A brief introduction to the National Data Bank for Rheumatic Diseases

F. Wolfe[1], K. Michaud[2,3]

[1]Arthritis Research Center Foundation, University of Kansas School of Medicine, Wichita, Kansas; [2]National Data Bank for Rheumatic Diseases, Wichita, Kansas; [3]Center for Primary Care and Outcomes Research, Stanford University, Stanford, California, USA.

Frederick Wolfe, MD; Kaleb Michaud, MS. Please address correspondence to: Frederick Wolfe, MD, National Data Bank for Rheumatic Diseases, Arthritis Research Center Foundation, 1035 N. Emporia, Suite 230, Wichita, KS 67214, USA. E-mail: fwolfe@arthritis-research.org

Clin Exp Rheumatol 2005; 23 (Suppl. 39): S168-S171.

Key words: Database, data, outcomes, research, rheumatic diseases.

ABSTRACT

The National Data Bank for Rheumatic Diseases (NDB) is a research data bank with a broad agenda and approach to important clinical, regulatory, epidemiological, safety, effectiveness, outcome and patient care questions that cannot be answered by conventional data banks. It has systematic, ongoing quality control programs that assure high quality data. NDB-developed programs result in very efficient analytic capabilities and rapid publication.

Introduction

The National Data Bank for Rheumatic Diseases (NDB) is a rheumatic diseases research data bank. Many of the details of the this data bank have been described previously in a detailed review (1) and in publications from the NDB (2-12). In this article we describe characteristics of the NDB data bank methodology, contrast it with other methods of data banking, and address factors relating to quality control.

Contrasting registries and data banks

Data banks and registries share a number of common features, but also differ in important ways. The term "registry" is applied to data banks that generally have narrowly defined tasks, most often relating to issues of drug safety and sometimes of efficacy. When administered directly or indirectly by pharmaceutical companies, registries often tend to be limited by the absence of a full range of control subjects and often by underlying commercial purposes. Government-sponsored registries may also have problems with appropriate comparison subjects unless the registry is broadly defined. Thus, a common feature of registries is narrowness of purpose.

The issues and outcomes of rheumatic disease, however, are much broader than those defined by the regulatory authorities, and they should serve to increase the knowledge of health care providers, policy makers, investigators, clinicians and persons who suffer from rheumatic diseases. When the broad issues of rheumatic diseases are the study substrate, the term "data bank" is more appropriate. Thus, the NDB was designed to address such broad issues. As an example of NDB work, Table I is a list of topics that have been addressed by the NDB as abstracts at American College of Rheumatology (ACR) or EULAR meetings and that have been published subsequently or are in the process of being published.

National Data Bank for Rheumatic Diseases

The NDB was founded in 1998 and is now completing its eighth year of data collection and publication. During this time 28,582 patients have completed 155,311 detailed semi-annual questionnaires. Although its major emphasis has been on outcomes in rheumatoid arthritis (RA), the NDB has recently established a separate data bank for systemic lupus erythematosus (SLE), and continues general data banks that include all other rheumatic diseases, including fibromyalgia and osteoarthritis. In 2003 the NDB established international data banks with assessments available in English, French, Spanish, Portuguese and Zulu. Additional details of the NDB, including its research questionnaires, can be found on its web site, www.arthritis.research.org.

Methods of data acquisition

The primary sources for NDB data are reports from patients and data from medical records. Patient self-report forms the basis of most outcome assessments in rheumatic diseases. Self-reports of pain, function, mood, fatigue and work limitations, for example, are measured by standard tools such as the

Table I. Abstracts and publications of the National Data Bank for Rheumatic Diseases by topic.

Genetics of RA	Retardation of disability
Disease activity assessment	Pain
Questionnaire development	Functional status
Direct medical costs	Clinically important differences
Indirect medical costs	Liver disease and drug toxicity
Cost effectiveness	GI ulceration, bleeding and perforation
Income and wage losses	Toxicity and safety of specific biologic, DMARDs and NSAID treatments
Work disability	Sinus disease
Lymphoma and other cancers	Cataract
Tuberculosis	Osteoporosis
Joint infection	Anemia
Stroke	Pneumonia
Myocardial infarction	Interstitial lung disease
Heart failure	Mortality
Adequacy of care	Treatment effectiveness
Ethnicity	Access to care
Fatigue	DAS
Fibromyalgia criteria	Fibromyalgia
X-ray progression	Poverty
Quality of life	

Health Assessment Questionnaire (HAQ) (13), HAQ-II (14), Medical Research Council (MRC) dyspnea index (15), SF-36 (16), work limitations questionnaire (WLQ) (17), visual analog scales and reports of individual symptoms. Similarly work disability, time lost from work, household income and wages are data that are most accurately derived from patients (2).

Side effects to treatment may be assessed in differing ways depending on the type of side effects. In almost all settings, patients are the best source of information regarding non-serious but bothersome side effects of treatment, as these side effects are rarely recorded by medical personnel in the course of ordinary patient care or even during interviews. For example, although diarrhea is a well-known side effect of leflunomide, it is rarely elicited in the clinical interview unless it is bothersome enough for the patient to report. Nor is the severity of side effects known to health professionals. As an example of the effectiveness of the NDB methodology to capture these types of events, the NDB estimates of diarrhea for leflunomide among 6,011 users was 17.0%. Of those, 43.9% discontinued the treatment because of the side effects, and 80.5% rated the side effect as moderate or severe. Similarly, among 12,081 users of prednisone, side effects were noted in terms of weight gain (9.2%), bruising (5.1%) and edema (5.6%). While it seems clear that detailed interviews with trained medical interviewers can capture such data, such interviews are expensive and unlikely to be conducted accurately in a clinical practice setting because of time demands. Thus, the NDB questionnaire methodology can provide a broad understanding of the impact of less serious side effects on the quality of life of persons with rheumatic diseases.

For events that are medically important or may not be fully comprehensible to the patient, physician-collected data and physician records can represent the gold standard for accuracy. However, such accuracy only applies when the data are collected as part of usual medical care and when the data are collected reliably. Physician data becomes unreliable when the method used is inherently unreliable. For example, a comment in a chart of "doing well" can never be extrapolated to a more accurate measure, nor can data completed at the end of the day when the patients have left. Remembering that the rheumatologist is most often not involved in the patient's non-rheumatic disease care, a comment from the patient that she had a "heart attack" or "pneumonia," which is then recorded in the chart, is no more reliable than a direct report from the patient. With the shortening of the patient-physician encounter, the usual rheumatology interview does not ordinarily record accurate, non-rheumatic disease symptoms or events. Even when rheumatology visits are concerned, vital data are most often not collected. Joint counts are collected routinely in less than 10% of rheumatology visits and reliable measures of global activity and severity rarely collected (18).

Can reliable data be collected in the physician's office? Over a 25-year period in the author's clinic, he collected such data on each patient, and at every visit. For research purposes, specially-trained nurses collected and reviewed the systems data. Analyses of such data showed it not to be accurate, as nurses faced with the usual office time constraints triaged the data collection form. We also experimented with detailed interviews conducted by trained assessors. When we were willing to devote 20 minutes of a highly paid staff member's time, accurate data were obtained, but it should be remembered that the data were little different than that obtained by self-report. The "in clinics" system tended to weaken when multiple, frequent visits occurred, and treatment changes – particularly with the respect to changes in non-rheumatic treatments – were often overlooked. Additionally, we found that rheumatology staff was often unfamiliar with non-rheumatic therapies and doses.

Data collection in the clinic has the advantage that it can provide accurate joint counts and access to laboratory data. However, joint count data are only as good as the effort put in to collect them, and symptom, adverse event and comorbidity data will be good only when considerable time and expense are used to collect them. Finally, the essential outcomes of rheumatic diseases are patient-based: work disability, joint replacement, direct and indirect costs, quality of life and mortality.

The NDB data collection model

Experience with physician data led us to adopt and perfect the model of data collection originally suggested by James Fries in the 1970s. The model uses self-report data from patients and validates important medical data by medical record review and physician contact. The NDB assesses patients every 6 months by mailed questionnaire, internet entry or telephone interview. NDB questionnaires are complex (28 pages) in order that the important rheumatic disease issues are covered. For patients who cannot or will not complete long questionnaires, shorter versions are available, and interviews of varying lengths are conducted by telephone for some patients. The NDB has different versions of questions for international sites and for illnesses such as lupus and osteoarthritis

Identification and validation of data

The essential issue regarding patient-reported events is their accuracy and reliability. NDB identifies important events (for example, myocardial infarction or cancer) from reports of adverse events, new symptoms, specific questions, hospitalization descriptions and hospital records. Upon identifying a putative event, patients are contacted directly (usually by telephone) by trained staff members who use specifically designed assessment questionnaires and algorithms. Approximately 60% of all patients are contacted for follow-up interviews. In these interviews, we obtain all patient-provided details as well as information regarding hospitals and the physicians consulted. Interviews are stored electronically as part of the NDB data base. Depending on the specific event, interviewing is followed up by obtaining medical records. The advantage of gathering data centrally, as opposed to doing it at peripheral sites, is that highly trained interviewers working from protocols can obtain high quality data in contradistinction to the usual office staff. NDB routinely and randomly reviews and validates the work of assessors and interviewers, including their coding and adherence to protocols. NDB routinely contacts patients who withdraw from the study to conduct exit interviews to identify possible illnesses that might have been related to study withdrawal. In the event that patients cannot be reached, we contact relatives and physicians. We conduct annual searches of the National Death Index (19) regarding patients who have discontinued participation and whose current status is not known.

Data processing

The primary method of NDB data entry is via multi-station scanning of specially designed questionnaires with sophisticated scanning software. The NDB programming staff has enhanced available software with extensive programming additions so that entry is appropriate, to ensure the efficient entry of rheumatic disease data. In addition, quality control measures are built into the NDB software enhancements. Questionnaires are processed by a team of trained 'verifiers.' Entry via the Internet has additional software data controls, but follows the same path of verification. At the time of processing, important 'events' are identified, and thus begins the data validation process. The NDB relies further on teams of telephone interviewers to follow-up on the validation process. All questionnaires are also stored as image files. Such files are instantaneously available using NDB software so that past questionnaires can be assessed for comparison when required.

Newly captured data are merged into statistical data bases daily using Stata statistical software (Stata Corporation, College Station, TX). Although the first line of quality control and error checking occurs at the time of data entry and processing, the NDB has developed extensive error checking and quality control programs using Stata. In Stata, extensive logical checks and consistency checks are employed, and formal reports are generated. Each day a new 'test'data base is produced and subjected to quality control checks. As the natural NDB data cycle is six months, we produce a new 'final' data base at the end of each 6-month period.

The basis of rheumatology quality control and good data

All data bases contain missing data and data errors. Data may be missing either because patients don't answer some questions or because new questions have been added to the questionnaire. In addition, some questions may turn out to have been "bad questions." The traditional data base relies on programmers and analysts. By contrast, the NDB data base was developed, tested and programmed by rheumatologists and rheumatology epidemiologists. We believe that the only real way problems regarding data inference can be identified and the data base improved is when end users (rheumatology analysts) are involved in the development, programming, quality control and use of the data. This passage through development and analysis is unique to the NDB.

The NDB analytic software

Based on Stata statistical software, the NDB has developed more than a thousand NDB programs designed specifically for the analysis of NDB data. Traditionally, data are presented to the analyst who then checks the data set, creates programs to analyze the data, and then analyzes it. By contrast, the NDB programs are developed in advance for the analyst. The NDB preprogramming guarantees high quality data and simple, effective tools for analysis.

What may take months to analyze using ordinary data base structures and programs can often be analyzed in minutes using NDB software. For example, the NDB command – *mkdmard, prior* – merges in all treatment data, doses and lifetime history of treatments in 15 seconds. The program – *gettotcosts* – makes available all direct medical cost data, adjusted to the most recent calendar year, in less than 20 seconds. In addition, NDB programs format the data for presentation tables.

In short, the NDB data bases and programs are a unique, rapid approach to using validated data that allows the analysts to concentrate on analytic issues, not programming issues.

Conclusion

The NDB is a research data bank with a broad agenda and approach to important clinical, regulatory, epidemiological, safety, effectiveness, outcome and patient care questions that cannot be answered by conventional data banks. It has systematic, ongoing, quality control programs that assure high quality data. NDB developed programs result in very efficient analytic capabilities and rapid publication.

References

1. WOLFE F, MICHAUD K: Data collection, maintenance, and analysis for rheumatic disease research. *Rheum Dis Clin North Am* 2004; 30: 753-68.
2. WOLFE F, MICHAUD K, CHOI HK, WILLIAMS R: Household income and earnings loss among 6,396 persons with rheumatoid arthritis. *J Rheumatol* 2005 (in press).
3. STERN R, WOLFE F: Infliximab dose and clinical status: results of 2 studies in 1642 patients with rheumatoid arthritis. *J Rheumatol* 2004; 31: 1538-145.
4. WOLFE F, MICHAUD K: Heat failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. *Am J Med* 2004; 116: 305-11.
5. WOLFE F, MICHAUD K: Severe rheumatoid arthritis (RA), worse outcomes, comorbid illness, and sociodemographic disadvantage characterize ra patients with fibromyalgia. *J Rheumatol* 2004; 31: 695-700.
6. WOLFE F, MICHAUD K: Lymphoma in rheumatoid arthritis: the effect of methotrexate and anti-tumor necrosis factor therapy in 18,572 patients. *Arthritis Rheum* 2004; 50: 1740-51.
7. WOLFE F, ZHAO S, PETTITT D: Blood pressure destabilization and edema among 8538 users of celecoxib, rofecoxib, and non-selective non-steroidal anti-inflammatory drugs (NSAID) and non-users of NSAID receiving ordinary clinical care. *J Rheumatol* 2004; 31: 1143-51.
8. WOLFE F, MICHAUD K, ANDERSON J, URBANSKY K: Tuberculosis infection in patients with rheumatoid arthritis and the effect of infliximab therapy. *Arthritis Rheum* 2004; 50: 372-9.
9. WOLFE F, MICHAUD K: Fatigue, rheumatoid arthritis, and anti-tumor necrosis factor therapy: an investigation in 24,831 patients. *J Rheumatol* 2004; 31: 2115-20.
10. WOLFE F: Fatigue assessments in rheumatoid arthritis: comparative performance of visual analog scales and longer fatigue questionnaires in 7,760 patients. *J Rheumatol* 2004; 31: 1896-902.
11. WOLFE F, MICHAUD K, PINCUS T: Development and validation of the health assessment questionnaire II: A revised version of the health assessment questionnaire. *Arthritis Rheum* 2004; 50: 3296-305.
12. MICHAUD K, MESSER J, CHOI HK, WOLFE F: Direct medical costs and their predictors in persons with rheumatoid arthritis: a 3-year study of 7,527 patients. *Arthritis Rheum* 2003; 48: 2750-62.
13. FRIES JF, SPITZ PW, KRAINES RG, HOLMAN HR: Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980; 23: 137-45.
14. WOLFE F, MICHAUD K, PINCUS T: HAQ-II: Development and validation of a revised version of the Health Assessment Questionnaire (HAQ). *Arthritis Rheum* 2004; 50: 3296-305.
15. WARLEY AR, FINNEGAN OC, NICHOLSON EM, LASZLO G: Grading of dyspnoea and walking speed in cardiac disease and in chronic airflow obstruction. *Br J Dis Chest* 1987; 81: 349-55.
16. STEWART AL, HAYS RD, WARE JE: The MOS short-form general health survey. *Med Care* 1988; 26: 724-35.
17. WALKER N, MICHAUD K, WOLFE F: Work limitations among working persons with rheumatoid arthritis: results, reliability, and validity of the work limitations questionnaire in 836 patients. *J Rheumatol* 2005; 32: 1006-12.
18. WOLFE F, PINCUS T: Listening to the patient: a practical guide to self-report questionnaires in clinical care. *Arthritis Rheum* 1999; 42: 1797-808.
19. DOODY MM, HAYES HM, BILGRAD R: Comparability of national death index plus and standard procedures for determining causes of death in epidemiologic studies. *Ann Epidemiol* 2001; 11: 46-50.