalls. *J Health Life Sci Law.* 2009;2(4):171, 173-171, 239.

(86) Trifiro G, Pariente A, Coloma PM et al. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiol Drug Saf.* 2009;18(12):1176-1184.

(87) Coloma PM, Schuemie MJ, Trifiro G et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf.* 2011;20(1):1-11.

(88) Latourelle JC, Dybdahl M, Destefano AL, Myers RH, Lash TL. Estrogen-related and other disease diagnoses preceding Parkinson's disease. *Clin Epidemiol.* 2010;2:153-170.

(89) Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology.* 1991;2(4):244-251.

(90) Lash TL, Cronin-Fenton D, Ahern TP et al. CYP2D6 Inhibition and Breast Cancer Recurrence in a Population-Based Study in Denmark. *J Natl Cancer Inst.* 2011;103(6):489-500.

(91) Erichsen R, Lash TL, Hamilton-Dutoit S, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish National Pathology Registry and Data Bank. *Clin Epidemiol.* 2010;2:51-56.

(92) Mikkelsen EM, Hatch EE, Wise LA, Rothman KJ, Riis A, Sorensen HT. Cohort profile: the Danish Web-based Pregnancy Planning Study--'Snart-Gravid'. *Int J Epidemiol.* 2009;38(4):938-943.

(93) Huybrechts KF, Mikkelsen EM, Christensen T et al. A successful implementation of e-epidemiology: the Danish pregnancy planning study 'Snart-Gravid'. *Eur J Epidemiol.* 2010;25(5):297-304.

(94) Thwin SS, Clough-Gorr KM, McCarty MC et al. Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study. *BMC Med Res Methodol.* 2007;7:23.

(95) Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol.* 2003;158(9):915-920.

(96) Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf.* 2010;19(6):537-554.

(97) Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* 1987;40(5).

(98) Rassen JA, Solomon DH, Curtis JR, Herrinton L, Schneeweiss S. Privacy-maintaining propensity score-based pooling of multiple databases applied to a study of biologics. *Med Care.* 2010;48(6 Suppl):S83-S89.

(99) Wolfson M, Wallace SE, Masca N et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol.* 2010;39(5):1372-1382.

(100) Greenland S, O'Rourke K. Meta-Analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 652-82.