Antibiotic retreatment of Lyme disease in patients with persistent symptoms: A biostatistical review of randomized, placebo-controlled, clinical trials

Allison K. DeLong a,⁎, Barbara Blossom b, Elizabeth Maloney c, Steven E. Phillips d

a Center for Statistical Sciences, Department of Biostatistics, Brown University, Providence, RI, USA
b Department of Statistics, Colorado State University, Fort Collins, CO, USA
c Partnership for Healing and Health, Ltd., Wyoming, MN, USA
d Greenwich Hospital, Greenwich, CT, USA

A B S T R A C T

Introduction: Lyme disease (Lyme borreliosis) is caused by the tick-borne spirochete Borrelia burgdorferi. Long-term persistent illness following antibiotic treatment is not uncommon, particularly when treatment is delayed. Current treatment guidelines for persistent disease primarily rely on findings from four randomized, controlled trials (RCTs), strongly advising against retreatment.

Methods: We performed a biostatistical review of all published RCTs evaluating antibiotic retreatment, focusing on trial design, analysis and conclusions.

Results: Four RCTs met the inclusion criteria; all examined the efficacy of intravenous ceftriaxone versus placebo at approximately 3 or 6 months. Design assumptions for the primary outcomes in the two Klempner trials and two outcomes in the Krupp trial were unrealistic and the trials were likely underpowered to detect clinically meaningful treatment effects. The Klempner trials were analyzed using inefficient statistical methods. The Krupp RCT was well-designed and analyzed for fatigue, finding statistically significant and clinically meaningful improvement. Fallon corroborated this finding. Fallon also found improvement in cognitive functioning, a primary outcome, at 12 weeks which was not sustained at 24 weeks; improvements in physical functioning and pain were demonstrated at week 24 as an interaction effect between treatment and baseline symptom severity with the drug effect increasing with higher baseline impairment.

Discussion: This biostatistical review reveals that retreatment can be beneficial. Primary outcomes originally reported as statistically insignificant were likely underpowered. The positive treatment effects of ceftriaxone are encouraging and consistent with continued infection, a hypothesis deserving additional study. Additional studies of persistent infection and antibiotic treatment are warranted.

© 2012 Published by Elsevier Inc.

Keywords: Lyme disease, Neuroborreliosis, Randomized controlled trial, Statistical power, Minimum clinically important difference (MCID), Treatment guidelines

1. Introduction

Reporting bias in clinical trials, particularly with respect to publishing bias toward significant findings [1,2] and interpretive “spin” to overemphasize a possible benefit while de-emphasizing non-significant findings [3] is receiving increased attention within the statistical and medical communities. A variation on interpretive bias deserves concern as well, namely the interpretation of statistically insignificant findings from small, underpowered, or poorly executed clinical trials as evidence of treatment inefficacy. Such trials may lead to the premature and erroneous conclusion that the treatment is ineffective, constituting a type II error. Concerns about such
errors may arise when disagreement and uncertainty exists in the medical community, as is the case with Lyme disease (Lyme borreliosis).

Lyme disease, caused by the tick-borne spirochete *Borrelia burgdorferi* sensu lato, is classified as an emerging infectious disease by the U.S. National Institute of Allergy and Infectious Diseases (NIAID) due to the relatively recent discovery of its causal agent (1982) [4] and its rapidly increasing incidence over the last two decades in the U.S. [5] and much of Europe [6]. The infection is multi-systemic, resulting in diverse physical and neuro-psychiatric symptoms and manifestations and causing mild to severe disease [7–13]. Although many patients respond to antibiotic treatment regimens of 2 to 4 week duration [9], it is well recognized that long-term persistent illness can occur following a 30-day course of treatment, particularly when treatment is delayed [7,9,14,15]. Multiple randomized trials found significant morbidity in their study populations, similar to that of multiple sclerosis or congestive heart failure. Although the trials employed different entrance criteria, none required this degree of physical disability as a condition of enrollment [16,17].

The management of patients with ongoing debilitating symptoms following antibiotic treatment for Lyme disease has generated debate within the medical community. The primary questions concern whether or not infection persists after standard antibiotic treatment and whether additional antibiotic treatment is of benefit [18,19]. Until a sensitive laboratory test for active infection is clinically available, clinical trials evaluating retreatment in persistently symptomatic Lyme disease patients provide the cornerstone of treatment guideline recommendations. Most guidelines for the diagnosis and management of Lyme disease [20–23] direct clinicians to limit the duration of antibiotic treatment, even in cases where ongoing symptoms compatible with a *B. burgdorferi* infection are present. These publications base their recommendations on a similar interpretation of the four randomized, blinded, placebo-controlled antibiotic retreatment trials funded by the U.S. National Institutes of Health (NIH) for patients with ongoing symptoms following standard Lyme disease treatment [16,17,24].

For this reason, a rigorous, independent evaluation of the findings from these trials is needed. The present study is a biostatistical review of the four NIH-funded clinical trials. By focusing on the trial design and analyses of primary and secondary outcomes in each trial, the review demonstrates weaknesses which limit the ability to draw strong conclusions regarding retreatment. This review will likely be of broad interest to medical practitioners, researchers, medical ethicists, and treatment guideline developers in Europe and North America.

## 2. Methods

The four NIH-funded Lyme disease retreatment trials were initially selected for evaluation in January 2009 through a review of current Lyme disease treatment guidelines, which identify these trials as the only published RCTs relevant to the question of retreatment [21,22]. To ensure that other relevant RCTs to date were not missed, a Cochrane Library search of the published literature was conducted on September 10, 2010, setting the limits of study type to “clinical trial” and requiring the use of “Lyme” or “*Borrelia*” in the title, abstract or in the manuscript’s keywords. Additional studies were sought by searching ClinicalTrials.gov, a registry of both federally and privately funded clinical trials. The title and abstract of each selected publication were read by two authors (AKD and BB) and coded as a clinical trial and if it was a clinical trial evaluating retreatment of Lyme disease patients with persistent symptoms despite receipt of a standard course of antibiotics. The full text of all articles evaluating retreatment was read by all authors and eligibility was determined by consensus. All primary and secondary outcomes were tabulated for each clinical trial, including, where possible, the treatment effect and 95% confidence interval (CI) overall and by trial arm.

A review was conducted of each trial’s design, execution, statistical analysis and conclusions. For trial design, attention was paid to the enrolled patient population, the definitions and measurements of primary and secondary outcomes, and the definition of clinically meaningful changes in those outcomes which determine power of the sample sizes to detect clinically meaningful treatment effects. For trial execution, patient dropout, masking of study medication, and interim analyses were considered. We evaluated the appropriateness of the statistical method chosen to estimate the treatment effect and the handling of patient dropouts. Since our objective is to place the findings from these trials within the current framework of Lyme disease as of 2012, the present review is also informed by research conducted after the retreatment trials were designed, executed, and/or published. Three important statistical concepts are used throughout the review: statistical power, interim analysis and stopping rules, and non-inferiority trials.

### 2.1. Statistical power

When designing a clinical trial, the sample size can only be calculated after researchers determine an appropriate and plausible design treatment effect $\delta$, which is a hypothetical value of the effect of the treatment under investigation. In addition to selecting $\delta$, trial design also requires an acceptable probability of declaring treatment effectiveness if $\delta$ is true (i.e. power, typically 80–90%). For a fixed power, a smaller $\delta$ would necessitate a study design with a larger sample size, and vice versa. Ideally $\delta$ should correspond to the minimum clinically important difference (MCID) for the disease and outcome measure studied. If the true underlying treatment effect is greater than the MCID, yet less than the design treatment effect $\delta$, then the study is underpowered with an insufficient sample size, and thus inadequately designed to meet its stated goals, and the power may be far less than the nominal value set in the trial design. Such studies are likely to conclude an insignificant result although a true, clinically relevant treatment effect exists. Although MCID values are context-specific and difficult to ascertain, reasonable estimates are identified based on published knowledge of the disease studied or, when disease-specific data are not available, of studies of other similar diseases [25].

### 2.2. Interim analyses and stopping rules

Interim analyses are commonly used to gage the success of a clinical trial, by analyzing outcome data at pre-defined points during the study instead of waiting until all patients
have completed follow up. An interim analysis can trigger one of three possible actions: (1) conclude that the treatment is effective and stop the trial early, (2) continue the trial until the next interim ‘look’, and (3) stop the trial early for “futility”. If action (1) is triggered, trial findings can be published and disseminated quickly and effective treatments can be provided to patients sooner. Action (3) implies that at the study terminus, the authors will most likely fail to reject the null hypothesis that the outcomes in the two arms are the same. This action is often triggered when the designed sample size is too small to detect the true treatment effect, which may occur as a result of underestimation of patient variability in the study design, use of an unrealistically large design δ (greater than the MCID), or because the treatment is, indeed, ineffective. Many have argued that conducting underpowered trials is unethical; therefore stopping such trials is desirable. Stopping a trial for statistical insignificance or futility does not necessarily indicate that treatment is ineffective and it would be incorrect to conclude that this was the case.

2.3. Non-superiority trials

To examine whether a treatment is ineffective, statistical tests using non-superiority hypotheses are required. In such trials, the null hypothesis is that the treatments differ, with rejection of the null hypothesis indicating that the treatment effects in the two arms are similar, i.e. the difference lies within a certain small but acceptable window. None of the Lyme disease retreatment trials was designed as a non-superiority trial. However, if 95% confidence intervals (CIs) on the treatment effects exclude and are below the MCID, then the trial has essentially shown the treatment to be ineffective.

3. Results

The literature search found 105 clinical trials using the word “Lyme” or “Borrelia” in the title, abstract or keyword (Fig. 1). Of these, 100 papers were eliminated from consideration for the following reasons: did not assess antibiotic efficacy (49); evaluated antibiotic prophylaxis after a tick bite (4); evaluated first-line antibiotic treatment of early or late Lyme disease, including a study evaluating longer-term treatment which enrolled patients with and without a history of prior treatment [26] (39); evaluated treatment of coinfection of Lyme disease and babesiosis (1); and involved treatment of relapsing fever (7). The full text of the remaining 5 publications was read. One clinical trial was excluded because it did not present an intention-to-treat analysis of primary outcomes due to an excessive dropout rate in the placebo arm [27]. Klempner et al. [16] presented two primary and one secondary outcome from two trials which enrolled patients from two different populations. Kaplan et al. [28] presented an analysis of several additional secondary outcomes from the Klempner trials. Henceforth, these trials are collectively referred to as the Klempner trials. The publications by Krupp et al. [24] and Fallon et al. [17] present primary and secondary outcomes from two additional clinical trials. As a result, the primary outcomes from four clinical trials were presented in three publications.

Participants in all four trials had a confirmed history of Lyme disease for which they received at least one standard course of antibiotic therapy, and had persistent symptoms thought to be consistent with Lyme disease beginning at or within 6 months of disease onset, with symptoms persisting at least 4 months following the cessation of therapy. The studies enrolled different subpopulations of patients with persistent symptoms, but all examined intravenous (IV) ceftriaxone for a minimum of 4 weeks and evaluated various primary and secondary treatment effects at approximately 3 and/or 6 months as described (Table 1).

3.1. Klempner et al. trials [16]

3.1.1. Trial summary

Klempner et al. conducted two multicenter trials; the designs differed only in that one enrolled IgG-seropositive and the other IgG-seronegative patients. Patients received either IV placebo followed by 2 months oral placebo or 1 month of IV ceftriaxone followed by 2 months of oral doxycycline. Clinical inclusion criteria were broad, including any of: widespread musculoskeletal pain, cognitive impairment, radicular pain, and paresthesias that interfered with functioning per patient self-report. The primary outcomes were changes in SF-36 summary scores, which are commonly used subjective measures of health-related quality of life (HRQoL). The SF-36 physical component summary (PCS) and mental component summary (MCS) scores represent numeric composites of eight subcategories, scaled such that the means and standard deviations (SD) for the general U.S. population are 50 and 10 respectively with lower scores representing poorer health.

Klempner classified patients as “improved,” “worsened,” or “the same” based on changes in their summary scores from baseline to the 180-day evaluation. Positive and negative cutoffs for classification were set at 6.5 units for the SF-36 PCS and 7.9 for the SF-36 MCS; these values represent twice the standard error of measurement (SEM). A chi-square test of proportions was used to evaluate the treatment effect, which was taken to be whether the proportion of patients in each class differed by treatment arm. For sample size estimation, the researchers set the design treatment effects to be 25% and 35% for the difference in percent improved in the seropositive and seronegative trials, respectively. The calculated sample sizes were 194 participants in the seropositive trial and 66 in the seronegative trial. Interim analyses using O’Brien-Fleming boundaries were performed after 107 of 260 (41%) planned participants in both trials combined completed follow-up, and the trials were stopped for futility. No statistically significant treatment benefit was reported for either trial. The authors concluded that the trial regimen did not result in a significant treatment effect and also stated that other antibiotic regimens were unlikely to result in a different finding.

3.1.2. Trial critique

3.1.2.1. Design. In order to evaluate the Klempner trials’ design in light of all available evidence, a literature search for studies that determined MCIDs for SF-36 summary scores (PCS and MCS) was conducted. We were unable to find any studies evaluating MCIDs for the SF-36 in patients with Lyme disease. Studies evaluating MCIDs in patients with other chronic illnesses causing a level of disability similar to that of
the Klempner subjects were published after the Klempner trials were conducted. These studies identified clinically meaningful changes on the SF-36 summary scores to be in the range of 2 to 5 points (Table 2) [29–33]. Changes of this magnitude align with the SF-36 developers’ recommendations [34,35] and with studies identifying 1 SEM or 0.5 standard deviation in baseline scores as appropriate statistical benchmarks of clinical relevance for health-related quality of life (HRQoL) measures including the SF-36 [36–38].

The Klempner trial design assumed that a δ of absolute 25% or 35% difference between arms in the percent “improved” would correspond to a valid threshold for clinically relevant treatment effects. Since observed changes in SF-36 outcomes were not reported in the manuscript, we mapped Klempner’s δs to their corresponding δ∗ on the continuous SF-36 scale as follows. Let $s_i$ be the observed 6-month treatment effect in the PCS for participant $i$ in the placebo arm and let $t_j$ be the observed 6-month treatment effect for participant $j$ in the antibiotic arm and assume $s \sim N(\mu, \sigma)$ and $t \sim N(\mu + \delta^*, \sigma)$. We can use the observed percentages (quantiles) of patients classified as having “improved”, “stayed the same”, and “worsened” in the placebo arm to estimate $\mu$ and $\sigma$. The expected difference in the “% improved” on the PCS for pertinent values of $\delta^*$ can be estimated as $Pr(t > 6.5) - Pr(s > 6.5)$, using the estimates $\sigma$ and $\mu$. A similar calculation can be conducted for the MCS, with $\delta^*$ estimated as $Pr(t > 7.9) - Pr(s > 7.9)$.

A $\delta = 25\%$ corresponds to mean differences in SF-36 scores between the two arms of 6.7 and 9.1 points on the PCS and MCS, respectively, and a $\delta = 35\%$ corresponds to 9.3 and 12.8 points (Table 3). Thus the trials, as designed, called for treatment effects considerably larger than the 2 to 5 point MCIDs identified in other chronic illnesses, suggesting that the sample sizes were inadequate and the trials were very likely underpowered to detect the true underlying MCIDs. The importance of this finding becomes clear when one considers the scale of the SF-36 instrument. For example in the antibiotic arm of the seronegative trial, adding the estimated treatment effect of 12.8 points on SF-36 MCS to the baseline mean MCS score of 46.7 points would require the average participant to achieve a score essentially one standard deviation (SD) above...
the mean score for the U.S. general population. As such, the
treatment effects of 2 to 5 points correspond to expected
differences in the percent “improved” of 7 to 18% on the PCS
(Table 3). These differences are within the 95% confidence
intervals for both trials, indicating the trials did not show the
treatment to be ineffective.

Our estimated standard deviations, \( \sigma \), defined in an earlier
paragraph, were 10.1 for the PCS and 14.2 for the MCS. Although
one may speculate that this large variability is due to a “placebo
effect”, this terminology should be used cautiously; other
possible explanations for the large standard deviation could be
regression to the mean or higher variability in chronically ill
populations. Ware et al. [34], who used the same cutoffs as
Klepner et al. in their evaluation of the SF-36 in chronically ill
patients, called attention to the fact that outcome variations
in their categorical analysis were substantially larger than
expected.

Using these values for \( \sigma \), the sample sizes required per arm to
have 80% power to detect average treatment effects from 2 to 5
points using a t-test and a two-sided alpha of 0.05 are 66 and 128
patients for an assumed MCID of 5 points on the PCS and MCS,
respectively and 179 and 354 patients for an assumed MCID of 3
points. Detecting a treatment effect of 2 points on the PCS on
would require about 400 patients per arm and detecting
differences of 2 points on the MCS would require about 800
participants per arm. An analysis incorporating repeated,
longitudinal measurements per participant would require a
smaller number of participants.

3.1.2.2. Analysis. The trials’ use of a chi-square test on
categorized, continuous data collected at four study time
points is not an efficient use of data, yet its use can provide
unbiased results in certain circumstances. If the study design is
simple, if missing outcomes are non-informative (i.e. missing
completely at random), and if randomization is successful in
balancing patient arms by pertinent characteristics, the
chi-square test can be used. The Klepner studies did not meet
these criteria. The trials were multicenter, the
observed baseline outcomes differed by treatment arm, and
authors provided insufficient information about patient
dropout to determine whether or not missing outcomes
were uninformative; therefore, in this setting a chi-square
test is not recommended and its use may have produced
biased results. Lastly, combining the data from the two trials is
not valid without accounting for the stratification of patients in
the analysis, which was not done using a chi-square test.

The analysis of the secondary outcomes is compromised
in the same manner as the primary outcomes. In addition, the
results for the secondary outcomes were not presented by
trial. Instead, the seronegative and seropositive patients were
combined, disregarding the fact that these were designed as two
distinct trials (ClinicalTrials.gov Identifier No. NCT00001101 and
NCT0000938).

3.1.2.3. Interpretation. We found that the Klepner trials were
designed using excessive treatment effect sizes (much greater
than minimum clinically meaningful) making it likely that the
trials were underpowered to detect MCIDs. Although the trials
had adequate power to detect the large changes in SF-36 scores
used for outcome categorization (equal to 2*SEM for the general
population), it is important to note that while 2*SEM is an
appropriate benchmark to ensure statistical significance for an
individual, it is not necessarily the appropriate cutoff to identify
clinically meaningful and statistically significant differences at
a group level. Thus it is not surprising that the MCIDs for diseases
causing similar levels of disability, discussed above, and to
which \( \delta \) should correspond, are less than 2*SEM.

The authors noted in the discussion that their antibiotic
regimen did not lead to improved outcomes, and, given the
“in vitro and in vivo activity of both of these antibiotics against
B. burgdorferi” and experience with other chronic infections, they
concluded that it was unlikely that other antibiotic
regimes would be useful. These trials do not support such a
broad statement. Based on our findings, we conclude that
the Klepner trials are uninformative with regards to the
potential benefits of antibiotic retreatment utilizing 1 month of
ceftriaxone followed by 2 months of doxycycline (or any other
regimen) in patients with persistent symptoms of Lyme
disease.

3.2. Krupp et al. STOP-LD trial [24]

3.2.1. Trial summary

Krupp et al. enrolled 55 patients with a history of Lyme
disease and ongoing symptoms of severe fatigue validated by a
Fatigue Severity Scale (FSS-11) score ≥4.0. Patients were
randomized to receive 4 weeks of IV ceftriaxone versus placebo,
and three primary outcomes were evaluated: fatigue measured
by the FSS-11, mental speed using an alphabet arithmetic (A-A)
test, and clearance of outer surface protein A (OspA) from the
CSF. At 6 months follow-up, the authors found a significant
treatment effect on fatigue. Clinical improvement, defined as a
decrease ≥ 0.7 FSS-11 points, was seen in 15.8% on placebo
versus 64% on ceftriaxone (p < 0.01). Treatment effects on the
other two primary outcomes were not statistically significant.
The authors noted six significant adverse events. Four were
serious; three of these involved IV sepsis in placebo subjects
while the fourth was anaphylaxis in a ceftriaxone subject. The
other two events were minor allergic reactions in ceftriaxone
subjects.

Krupp et al. concluded that their findings did not support
antibiotic retreatment. The authors noted the positive effect
on fatigue but thought it may be due to unmasking of the
study medication. They also concluded that the beneficial
effect on fatigue was outweighed by the lack of effect on the
other primary endpoints and the high number of adverse
events.

3.2.2. Trial critique

3.2.2.1. Design. The trial was well-designed for the primary
endpoint of fatigue, with clearly defined inclusion criteria
and 80% power. However, it was inadequately designed with
regard to mental processing speed. The authors defined a
clinically meaningful change in the mental speed outcome as
a 25% improvement on the A-A test, and designed their study
with low (74%) power to detect a \( \delta = 25\% \) difference in the
percent improved between the arms. An earlier study by the
same author [39] found that patients with a history of Lyme
disease and continued fatigue or cognitive symptoms had an
overall deficit of less than 25% on 7 of the 8 measures.
Table 1

Available measures of treatment effects for each trial and outcome.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Measurement</th>
<th>Outcome</th>
<th>Primary or secondary outcome</th>
<th>Meas. time months</th>
<th>Effect or “Success” rate by arm</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klempner et al. Seronegative [16]: Antibiotic (n=25), Placebo (n=26)</td>
<td>SF-36 physical component summary (PCS)(a)</td>
<td>“Success” = change in PCS from baseline to 180 days &gt; 6.5</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 5/23 (22%)</td>
<td>Antibiotic 9/22 (41%)</td>
</tr>
<tr>
<td></td>
<td>SF-36 mental component summary (MCS)(a)</td>
<td>“Success” = change in MCS from baseline to 180 days &gt; 7.9</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 6/23 (26%)</td>
<td>Antibiotic 8/22 (36%)</td>
</tr>
<tr>
<td></td>
<td>Fibromyalgia impact questionnaire(b)</td>
<td>“Success” &gt; 25% improvement from baseline</td>
<td>Secondary</td>
<td>6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Medical outcome study symptom checklist(d)</td>
<td>Pain, cognitive functioning, performance of daily activities</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Neuropsychological tests(d)</td>
<td>Common battery</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Mood(d)</td>
<td>BDI and MMPI-2</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Krump et al. Seropositive [16]: Antibiotic (n=39), Placebo (n=39)</td>
<td>SF-36 physical component summary (PCS)(a)</td>
<td>“Success” = change in PCS from baseline to 180 days &gt; 6.5</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 10/35 (29%)</td>
<td>Antibiotic 11/35 (31%)</td>
</tr>
<tr>
<td></td>
<td>SF-36 mental component summary (MCS)(a)</td>
<td>“Success” = change in MCS from baseline to 180 days &gt; 7.9</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 16/35 (46%)</td>
<td>Antibiotic 11/35 (31%)</td>
</tr>
<tr>
<td></td>
<td>Fibromyalgia impact questionnaire(b)</td>
<td>“Success” &gt; 25% improvement from baseline</td>
<td>Secondary</td>
<td>6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Medical outcome study symptom checklist(d)</td>
<td>Pain, cognitive functioning, performance of daily activities</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Neuropsychological tests(d)</td>
<td>Common battery</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Mood(d)</td>
<td>BDI and MMPI-2</td>
<td>Secondary</td>
<td>3 and 6</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Fallon et al. [24]: Antibiotic (n=28), Placebo (n=27)</td>
<td>Fatigue severity scale (FSS-11)(b)</td>
<td>“Success” = Improvement of &gt; 0.7 points from baseline</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 5/22 (23%)</td>
<td>Antibiotic 18/26 (69%)</td>
</tr>
<tr>
<td></td>
<td>Alphabet arithmetic test</td>
<td>“Success” = improvement &gt; 25% from baseline</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 2/22 (9%)</td>
<td>Antibiotic 2/26 (8%)</td>
</tr>
<tr>
<td></td>
<td>Osp A antigen to Borrelia Burgdorferi</td>
<td>“Success” = clearance of Osp A antigen from baseline</td>
<td>Primary</td>
<td>6</td>
<td>Placebo 4/4 (100%)</td>
<td>Antibiotic 3/4 (75%)</td>
</tr>
<tr>
<td>Fallon et al. [17]: Antibiotic (n=23), Placebo (n=14)</td>
<td>Multivariate outcome measured across 6 cognitive domains(a)</td>
<td>Standardized to represent z-scores</td>
<td>Primary (efficacy)</td>
<td>3</td>
<td>Placebo 0.16 (−0.6, 0.38)</td>
<td>Antibiotic 0.43 (0.27, 0.61)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Primary (durability)</td>
<td>6</td>
<td>Placebo 0.31 (0.09, 0.53)</td>
<td>Antibiotic 0.35 (0.18, 0.53)</td>
</tr>
<tr>
<td></td>
<td>Fatigue severity scale (FSS-11)(b)</td>
<td>Continuous measure, interaction with baseline score</td>
<td>Secondary</td>
<td>3</td>
<td>Placebo −0.2 (−1.3, −0.1)</td>
<td>Antibiotic −1.3 (−2.1, −0.1) p &lt; 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>Placebo −0.4 (−1.1, −0.7)</td>
<td>Antibiotic −1.1 (−1.8, 0.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fatigue (FSS-11)</td>
<td>Krupp et al. analysis</td>
<td>Secondary</td>
<td>25%</td>
<td>Placebo 67%</td>
<td>Antibiotic 67%</td>
</tr>
<tr>
<td>Trial</td>
<td>Measurement Outcome</td>
<td>Secondary</td>
<td>Effect or “Success” rate by arm</td>
<td>Treatment effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>-----------</td>
<td>---------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Pain (McGill) VAS²</td>
<td>Continuous measure, interaction with baseline score</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Total pain²</td>
<td>Continuous measure</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>SF-36 physical component summary (PCS)²</td>
<td>Continuous measure, interaction with baseline score.</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>SF-36 mental component summary (MCS)²</td>
<td>Continuous measure</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td># joints with pain on exam²</td>
<td>Continuous measure</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Depression (Beck)²</td>
<td>Continuous measure</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Anxiety (Zung)²</td>
<td>Continuous measure</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Global Psycho-Pathology (GSI SCL-90)²</td>
<td>Continuous measure, interaction with baseline score</td>
<td>Secondary</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

NS effect not given, reported as not statistically significant.

“–”Within-arm effects were not reported for each trial.

² Higher score implies better health.

² Higher score implies worse health.

² All secondary outcomes presented as the mean of participants with worse baseline scores (75th percentile), estimated using values from the manuscript. Some estimated confidence intervals for statistically significant effects cross zero because data were available only to 1 decimal place.

² Presented in Kaplan et al. [28].
Table 2

SF-36 summary score changes found to be clinically and statistically significant for chronic diseases of similar severity to Lyme disease.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Disease</th>
<th>Increase in PCS</th>
<th>Increase in MCS</th>
<th>Verification of clinical significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kosinski et al. [29]</td>
<td>Rheumatoid arthritis</td>
<td>4.4, 4.3, 3.0, 2.6, 2.3</td>
<td>4.7, 3.1, 2.2, 3.1, 2.3</td>
<td>1 level of improvement across five clinical RA measures*</td>
</tr>
<tr>
<td>Angst et al. [30]</td>
<td>Osteoarthritis</td>
<td>2</td>
<td>4</td>
<td>Improvement in global health self-assessments</td>
</tr>
<tr>
<td>Coteur et al. [31]</td>
<td>Cohn's disease</td>
<td>4.1</td>
<td>3.9</td>
<td>IBDQ improvement*</td>
</tr>
<tr>
<td>Regensteiner et al. [32]</td>
<td>Peripheral artery disease</td>
<td>2</td>
<td>5</td>
<td>Increased maximum treadmill walking distance</td>
</tr>
<tr>
<td>Okamoto et al. [33]</td>
<td>Asthma</td>
<td>≤ 5.0</td>
<td></td>
<td>Increased FVE†</td>
</tr>
</tbody>
</table>

Table 3

Estimated differences in the proportion of patients expected to be classified as “improved” using Klempner et al.’s categorization for various mean treatment effects consistent with published MCIDs. Klempner et al.’s results are provided and confirm clinically meaningful mean differences of 2 to 5 points fall within Klempner et al.’s 95% confidence intervals [16].

| PCS (physical component) | Expected treatment effect (δ*) | Difference in % improved (treatment vs. placebo) | | MCS (mental component) | Expected treatment effect (δ*) | Difference in % improved (treatment vs. placebo) |
|--------------------------|--------------------------------|-----------------------------------------------|--------------------------|--------------------------------|-----------------------------------------------|
| 2                        | 7%                            |                                              | 2                        | 5%                            |                                              |
| 3                        | 10%                           |                                              | 3                        | 8%                            |                                              |
| 4                        | 14%                           |                                              | 4                        | 10%                           |                                              |
| 5                        | 18%                           |                                              | 5                        | 13%                           |                                              |
| 6.7                      | 25%                           |                                              | 9.1                      | 25%                           |                                              |
| 9.3                      | 35%                           |                                              | 12.8                     | 35%                           |                                              |
| PCS-observed results (95% CI) | 3% (−19 to 24) |                                              | MCS-observed results (95% CI) | 3% (−19 to 24) |                                              |
| *Seropositive trial      | 3% (<19 to 24)                |                                              | *Seropositive trial      | −14% (<37 to 8%)              |                                              |
| *Seronegative trial      | 19% (<7 to 46)                |                                              | *Seronegative trial      | 10% (<17 to 37%)              |                                              |

* Actual results reported by Klempner et al.

3.2.2.2. Analysis. With regard to fatigue, the authors performed a careful sensitivity analysis of loss to follow-up, demonstrating that the finding on fatigue was robust to patient dropout. After adjustment for baseline measures of psychiatric disorder, depressive symptoms, pain and age, the treatment benefit on fatigue remained significant.

Krupp et al. suggested that the finding of improved fatigue may have been biased due to unmasking of the study medication. This suggestion was based on their observation that the proportion of participants correctly guessing treatment assignment at 1 and 6 months was significantly higher in the antibiotic arm (p<0.05). This observation alone, however, is not indicative of unmasking. Consider an example in which patients were randomly assigned to a treatment or placebo group, and then guessed with equal probability of 0.8 in both arms that they were receiving treatment. In such a case, 80% of patients on treatment and only 20% on placebo would be expected to correctly guess their treatment assignment, yet masking was not corrupted. Instead of comparing the proportion in each arm that correctly guessed assignments, Krupp et al. should have compared the proportions that guessed they were on active therapy. In the STOP-LD trial, this proportion did not differ by arm at 1 month (57% placebo, 71% antibiotic, p = 0.37, Fisher exact test) or at 6 months (68% and 69%, p = 1.0).

Therefore, there is no evidence demonstrating that masking was compromised.

3.2.2.3. Interpretation. The benefits of retreatment were significant and clearly demonstrated for fatigue, the sole outcome for which the study was properly designed and analyzed; the authors’ suggestion that this positive finding was due to unmasking is unfounded.

Flaws in the trial’s design with regard to the clearance of OspA from the CSF and improvements in mental processing speed made it unlikely that a positive treatment effect on these endpoints would be found. Thus, the lack of demonstrable benefits on these endpoints is uninformative and the

Please cite this article as: DeLong AK, et al, Antibiotic retreatment of Lyme disease in patients with persistent symptoms: A biostatistical review of randomized, pla..., Contemp Clin Trials (2012), http://dx.doi.org/10.1016/j.cct.2012.08.009
Table 4

<table>
<thead>
<tr>
<th>Question type</th>
<th>Lyme patients (msec)</th>
<th>Healthy participants</th>
<th>% faster for healthy participants vs. Lyme patients</th>
</tr>
</thead>
</table>
| Letter match (true) | 1012 | 896 | 11.5%
| AA+2 (true) | 3022 | 2256 | 25.3%
| AA+3 (true) | 3631 | 2813 | 22.5%
| AA+4 (true) | 4180 | 3256 | 22.1%
| Letter match (false) | 1088 | 990 | 9.0%
| AA+2 (false) | 3572 | 2696 | 24.5%
| AA+3 (false) | 4074 | 3178 | 22.0%
| AA+4 (false) | 4324 | 3588 | 17.0%

1. In the study design, Krupp et al. assumed a 25% improvement as the MCID.  
2. Age- and education-matched controls.

500 authors were wrong to recommend against retreatment on this basis.  
505 Given the clear benefit on severe fatigue, the uninformative findings on OspA clearance and mental processing speed, and despite the potential for significant antibiotic-associated adverse events, we conclude that the trial by Krupp et al. demonstrates that retreatment with ceftriaxone may be helpful for patients with ongoing severe fatigue after a standard course of Lyme disease treatment.

3.3. Fallon et al. trial [17]

3.3.1. Trial summary

The Fallon et al. trial enrolled subjects who had memory impairment on subjective and objective assessment tools (Wechsler Memory Scale-III) despite having previously received a minimum of 3 weeks of IV ceftriaxone; IgG seropositivity on entrance was an inclusion criterion. Thirty-seven subjects were randomly assigned 2-to-1 to receive 10 weeks of ceftriaxone or placebo. The primary outcome was cognitive change over time as measured across six domains to assess multiple aspects of cognition, with memory being the domain hypothesized as showing greatest change. The 3-month outcome measured treatment efficacy and the 6-month outcome measured treatment durability. Secondary measures included the SF-36 PCS and MCS scores, fatigue (FSS-11), pain (VAS), depression (Beck), anxiety (Zung), and global psycho-pathology (GSI SCL-90). For the primary outcome, a healthy control group was also enrolled.

Fallon found improvement in cognitive functioning at 12 weeks, with a significance level of 0.053 that falls just above the margin of significance demonstrating treatment efficacy; however, it was unsustained at 24 weeks. Among the secondary outcomes, none of the psychiatric or mental outcomes were significant. However, there was a significant interaction effect between treatment and baseline scores, confirming that those with worse baseline scores had sustained improvement in the physical component score (SF-36 PCS) and decreases in VAS pain score to 24 weeks. In addition, a post hoc analysis of the subgroup meeting Krupp’s STOP-LD enrollment criteria and using the same definition for a positive treatment response on the FSS-11 as Krupp found that retreatment was beneficial (66.7% in the ceftriaxone arm vs 25% in the placebo arm). There were 7 significant treatment-related adverse events (18.9%); 6 occurred in subjects on active treatment.

Due to the lack of durable cognitive improvement and the risk of adverse events, Fallon et al. concluded that 10 weeks of ceftriaxone was not an effective strategy; the authors encouraged searching for more effective and safer retreatment strategies.

3.3.2. Trial critique

3.3.2.1. Design. The trial was designed with a planned enrollment of 45 participants but recruited only 37 subjects; 23 randomized to active treatment and 14 to placebo. Under-enrollment could have resulted in the cognitive functioning outcome becoming underpowered.

3.3.2.2. Analysis. While the study was under-enrolled, 32/37 (86%) of enrolled patients completed the protocol at 12 and 24 weeks. Detection of significance in pain and physical functioning among those with worse baseline scores can likely be attributed to an efficient statistical analysis incorporating monthly measures of these secondary outcomes, and incorporating effect modification due to baseline disease severity.

3.3.2.3. Interpretation. This trial, with its small sample size and extensive secondary outcome analysis, is more reminiscent of a pilot study than a definitive clinical trial. The conclusions were fittingly cautious. Noting a positive treatment effect on fatigue, similar to that seen in the Krupp trial, and a high rate of adverse events, the authors highlighted the need for additional studies and safer antibiotic regimens.

4. Discussion

This biostatistical review of the four NIH-sponsored Lyme disease retreatment trials highlights the need for close scrutiny of all clinical trials, including those which emphasize findings of insignificant treatment effects. Our careful examination of the trials suggests that, for some patients with Lyme disease, retreatment can, in fact, be beneficial. Krupp’s study was properly designed and analyzed with regard to fatigue, detecting significant, sustained and clinically meaningful improvement in this primary endpoint, and the Fallon trial demonstrated treatment efficacy on cognition at the margin of statistical significance at 3 months. And, although these were secondary outcomes, the Fallon trial corroborated Krupp’s finding on fatigue and, further, found that patients with worse baseline pain and physical functioning had significant and sustained improvement in these measures.

Unfortunately, misinterpretation of insignificant findings from underpowered or poorly designed trials can have profound ramifications on treatment guideline recommendations, patient care and the direction of future research. In Lyme disease, the lack of demonstrable improvement in persistent symptoms in the Klempern trials and the absence of an antibiotic effect on mental processing speed in the Krupp trial do not provide evidence against the efficacy of antibiotic retreatment. Our analysis reveals that these outcome measures were not well designed and lacked statistical power. Therefore, contrary to the
conclusions of some Lyme disease guidelines panels [20,21], the 617 inability of these trials to demonstrate a statistically significant 618 finding provides neither proof of the absence of a clinically 619 meaningful treatment effect nor evidence that patients with 620 persistent symptoms suffer from a post-infectious syndrome. 621
The findings from the Krupp and Fallon trials imply a causal 622 link between antibiotic treatment and physical improvement, 623 which would be consistent with the hypothesis that persistent 624 symptoms may be the result of a persistent B. burgdorferi 625 infection. A recent uncontrolled, observational study of Lyme 626 disease patients treated with ceftriaxone appears to support 627 such a link, with patients experiencing long-term benefits in 628 fatigue, pain and cognition [41]. While one may speculate 629 about the potential neuro-protective effects of ceftriaxone in 630 Lyme disease [17], the presence of such effects would not 631 explain findings of sustained benefits on fatigue or pain, or 632 dispel the existence of persistent infection.

Conclusions favoring post-infectious processes as the explana- 633 tions for persistent symptoms may be premature. Several 634 authors, using a variety of accepted laboratory tests, conclusively 635 demonstrated persistent B. burgdorferi infection in humans and 636 animals following antibiotic therapy appropriate for their stage 637 of illness [42–51]. The most recent of these was a primate study, 638 in which investigators recovered intact B. burgdorferi spirochetes 639 by xenodiagnosis from rhesus macaques treated with the 640 Infectious Diseases Society of America (IDSA)-recommended 641 regimen for disseminated infection [48]. Additionally, spirochet- 642 al DNA and RNA were detected post-treatment in this group and 643 in a second group of rhesus macaques that were given the 644 treatment used in the trial by Klemptner et al.: “the animals were 645 treated at the late disseminated phase of infection and the 646 treatment regimen was chosen to correspond to the regimen 647 used to treat human PTLDs patients in a clinical evaluation of 648 treatment for this population” [48].

It is incorrect to draw strong conclusions regarding 649 antibiotic retreatment in patients with persistent symptoms 650 of Lyme disease based on the four NIH-sponsored random- 651 ized controlled trials discussed in this review. Inadequacies 652 in trial designs and the small sample sizes leave many 653 questions unanswered, and underscore the need for addi- 654 tional clinical research on this question. Those who wrongly 655 conclude that the trials found no benefit from retreatment 656 commit an even greater error, as such a statement is 657 demonstrably false.

Future RCTs should investigate oral antibiotic regimens 658 which are likely to be safer [52–54] and less costly than 659 ceftriaxone. Such trials should also avoid enrolling patients 660 according to a proposed definition of “post-Lyme syndrome” 661 [21], not only because the terminology prematurely assumes a 662 post-infectious process, but also because this broader grouping 663 may mask significant treatment effects in specific patient 664 subsets, such as the fatigue subset identified in the Krupp and 665 Fallon trials. Instead, future trials should consider a stratified 666 design ensuring good balance by arm with respect to disease 667 symptom clusters (such as joint or CNS involvement) and with 668 sufficient power to detect realistic treatment effects within 669 strata. Until evidence from such trials becomes available, it 670 would be wise for clinicians to disregard generalized and 671 unsupported recommendations against retreatment and in- 672 stead rely on their clinical judgment to manage patients with 673 persistent symptoms of Lyme disease.

References

health-related quality of life: the remarkable universality of half a

[38] Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. Med Care 1999;37:469-78.

relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1998;36:1589-98.


[30] Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their biostatistical review of randomized, pla... Contemp Clin Trials (2012), http://dx.doi.org/10.1016/j.cct.2012.08.009
