

THE POTENTIALS AND CHALLENGES OF PATIENT REGISTRIES

Elizabeth Benito Garcia,* Frederick Wolfe**

Patient registries organize scattered data to be analyzed to answer one or more scientific questions and are a valuable contribution to research.

lidity, have the advantage of capturing real life situations in unselected patients and can evaluate multiple questions, not merely drug-related ones.

Randomized controlled trials versus observational studies

Randomized controlled trials (RCTs) are the most rigorous way of scientifically studying drug efficacy.¹⁻³ Through randomization, they have the advantage of controlling both recognized and unrecognized covariates. This unique feature makes RCTs a powerful design in epidemiology, providing data of such high quality that they closely resemble the controlled experiment done by basic science researchers.^{1,3}

The main disadvantages of RCTs include high costs; limitations in the number of questions analyzed; feasibility due to the necessity of large numbers of subjects that are willing to forego a treatment or practice believed to be beneficial for the duration of a trial; and limitations to the generalizability of results and their application in actual clinical practice because of strict inclusion and exclusion criteria.^{1,4-7} In addition, ethical concerns preclude the allocation of treatments that are known or suspected to be hazardous, frequently limiting the evaluation of many treatments or procedures in RCTs.^{1,4,5} RCTs are also prone to various selection biases, for example, patients who are unsatisfied with current treatment strategies may volunteer to enter a trial more often than those who are not.¹ Lastly, RCTs are not useful for detecting long-term therapeutic side effects and effectiveness because of insufficient length of follow-up time of the majority of trials.^{6,8}

Observational studies, although associated to higher scientific uncertainty and requiring more complex methodology to deal with threats to va-

Importance and potentials of patient registries

Registries and longitudinal observational studies have the potential to overcome the limitations observed in RCTs as well as in retrospective, cross-sectional and observational studies that collect data in a specified time point or interval. The basic idea of a registry is to systematically follow rheumatic disease patients without intervening in any way in their treatment.

Registries may study a wide variety of important rheumatology questions. These include 1) disease prevalence among specific populations; 2) comorbidities; 3) family history; 4) environmental exposures (i.e. diet, climate, etc.); 5) longitudinal disease outcomes; 6) cost-effectiveness and quality of care, and 7) management strategies. Additionally, they may study questions that for ethical reasons may not be considered in RCTs. Registries have the capability of studying these issues because the investigator is a passive observer.

Behind the ostensible simplicity of registries, however, is the requirement for difficult, costly methodology, as well as specialized knowledge in many areas of epidemiology, biostatistics, medicine and other sciences as well as multidisciplinary teams that must collaborate to achieve an ultimate common goal.

Methodology and difficulties

The first step in creating a registry, as in all studies, is the formulation of the underlying clinical question(s). The objective of the registry should be defined as well as who will use the registry and what the data will be used for. Sometimes it helps to start from the end. What do I want the data to tell me?

*BioEPI, Clinical & Translational Research Center, Oeiras, Portugal

**National Data Bank for Rheumatic Diseases, Wichita, Kansas, USA

Am I collecting the most appropriate data to answer my question?

The next step is defining the variables. Seek the help of an epidemiologist and biostatistician in this very early stage. Valid results depend on the correct definition of variables and their adequate measurement.

The data collection instrument should be designed with the help of experienced professionals. Questions, and the variables that define them, should be adequate, clear, simple, and formulated in such a way that little doubt remains as to interpretation of responses. The formulation of questions is important, and much science is available to guide question development.

The method of data collection is determined in part by feasibility. In subjects with limited literacy, it is useful to collect data through personal interviews, whereas auto-administered questionnaires might work in subjects who have higher levels of education.

Standardizing data collection methods and training interviewers to ask questions in a constant way is essential. The way a question is read may influence a subject's response, therefore an interviewer must ask questions in a very clear, nonsubjective manner ideally blinded to study objectives to avoid data collection biases.

Investigators must work diligently to minimize missing data. Objective data should be confirmed and validated through medical charts and test results where applicable.

The creation of a data dictionary (a book of variable codes that permit the interpretation of collected data) should be assembled and contain information on the type of variable (i.e. continuous, categorical, ordinal, etc.), values for the each variable (i.e., "Yes", "No", "Missing"), and permitted values (i.e., all fields should only contain dates after Jan 1, 2004).

An extremely important task preceding statistical analyses consists in the coding of non-numeric into numeric variables. Coding is an act of translation and summarizing. For example, "Have you ever smoked?" (1) 1= "Yes"; (2) 0= "No"; (3) 9= "Don't know"

Quality control must be thoroughly planned. Some responses in the registry should include logical relationships with others. For example, subjects whose response to sex was "Male" should not have answered the question pertaining to whether they were ever pregnant. Another example of qua-

lity control consists in only accepting values that are permissible. For example, for age, it should not be permissible to enter 0.

Data entry consists of efficiently entering data in a reliable way, minimizing time spent, errors, and simplifying the work for the biostatistician. There are several methods to enter data into the registry, from manual data entering and optical scanning, to electronic links between questionnaires and the registry, and web-based questionnaires. For example, in manual data entry, double data entry can be accomplished. This consists in entering data in 2 copies of the same data base, either by two people or the same person twice, and then comparing the discrepancies and correcting them. Checking data after optical scanning is another way to control errors.

Training of all personnel involved in data collection and data entry and having a project manager throughout the process to coordinate the project is important.

The most tedious task in data management is data cleaning. Before statistical analysis begins, errors, unclear data and missing data should be searched for and corrected in the registry. This may require going back to the instrument of data collection, or may require re-contacting subjects and searching other sources of information. Sometimes it is impossible to improve all variables and some data will be missing or contain errors.

It is useful to test methods before the start of the study: the instrument, data collection and how to register, code it and enter information it should be pilot tested. Validation strategies and corrections in the data dictionary should also be undertaken.

Other difficulties

Determining the periodicity and frequency of data collection should be defined in registry planning phase. For example, it might be adequate to collect DAS28 measures every 4 months and dietary intake every 4 years. Also, over time, data should be collected for all patients in the same time frames, for comparability reasons. It might be easier for a physician to collect data in his clinic each time his patients are seen, however, for research purposes, it is necessary to determine the time frames of data collection for all patients and not by the convenience of clinic visits. For example, a specific patient with rheumatoid arthritis may have been seen

by a rheumatologist 6 times in a specific year, in 2 month intervals and another only twice, in that same year, once at the beginning and once at the end of the year. For research purposes it should be defined that for each of these patients data collection should be performed in specific months.

Losses to follow-up or a subject's reluctance to participate in a study are common problems in registries. Motivation and techniques to avoid subject loss are necessary.

Data preparation for statistical analyses and exporting of the content of a registry should be undertaken into a statistical software package.

Legal technicalities, such as confidentiality agreements, informed consent, ethics committees and local laws should be taken into account. Often these procedures take more time than all others.

Finally, I give you the recipe for success. A successful registry needs the good will and collaboration of many individuals. Wanting to do it is not sufficient. One must have the time, the knowledge, the experience, and a professionalized team to assemble such a registry. Finally, one must not forget to focus on the science.

Correspondence to

Elizabeth Benito Garcia
BioEPI, Clinical & Translational Research Center
Taguspark, Núcleo Central, 232
2740-122 Oeiras - PORTUGAL

References

1. Simon SD. Is the randomized clinical trial the gold standard of research? *J Androl* 2001; 22:938-943.
2. Gough S. Post-marketing surveillance: a UK/European perspective. *Curr Med Res Opin* 2005; 21:565-570.
3. Olsson J, Terris D, Elg M, Lundberg J, Lindblad S. The one-person randomized controlled trial. *Qual Manag Health Care* 2005; 14:206-216.
4. Bugeja G, Kumar A, Banerjee AK. Exclusion of elderly people from clinical research: a descriptive study of published reports. *Bmj* 1997; 315:1059.
5. Gurwitz JH, Col NF, Avorn J. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *Jama* 1992; 268:1417-1422.
6. Vijan S, Kent DM, Hayward RA. Are randomized controlled trials sufficient evidence to guide clinical practice in type II (non-insulin-dependent) diabetes mellitus? *Diabetologia* 2000; 43:125-130.
7. Sokka T, Pincus T. Eligibility of patients in routine care for major clinical trials of anti-tumor necrosis factor alpha agents in rheumatoid arthritis. *Arthritis Rheum* 2003; 48:313-318.
8. Farahani P, Levine M, Gaebel K, Thabane L. Clinical data gap between phase III clinical trials (pre-marketing) and phase IV (post-marketing) studies: evaluation of etanercept in rheumatoid arthritis. *Can J Clin Pharmacol* 2005; 12:e254-263.