

# Methods to Analyze Real-World Databases and Registries

Hilal Maradit Kremers, M.D., M.Sc.

## Abstract

*Observational studies, administrative database studies, and registries offer a wealth of real-world information, if designed, maintained, and analyzed according to appropriate observational study methodology. This review summarizes basic observational study methods employed in rheumatology and highlights several notable examples. In response to growing interest in real-world effectiveness and safety data, registries are expected to proliferate in the near future. The availability of detailed clinical information in registries coupled with powerful tools for analysis offers promise for timely and accurate information on the safety and effectiveness of rheumatic treatments.*

**R**heumatic diseases are chronic inflammatory conditions characterized by pain, swelling, and limited movement in the joints and connective tissues of the body. Rheumatic diseases have a multifactorial etiology, and both genetic and environmental risk factors contribute to disease onset and progression. The natural history of most rheumatic diseases is characterized by a number of adverse outcomes, in particular, increased risk of infection, cardiovascular and gastrointestinal diseases, and mortality. The risk of adverse outcomes is influenced by various disease and treatment-related factors.

Traditionally, observational studies and disease-specific registries have played a significant role in understanding the risk factors, epidemiology, natural history, and outcome of rheumatic diseases. During the last 2 decades, there was an

upsurge of registries in rheumatic diseases, primarily rheumatoid arthritis (RA), psoriatic arthritis, and systemic lupus erythematosus (SLE).<sup>1-4</sup> Earlier registries were multicenter collaborations, created in response to the need to accumulate large cohorts of patients with uniform criteria. They have been invaluable for understanding the clinical manifestations, disease management, and genetic and environmental risk factors of these disorders.<sup>4,5</sup>

More recently, the growth of registries in rheumatic diseases has been driven largely by the growing portfolio of biological treatments and a lack of head-to-head comparisons on their effectiveness and safety.<sup>2,3</sup> In parallel, secondary use of large administrative databases has increased significantly in an effort to study rare outcomes in large representative populations.<sup>6-8</sup> Indeed, preapproval clinical studies are typically too short, too small, too simple, and too median-aged, and, therefore, they are not representative of real-world use.<sup>9</sup> The emergence and growth of new treatments coupled with the absence of relevant, underrepresented populations in clinical trials has resulted in consequent gaps in knowledge, for example: 1. the relative effectiveness and safety of alternative treatment strategies are largely unknown; 2. there is little evidence on effectiveness in patient populations who are underrepresented in clinical trials, such as elderly patients with multiple comorbidities; 3. surrogate end points typically applied in clinical trials do not necessarily equate with clinical outcomes; and 4. infrequent and long latency adverse outcomes are unknown. Therefore, observational studies and registries are essential in order to: 1. overcome the limitations of premarketing clinical trials, 2. address unresolved effectiveness and safety issues from premarketing studies, 3. evaluate specific safety concerns, 4. establish risk-benefit margins in selected patient subgroups, 5. understand quality of care and prescription and compliance patterns, and 6. adequately evaluate risk management programs.<sup>10</sup>

Hilal Maradit Kremers, M.D., M.Sc., is an Associate Professor of Epidemiology, in the Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota.

*Correspondence:* Hilal Maradit Kremers, M.D., Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, Minnesota 55905; maradit@mayo.edu.

Of note, despite some structural differences between traditional observational cohort studies and registries, they have more in common than not in terms methodology. For the most part, registries adopt observational study methodology “to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by particular disease, condition, or exposure, and that serve a predetermined scientific, clinical, or policy purpose(s).”<sup>10</sup> Therefore, methods described in this review refer to both registries and observational cohort studies in rheumatic diseases.

There are several different types of registries in rheumatic diseases, depending on the geographical coverage, type of registration, extent of clinical characteristics, frequency of follow-up, methods for data collection, and the diseases covered. The development of disease registries is most common where patients who have the same diagnosis, e.g., all RA or SLE patients or other rheumatic disease, are included in a registry. In contrast, the eligibility criteria for product registries are not the status of specific diseases, per se, but rather the initiation of the product(s) in question. Examples of product registries for rheumatic diseases include the biologics’ registries in Europe<sup>2</sup> and the recent adalimumab registry in the United States. Finally, health services’ registries enroll patients who have undergone a common procedure, such as the arthroplasty registries.<sup>11</sup> The following is a brief overview of various observational study methods used to analyze observational studies and registries, frequently referred to as “real-world” data sources.

### Ecological Studies—Time Series

Ecological studies are particularly useful when a drug is the predominant cause of a disease and the outcome and changes occur following an abrupt change in drug exposure, such as that from a policy or regulatory change, publications, media coverage. Although ecological studies have certain disadvantages, they still may provide important insights. Recent examples in rheumatic diseases include an exploratory ecological analysis that relied on data from the National Ambulatory Medical Care Survey [(NAMCS), National Center for Health Statistics (NCHS)] and the Surveillance, Epidemiology, and End Results [(SEER), National Cancer Institute (NCI)] for data to examine the association between NSAID use over time and colorectal cancer incidence.<sup>12</sup> The investigators found that the increased NSAID exposure preceded the decline in colorectal cancer incidence by about 7 years. Browstein and colleagues<sup>13</sup> recently showed a population-level rise in hospitalizations for myocardial infarction at two Boston hospitals between 1997 and 2006, approximately 1 year after the introduction of rofecoxib and celecoxib. Furthermore, there was some indication that the mean age of patients with myocardial infarction during this time period was also lower than the period following the withdrawal of rofecoxib.

Ecological studies are relatively easy to conduct using already available data, and trends are expected to emerge for

common exposures and strong associations. Nevertheless, ecological fallacy is a major concern in interpreting results, concluding that the average value of the drug exposure being studied and the average incidence and prevalence of an outcome apply to all the individuals in a population. On the other hand, some risk factors for diseases indisputably operate at the population level, either directly causing the disease or, more commonly, acting as effect modifiers or determinants of exposure to individual level risk factors. Therefore, ecological studies will likely continue in the future to monitor trends at the population level.

### Studies on Descriptive Epidemiology and the Natural History of Rheumatic Diseases

Descriptive studies in rheumatic diseases are plentiful and include incidence, prevalence, and mortality studies. Although a variety of data sources are used, such studies can be most reliably conducted using population-based data sources, such as the population registers in Nordic countries<sup>14</sup> or the Rochester Epidemiology Project (REP)<sup>15</sup> in the U.S. There are many reasons for this; however, most notably, characteristics or exposures that are associated with differential surveillance, diagnosis, referral, or case ascertainment can lead to biased estimates of incidence, prevalence, and mortality. In population-based incidence studies, all new cases of a particular disease in the community need to be ascertained, irrespective of disease severity. This provides a more generalizable and unbiased description of the clinical spectrum of the disease. In addition to minimizing bias, population-based data resources provide the unique opportunity as to how referral bias is able to distort research findings by comparing findings based on the local population with the same findings in the referral populations.

A noteworthy example in rheumatology is the wide variations observed in mortality studies in RA. In 2001, Ward concluded that the reported improvements in survival among RA patients were a function of differences in study designs, rather than a true improvement in mortality.<sup>16</sup> In a 2008 review, Sokka and coworkers reviewed 124 mortality analyses from 84 unique cohorts and highlighted how methodological differences can result in different results for risk and predictors of mortality in RA.<sup>17</sup> Unfortunately, methodological differences between disease cohorts and how they affect mortality findings in other rheumatic diseases are largely unexplored.

Secular trends and the natural history of diseases are also of considerable interest in rheumatology, as they provide clues to the etiology of diseases. A major challenge in such studies is secular trends in diagnostic criteria and practice patterns and, consequently, distorted findings due to reliance on physician diagnoses. Secular trend studies require cohorts that have been assembled over several decades using identical methodology. This is feasible in only a few places in the world, such as with the REP database. The availability of original medical records enables investigators to rereview all

data using contemporary standardized criteria. An excellent example is the study of secular trends in the epidemiology of RA, which demonstrated a progressive decline in the incidence of RA over the past 40 years.<sup>18</sup> Another example is the study of secular trends in the epidemiology of giant cell arteritis, which showed regular cyclic patterns of incidence that were suggestive of epidemic cycles of a putative infectious agent.<sup>19</sup>

### **Methods for Examining Outcomes in Rheumatic Diseases**

A significant advantage of observational cohort studies and registries is the ability to examine the risk and predictors of multiple outcomes or end points in the same patient population. Such observational studies are typically designed as either prospective or retrospective cohort studies, with either an internal or an external comparison cohort. These design options depend upon resources and data availability but can significantly impact internal and external validity. For example, rheumatic disease cohorts in REP are assembled retrospectively using the medical records linkage system.<sup>15</sup> Since the size and characteristics of the underlying population (Olmsted County, Minnesota) is well defined, it is also possible to identify internal comparison cohorts nested within the same source population. Over the years, a number of different outcomes have been examined in these cohorts, in particular, infections and cardiovascular events.<sup>20,21</sup> Observational cohorts and registries in Nordic countries are similarly population-based cohorts with an identifiable source population (i.e., whole population of the country) and internal comparison cohorts nested within the same source population. Cohorts formed in population-based settings offer the unique opportunity to assemble disease cohorts with the complete disease spectrum and representative internal comparison cohorts.

Most registries in rheumatic diseases, on the other hand, are designed as prospective cohorts, such as the National Databank for Rheumatic Diseases (NDB, Wichita, Kansas),<sup>5</sup> the Consortium of Rheumatology Researchers of North America (CORRONA),<sup>22</sup> and the European biologics registries.<sup>2</sup> The source population is typically unknown, as data are provided by multiple rheumatologists from a variety of geographical regions. Entry into each of these registries is open, in other words, subjects are recruited into the cohort on an ongoing basis throughout follow-up. Absence of an internal comparison cohort can be challenging in some instances.<sup>2</sup>

In contrast to clinical trials, ensuring the quality of data and interpretation of findings can be challenging in observational cohort studies and registries due to various avoidable and unavoidable biases, as discussed briefly below.

### **Measurement of Risk Factors and Drug Exposures**

One of the main measurement challenges is accurate measurement of risk factors and drug exposures. Depending on

the setting, a variety of methods are used, such as face-to-face interviews, phone or self-administered questionnaires, actual medical records, or pharmacy or claims records, each with its own strengths and weaknesses. Features that can have a significant impact on internal validity are completeness and accuracy, response rates, temporal change over time, details of the drugs, details of utilization, availability, and cost (reimbursement). A notable distinction between observational cohorts and registries versus administrative databases is the extent of clinical information, which is of paramount importance in rheumatic diseases. Various rheumatic disease characteristics are associated with both the outcomes<sup>23</sup> and the choice of drug therapy<sup>24</sup> in rheumatic diseases, introducing significant confounding by indication in examining drug effects.<sup>25,26</sup> Observational cohorts and registries are typically rich with serially evaluated clinical data; a limited number of studies capitalized on this to account for confounding by indication.<sup>27,28</sup> In contrast, a lack of clinical information on disease characteristics and severity in administrative databases is a major challenge, and some groups are attempting to overcome these challenges by developing surrogate severity measures based on utilization data.<sup>29</sup>

Confounding by indication in examining outcomes in rheumatic diseases comes in many shapes and sizes, but the typical ones are 1. confounding by contraindication in examining the risk of gastrointestinal adverse events associated with individual NSAIDs, where the physician's perception of a patient's tendency to develop bleeding is associated with an NSAID choice, and 2. confounding by disease severity in examining the cardiovascular risk associated with DMARD use, where disease severity is associated with the initiation of disease-modifying antirheumatic drugs (DMARDs) and an independent risk factor for cardiovascular diseases. Some rheumatology investigators successfully adopted novel methods to reduce confounding by indication. Most significantly, Choi and associates<sup>27</sup> adopted marginal structural models using a wealth of clinical data and demonstrated a substantial survival benefit with methotrexate in RA patients. Bukhari and colleagues<sup>30</sup> used propensity score-based methods to demonstrate the effectiveness of early DMARD treatment in reducing radiographic progression.

Additionally, irrespective of the data source, an important consideration is timing of the drug exposure in examining effectiveness or the safety of treatments. Studies that include both prevalent (ongoing) and incident (new) drug users are prone to bias if effectiveness or the risk of an outcome varies over time, and if the risk factors of interest are substantially affected by the treatment.<sup>31</sup>

### **Outcomes**

Content and methods for the ascertainment of outcomes in registries, observational cohorts, and administrative databases also vary considerably. Outcomes of interest in rheumatic

diseases can be broadly classified into two categories: beneficial outcomes or effectiveness (such as disease progression), and adverse outcomes, such as mortality, infection, and malignancy. Studies in recent years have taught us that each of these outcomes has unique challenges. In secondary database studies, the diagnostic information may be recorded incompletely, either randomly or systematically, and this may result in under ascertainment or over ascertainment of outcomes. Broadly, information on sensitivity and specificity of outcome ascertainment and, if possible, ability to validate self-reports or claims can significantly reduce measurement error and reduce bias in registries and cohort studies.

Another important but frequently ignored aspect of documenting rates of outcome in rheumatic disease is the competing risk of death, because the risk of death is particularly high in elderly rheumatic disease patients, and death prevents the occurrence of other events of interest. This is illustrated with an example in Figure 1, where the cumulative incidence of a cardiovascular event in patients with RA increased steadily from the time of RA incidence, and after 30 years it was about 55%, when follow-up was censored at death. However, mortality is high in RA patients, and when death was taken into account as a competing risk, the estimated cumulative incidence at 30 years was 40%, which is more reflective of what would actually be observed in practice.

In conclusion, although much has been done relative to examining adverse outcomes in rheumatic diseases, the research methodology for comparative effectiveness is still in its infancy. Nevertheless, there are several currently available approaches for improving the validity of effectiveness estimates from observational studies, including design options, such as restriction or crossover designs, or statistical methods that employ propensity scores or instrumental variable analyses.<sup>7,32</sup> Such studies are expected to proliferate in the near future with the growing interest in real-world effectiveness data and the advent of registries and large cohorts



**Figure 1** Cumulative incidence of cardiovascular disease in RA, with and without considering competing risk of death.

with rich longitudinal clinical data.

## Methods for Examining Risk Factors for Rheumatic Diseases

Studies of risk factors for rheumatic diseases typically employ case-control methods by examining exposure history among disease cases and controls. The greatest methodological challenge in case-control studies is identification of source population and appropriate controls. For example, the source population for REP based case-control studies is the Olmsted County population<sup>33</sup> or the whole nation in the case of studies in Nordic countries.<sup>34</sup> It is particularly difficult to define source population in hospital-based case-control studies. If exposure prevalence in controls differs from that of the source population, the estimates from case-control studies can be biased.

Other methodological issues to consider in case-control studies are the inclusion of only incident cases, as well as accurate and nondifferential information on exposures of interest and potential confounders. In case-control studies that include prevalent cases, the association between drug exposure and the disease in question reflects the association with a prognostic factor, rather than incidence. If only survivors of a disease received the drug, there will be a positive association between the drug and outcome. A typical example is NSAID use and dementia,<sup>35</sup> where the odd ratio was 0.51 in studies with prevalent dementia cases and 0.79 in studies with incident dementia cases. Special case-control designs, such as case-cohort, nested case-control, or case-crossover designs, may help in dealing with potential biases in case-control studies. Although there are many examples of case-control studies, so far, registry based case-control studies are limited.

## Disclosure Statement

Dr. Maradit Kremers' research is funded, in part, by unrestricted research grants from Amgen, Inc., Pfizer, and the National Institutes of Health (NIH).

## References

- Gladman DD, Rahman P, Krueger GG, et al. Clinical and genetic registries in psoriatic disease. *J Rheumatol*. 2008 Jul;35(7):1458-63.
- Zink A, Askling J, Dixon WG, et al. European Biologics Registers - Methodology, selected results, and perspectives. *Ann Rheum Dis*. 2008 Jul 22; Epub ahead of print.
- Kremer JM, Gibofsky A, Greenberg JD. The role of drug and disease registries in rheumatic disease epidemiology. *Curr Opin Rheumatol*. 2008 Mar;20(2):123-30.
- Lu LJ, Wallace DJ, Navarra SV, Weisman MH. Lupus registries: evolution and challenges. *Semin Arthritis Rheum*. 2008 Nov 6; Epub ahead of print.
- Wolfe F. A short history of data banking in the United States from 1974 to 2003. *J Rheumatol Suppl*. 2004 Mar;69:41-5.
- Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with

- cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet*. 2005 Feb 5-11;365(9458):475-81.
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005 Apr;58(4):323-37.
  8. Schneeweiss S, Setoguchi S, Weinblatt ME, et al. Anti-tumor necrosis factor alpha therapy and the risk of serious bacterial infections in elderly patients with rheumatoid arthritis. *Arthritis Rheum*. Jun 2007;56(6):1754-64.
  9. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *N Engl J Med*. Jun 22 2000;342(25):1907-9.
  10. Gliklich RE, Dreyer NA (eds): *Registries for Evaluating Patient Outcomes: A User's Guide*. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHS A290200500351 TO1.) AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality, April 2007.
  11. Sheng PY, Kontinen L, Lehto M, et al. Revision total knee arthroplasty: 1990 through 2002. A review of the Finnish arthroplasty registry. *J Bone Joint Surg Am*. 2006 Jul;88(7):1425-30.
  12. Lamont EB, Dias LE. Secular changes in NSAID use and invasive colorectal cancer incidence: an ecological study. *Cancer J*. 2008 Jul-Aug;14(4):276-7.
  13. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE*. 2007;2(9):e840.
  14. van Vollenhoven RF, Askling J. Rheumatoid arthritis registries in Sweden. *Clin Exp Rheumatol*. 2005 Sep-Oct;23(5 Suppl 39):S195-200.
  15. Maradit Kremers H, Crowson CS, Gabriel SE. Rochester Epidemiology Project: a unique resource for research in the rheumatic diseases. *Rheum Dis Clin North Am*. 2004;30:819-34.
  16. Ward MM. Recent improvements in survival in patients with rheumatoid arthritis: better outcomes or different study designs? *Arthritis Rheum*. 2001;44(6):1467-9.
  17. Sokka T, Abelson B, Pincus T. Mortality in rheumatoid arthritis: 2008 update. *Clin Exp Rheumatol*. 2008 Sep-Oct;26(5 Suppl 5):S35-61.
  18. Doran MF, Pond GR, Crowson CS, et al. Trends in incidence and mortality in rheumatoid arthritis in Rochester, Minnesota, over a forty-year period. *Arthritis Rheum*. 2002;46(3):625-31.
  19. Salvarani C, Crowson CS, O'Fallon WM, et al. Reappraisal of the epidemiology of giant cell arteritis in Olmsted County, Minnesota, over a fifty-year period. *Arthritis Rheum*. 2004;51(2):264-8.
  20. Doran M, Crowson C, Pond G, et al. Frequency of infection in patients with rheumatoid arthritis compared with controls: A population-based study. *Arthritis Rheum*. 2002;46(9):2287-93.
  21. Maradit-Kremers H, Crowson CS, Nicola PJ, et al. Increased unrecognized coronary heart disease and sudden deaths in rheumatoid arthritis: A population-based cohort study. *Arthritis Rheum*. 2005 Feb 3;52(2):402-11.
  22. Kremer JM. The CORRONA database. *Autoimmun Rev*. 2006 Jan;5(1):46-54.
  23. Maradit-Kremers H, Nicola PJ, Crowson CS, et al. Cardiovascular death in rheumatoid arthritis: A population-based study. *Arthritis Rheum*. 2005 Mar;52(3):722-32.
  24. Maradit-Kremers HM, Nicola P, Crowson CS, et al. Therapeutic strategies in rheumatoid arthritis over a 40-year period. *J Rheumatol*. 2004 Dec;31(12):2366-73.
  25. Walker AM. Confounding by indication. *Epidemiology*. 1996;7(4):335-6.
  26. Wolfe F, Flowers N, Burke TA, et al. Increase in lifetime adverse drug reactions, service utilization, and disease severity among patients who will start COX-2 specific inhibitors: quantitative assessment of channeling bias and confounding by indication in 6689 patients with rheumatoid arthritis and osteoarthritis. *J Rheumatol*. 2002;29(5):1015-22.
  27. Choi HK, Hernán MA, Seeger JD, et al. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *Lancet*. 2002;359(9313):1173-7.
  28. Wiles NJ, Lunt M, Barrett EM, et al. Reduced disability at five years with early treatment of inflammatory polyarthritis: results from a large observational cohort, using propensity models to adjust for disease severity. *Arthritis Rheum*. 2001;44(5):1033-42.
  29. Ting G, Schneeweiss S, Scranton R, et al. Development of a health care utilisation data-based index for rheumatoid arthritis severity: a preliminary study. *Arthritis Res Ther*. 2008;10(4):R95; 2008 Aug 21; Epub ahead of print.
  30. Bukhari MA, Wiles NJ, Lunt M, et al. Influence of disease-modifying therapy on radiographic outcome in inflammatory polyarthritis at five years: results from a large observational inception study. *Arthritis Rheum*. 2003;48(1):46-53.
  31. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003 Nov 1;158(9):915-20.
  32. Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clin Pharmacol Ther*. 2007 Aug;82(2):143-56.
  33. Doran MF, Crowson CS, O'Fallon WM, Gabriel SE. The effect of oral contraceptives and estrogen replacement therapy on the risk of rheumatoid arthritis: a population based study. *J Rheumatol*. 2004;31(2):207-13.
  34. Kallberg H, Jacobsen S, Bengtsson C, et al. Alcohol consumption is associated with decreased risk of rheumatoid arthritis: results from two Scandinavian case-control studies. *Ann Rheum Dis*. 2009 Feb;68(2):222-7.
  35. de Craen AJ, Gussekloo J, Vrijnsen B, Westendorp RG. Meta-analysis of nonsteroidal antiinflammatory drug use and risk of dementia. *Am J Epidemiol*. 2005 Jan 15;161(2):114-20.