

Hearing of the IDSA Lyme Disease Review Panel

Hemisphere A

Ronald Regan and International Trade Center

Washington, D.C. 20004

July 30, 2009

Challenge to IDSA Recommendations for Late Neurologic Lyme Disease Treatment and Post Lyme Syndrome: A Statistical Review of NIH- Funded Treatment Studies*

Allison DeLong, M.S.‡

‡Brown University

Center for Statistical Sciences

Department of Public Health, Division of Biology and Medicine

121 South Main Street, Providence, RI 02912

*Please refer to the written submission to the IDSA panel with the same title, co-authored by Tao Liu, Ph.D.‡, Allison DeLong, M.S.‡, and Barbara Blossom, B.A.

This challenge is to the following IDSA Recommendations¹

1) Treatment of Late Neurologic Lyme Disease—Recommendation 3, p. 1113

“Adult patients with late neurologic disease affecting the central or peripheral nervous system should be treated with ceftriaxone (2 g once per day intravenously for 2–4 weeks) (tables 2 and 3) (B-II). Cefotaxime or penicillin G administered intravenously is an alternative (B-II). Response to treatment is usually slow and may be incomplete. Re-treatment is not recommended unless relapse is shown by reliable objective measures. Ceftriaxone is also recommended for children with late neurologic Lyme disease (tables 2 and 3) (B-II). Cefotaxime or penicillin G administered intravenously is an alternative (B-III).”

2) Post-Lyme Disease Syndrome Treatment—Recommendation 2, p. 1120-1121

“To date, there is no convincing biologic evidence for the existence of symptomatic chronic *B. burgdorferi* infection among patients after receipt of recommended treatment regimens for Lyme disease. Antibiotic therapy has not proven to be useful and is not recommended for patients with chronic (≥ 6 months) subjective symptoms after administration of recommended treatment regimens for Lyme disease (E-1).”

Table 1. Infectious Diseases Society of America–U.S. Public Health Service Grading System for ranking recommendations¹

Strength of recommendation	
A	Strongly in favor
B	Moderately in favor
C	Optional
D	Moderately against
E	Strongly against

Quality of evidence	
I	Evidence from ≥ 1 properly randomized, controlled trial
II	Evidence from ≥ 1 well-designed clinical trial, without randomization; from cohort or case-controlled analytic studies (preferably from >1 center); from multiple time series studies; or from dramatic results from uncontrolled experiments
III	Evidence from opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees

E-I: Implies that one or more randomized controlled trial proved antibiotic treatment is ineffective

E-I: Essential that **no trial** has shown antibiotic treatment to be effective

Four NIH-funded studies examined retreatment of patients with a history of Lyme disease and persistent symptoms:

The trials:

Fallon BA, Keilp JG, Corbera KM, et al. A randomized, placebo-controlled trial of repeated IV antibiotic therapy for Lyme encephalopathy. *Neurology*. Mar 25 2008;70(13):992-1003.²

Krupp LB, Hyman LG, Grimson R, et al. Study and treatment of post Lyme disease (STOP-LD): a randomized double masked clinical trial. *Neurology*. Jun 24 2003;60(12):1923-1930.³

Klempner MS, Hu LT, Evans J, et al. Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease. *The New England Journal of Medicine*. Jul 12 2001;345(2):85-92.⁴

All participants had prior diagnosis of LD, had been treated with standard therapy, and had current symptoms that started with the onset of LD yet persisted 4-6 months following therapy for LD.

Fallon Trial: Well designed, executed and analyzed

Design:

Treated for 10 weeks ceftriaxone vs. placebo

Required memory impairment at baseline

Sample size = 37

Follow-up at 12 and 24 weeks

Statistical Review:

Small sample size, but no problems with statistics or interpretation

Results:

Primary outcome: **Cognition** – significant improvement at 12 wks ($p=0.053$)

Secondary outcomes:

Among those with worse baseline scores -

Improvement in **overall physical health** and **pain** (24 wks), **fatigue** (12 wks)

Using criteria of Krupp, sustained improvement in **fatigue** (24 wks)

Krupp: Well designed for fatigue but problems with interpretation

Inclusion Criteria: Severe fatigue (≥ 4.0 on the Fatigue Severity Scale, FSS-11)

Treatment: 1 month ceftriaxone (N=28) vs. placebo (N=27)

Primary outcomes: Fatigue (FSS-11), Mental speed (AA, Alphabet Arithmetic test),
Clearance of OspA antigen

Follow-up: 1 and 6 months

Clinical Improvement: Fatigue – decrease in FSS-11 ≥ 0.7
Mental speed – increase $\geq 25\%$

Power: 81% power for 35% difference in % of patients with improvement in fatigue
74% power for 25% difference in % of patients with improvement in AA test

Analysis: Chi-square, and Wilcoxon rank sum
Sensitivity analysis to assess drop out

Krupp: Significant improvement in fatigue

Effect of treatment on fatigue

Measure – FSS-11	Ceftriaxone	Placebo	P-value
Intention to treat: % with clinical improvement	64%	18.5%	<0.001
Observed: % with clinical improvement	69%	23%	<0.01
Observed: Mean change from baseline (SD)	-1.3 (1.4)	-0.5 (0.93)	0.01
Observed: Mean % change (SD)	-22.1 (24.8)	-9.1 (17.4)	0.01

Guidelines criticize Krupp's findings based upon

- Patient selection
- Loss to follow-up
- Potential unmasking of study medication

It is ok if a few participants did not meet entry criteria

An intention-to-treat analysis: *“the analysis that includes all randomised patients in the groups to which they were randomly assigned, regardless of their adherence with the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol”*¹³

13. Piantadosi S. Clinical Trials: A Methodologic Perspective. Wiley Series in Probability and Statistics. 1997. New York, NY.

Krupp: Results robust to loss to follow-up

Loss to follow up: 7/55 participants (13%) were lost: 2 abx, 5 placebo

Outcomes on fatigue by trial arm

Group	# Improved	# Not Improved	# Unknown
Ceftriaxone	18	8	2
Placebo	5	17	5

Each row is 1 of 3 potential full datasets for ceftriaxone arm

# Improved	# Not Improved
18	10
19	9
20	8

Each row is 1 of 6 potential full datasets for placebo arm

# Improved	# Not Improved
5	22
6	21
7	20
8	19
9	18
10	17

- Krupp examined the $3 \times 6 = 18$ potential datasets

- All tests remained significant ($p < 0.05$)

Loss to follow-up did not affect the trial findings on fatigue ⁹

Krupp: No quantitative evidence of problems with masking

Krupp: "...patients in the ceftriaxone group may have improved in fatigue because they were more likely to believe they were on active therapy, perhaps because they felt better."

- The percent thinking they were taking ceftriaxone did not differ by arm at either time point ($p > 0.05$)

Thinking taking active therapy

	Ceftriaxone group	Placebo group	p-value
1 Month	17/24 (71%)	13/23 (57%)	0.37
6 Month	18/26 (69%)	15/22 (68%)	1.0

Krupp made a mistake

- Compared the percent correctly guessing treatment assignment

Correctly guessing treatment group

	Ceftriaxone group	Placebo group	P-value
1 Month	17/24 (71%)	10/23 (44%)	0.04
6 Month	18/26 (69%)	7/22 (32%)	0.004

Krupp: Mental speed outcome - "Improvement" poorly defined

Treatment arms were not different. Why?

- Krupp: *"Although patients showed cognitive slowing compared to healthy controls, these deficits were relatively mild, which may have contributed to the lack of a treatment effect on cognition."*
- Designed with low power (74%)
- Clinical improvement: 25% improvement in the alphabet arithmetic test
- 25% improvement forces Lyme patients to perform better than healthy controls¹¹ (See written submission to panel)

Krupp: Interpretation for Guidelines

Krupp: *"[T]he improvement in fatigue could be considered an encouraging finding in that future studies exploring other less expensive and noninvasive methods for treating severe fatigue might be effective."*

Klempner's two trials: Substantial statistical problems

Treatment: Placebo vs. 30 days ceftriaxone + 60 days doxycycline

Measurement times: Baseline, 30, 90 and 180 days

Sample size: Enrolled/Designed=70/194 (seropos), 45/66 (seroneg)

Instrument: SF-36 Health Survey (www.qualitymetric.com)

- Klempner examines the SF-36 summary scores of Physical (PCS) and Mental (MCS) health
- Numeric value
 - Range 0-100
 - Average (SD) for U.S. population is 50 (10)
- Lower scores imply worse health

Klempner: The statistical analysis that should have been done

Value of and changes in PCS and MCS linked to disease severity and quality of life

Improvement of 5 points on the PCS translates into a 20% decrease in the percent of patients unable to work¹⁰

Recommended statistical analysis of PCS and MCS scores:

Use a regression analysis for continuous, longitudinal data that adjusts for baseline, uses data from all follow-up time points, and considers possible issues with non-random patient dropout^{12,13,14,15}

- Most efficient use of data (i.e. most likely to obtain a significant result)^{14,15}
- Reduce estimation bias of the treatment effect^{12,13,14,15}
- Treatment effect is an estimate of difference in change in PCS or MCS and has clinical meaning¹³

12. Daniels MJ, Hogan JW. *Missing Data in Longitudinal Studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC. 2008.

13. Piantadosi S. *Clinical Trials: A Methodologic Perspective*. Wiley Series in Probability and Statistics. 1997.

14. Taylor et al. *Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized*. 2006.

15. Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal data analysis*. Wiley Series in Probability and Statistics. 2004.

Klempner: Inadequate statistical analysis

Outcome: **Participants categorized using change from baseline to 180 days**

Category	Physical score (PCS)	Mental score (MCS)
Improved	Increase > 6.5* points	Increase > 7.9* points
Same	Changed < 6.5 points	Changed < 7.9 points
Worsened	Decrease > 6.5 points	Decrease > 7.9 points

* 6.5 and 7.9 = 2*(standard error of the measurement)

Treatment effect: **Difference between arms in the % "improved" or "worsened"**
Klempner Analysis: **Chi-square test**

Finding #1: The statistical analysis was
Inefficient:

- Omits 30 and 90 day measures
- Categorized a continuous outcome

Probably biased:

- Baseline scores differ by arm
- Loss to follow-up

Had unsatisfying measure of treatment effect

- No measure of mean changes on the scale of the SF-36
- Categorized outcomes difficult to interpret clinically

Mean changes in SF-36 smaller than 6.5/7.9 are clinically meaningful

SF-36 summary score changes found to be clinically and statistically significant

Reference	Disease	Increase in physical score	Increase in mental score	Verification of clinical significance
Kosinski et al. ⁶	Rheumatoid arthritis	4.4, 4.3, 3.0, 2.6, 3.2*	4.7, 3.1, 2.2, 3.1, 2.3*	1 level of improvement across five clinical RA measures*
Angst et al. ⁷	Osteoarthritis	2.0	‡	Improvement in global health self-assessments
Coteur et al. ⁸	Crohn's disease	4.1	3.9	IBDQ improvement**
Regensteiner et al. ⁹	Peripheral artery disease	2.0	§	Increased maximum treadmill walking distance
Okamoto et al. ¹⁰	Asthma	≤ 5.0	§	Increased FEV ₁ ***
Klempner et al. ⁴	Lyme disease	6.5 (expected)	7.9 (expected)	

* Values presented in order: Patient global assessment, physician global assessment, pain assessment, joint swelling, and joint tenderness; ** Inflammatory Bowel Disease Questionnaire (authors considered this the "best" MCID estimate among several clinical measures in this study because it correlated most closely with SF-36 scores); *** Forced expiratory volume in 1 second; ‡ Not determined; § Not significant

Finding #2: SF-36 changes of 2 to 5 are clinically meaningful, much smaller than 6.5 and 7.9

Klempner's expected treatment effects were too big

What happens when expected treatment effects are too big?

- The trial was designed with a sample size that was too small
- Small sample sizes resulted in inadequate power, or ability, to detect smaller, relevant treatment effects

As a result

- It would be unlikely that smaller, meaningful treatment effects would be statistically significant
- An interim analysis should, and did, terminate the trial for futility (i.e. lack of statistical significance)
- Does not mean that the treatment was shown to be ineffective

Klempner: Lyme patients are quite sick

Average Baseline SF-36 Scores

Trial	Seropositive	Seronegative	U.S. population
Physical score (PCS)	34.8	36.7	50
Mental score (MCS)	44.3	42.5*	50

* Antibiotic arm was 8.3 points higher than Placebo arm (46.7 vs. 38.4, $p < 0.05$)

Baseline PCS

- Similar to congestive heart failure and osteoarthritis
- Worse than type 2 diabetes or recent heart attack

Baseline MCS

- A change of 7.9 forces Lyme patients to perform better than population average

$$44.3 + 7.9 = 52.2 > 50$$

$$42.5 + 7.9 = 50.4 > 50$$

Finding #3: Expected treatment effect of 7.9 MCS is unrealistic

Klempner: Probable type II error

For each hypothesis we obtain an estimate of the treatment effect with confidence interval and p-value

Draw 1 of 3 conclusions:

1. p-value < 0.05: Treatment is effective (type I error rate 5%)
2. p-value > 0.05 and confidence interval excludes a clinically important difference: Treatment is clinically ineffective
3. p-value > 0.05 and Confidence interval includes a clinically important difference: Cannot conclude treatment was effective or ineffective

In case (3), stating a treatment is ineffective is very likely an error (type II)

Do Klempner's confidence intervals include clinically meaningful differences?

Klempner: Declaring treatment to be ineffective is likely a type II error

Observed treatment effects* (95% CI)

Trial	PCS	MCS
Seropositive	3% (-19 to 24%)	-14% (-37 to 8%)
Seronegative	19% (-7 to 46%)	10% (-17 to 37%)

The expected treatment effects under clinically meaningful changes in SF-36 fall in the observed confidence intervals

Expected** treatment effects* for clinically meaningful SF-36 changes

Difference in SF-36 change by arm	Physical score	Mental score
2	7%	5%
3	10%	8%
4	14%	10%
5	18%	13%

Finding #4: Clinically meaningful differences are not ruled out

* Difference in % "improved" abx vs. placebo. Baseline to 180 days.

**SD of change from baseline to 180 days in placebo group: PCS=10.1, MCS=14.2

Klempner: Sample size calculation was flawed

Klempner's sample size calculation

Trial	treatment effect chosen for sample size calculation	Required PCS score change for chosen treatment effect	Required MCS score change for chosen treatment effect
Seropositive	25%	6.7	9.1
Seronegative	35%	9.3	12.8

- SF-36 summary score changes of 2 to 5 are clinically meaningful
- The trial was designed to be able to detect mean changes greater than or equal to those in this table

Finding #5: The trial was underpowered, even under Klempner's expected treatment effects

Klempner: A statistical review found many shortcomings

In summary:

- Trial had insignificant results that did not rule out clinically meaningful treatment effects
- Underpowered to detect clinically meaningful effects
- Treatment was not shown to be ineffective
- Klempner's trials cannot be used to formulate treatment guidelines

Recommendation for Guideline Revision

The Klempner trials cannot be used to inform treatment guidelines. Two well designed and executed trials (Krupp and Fallon) have examined the effect of re-treatment among patients with persistent symptoms of Lyme disease following the IDSA's current recommended therapy. Neither trial has proven re-treatment to be ineffective. In fact, both trials have findings that indicate benefits of re-treatment in certain subpopulations: among individuals with worse fatigue (Krupp and Fallon), worse physical functioning (Fallon), and more severe pain (Fallon). Both studies were small. More research is needed.

References

1. Wormser GP, Dattwyler RJ, Shapiro ED, et al. The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. *Clin Infect Dis*. Nov 1 2006;43(9):1089-1134.
2. Fallon BA, Keilp JG, Corbera KM, et al. A randomized, placebo-controlled trial of repeated IV antibiotic therapy for Lyme encephalopathy. *Neurology*. Mar 25 2008;70(13):992-1003.
3. Krupp LB, Hyman LG, Grimson R, et al. Study and treatment of post Lyme disease (STOP-LD): a randomized double masked clinical trial. *Neurology*. Jun 24 2003;60(12):1923-1930
4. Klemmner MS, Hu LT, Evans J, et al. Two controlled trials of antibiotic treatment in patients with persistent symptoms and a history of Lyme disease. *The New England Journal of Medicine*. Jul 12 2001;345(2):85-92.
5. Weinstein A, Klemmner MS. Author Reply to Treatment of patients with persistent symptoms and a history of Lyme disease. *The New England Journal of Medicine*, Nov 8 2001;345(19):1425.
6. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE, Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis and rheumatism*. Jul 2000;43(7):1478-1487.
7. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis and rheumatism*. Aug 2001;45(4):384-391.
8. Coteur G, Feagan B, Keininger DL, Kosinski M. Evaluation of the meaningfulness of health-related quality of life improvements as assessed by the SF-36 and the EQ-5D VAS in patients with active Crohn's Disease. *Aliment Pharmacol Ther*. May 2009; 29(9):1032-1041.
9. Regensteiner JG, Ware JE, Jr., McCarthy WJ, et al. Effect of cilostazol on treadmill walking, community-based walking ability, and health-related quality of life in patients with intermittent claudication due to peripheral arterial disease: meta-analysis of six randomized controlled trials. *J Am Geriatr Soc*. Dec 2002;50(12):1939-1946.
10. Okamoto LJ, Noonan M, DeBoisblanc BP, Kellerman DJ. Fluticasone propionate improves quality of life in patients with asthma requiring oral corticosteroids. *Ann Allergy Asthma Immunol*. May 1996;76(5):455-461.

References Cont.

11. Pollina DA, Sliwinski M, Squires NK, Krupp LB. Cognitive processing speed in Lyme disease. *Neuropsychiatry Neuropsychol Behav Neurol*. Jan 1999;12(1):72-78.
12. Daniels MJ, Hogan JW. Missing Data in Longitudinal Studies: Strategies for Bayesian modeling and sensitivity analysis. Chapman and Hall/CRC. 2008. Boca Raton, FL.
13. Piantadosi S. Clinical Trials: A Methodologic Perspective. Wiley Series in Probability and Statistics. 1997. New York, NY.
14. Taylor AB, West SG, Aiken LS. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*. 2006; 66:228-239.
15. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Data Analysis. Wiley Series in Probability and Statistics. 2004. New York.
16. Shelledy DC, Legrand TS, Gardner DD, Peters JI. A randomized, controlled study to evaluate the role of an in-home asthma disease management program provided by respiratory therapists in improving outcomes and reducing cost of care. *Journal of Asthma*. March 2009;46:194-201.
17. Cameron D. Severity of Lyme disease with persistent symptoms. Insights from a double-blind placebo-controlled clinical trial. *Minerva Med*. 2008;99:489-96.
18. Clauw DJ, Mease P, Palmer RH, Gendreau RM, Wang Y. Milnacipran for the treatment of fibromyalgia in adults: a 15-week, multicenter, randomized, double-blind, placebo-controlled, multiple-dose clinical trial. *Clinical Therapeutics*. Nov 2008; 30(11):1988-2004.
19. de Andrade SC, de Carvalho RFPP, Soares AS, de Abreu Freitas RP, de Medeiros Guerra LM, Vilar MJ. Thalassotherapy for fibromyalgia: a randomized controlled trial comparing aquatic exercise in sea water and water pool. *Rheumatol Int*. 2008; 29:147-152.
20. Keystone E, Burmester GR, Furie R, Loveless JE, Emery P, Kremer J, Tak PP, Broder MS, Yu E, Cravets M, Magrini F, Jost F. Improvement in patient-reported outcomes in a Rituximab trial of patients with severe rheumatoid arthritis refractory to anti-tumor necrosis factor therapy. *Arthritis and Rheumatism*. June 2008; 59(6): 785-793.

Hearing of the IDSA Lyme Disease Review Panel

Hemisphere A

Ronald Regan and International Trade Center

Washington, D.C. 20004

July 30, 2009

Challenge to IDSA Recommendations for Late Neurologic Lyme Disease Treatment and Post Lyme Syndrome: A Statistical Review of NIH- Funded Treatment Studies*

Allison DeLong, M.S.‡

‡Brown University

Center for Statistical Sciences

Department of Public Health, Division of Biology and Medicine

121 South Main Street, Providence, RI 02912

*Please refer to the written submission to the IDSA panel with the same title, co-authored by Tao Liu, Ph.D.‡, Allison DeLong, M.S.‡, and Barbara Blossom, B.A.

Krupp et al: Mental Speed (AA Test)

Treatment arms were not different. Why?

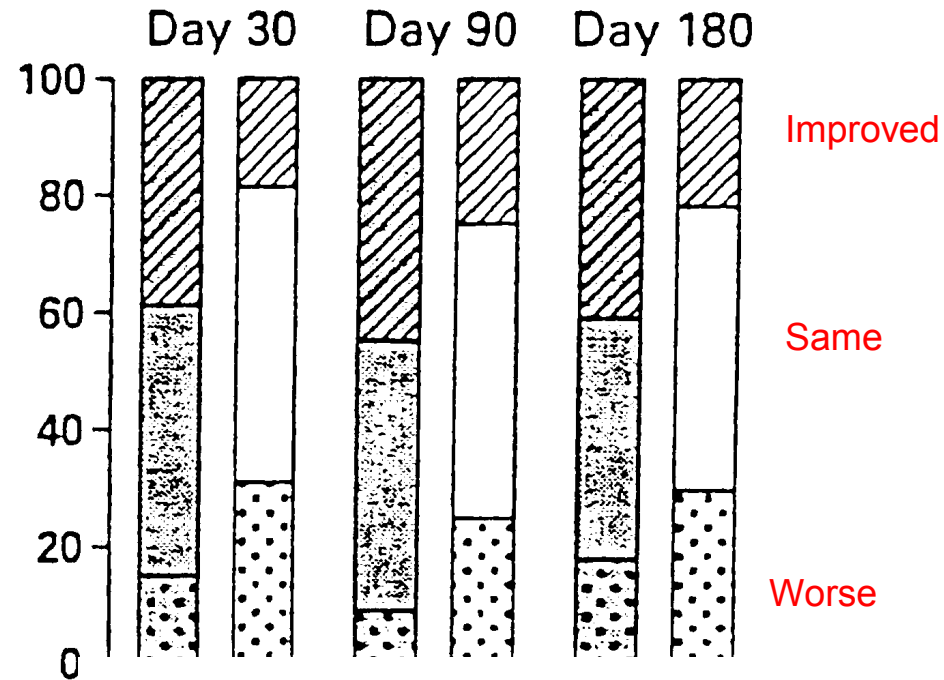
- No patients had baseline objective abnormalities on neurologic exam
- Krupp: "Although patients showed cognitive slowing, the deficits were relatively mild."
- Designed with low power (74%) to detect 25% improvement
- Prior work implies 25% improvement forces Lyme patients to be faster than healthy controls

Mean response times of Lyme patients vs. controls copied directly from Table 3 in Pollina et al.¹¹

Question type	Lyme patients (msec)	Healthy participants* (msec)	% improvement required for average Lyme patient to achieve mean healthy score**
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%

* Age- and education-matched controls; ** Percentage represents the mean score difference as a proportion of Lyme patient mean score

H SF-36 Physical-Component Outcome, Seronegative Patients



Difference in % "improved":
 19% (-7 to 46)
 chi-square $p=0.34$

Row Mean Score test uses ordinal relationship between worse, same, improved ($p=0.16$).

Comparison of the standard deviation (SD) of SF-36 score changes calculated for the Klempner study compared to other studies in patients with chronic illnesses

Observed standard deviation (SD) in the change in PCS and MCS from baseline to follow-up

Study	Disease	SD change PCS	SD change MCS	Follow-up time
Shelledy et al. ¹⁶	Asthma	15	25	26 weeks
Coteur et al. ⁸	Crohn's disease	16.8	-	26 weeks
Regensteiner et al. ⁹	Peripheral artery disease	9	-	12-24 weeks
Angst et al. ⁷	Osteoarthritis	8.0	-	12 weeks
Keystone ²⁰	Rheumatoid Arthritis	7.3	12.2	24 weeks
Cameron ¹⁷	Lyme	7	11	26 weeks
Clauw et al. ¹⁸	Fibromyalgia	8	12	15 weeks
Klempner ⁴	Lyme	10.1	14.2	26 weeks

Kaplan et al: Cognitive function in post-treatment Lyme disease. Do additional antibiotics help?

- Companion paper to Klempner et al.
- Analysis of secondary (unpowered) outcomes from a study that was stopped early
- Participants had low rates of cognitive impairment:
“Although there was no healthy control group, the z-scores allow for comparison between our patients with Lyme disease and age-referenced normative data. There were no significant differences (z-scores of ≤ -1) for any of the neuropsychological tests.”

Lack of statistical significance in the tests of treatment effects does not prove the treatment was ineffective

Klempner et al: Effect Size Simulation

Estimated power at various SF-36 effect sizes

SF-36 change	Seronegative trial (N=66)		Seropositive trial (N=194)	
	PCS	MCS	PCS	MCS
3	10%	6%	30%	16%
4	16%	10%	50%	28%
5	26%	14%	71%	42%
6.5	45%	-	92%	-
7.9	-	35%	-	82%

Klempner's trial, as designed, was underpowered to detect clinically meaningful effect sizes