

Towards an epidemiology of rheumatoid arthritis outcome with respect to treatment: randomized controlled trials overestimate treatment response and effectiveness

F. Wolfe^{1,2} and K. Michaud^{1,3}

The results and outcomes of randomized clinical trials of leflunomide and anti-TNF therapy are much better than are seen in rheumatoid arthritis patients in the community. This appears to be an effect of the clinical trial system. The consequence of deriving effectiveness estimates from clinical trials is to overestimate the effectiveness and thereby the cost-effectiveness of rheumatoid arthritis treatments.

KEY WORDS: Randomized controlled trials, Observational studies, Bias, Outcome, Rheumatoid arthritis.

Treatment effectiveness

Randomized clinical trials (RCTs) and observational studies often provide similar evidence of comparative efficacy in most disorders [1, 2]. However, there are features of RCTs in rheumatoid arthritis (RA), such as disease activity-related entry and exclusion criteria, restriction of concomitant medications and participation in the trial itself, that limit the generalizability of trial results. In addition, RCTs only provide hints as to how well drugs will work in the community. While RCTs provide evidence of efficacy, what is really needed is information on effectiveness, or on the extent to which treatments work in the community of patients who actually use them. In the data that follow, we provide evidence regarding differences in results between RCTs and clinical practice.

A model for treatment effectiveness

We propose a simple model to measure treatment effectiveness in RA. At its core is the hypothesis that if an effective treatment is introduced into the general population of persons with RA, it will reduce symptoms and improve function in a measurable way in that population. Symptom severity is usually measured by visual analogue (VAS) pain and global severity scales, and function is usually measured by the Health Assessment Questionnaire (HAQ). These simple measures are based on patient self-report and are inexpensive and relatively easy to obtain.

Let us assume that a treatment reduces HAQ scores by a measurable percentage or by an absolute amount. For the purpose of simplicity we will assume that RCT data for anti-tumour necrosis factor (TNF) therapy indicate a reduction in HAQ score of 0.5 units and that this result is not artefactual. In the National Data Bank for Rheumatic Diseases (NDB) [3–6], a large longitudinal outcomes data bank, the average HAQ score is 1.1, a result that is remarkably similar to that in the Norfolk Arthritis Register (NOAR) [7] after accounting for disease duration [8]. If 30% of patients receive this treatment, the HAQ score in the community will be reduced to 0.95 [$0.7 \times 1.1 + 0.3 \times (1.1 - 0.5)$]. However, if the final HAQ score is greater than 0.95 then the effect of treatment in the community is less than the effect noted in RCTs.

One might apply this method by annual or semi-annual sampling of RA patients in the general population. If the current and past treatment history is obtained, a direct measurement of treatment effectiveness can be obtained.

This method of measuring treatment effect has certain advantages. It does not matter whether channelling bias results in the patients with the most severe arthritis receiving the medication [9], as we are only measuring the population effect. As long as we sample in large general populations our results will be robust. Although we have used the HAQ functional score in this example [10, 11], we can also evaluate the level of symptoms such as pain and fatigue.

Differences between RCT results and those of observational studies

It is desirable to know a therapy's actual effectiveness. Cost-effectiveness analyses, for example, base their results on presumed HAQ scores and HAQ improvement in the community from data that is extrapolated from RCTs [12, 13]. There is reason to believe, however, that estimates of effectiveness derived from RCTs may be biased [14].

Results of anti-TNF therapy clinical trials have shown these treatments are substantially better than placebo [15–18] and that they appear to produce profound improvement when measured by ACR improvement criteria [19] and/or the Disease Activity Scale (DAS) [20]. This degree of improvement is also seen with leflunomide [21]. In the analyses that follow we have selected a well-done adalimumab RCT as the prototype for clinical trial responsiveness and effectiveness [16].

Tables 1 and 2 present data for pain (Table 1) and HAQ (Table 2) from the adalimumab RCT [16]. These results are compared with results of anti-TNF therapy in the NDB [22], where data represent the ordinary usage of anti-TNF therapy by RA patients in the clinical setting. The NDB observes treatment and outcome but does not influence treatment in any way. Data from the RCT represent baseline values and results after 52 weeks of therapy. Results from the NDB in the tables are for a cohort of 2211 patients assessed approximately 3 months before and at

¹National Data Bank for Rheumatic Diseases, Wichita, KS, ²University of Kansas School of Medicine, Wichita, KS and ³Center for Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA.

Correspondence to: F. Wolfe, 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

TABLE 1. Pain outcomes of anti-TNF therapy after 1 yr in an adalimumab clinical trial compared with observational data from the National Data Bank for Rheumatic Diseases

Measurement	Adalimumab Prescription	NDB prescription At prescription (estimated)	NDB standard ~3 month before prescription
Baseline time			
<i>N</i>	419	2211	2211
Estimated DAS	4.7	4.7	4.3
Baseline pain (0–10)	5.5	5.3	4.5
Pain at 1 yr (0–10)	3.0	3.9	3.9
Pain difference (units)	2.5	1.4	0.7
Pain group improvement (%)	~50%	26.5%	14.4%
Pain 20 (%) improvement (%)		(28.5%) ^a	(14.9%) ^a
Overall 20 (%) improvement (%)	~57%	47.4%	41.9%
Improvement (~ACR-N) (%)		53.5%	45.3%
Overall improvement (%)		21.1%	10.3%

^aAdjusted to an estimated DAS of 4.7.

TABLE 2. HAQ outcomes of anti-TNF therapy after 1 yr in an adalimumab clinical trial compared with observational data from the National Data Bank for Rheumatic Diseases

Measurement	Adalimumab Prescription	NDB prescription At prescription (estimated)	NDB standard ~3 months before prescription
Baseline time			
<i>N</i>	419	2211	2211
Estimated DAS	4.7	4.7	4.3
Baseline HAQ (0–3)	1.45	1.34	1.25
HAQ at 1 yr (0–3)	0.85	1.19	1.19
HAQ difference (units)	0.60	0.14	0.06
HAQ group improvement (%)	~41%	11.2%	5.1%
HAQ 20 (%) improvement (%)		(11.6%) ^a	(5.7%) ^a
Overall 20 (%) improvement (%)	~57%	37.6%	33.9%
Improvement (~ACR-N) (%)		47.4%	41.8%
Overall improvement (%)		19.2%	10.3%

^aAdjusted to an estimated DAS of 4.7.

the start of an anti-TNF agent and again 9–12 months later. RCT and NDB timing and results are shown graphically in Fig. 1 [22].

Patients on adalimumab experienced a 50% reduction in pain compared with reductions in pain of 28.5 and 14.9% for the NDB cohort, depending on whether the starting time was calculated from the time of prescription or 3 months before prescription. The RCT demonstrated not only a greater reduction in pain but also lower final values than did the NDB study (3.0 vs 3.9). The overall ACR 20% improvement for the RCT was 57% compared with the NDB determination of 47.4 and 41.9% (based on patients' scores for pain, HAQ and global, ACR-N). When an average level of treatment improvement was calculated based on the mean improvement in pain, HAQ and patient global in the NDB, improvement was 21.1% judged from the time of prescription and 10.3% when the baseline observation was 3 months prior to the start of anti-TNF therapy.

When HAQ was considered (Table 2), adalimumab RCT patients' HAQ scores improved by 41% compared with improvements for the NDB subjects of 11.6 and 5.7%. These data show that, compared with patients in the community assessed by the NDB, adalimumab RCT patients had much greater improvement

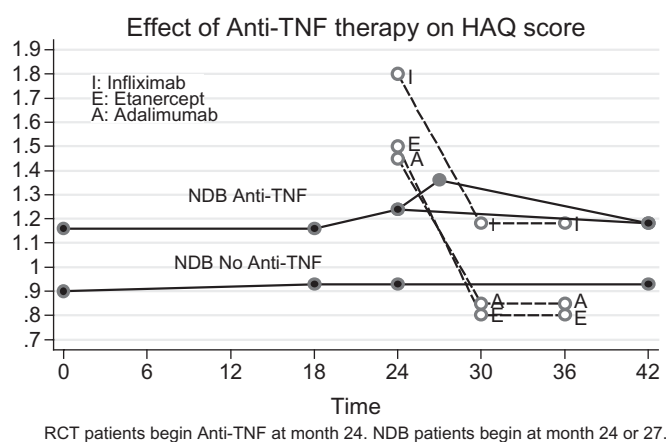


FIG. 1. Anti-TNF therapy in RCTs and observational data from the National Data Bank for Rheumatic Diseases (NDB). The lower solid line represents the HAQ course of NDB patients who did not receive anti-TNF therapy. The upper solid line represents the course of NDB patients ($n=1494$) who received anti-TNF clinical trials. Open circles represent results from three anti-TNF clinical trials. Note the absence of pre-RCT data on clinical trial patients. The HAQ level at 27 months represents the estimated level of HAQ at the moment of starting anti-TNF therapy among NDB patients. E, an etanercept trial; A, an adalimumab trial; I, an infliximab trial. (From [22].)

in pain and HAQ, and had final scores that were considerably lower than found in the NDB community surveys.

How should we interpret RCT results in terms of effectiveness?

There is some reason to be confused by the clinical trial results. Should we be reporting the overall study improvement of 57% or should we consider the improvement to be 33%, the result that would be obtained if we subtracted the reported results of the placebo group [16]? If persons are in remission on active drug and we subtract the results of the placebo group, does that mean that you can be in a false remission, with a remission rate that is dependent on handling of placebo rates? If advertising reports the percentage of patients on active drug with 20, 50 and 70% responses, should we accept that as being true or should we somehow adjust it for placebo effect and other factors that might bias the response? The differences between the improvements achieved by those on active drug vs those on placebo are critical if we are to extrapolate the results to effectiveness in the community. How would we handle the results of the RCT in which the two arms tested treatments that were equally effective? It would seem that improvement in an RCT, given the placebo/comparator response, is not commensurable with effectiveness in the community.

Do high rates of ACR 20% improvement mean high rates of improvement for the group?

Another factor of importance is that high rates of ACR response do not necessarily mean high rates of overall response (percentage improvement). The rate of ACR 20% improvement for pain (Table 1, column 3) was 45.3%. This corresponds to a cohort improvement of only 14.4% because improvement scores do not consider patients who do not improve or even worsen.

Current methods of assessing treatment effect in clinical trials generally provide an overall improvement score, such as the ACR

20, 50 and 70% scores, or improvement based on the DAS score. Another problem with interpreting clinical trial results is that the percentage improvement, as well as the final value for the individual components (pain, HAQ), is a function of the starting score and the severity of the study populations [22].

What is the real baseline from which we should infer effectiveness?

In biological registries the degree of improvement observed is dependent on when the baseline assessments are performed as well as the selection criteria for treatment. Where criteria for receiving and/or continuing a biological agent is dependent on an index like the DAS or on its individual components (e.g. UK Biologics Registry), it is almost certain that the strictness of the criteria will lead to greater apparent severity at the time of prescription than actually exists and greater improvement than actually exists. There have been no direct studies of this assertion, to our knowledge, but clinicians who have ever filled out a disability form, corresponded with an insurance company or participated in a clinical trial know of this phenomenon. If true, this ‘gaming’ of the system will distort to some extent the response and final measured result of treatment.

Both registries and clinical trial improvement scores are sensitive to the timing of baseline HAQ and activity measures. What is a proper and appropriate baseline time? Among the courses of RA, there are three that are of interest in the setting of biological prescription. Course 1 describes a patient trajectory in which HAQ and activity scores are constant over time. A new, better treatment becomes available and is prescribed. The baseline HAQ/activity at the time of prescription (or before) therefore accurately reflects the patient’s true pretreatment status.

Course 2 may be more common. It is the same straight-line course as Course 1 except that it is punctuated by exacerbations and improvements (like a sine curve). How often do these fluctuations occur? Using 1869 patients and 34 573 pain scores available in the Wichita clinical data bank section of the NDB, we determined that the mean pain score was 4.2 and the within-patient standard deviation was 1.9, while the pain score changed only slightly with time (0.02 units per yr on 0–10 scale). This means that increases in activity of 25–50% are common in individual patients. Based on the author’s clinical experience and analyses (and that of others), treatment changes tend to occur at times of flare and exacerbation. The Course 2 model suggests that a ‘baseline’ score does not reflect the overall patient’s activity; instead it reflects a temporary flare state. If this assumption is correct, the improvement seen should not be measured at the time of flare but instead should reflect the patient’s overall HAQ and activity, perhaps using a smoothed estimate, such as is commonly used in economic studies, in order to remove daily or monthly fluctuations. The upper horizontal line of Fig. 1 suggests that baseline values that are 3 or more months previous to the start of therapy may be the most appropriate point for baseline measurement. These two courses are illustrated in columns 2 and 3 of Tables 1 and 2.

Course 3 differs from Course 2 in that the flare is not temporary; instead the increase in HAQ and disease activity reflects a permanent change. In this instance, as in the case of Course 1, the baseline value should be the value observed at the time of anti-TNF administration. Course 3 is particularly important because it implies that RA would end up being much more severe if there were no intervention. Courses 1 and 2, by contrast, improve clinical status but are not involved in preventing worsening. In understanding courses and the result of treatment in the clinic, we should recognize the treatments are added to all other treatments that might be given in the clinic, including joint injections, steroids, NSAIDs, analgesics and other physical and social support measures.

In contrast to clinical practice, entry criteria for RCTs generally ensure that participants will have more activity than is found in ordinary clinical practice. Trials also have a strong flare component, as the ordinary clinical interventions that would reduce activity are usually prohibited within a month (sometimes more) of entry into the trial. There is also evidence that the placebo effect is greatest among physician measures [23], indicating an over-rating of activity by physicians compared with patient and laboratory measures. In addition, patients usually know they have to be ‘severe enough’ to enter trials. These factors ensure that baseline measures will be more abnormal in RCTs than in the community. However, even if such factors occur they are applied equally to all treatment arms and so do not imperil the validity of the trial.

Participation in a trial may also increase response to treatment. Physician and patient factors that lead to increased apparent severity at the time of trial entry may also lead to apparent improvement once the trial has begun. When patients drop out of trials their final results are most often estimated using the last value carried forward (LVCF) method. When we have analysed data from patients in the NDB, we found that patients discontinuing anti-TNF therapy had subsequent HAQ and severity scores that were considerably worse on average than at the time of study entry. In addition, merely participating in a study may be beneficial. The improvement seen in a clinical trial, therefore, is a mixture of treatment effect, measurement error, regression to the mean and participation effect.

Our purpose in raising these issues is not to vitiate RCTs or their positive results, but to call attention to the problem of generalizing the degree of apparent improvement to the overall level of effectiveness in the community. These issues occur in all trials, and the problems are not restricted to anti-TNF therapy.

How we should measure effectiveness?

So far, we have suggested that clinical trials are biased in favour of increased response and that registries that document the clinical status of patients who are dependent upon severity criteria to receive and continue a treatment are also biased towards increased response. The ideal way of evaluating drugs in the community is to evaluate patients continuously and to observe the effect of treatment. This implies following patients who will not receive treatment as well as those who will receive such treatment. In addition, it means that patients should be followed in the years before and the years after treatment begins.

This is not an impractical solution. The cost of just one of the many major clinical trials in anti-TNF therapy could sustain data collection in a very large number of patients for a decade. The outcomes of RA that tend to matter most—functional ability, employment, symptoms and costs—do not require expensive physician participation. What is required is a paradigm shift away from the idea that only RCTs can produce reliable and expectable evidence to the idea that RCTs cannot provide evidence regarding the degree of effectiveness of treatment, and that the best evidence of effectiveness can most often come from patient data. One force that can move us in this direction is when payers and regulatory authorities ask for actual evidence of effectiveness rather than evidence that is extrapolated from clinical trials. There is some indication that regulatory authorities are becoming interested in community outcomes [24]. The forces that work against this direction are commercial interests and a community of academic and semi-academic rheumatologists who are dependent on the RCT model. An additional important factor that works against this direction is that very few academic or regulatory physicians have any experience with observational data, with the result that, while problems with observational data

may be understood, the proper use and value of observational data are truly not understood.

An example from the NDB

We have suggested that observational data might provide a more realistic measure of effectiveness. We also pointed out that apparent improvement (and effectiveness) depends upon the timing of the baseline measure. When baseline is the moment of prescription, improvement will be greater than if it occurs at a prior time. The NDB makes observations at 6-month intervals. Therefore, baseline assessments occur an average of 3 months prior to anti-TNF start. Figure 1 provides insights regarding the issues of improvement and status. On average in this figure, NDB patients were not on anti-TNF therapy from month 0 to month 24, and were noted to be on therapy at month 30. The time of starting therapy was between months 24 and 30. It is indicated on the figure as month 27. Patients who would not receive anti-TNF therapy ($n = 2406$) had mean HAQ scores that were relatively constant and were always lower than those who would receive anti-TNF therapy. The HAQ score in this group was approximately 0.90. The HAQ score of the 1494 patients who would receive anti-TNF therapy was approximately 1.16 to month 18. We observed it to be 1.24 at month 24 which, on average, is 3 months before they started anti-TNF therapy. From an additional 2018 patients whose HAQ data were available at the moment of prescription, we know the mean HAQ score at prescription time was 1.40. From other available data we estimate the DAS 28 scores for HAQ values of 1.16, 1.24 and 1.40 to be 4.3, 4.4 and 4.6. The open circles in the graph represent three clinical trial HAQ trajectories for infliximab, etanercept and adalimumab trials. From these data, we can see that patients in RCTs had worse starting scores and more improvement compared with patients in the NDB. Note that we have not subtracted a placebo effect for the clinical trials data.

Which baseline is appropriate for observational data? If we believe that the moments of prescription truly represent the persistent status of the RA patients treated with anti-TNF therapy then we should choose the 1.40 value. However, an increase in HAQ scores from 1.16 to 1.40 in 2–3 years is inconsistent with published progression rates in RA of ~ 0.03 units per year [11]. If we choose the value at month 24 of 1.24 we might be more accurate, as we will be discounting a flare effect to some extent. These observations from the NDB compared with those from clinical trials have important implications for cost-effectiveness studies that rely on clinical trial data. It is clear that improvement is less if we posit that the baseline should be a real clinical state rather than one identified in RCT participants.

Measuring outcome of therapy in the community

Both RCTs and observational studies run into problems regarding timing of the baseline and (for RCTs and registries) selection of patients. We suggested earlier a different approach. Instead of measuring within-patient change, we could be measuring symptoms and function in the general population. The introduction of a new treatment into a community of RA patients should result in a reduction of the burden of RA that is independent of patient selection and timing of therapy. While this is the model of an observational study in which cohorts are followed over time, we suggest an additional model: annual random sampling of RA patients in the community with a brief survey. Such a model would certainly allow measurement and documentation of the true effectiveness of therapy in the community.

F.W. is an employee of the non-profit research corporation, The National Database for Rheumatic Diseases. This agency has received research grants from centres Bristol-Myers-Squibb and Amgen.

Acknowledgements

The National Data Bank for Rheumatic Diseases has received support from the pharmaceutical companies Centocor, Amgen, Abbott and Aventis.

References

1. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–86.
2. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–92.
3. Wolfe F, Michaud K. Heart failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. *Am J Med* 2004;116:305–11.
4. 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.
5. 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.
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. Wiles NJ, Scott DG, Barrett EM *et al.* Benchmarking: the five year outcome of rheumatoid arthritis assessed using a pain score, the Health Assessment Questionnaire, and the Short Form-36 (SF-36) in a community and a clinic based sample. *Ann Rheum Dis* 2001;60:956–61.
8. Wolfe F, Choi HK. Benchmarking and the percentile assessment of RA: adding a new dimension to rheumatic disease measurement. *Ann Rheum Dis* 2001;60:994–5.
9. Petri H, Urquhart J. Channeling bias in the interpretation of drug effects. *Stat Med* 1991;10:577–81.
10. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;23:137–45.
11. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. *Arthritis Rheum* 2000;43:2751–61.
12. Kobelt G, Eberhardt K, Jonsson L, Jonsson B. Economic consequences of the progression of rheumatoid arthritis in Sweden. *Arthritis Rheum* 1999;42:347–56.
13. Brennan A, Bansback N, Reynolds A, Conway P. Modelling the cost-effectiveness of etanercept in adults with rheumatoid arthritis in the UK. *Rheumatology* 2004;43:62–72.
14. Wolfe F, Michaud K, Pincus T. Do rheumatology cost-effectiveness analyses make sense? *Rheumatology* 2004;43:4–6.
15. Bathon JM. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *N Engl J Med* 2000;343:1586–93. Erratum in: *N Engl J Med* 2001;344:240 and *N Engl J Med* 2001;344:76.
16. Keystone EC, Kavanaugh AF, Sharp JT *et al.* Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: a randomized, placebo-controlled, 52-week trial. *Arthritis Rheum* 2004;50:1400–11.

17. Lipsky PE, van der Heijde DMFM, St Clair EW *et al.* Infliximab and methotrexate in the treatment of rheumatoid arthritis. *N Engl J Med* 2000;343:1594–602.
18. Maini RN, Breedveld FC, Kalden JR *et al.* Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. *Arthritis Rheum* 2004;50:1051–65.
19. Felson DT, Anderson JJ, Boers M *et al.* The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. *Arthritis Rheum* 1993;36:729–40.
20. van der Heijde DMFM, Van 't hof MA, van Riel PLCM, van Leeuwen MA, van Rijswijk MH, van de Putte LBA. Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis. *Ann Rheum Dis* 1992;51:177–81.
21. Strand V, Tugwell P, Bombardier C *et al.* Function and health-related quality of life: results from a randomized controlled trial of leflunomide versus methotrexate or placebo in patients with active rheumatoid arthritis. *Leflunomide Rheumatoid Arthritis Investigators Group. Arthritis Rheum* 1999;42:1870–8.
22. Wolfe F, Michaud K, Dewitt EM. Why results of clinical trials and observational studies of antitumour necrosis factor (anti-TNF) therapy differ: methodological and interpretive issues. *Ann Rheum Dis* 2004;63(Suppl. 2):ii13–ii17.
23. Pincus T, Strand V, Koch G *et al.* An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR 20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. *Arthritis Rheum* 2003;48:625–30.
24. Kolata G. Medicare covers new treatments with a catch. *New York Times*, November 5, 2004.