

The Serano Group

White Paper

Analysis Of The Klempner Trial Of Ceftriaxone Followed By Doxycycline For Persisting Symptoms Of Lyme Borreliosis

Study analyzed: Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease, New England Journal of Medicine, 2001. 345(2):85-92.

March 5, 2009

Study Analyzed:

Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease, New England Journal of Medicine, 2001. 345(2):85-92.

Klempner, M. S., Hu, L. T., Evans, J., Schmid, C. H., Johnson, G. M., Trevino, R. P., Norton, D., Levy, L., Wall, D., McCall, J., Kosinski, M., Weinstein, A.

Summary

Question posed: Would 2 grams daily of IV ceftriaxone for 30 days, followed by 100 mg of doxycycline twice daily for 60 days benefit patients previously treated for Lyme disease who continue to exhibit symptoms?

Methods: Patients with persisting symptoms, previously treated for Lyme disease, were selected and treated either with antibiotics or placebo. Patients' mental and physical health status were self-evaluated with the SF-36, a patient-completed questionnaire, and the Fibromyalgia Impact Questionnaire.

Results: After reducing data to "Improved", "Worse", or "Unchanged", no significant differences were detected between the treated and placebo groups.

Implication: This study did not provide evidence that the ceftriaxone followed by doxycycline treatment was effective in treating patients with persisting symptoms of Lyme disease who had already received some antibiotic treatment.

Impression

Little was contributed to the existing knowledge of borreliosis in humans. Researchers had a strong bias toward not detecting a treatment effect. They completed the minimal requirements necessary to present their study as a valid randomized, double-blind clinical trial but provided minimal post-treatment data, using grossly insensitive measures.

Background

History

Lyme disease, like syphilis, is caused by a spirochetal (spiral-shaped) bacteria. In November, 1976, Drs. William Mast and William Burrows, at the U.S. Navy Submarine Medical Center nineteen miles from Lyme, Connecticut, published the first description of an outbreak of unusual symptoms in southeast Connecticut.¹ They described the history of similar diseases and their successful treatment with antibiotics.

By 1977, Yale rheumatologists and state epidemiologists reported on the same disease outbreak in Connecticut where they selected 51 patients with a various symptoms, 25% reporting an expanding round rash later referred to as erythema migrans (EM)². Although residents had reported unusual clustering of neurological problems, extreme fatigue, and recurring skin rashes in the locale for at least twenty years, investigators concentrated on the two most visible symptoms, the EM rash and arthritis, and treated the outbreak as a new disease entity. The EM rash eventually became the predominant defining characteristic of the disease³, although there has never been an accurate assessment of how often infection occurs in absence of the EM rash.

Borrelia burgdorferi (Bb) was identified as the causative bacteria in 1981⁴ and Ixodes ticks (deer ticks) were found to carry and transmit the bacteria to humans. A similar disease had been described in Europe at least fifty years earlier⁵. Some physicians assumed that the common antibiotic treatments used for treating syphilis would be effective for Lyme disease, even though they are largely unproven in efficacy for eliminating the syphilis spirochete. There is much evidence that Lyme disease symptoms return and worsen after shorter antibiotic treatments⁶. There has been some speculation that this is an autoimmune condition that persists after Bb is eliminated, although there is little, if any, evidence to support this.

Tests

The absence of an adequate test for Bb infection complicates Lyme borreliosis research. The study analyzed here, commonly referred to as the “Klempner study”, reinforces the failure of Lyme antibody tests to discriminate. The “Two controlled trials” in the study title refers to separate trials with participants who tested

positive and participants who tested negative in an antibody test. There were no significant differences observed between the seropositive and seronegative groups before or after treatment.

The bacteria Bb is difficult to culture⁷ and is often present in the bloodstream at extremely low counts. It typically sequesters in poorly oxygenated fibrous tissues such as nerves, ligaments, and the skin. Common tests for antibodies to Bb (usually ELISAs or Western blots) present many problems⁸. Detailed studies of antibody response in humans and animal models show a confusing host immune response difficult to relate to commonly accepted immune response in other infections⁹. Further, the tests detect only free antibodies, problematic because an infected patient, antibodies may be entirely, or nearly entirely, bound to bacterial antigen, making the antibodies undetectable by the available tests¹⁰.

Estimates of the accuracy of the antibody tests are highly speculative. Existing studies evaluating test accuracy typically compare patients who have had the EM rash to patients without a history of the rash. Because the percentage of infected patients who never exhibit EM has never been determined, the calculations of accuracy of the antibody tests in detecting infected individuals are highly suspect. The studies only evaluate how well the tests correlate to an EM rash not how well they correlate to Bb infections. Basing diagnosis on other symptoms is also difficult because symptoms are variable and can appear and reappear unpredictably. No set of symptoms is definitive for all infected individuals.

PCR testing for Bb detects nucleic acid sequences in DNA or RNA strands. The small number of Bb in the bloodstream and difficulties in technique make PCR testing insensitive. Best estimates are that PCR testing of serum or whole blood will detect less than 30% of infected individuals¹². There is no general consensus as to which nucleic acid sequences best indicate the presence of Bb. Probably because of the influence of medical politics in Lyme disease stressing under-diagnosis, PCR testing is not routinely accepted as a definitive test for Lyme disease as it is in other infectious diseases.

Symptoms

The highly variable symptoms with much overlap with other diseases make diagnosis subjective. Experienced diagnosticians recognize that uncommon symptom combinations, such as arthritis and neurological problems, are highly

suggestive of Bb infection. Partial or complete remission of multiple symptoms after antibiotic treatment also strongly indicates Bb as the causative agent. These subtleties of diagnosis are not typically emphasized in medical literature. Instead, great emphasis is placed on the EM rash and a history of tick bite.

Another complication in evaluating symptoms and treatment results of Lyme disease is that an effective treatment can produce a temporary worsening of symptoms. This variable reaction, known as a Jarisch-Herxheimer reaction, or simply Herxheimer reaction is recognized in several infectious diseases, particularly those caused by spirochetal bacteria, such as Lyme disease and syphilis (see Supplemental Information).

Politics

The political climate around borreliosis has resulted in much research being censored. Speakers with views not adhering to narrow viewpoints are routinely not permitted at conferences and their articles are usually rejected by journals.

The study received more attention than its data or methodology would warrant. The New England Journal of Medicine, made a pre-publication version available on the Internet, implying the study had information that was immediately important to clinicians. The study's limited measure of one treatment protocol is often quoted as evidence that long-term antibiotics are ineffective for treating chronic Lyme disease. Actually, the researchers were unable to discern anything about one studied treatment: 30 days of IV ceftriaxone followed by 60 days of oral doxycycline. Overall, nothing was discerned. This leaves only the question of how sensitive were the measures used in post-treatment evaluation and were enough treated subject studied. Unfortunately, the measure used was a grossly insensitive patient self-evaluation and the number of subjects inadequate to detect anything less than a dramatic change. The study, as presented, contributed very little to our understanding of Lyme borreliosis.

Analysis

Contributions to Existing Knowledge

The Klempner study did indicate the participants were seriously impaired: their pre-treatment physical component summary scores were comparable to patients with congestive heart failure and participants had greater impairment than patients with a recent myocardial infarction (heart attack).

Problems in Methods and Analysis

The authors attempted to attribute significance to the fact that none of the participants had positive cultures or PCR (genetic) tests for Bb. The original paper stated that subjects testing positive in these tests were excluded from the study, making their absence less than surprising. When later challenged on this point, the authors stated in a response that they conducted over 1800 cultures and PCR tests, both for participant selection and for assessment during the study. All 1800+ results were reported as negative 1. This calls into question the techniques used in performing these tests, techniques not specifically described in the case of the PCR and poorly described for the culture. Performing more than 1800 tests on participants from locales highly endemic for Lyme disease would be expected to produce some positives, even if only a few, from new infections. Additionally, there was no statement in the study that positive controls were used to validate test procedures.

Imprecise Measurement, Summarized before Analysis

As mentioned, the study reported no post-treatment objective measures or results of physician examination, only patient self-assessments of how they perceived their quality of life. Overall, there is a significant lack of post-treatment information of any type in the study, even though many tests and assessments were made prior to treatment. Post-treatment results were limited to participants' questionnaire responses of how they perceived their health, many questions dealing with mental outlook. Scores were highly summarized to an extreme degree before any statistical analysis. In addition, the primary questionnaire used, the SF-36, is commonly recognized as insufficient for measuring improvement in a specific health condition^{11,12}.

Instead of reporting the eight subscales of the SF-36, each which produces a 0 to 100 score, the researchers use only the SF-36's two summary scores: the physical component summary (PCS) and mental component summary (MCS). Many researchers question the validity of using summary scores as they are highly derivative from the actual 0 to 100 scores. The authors further diminished accuracy of measurement by changing numerical 0 to 100 PCS and MCS scores into "Improved", "Worse", and "Unchanged" rating. Essentially, for each participant, the eight 0 to 100 scores produced by the questionnaire were reduced to two 0 to 100 scores, which were then changed to a 1 (improved), 2 (worse), or 3 (unchanged) for mental and physical change. At that point, statistical analysis was performed.

If the researchers were genuinely trying to detect a change produced by the tested antibiotic treatment, they would have used the most sensitive measures available. Instead, they started with a grossly insensitive questionnaire and grossly reduced its measurements before performing any analysis. It is hardly surprising that both the treated and placebo groups produced analyses that found about one-third improved, one-third worsened, and one-third remained unchanged. This would be the expected results in a study producing random results disassociated from any real-world phenomena.

Another self-rating instrument, the Fibromyalgia Impact Questionnaire, was also administered, but scores were not reported; only that changes were not significant, and again. after summarization.

Problems with Methodology

Beyond the gross summarization of data before analysis, there were other problems in methodology.

There was a strong bias toward selecting only participants that were highly likely to have already failed a treatment similar to the one studied. Specifically, the participants were required to have already had at least one course of antibiotics and to have continuing symptoms.

Neither of the two antibiotics used in the treatment studied, IV ceftriaxone or oral doxycycline, have intracellular penetrability. There is a great deal of evidence that Bb is a invades cells (including one report by Klempner in 1993)¹³ where it is protected from extracellular antibiotics. Also, 2 grams daily of IV ceftriaxone for 30

days followed by 100 mg of doxycycline twice daily for 60 days is not definitive or aggressive treatment for patients who have had earlier treatment failures.

Additional Study Weaknesses

Although there were a number of baseline pre-treatment evaluations and tests performed, none of these were reported post-treatment.

At baseline, there were assessments of six symptom classes, plus fibromyalgia tender points, and objective measures of white-cell counts and protein levels in cerebrospinal fluid. None of these assessments were reported post-treatment, although the authors state additional clinical and laboratory evaluations were performed on days 3, 5, 13, 21, 30, 45, 75, 90, and 180.

Effective treatment of Lyme borreliosis, particularly in persisting cases, usually requires antibiotic selection tailored to individual patients and often requires mixes of different antibiotics, selected and modified after evaluation of where a patient is in their disease course 5. This study applied the same treatment to all participants regardless of clinical status and response.

Another factor obscuring treatment success was that subjects were required to have only 75% compliance with the medication protocol.

There are indications that at baseline the researchers administered the complete 116-question Medical Outcomes Study (MOS) Core Survey questionnaire upon which the SF-36 is based, because a baseline "Cognition" score is reported, a result not produced by the SF-36. No post treatment results of the 116-question MOS were reported.

Statistical Problems

This study did not report mean (average) changes in scores after treatment, highly unusual in a study of this type 4. While reporting scores for each subject as a simple change up or down can add an additional aspect to result analysis, it should not replace reporting group mean change. Because mean change was not reported in this study, there cannot be an estimate of whether the difference in mean change between treated and placebo groups was significant. The most basic statistical test for determining the significance of difference between two samples was not possible nor was it performed.

Researchers made arbitrary decisions about scoring. Any participant who withdrew from the study was recorded as having “Worse” health, regardless of the withdrawal reason. If a participant had a significant decline in either mental or physical health, their overall health was recorded as “Worse”. (In other words, a “Worse” score on either mental or physical health trumped an “Improved” score on the other measure.)

The small sample of participants in this study indicates a deficiency in statistical power (see “Supplemental Information”). Using the purely theoretical assumption that the SF-36, when scored according to the authors’ criteria, is a perfect measure of improved health, an estimate of statistical power would indicate how likely the study produced the right conclusion based on the number of participants.

The authors do not make an estimate of statistical power of the group actually studied. Instead, they calculate the statistical power would have been if they had enrolled their full goal of 194 seropositive patients and 66 seronegative patients. As executed, the study only enrolled 70 seropositive and 45 seronegative patients: their goal of 90% power for the seropositive patients and 80% for the seronegative patients was substantially unrealized.

Independently estimating the study’s true statistical power requires making some assumptions. What follows is one example: if in the real-world population of participants meeting criteria for study, 25% would improve on placebo and 50% improve on treatment, this study would have a statistical power of 76% (Fisher Exact Test). In other words, we would expect testing this small of group to produce the correct conclusion only 76% of the time if all methods and measurements were perfect, solely because there were not enough participants.

Misinterpretation By Others

Any argument saying this study is evidence the antibiotic treatment did not work violates one of the most elementary tenets of experimental research. Basically, when two samples are evaluated, this is a test of the null hypothesis: meaning the hypothesis there is no difference between the treated and placebo groups. Experimental trials are performed on a sample taken from a much larger, theoretically limitless, qualifying population. In this study, the researchers took a post-treatment sample of 64 from a theoretically unlimited population of antibiotic-treated patients who could meet inclusion criteria. They also selected a sample of 65

from a theoretically unlimited population of placebo-treated patients who could meet inclusion criteria.

The study states these two particular individual samples did not show a statistically significant difference, leaving aside all the problems introduced by using summarized SF-36 scores. Basic statistical analysis produced the conclusion that, “the two samples did not indicate a high probability they came from populations that differed”, in other words, the antibiotic and placebo samples as tested don’t show probability of being different. Elementary hypothesis testing says this should never be construed as evidence that, “the two populations are the same,” which is how this study has been presented. It is much more difficult to produce evidence that two samples come from populations that do not differ, or differ very little. To do this, many different samples of appropriate size would need to be tested many times and even then conclusions would be suspect because the evaluation tool could be insensitive. For anyone to interpret this study as evidence that antibiotics produced no improvement in patients is scientifically fallacious. The authors are careful not to cross this line, but virtually all editorials and press accounts do.

Probable Reasons Treatment Effect Was Not Detected

In summary, there are several possible, most likely highly probable, reasons the Klempner study did not show significant differences between placebo and treatment groups:

- The primary measurement tool, the SF-36, probably does not have required sensitivity to detect treatment difference even if results had been completely reported 5
- Reducing the SF-36's eight scales (each scored 0-100) to two summary scores (ranging only from 20-58 for PCS and 17-62 for MCS in about 85% of the general population 6) further reduced sensitivity to sample differences
- Reducing the PCS and MCS summary scores to three values ("Improved", "Unchanged", or "Worse") further reduced sensitivity
- Combining the SF-36's two summary scores to a total score still further reduced sensitivity
- The small number of participants reduced chances of detecting a difference
- Selecting participants who had a history of already failing similar treatments reduced sensitivity to detect treatment difference
- Testing a treatment considered marginal, at best, for patients exhibiting these symptoms made detecting improvement unlikely

Conclusion

This study reported no post-treatment results other than the results of a generic quality-of-life self-assessment.

The results were repeatedly summarized before analysis, losing sensitivity to detect change at each step.

The treatment tested would rarely be used by a clinician treating a patient who had already failed one or more previous courses of antibiotics.

The participants selected for study did not reliably represent the population potentially needing additional antibiotics for resolving continuing symptoms of Lyme disease.

Published Abstract of Study

Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease.

Klempner MS, Hu LT, Evans J, Schmid CH, Johnson GM, Trevino RP, Norton D, Levy L, Wall D, McCall J, Kosinski M, Weinstein A.

New England Medical Center and Tufts University School of Medicine, Boston, MA, USA. klempner@bu.edu

BACKGROUND: It is controversial whether prolonged antibiotic treatment is effective for patients in whom symptoms persist after the recommended antibiotic treatment for acute Lyme disease. **METHODS:** We conducted two randomized trials: one in 78 patients who were seropositive for IgG antibodies to *Borrelia burgdorferi* at the time of enrollment and the other in 51 patients who were seronegative. The patients received either intravenous ceftriaxone, 2 g daily for 30 days, followed by oral doxycycline, 200 mg daily for 60 days, or matching intravenous and oral placebos. Each patient had well-documented, previously treated Lyme disease but had persistent musculoskeletal pain, neurocognitive symptoms, or dysesthesia, often associated with fatigue. The primary outcome measures were improvement on the physical- and mental-health-component summary scales of the Medical Outcomes Study 36-item Short-Form General Health Survey (SF-36)--a scale measuring the health-related quality of life--on day 180 of the study. **RESULTS:** After a planned interim analysis, the data and safety monitoring board recommended that the studies be discontinued because data from the first 107 patients indicated that it was highly unlikely that a significant difference in treatment efficacy between the groups would be observed with the planned full enrollment of 260 patients. Base-line assessments documented severe impairment in the patients' health-related quality of life. In intention-to-treat analyses, there were no significant differences in the outcomes with prolonged antibiotic treatment as compared with placebo. Among the seropositive patients who were treated with antibiotics, there was improvement in the score on the physical-component summary scale of the SF-36, the mental-component summary scale, or both in 37 percent, no change in 29 percent, and worsening in 34 percent; among seropositive patients receiving placebo, there was improvement in 40 percent, no change in 26 percent, and worsening in 34 percent ($P=0.96$ for the comparison between treatment groups). The results were similar for the seronegative patients. **CONCLUSIONS:** There is considerable impairment of health-related quality of life among patients with persistent symptoms despite previous antibiotic treatment for acute Lyme disease. However, in these two trials, treatment with intravenous and oral antibiotics for 90 days did not improve symptoms more than placebo.

Supplemental Information

Double-blind Trials

This trial was conducted with the procedures generally used in clinical trials to provide evidence for the effectiveness and safety of new drugs. “Double-blind” indicates medication containers were labeled with a code known only to personnel not involved with drug administration or patient interaction. The code, when translated post-execution, indicates which containers had medication and which had placebo.

Herxheimer reaction

Lyme disease antimicrobial treatment has long been identified as producing a temporary worsening of symptoms known as a Jarisch-Herxheimer, or simply, Herxheimer reaction. This reaction has been extensively reported in other spirochetal diseases such as syphilis and relapsing fever and obviously complicates evaluation of treatment results.

Although the cause is somewhat speculative, this temporary worsening of symptoms or manifestation of new symptoms, is thought to be due to toxins released from bacterial cell walls during cell death caused by the antibiotics. The Herxheimer reaction can cause confusion for the less-experienced clinician or, for the more experienced and perceptive clinician could be used as a diagnostic tool. The time lapse between treatment initiation and indications of a Herxheimer reaction with Lyme disease is highly variable, anywhere from a few hours to a few weeks. The duration and intensity of a Herxheimer reaction is also highly variable.

Measuring health status during treatment cycles for patients with Lyme disease could present a confusing picture to researchers as effects of the Herxheimer reaction could wax and wane.

Statistical Power

Statistical power estimates how likely a study would produce the “right” conclusion based on the likelihood the size of the sample represents the population as a whole. Several factors affect statistical power. For example, larger samples

increase statistical power, smaller samples decrease statistical power. Designing a trial to use a large enough sample is the researcher can control. Greater differences in the populations from which the samples are selected also increase statistical power, but this might be a basic reality the researcher cannot control.

The SF-36 Questionnaire

The SF-36 is probably the world's most popular quality of life questionnaire. Patients check a box in response to 36 questions. They have from three to five response choices, generally of the nature of "Excellent", "Very Good", "Good", "Fair", or "Poor". The responses are scored and the results are eight subscale scores in the range of 0-100:

- Physical Functioning (PF)
- Role-Physical (RF)
- Bodily Pain (BP)
- General Health (GH)
- Vitality (VT)
- Social Functioning (SF)
- Role-Emotional (RE)
- Mental Health (MH)

These eight subscale scores are then summarized into two summary scores, the Physical Component Summary (PCS) and Mental Component Summary (MCS) which are scored using fairly elaborate techniques to produce norm-based scores (bell-shaped distribution) with a mean of 50 and standard deviation of 10. In the general population about 85% of the population scores between 20 and 58 for PCS and 17 to 62 for the MCS.

The eight subscale scores are straightforward but the PCS and MCS are subjects of continuing debate as to how well they measure physical and mental health and what score changes mean clinically. John Ware and Mark Kosinski, primary refiners of the SF-36, best summarize how the SF-36 scores should be used, "Because of the potential for information loss with summary health measures, we encouraged those who use them to interpret their results in parallel with the profile of SF-36 subscales..."¹⁴.

[John Ware and Mark Kosinski are executives at QualityMetrix Corporation, a private company that licenses use of the SF-36. Mark Kosinski is a coauthor of the Klempner study, making the questionable use of the SF-36 results especially surprising.]

References

- 1 Mast WE, Burrows WM. Erythema chronicum migrans and "lyme arthritis". JAMA. 1976 Nov 22;236(21):2392.PMID: 98984
- 2 Steere, A.C., J.A. Hardin, and S.E. Malawista, Erythema chronicum migrans and Lyme arthritis:cryoimmunoglobulins and clinical activity of skin and joints. Science, 1977. 196(4294): p. 1121-2.
- 3 *Case definitions for infectious conditions under public health surveillance*. MMWR Morb Mortal Wkly Rep, 1997. **46**(RR-10): p. 20-22.
- 4 Burgdorfer, W., et al., *Lyme disease-a tick-borne spirochetosis?* Science, 1982. **216**(4552): p. 1317-9.
- 5 Lipschutz, B., *Uber eine seltene Eythemform (erythema cronicum migrans)*. Arch Dermatol Syph, 1913. **118**: p. 349-356.
- 6 Oksi, J., et al., *Borrelia burgdorferi detected by culture and PCR in clinical relapse of disseminated Lyme borreliosis*. Ann Med, 1999. **31**(3): p. 225-32.
- 7 Wormser, G.P., et al., *Improving the yield of blood cultures for patients with early Lyme disease*. J Clin Microbiol, 1998. **36**(1): p. 296-8.
- 8 Tylewska-Wierzbanowska, S. and T. Chmielewski, Limitation of serological testing for Lyme borreliosis: evaluation of ELISA and western blot in comparison with PCR and culture methods. Wien Klin Wochenschr, 2002. 114(13-14): p. 601-5.
- 9 Salazar, J.C., et al., *Coevolution of markers of innate and adaptive immunity in skin and peripheral blood of patients with erythema migrans*. J Immunol, 2003. **171**(5): p. 2660-70.
- 10 Brunner, M. and L.H. Sigal, *Immune complexes from serum of patients with lyme disease containBorrelia burgdorferi antigen and antigen-specific antibodies: potential use for improved testing*. JInfect Dis, 2000. **182**(2): p. 534-9
- 11 Hobart, J.C., et al., *Quality of life measurement after stroke: uses and abuses of the SF-36*. Stroke, 2002. **33**(5): p. 1348-56.
- 12 Leong, K.P., et al., *Why generic and disease-specific quality-of-life instruments should be usedtogether for the evaluation of patients with persistent allergic rhinitis*. Clin Exp Allergy, 2005.**35**(3): p. 288-98.
- 13 Klempner, M.S., R. Noring, and R.A. Rogers, *Invasion of human skin fibroblasts by the Lymedisease spirochete, Borrelia burgdorferi*. J Infect Dis, 1993. **167**(5): p. 1074-81.
- 14 Ware, J.E., Jr., et al., *Differences in 4-year health outcomes for elderly and poor, chronically illpatients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study*.Jama, 1996. **276**(13): p. 1039-47.