Published in final edited form as:
*Ann Rheum Dis*. 2009 August ; 68(8): 1255–1259. doi:10.1136/ard.2008.092460.

# The Sensitivity to Change for Lower Disease Activity is Greater than for Higher Disease Activity in Rheumatoid Arthritis Trials

**Bin Zhang, PhD.**, **Michael Lavalley, PhD.**, and **David T. Felson, MD, MPH.**
Boston University Clinical Epidemiology Unit, Boston, MA

## Abstract

**Objective—**To test whether RA trials treatment efficacy vs. control is better detected for patients with lower tender or swollen joint counts than for higher counts.

**Methods—**Using data from 6 large multicenter trials (N = 2002) and an intent to treat approach at 6 months, we created within each trial two subtrials, the lower disease activity group(defined by TJC <= overall median) and the higher disease activity group (TJC above median). The same approach was used for SJC. We tested for active treatment - control differences using several RA trial outcome measures: ACR20, EULAR response, ACRHybrid. We compared sample sizes needed for higher TJC and SJC RA trials vs. lower TJC and SJC trials, and explored consistency of results across trials and explanations for results by examining active treatment and control responses.

**Results—**We found that subtrials of subjects with lower TJC had much higher sensitivity to change than those of subjects with higher TJC across all trials and outcome measures. A trial with lower TJC patients would require a smaller sample size than ones with higher TJC patients. Results were not consistent for SJC subgroups. We found 3 reasons for sensitivity to change of lower TJC: 1) Compared to higher TJC, those with lower TJC showed greater response to active treatment. Subjects with higher TJC on control treatment had 2) greater % improvement and 3) more variable responses than those in the lower TJC group.

**Conclusions—**In RA trials, patients with lower disease activity within the range of current trial eligibility are more likely to show treatment efficacy than patients with higher disease activity. Lowering thresholds especially for TJC in trials may make it easier to detect treatment effects in RA.

In trials of treatments for RA, it is customary for high levels of tender and swollen joint count to be required for patient eligibility. Most trials require at least 10 tender joints and 8 swollen joints, and the mean number of active joints in patients enrolled is far higher than this. The disease activity of patients with RA in the US and in some Western European countries has fallen[1] and, at least as measured by DAS scores, is now quite low on average[2]. Thus, disease activity has fallen to a point where most RA patients are likely not eligible for most trials testing treatment. Increasingly, trial patients originated from South America, Eastern Europe, and elsewhere, where evidence suggests[2] that the disease on average is still very active. If new treatments are tested outside of the US and of Western Europe because of the absence of high disease activity patients, this may suggest that new treatments might not be generalizable to

Corresponding author: David T Felson, Boston University, Address: 715 Albany Street A203 Boston, United States, Zip/Postal Code: MA 02118, Phone: 001-617-6385180, Fax: 001-617-6385239, dfelson@bu.edu.
Address reprint requests to Dr. Bin Zhang at Suite 200, 650 Albany St, Boston U. School of Medicine, Boston, MA 02118 (binzhang@bu.edu)

those with lower disease activity. Further, because trials provide the central evidence on treatment efficacy, our failure to test treatments in patients like those we see in practice in the U.S. and Western Europe raises concerns about whether this evidence is relevant to our patients.

One major assumption behind the high threshold for eligibility in trials is that effective treatments suppress very active disease and persons with more active disease would be likely to improve more than persons with less disease activity, and therefore they would be better subjects for trials. However, the relative response of persons with higher versus those with lower disease activity has not been examined in trials of patients with RA. At any rate, the efficacy of treatment in a trial is tested not by the response of patients to active treatment, but by the comparative response of patients versus placebo or control treatment. The efficacy of treatment in a trial is a function both of the response to active treatment and the response to placebo. So, even if subjects with lower disease activity are less likely to respond to active treatment, they may also be less likely to show a response to placebo, making their comparative response profile similar to, or even better than, a subject with greater disease activity.

With these considerations in mind and using a dataset of multiple large randomized trials in RA, we tested whether treatment efficacy in RA trials would be better detected if trials were conducted in patients with higher disease activity versus if they were conducted of patients with lower disease activity.

## Methods

The data sets used were made available through the American College of Rheumatology's effort to redefine and reevaluate response criteria for RA, an effort which has led to the promulgation of the hybrid ACR measure [3, 4] There were eleven multicenter randomized trials used in that effort, most of them of TNFα inhibitors with some of conventional disease modifying antirheumatic drugs (DMARD'S).

To define lower and higher disease activity, we identified the overall median swollen joint count across all trial, we then created two subtrials for each trial, one limited to those subjects with higher disease activity (top 50%) and the other limited to those subjects with lower disease activity (bottom 50%). We alternatively defined higher disease activity as patients with higher tender joint counts (at or above the overall median for all trials) and patients with lower tender joint counts (below the overall median for all trials). Because some of these trials did not have sufficient numbers of subjects in both the higher and lower disease activity subtrials to make a valid comparison of treatments, we restricted our analyses of higher versus lower disease activity to trials with at least 200 subjects, sufficient numbers in our view for the evaluation of higher versus lower disease activity subgroups. In our 11 data set, 6 trials met these criteria. Agreements with the industry sources that provided these data prohibit us from identifying specific trials.

To test sensitivity to change, we used three different candidate measures of response in rheumatoid arthritis trials that are either widely used or which permitted us to readily test the relative sensitivity for detection of treatment effects: the ACR 20, EULAR good response (present if DAS28<=3.2 and Change in DAS or DAS28>1.2 from baseline) and the ACRHybrid, The first two are dichotomous measures (yes/no for each patient) and the third is a continuous measure of improvement taking on values between 0 (no improvement) and 1 (100% improvement in core set measures). To evaluate the sensitivity to change of higher vs. lower disease activity, we compared active vs. control in each of the subtrials using an intent to treat approach at 6 months of follow-up with a Student t-test (for ACRHybrid) or chisquare test (ACR20 and EULAR good response) to compare treatments. We translated these findings to estimates of sample sizes needed to undertake clinical trials conducted in the lower and higher disease activity subtrials[5-7]. For the dichotomous outcome measures (ACR20 and EULAR good response), to estimate sample size, we used a hypothetical response rate of 30% in the control group and the observed odds ratio of treatment response from the clinical trials for treatment vs. control. For the ACRHybrid we used a difference of 30% (delta=0.3) between treatment and control and the observed standard deviation from the clinical trials. All sample size estimates are based on tests with two sided alpha=0.05 and 80% power. We tried additional assumptions for control response rate and treatment control differences, and results were similar.

To better understand our results, we investigated for each of the higher and lower disease activity subtrials response in the active treatment and placebo groups and the variation around that response. For example, if the higher disease activity groups needed a larger sample size, we anticipated two explanations: the difference in response between active and placebo treatment would be greater in that disease activity subgroup and/or the variability in response would be greater in the placebo group and less in the active treatment group. Thus, the difference and the variation would determine the sample size requirement. The effect size, which equals the difference between treatments divided by the pooled standard deviation of the differences, is an excellent tool to measure the effect of a treatment vs. control. We used it to compare the efficacy of treatment vs. control in each subtrial within each trial to get a sense of the consistency of our findings across trials.

## Results

Of the 6 multicenter trials, 2 were trials of TNFα inhibitor vs. placebo, two were trials of conventional second line drugs vs. placebo and two compared combinations of drugs vs. single drugs. All were reported as positive for either the active treatment or the combinations of treatments being tested.

All of these trials excluded patients with joint counts below a certain threshold, a threshold which varied by trial. The minimum tender joint count permitted ranged from 5 to 12 and the minimum swollen joint count ranged from 3 to 10 and agreed with the trial protocols. We normalized joint counts to 68 tender and 66 swollen (some trials had potential joint counts less than this) Using the 68 joint tender joint count and a 66 joint swollen count, median counts for patients in trials were 27 tender joints and 18 swollen joints Subtrials were created with subjects below and above these normalized medians in each trial.

When we divided subjects in the trials according to the median tender joint count (see table 1), we found that, regardless of how we assessed outcome, trials of those with tender joint counts below the median required smaller sample sizes than those with counts at or above the median. For example, if we used the ACRHybrid as our outcome measure, restricting trials to those with tender joint counts below the median, a trial with 80% power to detect a treatment difference of 0.30 on the ACRHybrid score would require a sample size per treatment arm of 30, whereas for patients with higher tender joint counts, the sample size requirement would be 53 per treatment arm. (see table 2) We found similar results when we used dichotomous outcomes---EULAR good response and ACR20. For all the outcomes we tested, trials had greater sensitivity to change and required smaller sample sizes if patients with lower tender joint counts were studied.

We explored why lower tender joint count increased sensitivity to change by examining treatment and control responses in each of the trials. For the ACRHybrid, a continuous measure of response, we found that the control group with higher tender joint counts showed a greater median response than the control with lower tender joint counts for 5/6 trials. (see figure 1) and for active treatment groups, the median treatment response was actually higher for those with lower tender joint counts in 5/6 trials.

We calculated effect size of treatment in each subtrial, and in 4 out of 6 trials the effect sizes of the lower were greater than those of the higher tender joint count subgroups and in one of the other two, effect size was almost identical (Table 3).

For ACR20 and EULAR good response, response rates for active treatment were higher in the lower than the higher tender joint count group in most trials (see figures 2a and 2b). This created a larger difference in response rate comparing subjects on active treatment to control subjects.

For swollen joint count, the findings were, by no means, as clear (see table 3). The sensitivity to change did not differ greatly in those with swollen joint counts above vs. below the median. For EULAR good response, sample size requirements were slightly higher for the subgroup with lower swollen joint count, whereas for ACR20, sample size needs were slightly greater for the subgroup with higher counts. There were no consistent trends in trials of lower vs. higher swollen joint counts (see figures 3; 4a and 4b). In some subgroups, patients with higher swollen joint counts had a greater active treatment response rate whereas, in others, those with lower swollen joint counts did better on active treatment. Control group responses also varied across trials.

## Discussion

In an analysis of multiple rheumatoid arthritis trials, we found that if trials were restricted to patients with lower tender joint counts, treatments could be more easily detected as statistically significantly superior to placebo or to a comparator than if trials were restricted to those with higher tender joint counts. We did not find the same trend with swollen joint count. The reasons behind this result include a higher response rate to the control treatment (usually placebo) in persons with higher tender counts and greater variability of response for the placebo treated patients in those with higher tender counts. Even the response of the active treatment group tended to be greater among persons with lower tender joint counts, regardless of whether this response was assessed as percentage change in disease activity (e.g. ACRHybrid or ACR20) or as absolute change in activity (EULAR good).

While patients with higher disease activity might be expected to have a greater absolute response in their disease activity when treated with active treatments, this was not consistently true. Further, we suggest that those with lower disease activity may be better candidates for trials. The reason is not intuitively obvious and does not necessarily derive from their better or worse response to active treatment. Rather, it derives in part from the placebo (or control) group's response in these situations. For ACRHybrid and ACR20, placebo groups tended to show greater response if patients started with higher disease activity. That makes it harder in higher disease activity patients to distinguish between active treatment and placebo. On the other hand among persons with lower disease activity, placebo responses were worse (4/6 trials for ACR20) and treatment responses were often better.

For a patient to achieve a EULAR good response, they must experience both an absolute decrease in DAS score and reach a low DAS score threshold. EULAR good response rates were uniformly higher in those with lower tender joint counts than those with higher counts (see figure 2b) regardless of whether they were in the active treatment or placebo groups (figure 2b). Because patients with lower joint counts start closer to that threshold than those with higher counts, the high EULAR good response rates in those with baseline lower joint counts may reflects this. The net decrease in counts was greater in active treatment than placebo, and the consequence was that, for EULAR good response also, sample size requirements were less for those with lower tender joint counts than for those with higher counts.

One reason underlying our results may be regression to the mean. Patients with higher disease activity may not consistently have such higher levels of disease activity, but rather may enter a trial when they are at an apex in terms of their disease activity level. Their natural course may be to regress to their own mean and have lower disease activity. This will occur whether they are treated with placebo or active treatment and will make it hard to detect the added effect of active treatment on their improvement. The variability of placebo response suggests that some patients with higher disease activity are experiencing regression to the mean while others are not.

Why do our results differ for swollen and tender joints? Swollen joints are more stable and vary less with improvement [8]. That may make them less susceptible to regression to the mean.

We also explored whether the subgroup analyses were valid, whether differences observed between treatment and control in the two subgroups were due purely to sample size differences (we chose the overall median count but not the trial specific median), and whether the initial randomizations were preserved to make the statistical comparison valid. Even though we used an overall median count for all trials, there were almost equal numbers of subjects in each trial in the TJC higher vs lower subgroups and in the SJC higher vs. lower subgroups Further, the treatment-control assignment ratios almost matched the original assignment ratios. showing little evidence of violating randomization.

In an era when rheumatoid arthritis in the US is becoming milder (perhaps due to better treatments), our results may have important implications for both practice and the design and conduct of RA trials. Our results suggest that recruiting more patients with milder disease (defined as tender joint counts in the lower end of current inclusion criteria) will make it easier to detect treatment effects.

The major limitation in our ability to address the relation of low joint counts to treatment response definitively is that trials that we analyzed were done with restrictive inclusion criteria, prohibiting us from testing the sensitivity to change of treatment versus placebo in patients with substantially lower disease activity than was present in these trials. Further, even though we took large trials and divided them in half, because of relatively small numbers, we are unable to subdivide this activity level into smaller increments to extrapolate our results to even lower disease activity levels. such as tender joint count less than six.

In conclusion, rheumatoid arthritis trials would likely be more efficient in detecting the efficacy of treatments if they included patients with lower, not higher disease activity, especially lower tender joint counts. This intuitively paradoxical finding is based on the high variability in placebo response in those with higher disease activity and to a higher placebo response rate in this group. However, even response to active treatment is as robust in those with lower joint counts as in those with higher counts.

## Acknowledgments

Supported by NIH AR47785 and by a grant from the American College of Rheumatology to reevaluate response criteria in RA.

## References

1. Pincus T, Sokka T, Chung CP, Cawkwell G. Declines in number of tender and swollen joints in patients with rheumatoid arthritis seen in standard care in 1985 versus 2001: Possible considerations for revision of inclusion criteria for clinical trials. Ann Rheum Dis 2006;65:878–83. [PubMed: 16339290]

2. Sokka T, Pincus T, Toloza S, Baecklund E, Horslev-Petersen K, Maillefert JF, et al. Variation in the proportion of patients in standard care in 58 clinical sites in 21 countries who meet DAS28>5.1 criteria

for high disease activity; data from the QUESTRA database. Arthritis Rheum 2007;56(suppl):9. S178. [PubMed: 17195185]

3. Felson DT. A proposed Revision to the ACR20: The Hybrid Measure of American College of Rheumatology Response. Arthritis Rheum 2007;57:193–202. [PubMed: 17330293]
4. Felson DT. A proposed Revision to the ACR20: The Hybrid Measure of the American College of Rheumatology Response. Arthritis Rheum 2007;57:193–202. [PubMed: 17330293]
5. Fleiss JL, Tytun A, Ury SHK. A simple approximation for caculation sample sizes for comparing independent proportions. Biometrics 1980;36:343–6.
6. O'Brien, RG.; Muller, KE. Applied Analysis of Variance in Behavioral Science. New York: Marcel Dekker; 1983. p. 297-344.
7. Dixon, WJ.; Massey, FJ. Introduction to Statistical Analysis. Vol. 4th. New York: McGraw-Hill; 1983. McGraw-Hill
8. Anderson JJ, Felson DT, Meenan RF, Williams HJ. Which traditional measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1989;32:1093–9. [PubMed: 2505779]

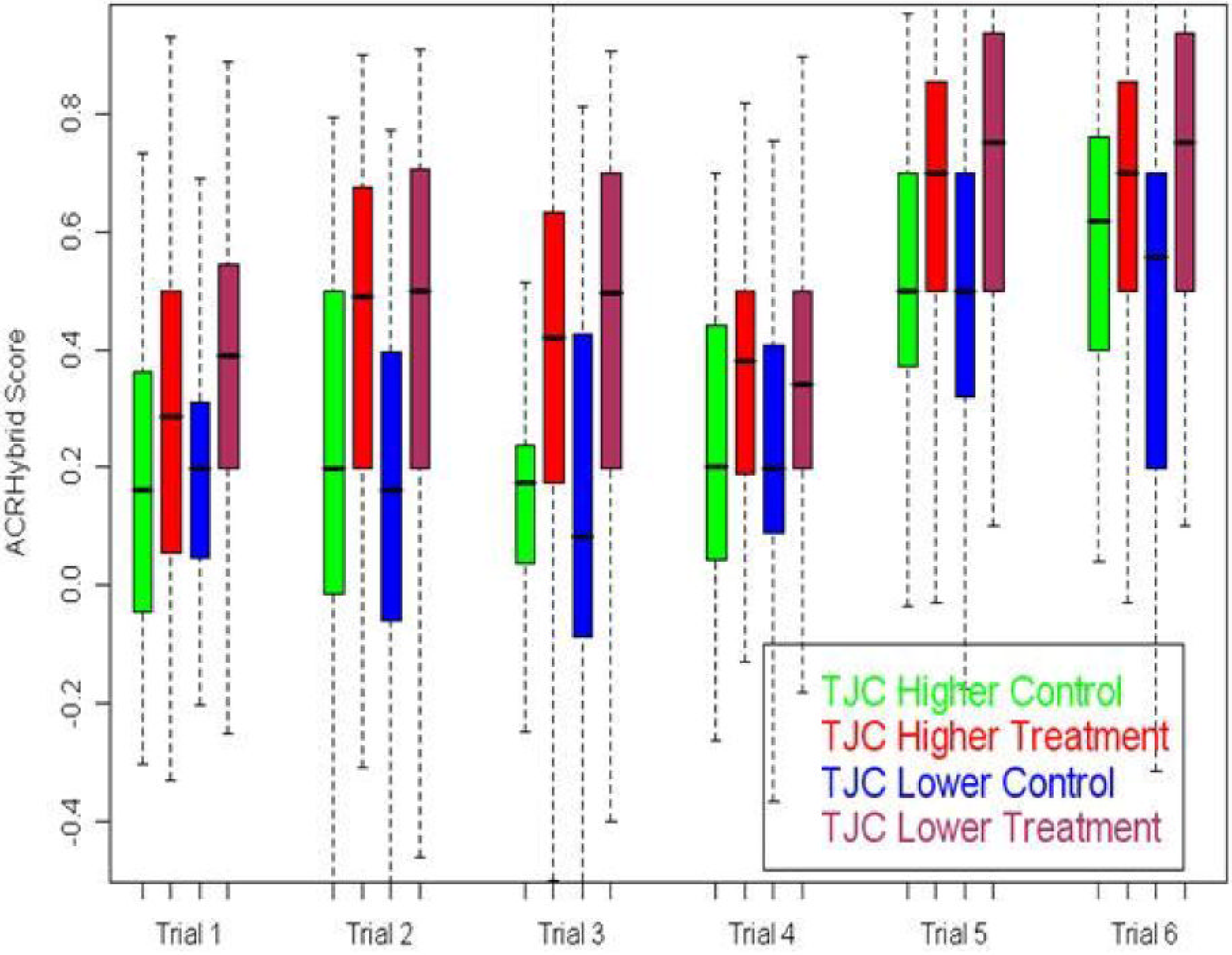

**Fig 1.**

Boxplot of ACRHybrid Score for Higher and Lower TJC subgroups by trial The lower edge of the box shows the 25th percentile of the data, the upper edge shows the 75th percentile, the line within each box shows the median (50th percentile) and the whiskers extend down to the minimum and up to the maximum. Five out the six trials the treatment-control median score differences are higher in the TJC lower subgroup than that of the TJC higher subgroup.

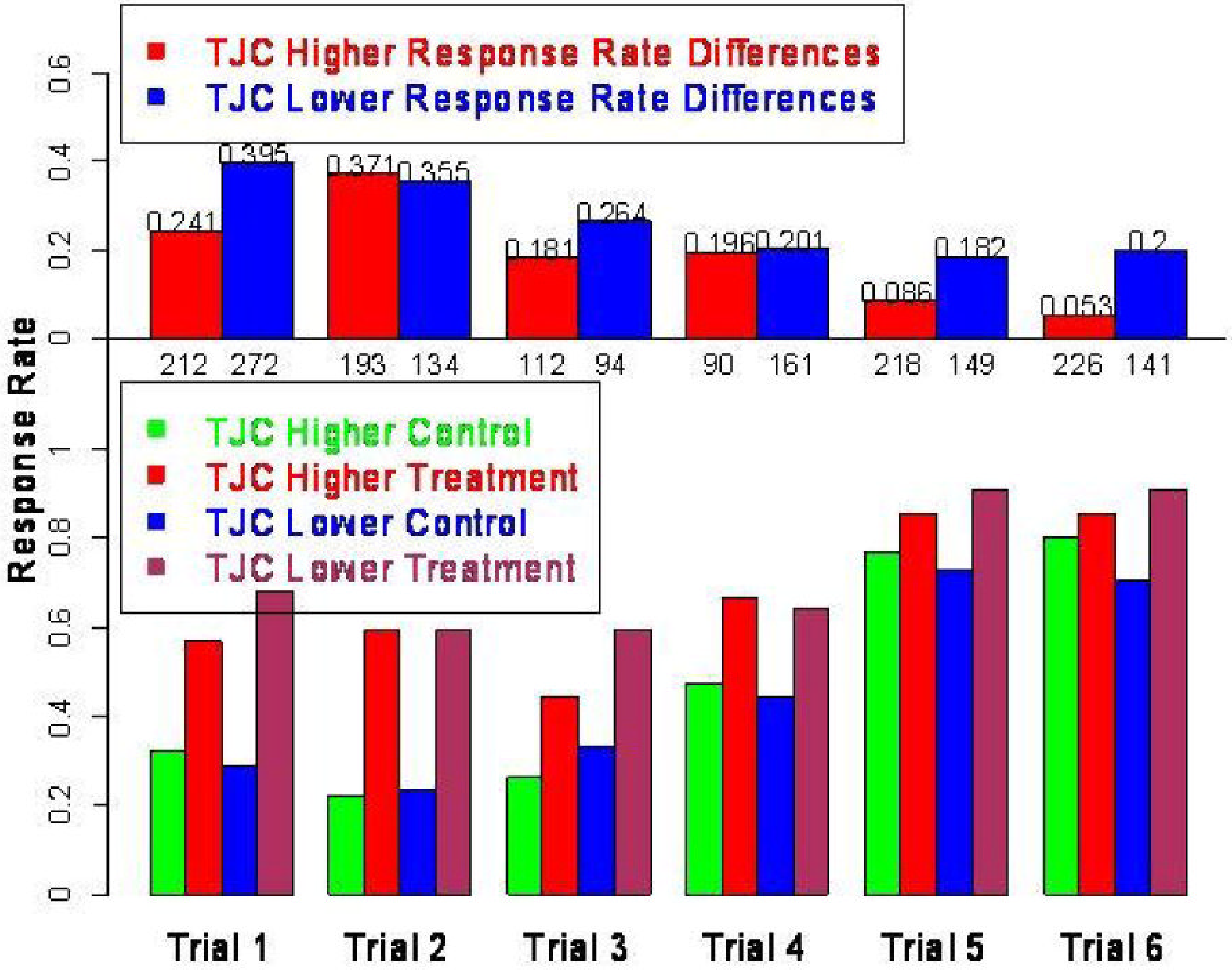
TJC Higher Response Rate Differences
TJC Lower Response Rate Differences
0.6
0.4
0.2
0
0.241
0.395
0.371
0.355
0.181
0.264
0.196
0.201
0.086
0.182
0.053
0.2
212
272
193
134
112
94
90
161
218
149
226
141
Response Rate
TJC Higher Control
TJC Higher Treatment
TJC Lower Control
TJC Lower Treatment
1
0.8
0.6
0.4
0.2
0
Trial 1
Trial 2
Trial 3
Trial 4
Trial 5
Trial 6

Figure 2a

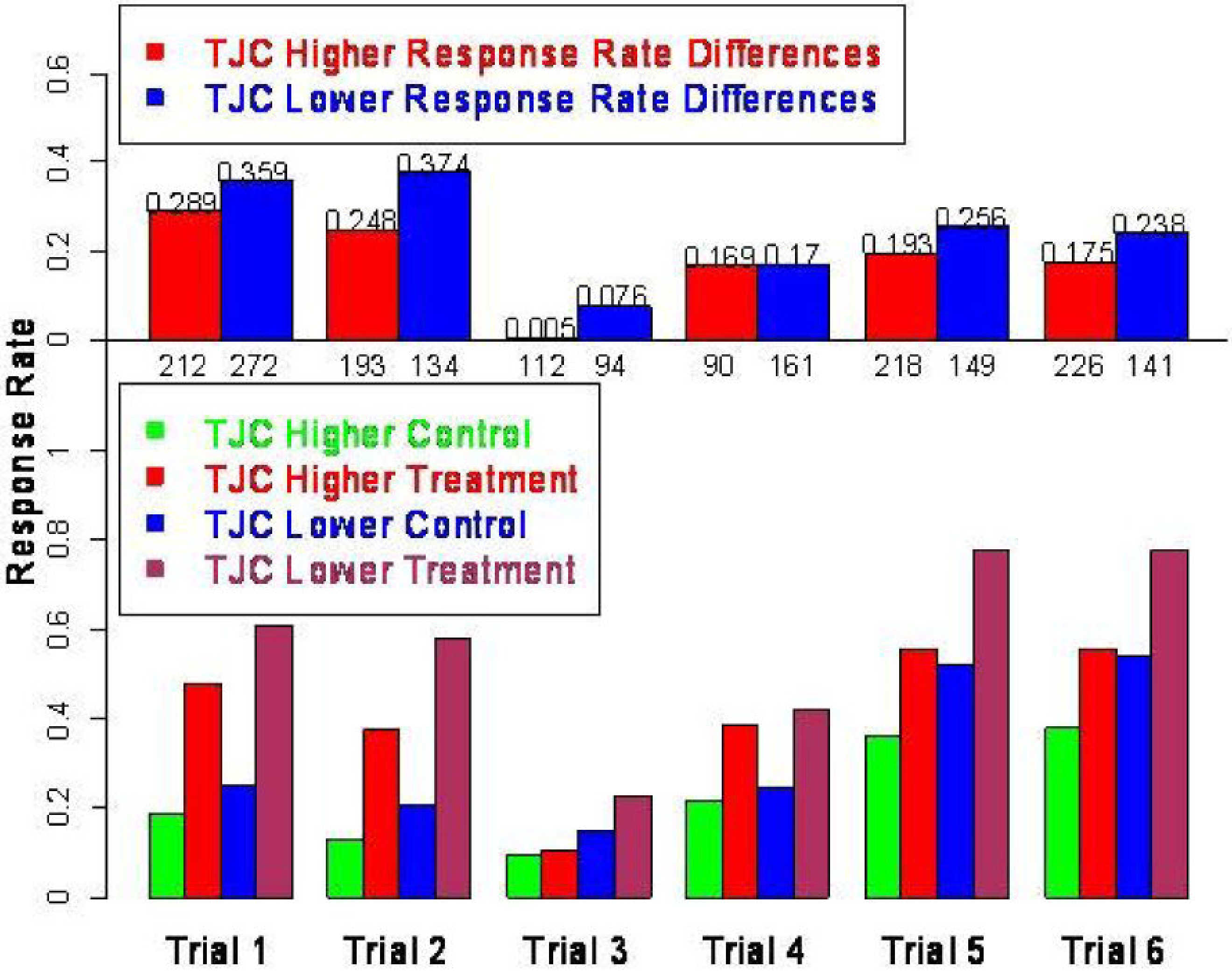

Figure 2b

**Fig 2.**

Fig 2a: Bottom histogram: ACR20 response rates in subgroups with higher and lower TJC by trial Treatment-placebo difference is greater in TJC lower disease activity subgroup because of higher treatment response rates (4/6 trials) and/or lower control response rates (4/6 trials). In five out the six trials (see histogram above blue bars) the treatment-control response rate differences are higher in the TJC lower subgroup than those in the TJC higher subgroup. The numbers beneath the upper 0 line are the numbers of subjects in each of the subtrials by the overall median TJC cut.

Fig 2b: Bottom histogram: EULAR good responses in lower TJC treatment and placebo subgroups by trial. Treatment-placebo difference is greater in TJC lower disease activity subgroup because of higher treatment response rates (5/6 trials) and/or lower control response rates (4/6 trials). As shown in top histogram, in all six trials the treatment-control response rate differences are higher in the TJC lower subgroup than those in the TJC higher subgroup.

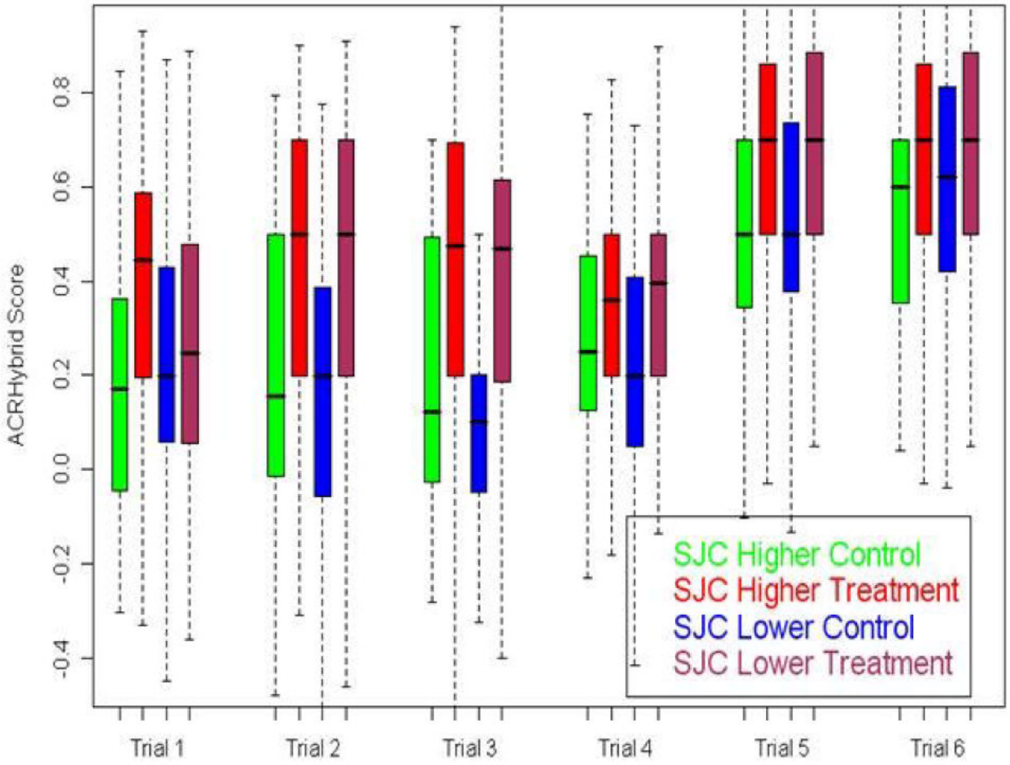

**Fig 3.**

Boxplot of ACRHybrid Score Comparisions in Higher and Lower SJC groups by trial: No consistent findings. The lower edge of the box shows the 25th percentile of the data, the upper edge shows the 75th percentile, the line within each box shows the median (50th percentile) and the whiskers extend down to the minimum and up to the maximum.

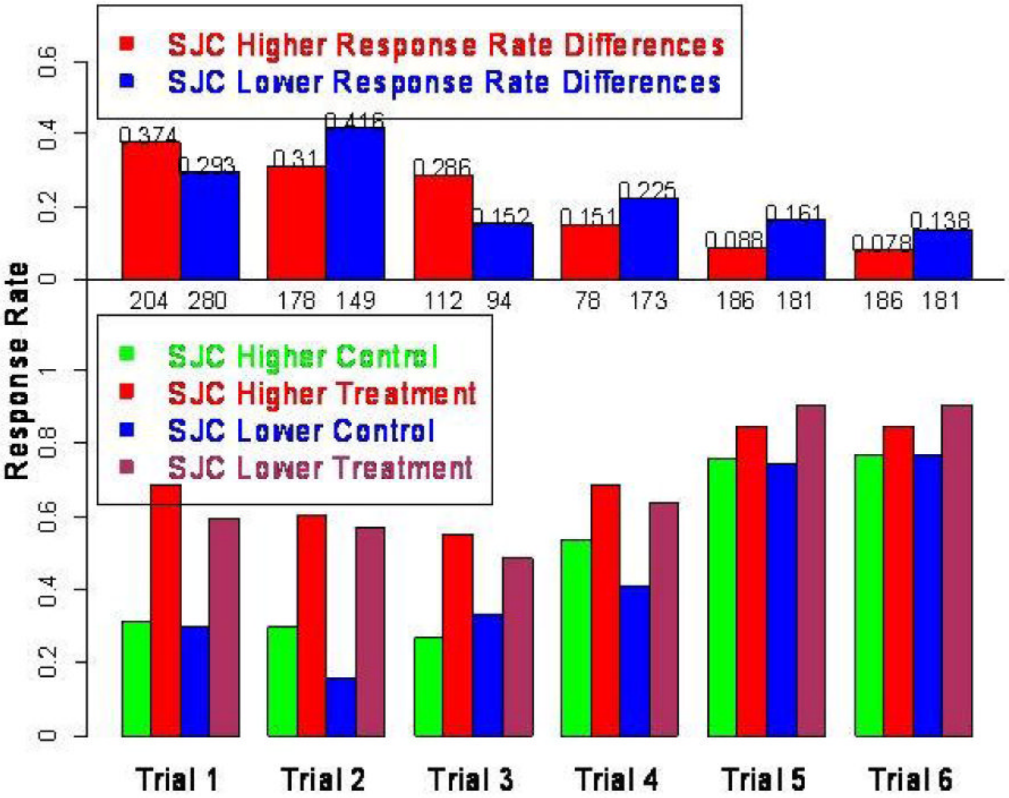
🟥 SJC Higher Response Rate Differences
🟦 SJC Lower Response Rate Differences
0.374
0.293
0.31
0.416
0.286
0.152
0.151
0.225
0.088
0.161
0.078
0.138
204 280
178 149
112 94
78 173
186 181
186 181
🟩 SJC Higher Control
🟥 SJC Higher Treatment
🟦 SJC Lower Control
🟪 SJC Lower Treatment
Response Rate
0
0.2
0.4
0.6
0.8
1
Trial 1
Trial 2
Trial 3
Trial 4
Trial 5
Trial 6

Figure 4a

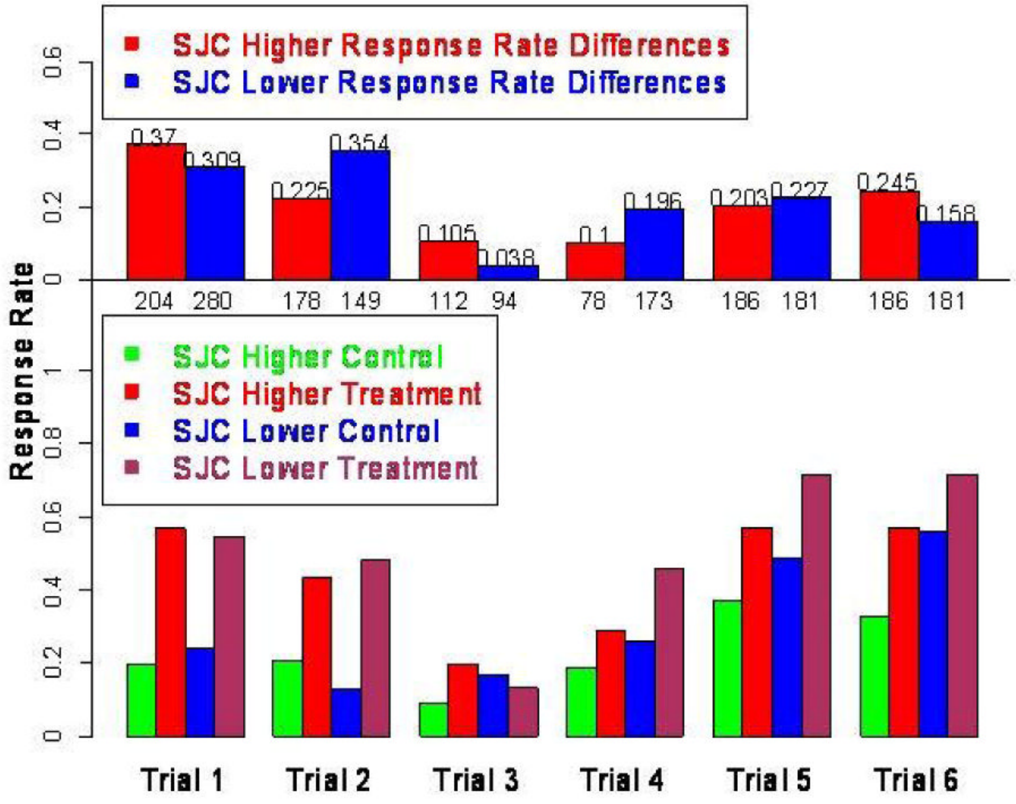

Figure 4b

**Fig 4.**

Fig 4a: ACR20 response rates in subgroups with higher and lower SJC by trial Active treatment response higher in higher SJC subgroup (4/6) trials but control treatment response also higher in this same subgroup (4/6 trials). The numbers just beneath the upper 0 line are the numbers of subjects in each of the subtrials by the overall median SJC median cutoff.

As shown in top histogram: there were no consistent findings with respect to active treatment-control differences.

Fig 4b: EULAR good responses in lower SJC treatment and placebo subgroups by trial. Active treatment response higher in higher SJC subgroup (4/6) trials but control treatment response also higher in this same subgroup (4/6 trials). As shown in top histogram, there were no consistent findings with respect to active treatment-control differences

**Table 1**

Median Tender Joint Count (TJC) and Swollen Joint Count (SJC) in the higher and lower disease group

| | **Median** |
|---|---|
| TJC Lower | 18 |
| TJC Higher | 38 |
| SJC lower | 13 |
| SJC Higher | 26 |

## Table 2

Sample size for each treatment group in a trial needed to have an 80% likelihood of showing statisitical significance (P<.05) using different response criteria with a fixed rate (30%) of response in control group. **Lower vs. Higher Tender Joint Counts Subgroups and Lower vs. Higher Swollen Joint Counts Subgroups**

| | **Numbers of Subjects Needed** | | | |
|---|---|---|---|---|
| | **Lower Tender Joint Count** | **Higher Tender Joint Count** | **Lower Swollen Joint count** | **Higher Swollen Joint Count** |
| ACRHybrid | 30 | 53 | 42 | 35 |
| ACR20 | 53 | 125 | 76 | 86 |
| EULAR Good Response | 62 | 94 | 84 | 70 |

## Table 3

Effect Size of treatment vs. control comparisons in higher vs. lower counts in each of the six trials: In 4/6 trials, lower tender joint count had higher effect size; in one trial (3) these were nearly identical and in one trial (4), higher tender joint count had a greater effect size. For swollen counts, there were no consistent trends.

| | TJC Lower | TJC Higher | SJC Lower | SJC Higher |
|---|---|---|---|---|
| Trial 1 | 0.91 | 0.70 | 0.75 | 0.93 |
| Trial 2 | 0.96 | 0.73 | 0.90 | 0.78 |
| Trial 3 | 0.43 | 0.44 | 0.20 | 0.67 |
| Trial 4 | 0.44 | 0.57 | 0.52 | 0.41 |
| Trial 5 | 0.77 | 0.53 | 0.64 | 0.63 |
| Trial 6 | 0.73 | 0.33 | 0.49 | 0.48 |