Shifting Our Thinking About Uncommon Disease Trials: The Case of Methotrexate in Scleroderma

SINDHU R. JOHNSON, BRIAN M. FELDMAN, JANET E. POPE, and GEORGE A. TOMLINSON

ABSTRACT. ***Objective.*** Randomized trials for uncommon diseases suffer from methodological challenges: difficulty in recruiting sufficient numbers of patients and low power to detect important treatment effects. Using traditional (frequentist) analysis, p values > 0.05 mean investigators are unable to reject the null hypothesis (of no treatment effect). The medical community often labels trials with p values > 0.05 as "negative." Our study demonstrates how Bayesian analysis conveys more relevant information to clinicians — using the example of methotrexate (MTX) in systemic sclerosis (SSc).

Methods. Data from 71 patients with diffuse SSc (n = 35 MTX, n = 36 placebo) in the trial were reanalyzed using Bayesian models. We examined 3 primary outcomes: modified Rodnan skin score (MRSS), University of California Los Angeles (UCLA) skin score, and physician global assessment of overall disease activity. Using noninformative prior probability distributions, the probability of beneficial treatment effects for each outcome and the probability of simultaneous benefit in outcomes were computed.

Results. The probability that treatment with MTX results in better mean outcomes than placebo was 94% for MRSS, 96% for UCLA skin score, and 88% for physician global assessment. There was 96% probability that at least 2 of 3 primary outcomes were better on treatment.

Conclusion. Bayesian analysis of uncommon disease trials allows for more flexible and clinically relevant interpretations of the data. From the trial data, clinicians can infer that MTX has a high probability of beneficial effects on skin score and global assessment. (First Release Dec 1 2008; J Rheumatol 2009;36:323–9; doi:10.3899/jrheum.071169)

Key Indexing Terms:
RARE DISEASE TRIALS BAYESIAN ANALYSIS METHOTREXATE SCLERODERMA

Randomized trials studying therapy for uncommon diseases suffer from a number of methodological challenges. Due to the rarity of the disease under study, small numbers of patients are available for recruitment. Consequently, well designed trials may have insufficient power against important treatment effects and therefore be unable to yield definitive conclusions about an intervention's effect[1]. The principles of significance and power are derived from the school of statistical inference referred to as "frequentist" statistics. When designing a trial under this paradigm, a null hypothesis of no treatment effect is usually specified, and an alternative hypothesis (usually, that a treatment effect of some magnitude exists) is also specified. A level of significance (allowable false-positive rate), indicated by the Greek letter α, is most often set — without much thought — at 0.05. If a p value > 0.05 is found, investigators must conclude that "there is insufficient evidence to reject the null hypothesis"[2]. However, such studies are often inappropriately labelled as "negative" studies[3] and over time this precipitates a belief in the medical community that the treatment "has no effect" and/or "doesn't work"[4]. Although this has the positive effect of screening out most completely ineffective treatments, it can result in trials having low power against important effect sizes[5]. This can have important implications when balancing the seriousness of statistical error[6]. Investigators commit type I error (false-positive) when they state a treatment effect exists when in truth it does not[6]. Investigators commit type II error (false-negative) when they state that a treatment effect does not exist when in truth it does[6]. In the setting of a relatively safe treatment with no alternative treatment options, the seriousness of type I error is relatively small,

From the Division of Rheumatology, University Health Network, The Hospital for Sick Children, Toronto; Departments of Paediatrics, Health Policy Management and Evaluation, and Public Health Sciences, University of Toronto, Toronto; University of Western Ontario, London; and Division of Clinical Decision Making and Health Care, Toronto General Research Institute, Toronto, Ontario, Canada.

Dr. Johnson has been awarded an Abbott Scholar Award for Rheumatology Research and Canadian Arthritis Network Fellowship. Dr. Feldman is supported by a Canada Research Chair in Childhood Arthritis.

S.R. Johnson, MD, FRCPC, Division of Rheumatology, University Health Network, Department of Health Policy Management and Evaluation, University of Toronto; B.M. Feldman, MD, MSc, FRCPC, Division of Rheumatology, The Hospital for Sick Children, Departments of Paediatrics, Health Policy Management and Evaluation, and Public Health Sciences, University of Toronto; J.E. Pope, MD, MPH, FPCPC, Division of Rheumatology, University of Western Ontario; G.A. Tomlinson, PhD, Departments of Health Policy Management and Evaluation, and Public Health Sciences, University of Toronto, Division of Clinical Decision Making and Health Care, Toronto General Research Institute.

Address reprint requests to Dr. S. Johnson, Division of Rheumatology, Toronto Western Hospital, Ground Floor, East Wing, 399 Bathurst Street, Toronto, ON M5T 2S8. E-mail: Sindhu.Johnson@uhn.on.ca

Accepted for publication September 22, 2008.

whereas the seriousness of type II error is relatively large. This can result in patients being denied treatments with beneficial effects, and may suggest that further research is not required[4]. An example is the use of fibrinolytic agents in myocardial infarction. The lack of a significant p value in individual trials led to a delay in the clinical acceptance of this therapy due to a belief in its lack of efficacy[3]. Only through the evaluation of 33 trials was the reduction in mortality appreciated, and the belief within the medical community revised[7].

Increasingly in the literature, there is a move away from the p value towards looking at confidence intervals (CI). However, CI also suffer similar disadvantages[8]. The CI is often used as a de facto significance test when users examine whether its endpoints overlap the null value[6]. Consumers of the medical literature often mistakenly conclude that there is a 95% probability that the true treatment effect lies with the 95% CI[9]. However, a 95% CI indicates that if the same study was repeated many times, 95% of CI would include the true treatment effect.[9] In addition, the CI does not report the information that clinicians are interested in, namely, the probabilities for clinically important benefits. Clinical care involves making decisions with less than 95% confidence[10]. Clinicians may still be interested in knowing whether an intervention has a 70% or 80% probability of a clinically important effect.

As an example of an uncommon disease, we have chosen to examine systemic sclerosis (SSc), a chronic disease with an incidence of 1–2 per 100,000 a year[11], and a prevalence of 276/100,000[12]. It is characterized by fibrosis, vasculopathy, and immune activation. Patients with SSc have pain, impaired physical function, and poor quality of life[13]. There is no cure. Investigators have evaluated treatment options, but this has been a frustrating effort. Two studies (an observational study and a 24-week randomized trial followed by a 24-week observational study) suggested that methotrexate (MTX) conferred beneficial effects[14,15]. This led to a placebo-controlled trial of MTX in early diffuse SSc[16]. The investigators reported a nonsignificant ($p \geq 0.05$) benefit of MTX versus placebo in 2 primary outcomes [modified Rodnan skin score (MRSS) and University of California Los Angeles (UCLA) skin score] and a statistically significant benefit ($p = 0.04$) in the physician global assessment of overall disease activity. In the design of the study, an optimistic difference in skin score (35% difference) was used in the sample-size calculation. As a result the study was underpowered to "detect" smaller, but clinically important treatment effects under the frequentist paradigm. The investigators (correctly) concluded that there was insufficient evidence to reject the null hypothesis of no treatment effect. However, over time this trial has been labelled a "negative" study[17]. This has precipitated a belief by some that "methotrexate doesn't work in SSc." A survey of both general rheumatologists and scleroderma experts found that only 22% frequently use MTX in the treatment of skin involvement[17].

The interpretation of the p value and these study results is likely related to discordance between the meaning of the p value and the way clinicians think. A p value is the probability that one would observe an effect (test statistic) as extreme or more extreme than the one observed if the null hypothesis were true. The discordance occurs because clinicians think in terms of conditional probabilities, not in terms of long-run frequencies[18]. When confronted with a patient, the clinician (implicitly) estimates a pretest probability that the patient has a disease based on the constellation of signs and symptoms. If a diagnostic test yields a positive result, a post-test probability of disease is (again implicitly) revised based on the test result. Clinicians do not think of the patient in terms of long-run repeated experiments with a p value[6]. Clinicians are less interested in the long-run frequency of data at least as unusual as the observed study data, given that an intervention is ineffective, but rather, clinicians are more interested in the probability that an intervention is effective given the current data.

The Bayesian school of statistical inference provides an alternative paradigm in which to analyze data and present results. Bayes' Theorem formalizes the way that preexisting knowledge or belief can be combined with new data to inform clinicians about the actual probability of a treatment effect[19]. Thus, the objectives of our study are to demonstrate how the use of Bayesian analysis makes efficient use of the available data and presents the results in a format that is more clinically relevant to consumers of the medical literature using data from a trial of uncommon disease (systemic sclerosis).

MATERIALS AND METHODS

This study is a reanalysis of the MTX in SSc trial[16].

Patients. Subjects were recruited from 8 centers, were ≥ 18 years of age, fulfilled criteria for SSc[20], had diffuse disease[21], and had a disease duration < 3 years.

Study design. The study was a 1-year, parallel-groups, double-blind, randomized, placebo-controlled trial. Details of the inclusion and exclusion criteria, methods of randomization, and allocation concealment are as described[16].

Outcome measures. The primary outcomes were the 12-month skin score measured by 2 methods (UCLA skin score[22] and MRSS[23]) and physician global assessment of overall disease activity. The UCLA skin score evaluates 10 sites with a maximum score of 3 at each site, giving a possible score of 0–30[22]. The MRSS evaluates 26 sites with a maximum score of 3 at each site, giving a possible score of 0–78[23]. The physician global assessment of overall disease activity was recorded on a 10-cm visual analog scale, anchored by 0 indicating no disease and 10 indicating worst disease.

Statistical analysis. Raw data were provided by the principal investigator of the original study (JEP). In the original analysis, the investigators analyzed only the data for patients who completed the study at 12 months. Using the same data, we used Bayesian methods to calculate the probability of a beneficial treatment effect for each outcome measure at 12 months. We used a 2-group comparison, analogous to a 2-sample t-test, but set up as a linear model. Individual outcomes Y_i were assumed to come from a normal dis-

tribution with a common between-subject variance σ^2. The mean of that distribution was determined by the group the subject was in:

$$Y_i \sim N(\alpha + ß * group_i, \sigma^2)$$

Here, $group_i = 1$ if a subject received MTX and $group_i = 0$ if a subject received placebo, so that ß is the difference in the mean of the outcome between the MTX and placebo groups. These models were fitted using WinBUGS and the posterior distribution of ß was used to compute the probability that MTX had a beneficial effect. The phrase "doesn't work" was assumed to mean that MTX has no beneficial effect on an outcome measure, so a beneficial effect was defined as any improvement in score greater than zero. This is comparable to the investigators' original frequentist analysis, setting a null hypothesis of "no effect" and not specifying minimal clinically important differences. With a particular prior distribution for the variance, the posterior probability of a beneficial effect computed this way will be exactly $100 \times (1 - p)\%$, where p is the p value from a 1-tailed t-test of treatment efficacy. Our results are slightly different from this because we used a uniform prior distribution over the range 0 to the maximum possible standard deviation (SD) for the instrument in question.

We repeated the analysis with multiple imputation of missing 12-month data. Evaluation of the raw data indicated a potential bias in the complete case analysis towards no apparent treatment benefit: more patients in the placebo group dropped out of the study, and patients who were sicker at randomization tended to drop out. Missing 12-month outcomes were multiply imputed from within-subject longitudinal data, and the Bayesian model above was fitted to the resulting imputed datasets. Results for the estimated treatment effect were combined across multiple imputed datasets so that they reflected both within-dataset and between-imputation variability.

In addition, we fitted a multivariate Bayesian regression model to all 3 outcomes, treating them as a sample from a 3-dimensional multivariate normal distribution, again in the imputed datasets. This approach allowed for the correlation between the 3 outcomes on the same individual and meant that we could make statements about the probability of a simultaneous benefit in these outcomes, for example, that all outcomes showed an effect or that 2 or more showed an effect.

Noninformative versus informative prior. The inferences from a Bayesian analysis of a clinical trial combine information in the data from the trial with information external to that trial. When there is no information external to the trial, the point and interval estimates of the treatment effect will often be similar to those from a frequentist analysis. To illustrate how the Bayesian approach can formally include external information and how this can affect inferences, we reanalyzed the MRSS data using a prior distribution for the treatment effect derived from an earlier study of MTX in this population[15]. The outcomes in the earlier work were reported in units of Total Skin Score (TSS) and required that a few intermediate calculations be made to construct a prior distribution on the scale of the MRSS. We rescaled the result from the TSS to the MRSS by first rescaling the TSS in terms of an effect size and then converting this to the MRSS by using the SD for MRSS. This is an approach commonly seen in metaanalysis, when studies report different continuous outcomes measuring the same underlying construct. The standardized mean difference was computed as the treatment effect divided by the pooled SD. This was then multiplied by the pooled SD of the MRSS to create a treatment effect on that scale. The result was a prior distribution with a mean of –2.5 and SD of 2.4 MRSS points. Based on the previous study, we would go into the present one with a belief that MTX decreased the MRSS by 2.5 points at 24 weeks compared to placebo. This prior distribution corresponds to about an 85% probability that MTX improves the MRSS.

Computational details. Noninformative prior probability distributions were used for all parameters in our analyses, representing ignorance about placebo group scores, treatment effects, and between-subject SD. In particular, in the univariate models, priors for the treatment effects were centered at zero with SD of 1000 and priors for between-subject SD were uniform on (0, W), where W is the range of the observed values for the outcome. The multivariate model used a diffuse conjugate multivariate normal-inverse Wishart prior. Posterior means, quantiles, and probabilities were computed from Monte Carlo Markov Chain (MCMC) sampling of the posterior distributions. Where appropriate, convergence was evaluated using the Brooks-Gelman-Rubin convergence statistic. The reporting of the analysis and results are in accord with the ROBUST criteria[24]. Analyses were performed using MCMCpack[25], bayesm[26], and R2WinBUGS[27] in R 2.4 (R Foundation for Statistical Computing, Vienna, Austria) and WinBUGS v1.4 (Imperial College and Medical Research Council, London, UK). The code for all analyses is available from the authors upon request.

RESULTS

Baseline characteristics. Seventy-one patients were enrolled in the study, with 35 patients randomized to receive MTX and 36 to receive placebo. Twenty-two patients in the MTX group and 25 in the placebo group completed the study. In both groups, most patients dropped out prior to study completion due to treatment inefficacy. One patient in the MTX group dropped out due to development of oral ulcers.

Frequentist analysis. In patients who completed the study, the investigators found improvement in all primary outcomes at 12 months in the MTX group compared to the placebo group. Neither of the skin score measurements achieved a level of statistical significance of $p < 0.05$ using the original frequentist analysis. (Table 1).

Bayesian analysis. In patients who completed the study at 12 months, the probability of a beneficial effect compared to placebo on the MRSS is 90%, on the UCLA skin score 92%, and on the physician global assessment 98% (Table 2). Comparisons are relative to the placebo group, so improvements are those over and above what were seen in the placebo group. The posterior probability density curve for the mean difference in modified Rodnan skin score between MTX and placebo patients is shaded to illustrate the probability that MTX results in better mean skin score than placebo (Figure 1).

When values were imputed for patients who did not complete the study, the probability of a beneficial effect on the MRSS is 94%, on the UCLA skin score 96%, and on the physician global assessment 88% (Table 2). From the multivariate analysis, there is a 96% probability of benefit in at least 2 of 3 primary outcomes.

Noninformative versus informative prior. The informative prior based on van den Hoogen, *et al*[15] is centered at less of a treatment effect than the one we observed. The result is that the posterior for the analysis with the informative prior is shifted slightly towards no treatment effect; the posterior mean and credible interval (CrI) were –3.4 and –7.3 to 0.4, respectively. This is a substantial reduction in the uncertainty around the treatment effect than we obtained with the uninformative prior, which had a 95% CrI of –11.8 to 1.3. The combination of a smaller estimate of the treatment effect and the narrower CrI resulted in a slight increase in

Table 1. Twelve-month outcomes among those who completed the study using a frequentist approach.

Outcome	Methotrexate, n = 35; Randomized	Placebo, n = 36; Randomized	p	Treatment Effect (95% CI)
Modified Rodnan skin score, range, 0–78	21.4 ± 2.8 (n = 27)	26.3 ± 2.1 (n = 24)	0.18	–4.9 (–11.9, 2.2)
UCLA skin score, range 0-30	8.8 ± 1.2 (n = 27)	11.0 ± 0.9 (n = 24)	0.15	–2.2 (–5.2, 0.8)
MD global assessment, 10 cm VAS	4.2 ± 0.5 (n = 25)	5.5 ± 0.4 (n = 29)	0.04	–1.3 (–2.7, –0.1)

Values are mean ± standard error of the mean. n is the number with 12-month outcomes. UCLA: University of California Los Angeles; MD: physician; VAS: visual analog scale.

Table 2. Probability and odds of a beneficial treatment effect using 12-month scores.

Outcome	Excluding Missing Data Probability of Beneficial Effect Compared to Placebo, % (odds)	Imputing Missing Data Probability of Beneficial Effect Compared to Placebo, % (odds)	Imputing Missing Data Estimated Treatment Effect (95% credible interval)
Modified Rodnan skin score	90 (9:1)	94 (15.7:1)	–5.3 (–11.8, 1.3)
UCLA skin score	92 (11.5:1)	96 (24:1)	–2.5 (–5.1, 0.2)
MD global assessment	98 (49:1)	88 (7.3:1)	–0.77 (–2.00, 0.46)

UCLA: University of California Los Angeles; MD: physician.

Table 3. The probability that combinations of outcomes are beneficial.

Outcomes	Probability, %
Exactly 0 outcomes are positive	1
Exactly 1 outcome is positive	3
Exactly 2 outcomes are positive	11
Exactly 3 outcomes are positive	85

the probability of a treatment benefit, from 94% to 96% (Figure 2).

DISCUSSION

This study illustrates how Bayesian analysis can convey a more clinically relevant interpretation of trial data. In the setting of uncommon disease, it allows for inferences to be made with the data at hand. Unlike the frequentist paradigm, it informs clinicians directly on the probability of beneficial treatment effects (for one or more outcomes) for use in clinical practice.

Our reanalysis refutes the belief that MTX is ineffective in SSc[17]. Our study results indicate that treatment with MTX has favorable odds of beneficial treatment effects on skin score and physician's global assessment of overall disease activity compared to no treatment (placebo). In the setting of a disease with significant morbidity and mortality[13,28], no curative treatment options[29], and an intervention that is safe, inexpensive and readily available, many clinicians would be willing to accept a > 50% probability (greater than 1:1 odds) of a beneficial treatment effect; in our analysis, MTX conferred greater than 9:1 odds of benefit. Further, any one measure of the severity of SSc may give an incomplete picture. Within the Bayesian multivariate model, it was possible to compute the probability that MTX produced a benefit in 1, 2, and 3 of the measures. This meant that we are able to make such clear statements as there is an 85% probability that improvement occurs in all 3 outcomes.

The Bayesian paradigm does not create positive results from a negative trial. If a treatment effect has no effect, then on average the posterior probability of treatment being better than placebo in a Bayesian analysis will be 50%. That is, a truly useless treatment (not helpful and not harmful) would result in 50% of the shaded area in Figure 1 lying above zero and 50% lying below zero. Thus the Bayesian approach should not be viewed as a statistical alternative for investigators trying to demonstrate a "treatment effect." Rather, use of the Bayesian approach should be a conscious choice for analyzing one's data. In the MTX in SSc trial, the raw and aggregate data indicate benefit in all outcome measures. The frequentist analysis provides an incomplete summary of what the data are telling us. Further, incorrect interpretation of the p value and 95% confidence interval has previously led to a misguided clinical interpretation of the study results. Indeed, many people would interpret the frequentist 95% confidence interval, which contains 0 as an indication, that there is insufficient evident to reject the hypothesis of no treatment effect. The study interpretation would stop there. No additional inferences could be made. The advantage of the Bayesian approach is that we can not

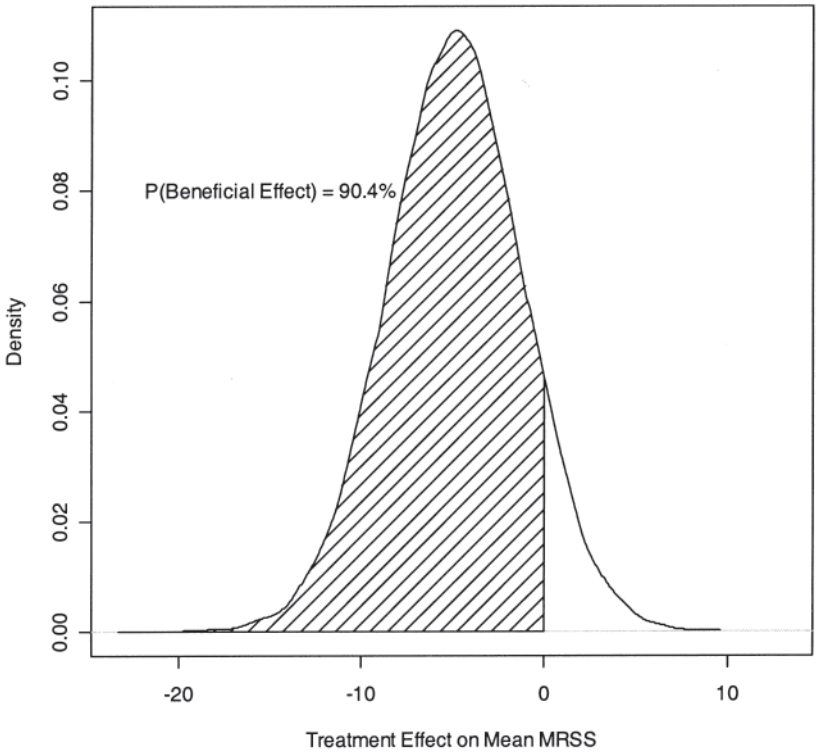

Figure 1. Probability density curve of mean difference in modified Rodnan skin score (MRSS) between MTX and placebo patients using non-imputed data. Negative scores favor MTX. This graph illustrates the probability that MTX results in better mean skin score than placebo.

only state that there is 95% probability of the treatment effect lying in a certain interval (the 95% credible interval) but we are also able to state the probability that the treatment effect lies in any interval we choose. In this case, we can compute the probability of a beneficial treatment effect given these data. Thus the Bayesian approach allows for more efficient use of the data at hand. Although it is true that the 95% credible interval contains 0, there is still a high probability of a beneficial treatment effect. In the case of MTX for SSc, the fairly low probability of harm may justify its use.

The Bayesian analysis allows for more clinically relevant interpretations of the data to be presented to consumers of the medical literature. Traditionally, trials report a "treatment effect" in terms of a statistically significant difference in mean scores between control and intervention group. Clinicians often struggle with how they can apply the results of a trial to their practice. The Bayesian framework informs clinicians about the probability of a beneficial treatment effect for their patients based on the data. Second, the presentation of a probability density curve allows clinicians to exercise clinical judgement about what treatment benefit is needed to offset the risks of treatment. A clinician may believe there is no utility in values around zero that are clinically indistinguishable from zero. A clinician may believe that a 3-point reduction in skin score is required to be clinically meaningful. It is a simple matter to use the output of the Bayesian model to compute the area under the curve to the left of –3 in the probability density curve in Figure 1, which corresponds to the posterior probability that the treatment reduces the MRSS by 3 or more points; for the non-imputed data, this probability is 70% (2.3:1 odds of > 3 MRSS points improvement with MTX). Third, Bayesian analysis allows for determination of the probability of simultaneous benefit in multiple outcomes. In the posterior distributions from the multivariate model, the Bayesian analyst has access to the probabilities of any combination of outcomes showing a benefit. For example, a skeptical clinician may require improvement in both measures of skin involvement (which we find to have a posterior probability of 92%) or require improvement in at least one skin score as well as the global measure of health (which we find to have a posterior probability of 88.6%). We are able to report the probability of a beneficial effect in multiple outcomes without being limited by the frequentist issues of adjustment for multiple comparisons.

The presentation of results in terms of probabilities, odds, and a probability density curve differs from the concept of "number needed to treat" (NNT). The widely used NNT indicates the benefit of active treatment over control, and is expressed as the reciprocal of the absolute risk reduction[30,31]. In this particular case, since the outcomes are con-

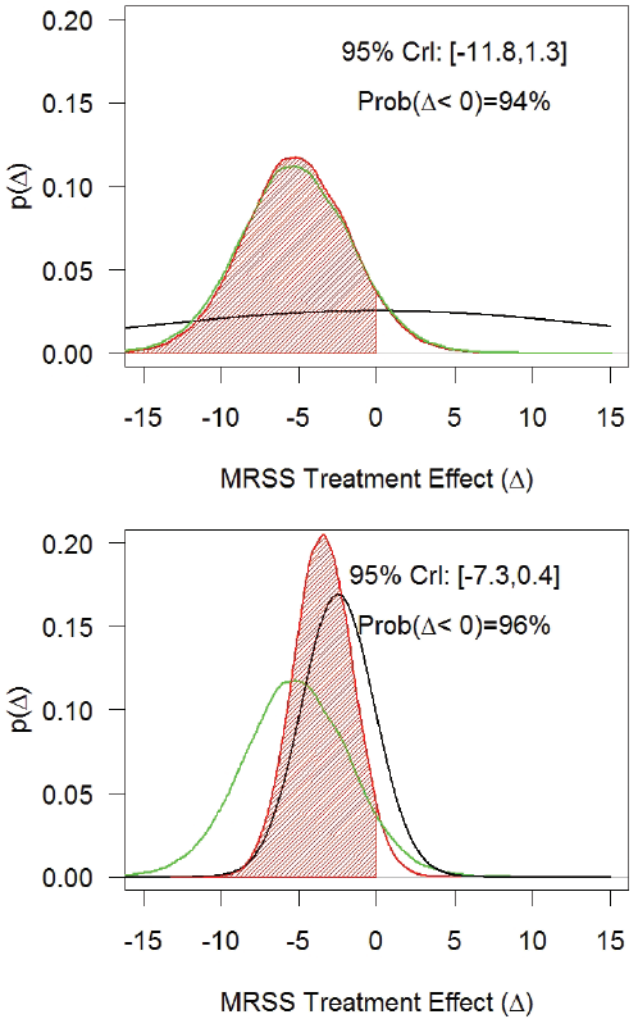

Figure 2. Triplots for probability of a beneficial effect of MTX on modified Rodnan skin score (MRSS) using an uninformative (top) and informative (bottom) prior probability distribution. 95% CrI: 95% credible interval. Black line: prior probability distribution; green line: likelihood; red line: posterior probability distribution. Red shading indicates area of treatment benefit.

tinuous measurements, they would first have to be dichotomized so that a difference in proportions and NNT could be computed. NNT has been used for summarizing trial results and medical decision-making[32]. Unfortunately, its application in the clinic for individual patients has been limited by the need for complex calculations[33] or reliance on nomograms[32]. Although NNT and Bayesian reporting of data can facilitate the interpretation of study data in terms of individual patients, the NNT expresses the number of patients who need to be treated to prevent one additional event; Bayesian analysis can be used to express the probability of a beneficial effect in any patient.

Finally, Bayesian analysis allows for the inclusion of previous knowledge into study analysis. We contrast the inferences that are made by using various priors: one noninformative prior and an informative prior. The use of an informative prior allowed the inclusion of the estimated treatment effect from a previous trial with the current trial. The combination resulted in an increase in the probability of a beneficial treatment effect with less uncertainty.

Limitations to the wide-scale use of Bayesian analysis do exist[34], but are not insurmountable. First, computational complexity of the analysis has previously limited its use; however, the availability of faster computers and relatively easy to use software has made Bayesian statistical analysis more accessible to clinical researchers. Indeed, the run-time per analysis in our study ranged from 6–10 seconds (for the complete case analysis) to 1–2 minutes (for the analyses using multiple imputation). Second, critics of Bayesian analysis are concerned that interpretation of results is subjective, as the interpretation is influenced by prior beliefs. We argue that clinicians implicitly interpret the results of a trial within the context of their clinical experience and beliefs, no matter how the trial is analyzed[35,36]. A strength of Bayesian analysis is that it allows formal and explicit incorporation of a spectrum of prior beliefs (skeptical versus optimistic views of the treatment effect) in the analysis. Here, to be most objective, we used priors that were consistent with no prior belief about the effectiveness of MTX; in practice this meant that every possible size of treatment effect was a priori considered equally likely. There has been concern that the subjective nature of the Bayesian approach as incorporated in the prior belief may lead to a preferential selection of the new treatment. This may be problematic when clinicians are confronted with findings from small clinical trials. In the setting of a small trial with weak evidence, a Bayesian analysis would indicate in a transparent manner that is evident to the reader, that the study interpretation is heavily influenced by subjective belief. This transparency in study interpretation is an important advantage.

Our study illustrates how Bayesian statistics can convey a more flexible and clinically relevant interpretation of clinical trial data of an uncommon disease. Using the data at hand, we are able to report the probability of a beneficial treatment effect and probability of a beneficial effect in multiple outcomes. As an example of the use of Bayesian inference in the setting of uncommon disease trials, we have demonstrated that MTX has a high probability of a beneficial effect on skin score and physician global assessment of overall disease activity.

REFERENCES

1. Cox SR, Walker JG, Coleman M, et al. Isolated pulmonary hypertension in scleroderma. Intern Med J 2005;35:28-33.
2. Fisher RA. Statistical methods and statistical inference. Edinburgh: Oliver and Boyd; 1956.
3. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
4. Alderson P, Chalmers I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ 2003;326:475.
5. Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Statistical

power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. J Clin Oncol 2007;25:3482-7.
6. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995-1004.
7. Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. Eur Heart J 1985;6:556-85.
8. Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2008 Oct 21. [E pub ahead of print]
9. Burton PR. Helping doctors to draw appropriate inferences from the analysis of medical studies. Stat Med 1994;13:1699-713.
10. Salsburg D. The religion of statistics as practiced in medical journals. Am Stat 1985;39:220-3.
11. Lawrence RC, Helmick CG, Arnett FC, et al. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. Arthritis Rheum 1998;41:778-99.
12. Mayes MD, Lacey JV Jr, Beebe-Dimmer J, et al. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. Arthritis Rheum 2003;48:2246-55.
13. Johnson SR, Gladman DD, Schentag CT, Lee P. Quality of life and functional status in systemic sclerosis compared to other rheumatic diseases. J Rheumatol 2006;33:1117-22.
14. Bode BY, Yocum DE, Mann CC, Ko M, Boyer J. Methotrexate (MTX) in scleroderma: experience in ten patients [abstract]. Arthritis Rheum 1990;33 Suppl:S66.
15. van den Hoogen FH, Boerbooms AM, Swaak AJ, Rasker JJ, van Lier HJ, van de Putte LB. Comparison of methotrexate with placebo in the treatment of systemic sclerosis: a 24 week randomized double-blind trial, followed by a 24 week observational trial. Br J Rheumatol 1996;35:364-72.
16. Pope JE, Bellamy N, Seibold JR, et al. A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma. Arthritis Rheum 2001;44:1351-8.
17. Pope JE, Ouimet JM, Krizova A. Scleroderma treatment differs between experts and general rheumatologists. Arthritis Rheum 2006;55:138-45.
18. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. J Epidemiol Community Health 1998;52:318-23.
19. Spiegelhalter DJ, Abrams KR, Myles JP. An overview of the Bayesian approach. In: Spiegelhalter DJ, Abrams KR, Myles JP, editors. Bayesian approaches to clinical trials and health care evaluation. Chichester: John Wiley & Sons; 2004:49-137.
20. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum 1980;23:581-90.
21. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202-5.
22. Furst DE, Clements PJ, Hillis S, et al. Immunosuppression with chlorambucil, versus placebo, for scleroderma. Results of a three-year, parallel, randomized, double-blind study. Arthritis Rheum 1989;32:584-93.
23. Kahaleh MB, Sultany GL, Smith EA, Huffstutter JE, Loadholt CB, LeRoy EC. A modified scleroderma skin scoring method. Clin Exp Rheumatol 1986;4:367-9.
24. Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a Bayesian analysis of a clinical study. J Clin Epidemiol 2005;58:261-8.
25. Martin AD, Quinn KM. MCMCpack: Markov chain Monte Carlo (MCMC) package. R package version 0.8-1 2007. [Internet. Accessed Sept 30 2008.] Available from: http://mcmcpack.wustl.edu
26. Rossi P, McCulloch R. bayesm: Bayesian inference for marketing/micro-econometrics. R package version 2.0-9 2007. [Internet. Accessed Sept 30 2008.] Available from: http://faculty.chicagogsb.edu/peter.rossi/research/bsm.html
27. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. J Statistical Software 2005;12:1-16.
28. Ioannidis JP, Vlachoyiannopoulos PG, Haidich AB, et al. Mortality in systemic sclerosis: an international meta-analysis of individual patient data. Am J Med 2005;118:2-10.
29. Charles C, Clements P, Furst DE. Systemic sclerosis: hypothesis-driven treatment strategies. Lancet 2006;367:1683-91.
30. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452-4.
31. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988;318:1728-33.
32. Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. BMJ 1996;312:426-9.
33. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ 1999;319:1492-5.
34. Moye LA. Bayesians in clinical trials. Asleep at the switch. Stat Med 2008;27:469-82; discussion 483-9.
35. Knipschild P. Changing belief in iridology after an empirical study. BMJ 1989;299:491-2.
36. Rovers MM, van der Wilt GJ, van der Bij S, Straatman H, Ingels K, Zielhuis GA. Bayes' theorem: a negative example of a RCT on grommets in children with glue ear. Eur J Epidemiol 2005;20:23-8.