Published in final edited form as:

Contemp Clin Trials. 2010 September ; 31(5): 423–428. doi:10.1016/j.cct.2010.07.003.

Group analysis versus individual response: The inferential limits of randomized controlled trials

John M. Kelley[a,b,*] and **Ted J. Kaptchuk**[c]

[a]Psychology Department, Endicott College, 376 Hale Street, Beverly, MA 01915, USA

[b]Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street, Boston, MA 02114, USA

[c]Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, MA 02215, USA

Abstract

The randomized controlled trial (RCT) is the gold standard for assessing the efficacy of medical treatments. Over the past 50 years, RCT methodology has proven to be quite successful in identifying effective treatments and weeding out ineffective ones, thus transforming medicine from an intuitive art into an empirical science. However, the enormous success of the RCT has inadvertently contributed to a common inferential error that is insufficiently appreciated by some clinicians and researchers. Although RCTs can effectively distinguish between placebo and active treatment effects at the level of the group, contrary to intuition, this same disentanglement is much more difficult to achieve at the level of the individual. For individual patients it is surprisingly difficult to determine who is a treatment responder and who is not. Using data from a recent RCT, we illustrate the problem and detail its negative effects for research and clinical practice. Finally, we suggest strategies for minimizing these negative effects.

Keywords

Randomized controlled trial; Clinical trial; Placebo effect; Treatment responder

> "If it were not for the great variability among individuals, medicine might as well be a science and not an art."
>
> Sir William Osler, 1892

1. Introduction

Randomized controlled trials (RCTs) have been remarkably effective at separating the specific effects of an active treatment from the non-specific effects associated with a matching placebo intervention. Consequently, RCTs have been successful both in identifying effective treatments and in weeding out ineffective ones. Prior to the development of the RCT, medical treatments were frequently ineffective – harmless placebos at best (e.g., bread pills) or dangerous interventions at worst (e.g., mercury compounds). Thus, the development of the RCT marked an important transformation of medicine from an intuitive, theoretical, and anecdotal art to a

[*]Corresponding author. Psychology Department, Endicott College, 376 Hale Street, Beverly, MA 01915, USA. Tel.: +1 978 232 2386; fax: +1 978 232 3100. JohnKelley@Post.Harvard.Edu (J.M. Kelley).

Competing interests JMK has no competing interests. TJK is a consultant for Kan Herbal Company, Scotts Valley, CA.

science based on empirical data. Evidence-based medicine has profoundly improved the quality of patient care. We argue, however, that the extraordinary success of RCTs in improving medical treatment has resulted in an over-generalization of ideas drawn from RCT methodology to individual patients.

The general problem we discuss in this article has been described previously. Kent [1] presents several hypothetical examples that illustrate how group level data may not apply to individual patients. He also shows how the summary results of an RCT might not even apply to most of the patients in the trial. Rosser [2] decries the fact that public health agencies and insurance companies rely on population level data to justify monetary incentives encouraging physicians to provide preventive treatment to all patients, regardless of “the patient’s personal context and values” (p. 663) [2]. Van Weel and Knottnerus [3] discuss the problem of complex treatment packages that are not typically tested in standard RCTs. And Mant [4] describes a number of limitations in generalizing from RCTs to individual patients (e.g., participants in RCTs typically have less serious disease and lower co-morbidity than many patients seen in clinical practice). Although others have addressed this problem, we believe that it merits consideration once again because the inferential error is common and apparently resistant to extinction (a recent high profile example published in the *New England Journal of Medicine* is presented below). In addition, in this article we focus particularly on the responder/non-responder distinction, and we illustrate how difficult it is to make this distinction using group and individual data from an RCT.

RCT methodology is effective at drawing reliable inferences at the level of the group, but inferences at the level of the individual cannot be made with similar confidence. Specifically, we argue that the distinction drawn between treatment responders and non-responders in research and clinical practice may often be erroneous due to the misapplication of a group methodology (the logic of the RCT) to individual patients.[1] In the sections to follow, we discuss the various components of treatment response; the distinction between group and individual effects (as illustrated by data from a recently completed RCT performed by our team); the difficulty of separating treatment responders from non-responders; the implications of this problem for research and clinical practice; and strategies for minimizing the negative effects of the problem.

2. Components of the treatment response

Randomized controlled trials attempt to separate the specific effects of an active treatment from the non-specific effects associated with a matching placebo intervention. Responses in the placebo condition of an RCT may be attributable to an array of non-specific effects including: (1) natural history (including co-interventions); (2) regression to the mean; (3) response bias; and (4) the placebo effect. Natural history refers to the natural waxing and waning of an illness, including any additional treatment a patient might utilize on his or her own. Some patients will show improvement because of a spontaneous reduction in symptoms or in response to an outside intervention, rather than as a result of the treatment under study. Regression to the mean is a statistical artifact that occurs whenever an inclusion criterion requires that baseline symptoms exceed a certain threshold. Since all symptom assessment instruments include errors in measurement, the imposition of threshold criteria necessarily results in over-estimates of symptom severity at baseline, which inflate symptom improvement estimates at later assessments. Response bias is the tendency for patients to give responses that they believe will

[1] We note here that the definition of “responder” may differ between the clinical and research settings. When dealing with a continuous outcome variable (e.g., hypertension), researchers often develop an *a priori* definition (e.g., 50% reduction in symptoms) that dichotomizes the continuous variable into responders and non-responders. In contrast, clinicians may not employ a strict dichotomization, but they still must make a judgment as to whether the treatment is effective for a particular patient. For simplicity, we will refer to this judgment as determining whether the patient is a “responder,” even though a clinician might not use this particular term.

please the investigators. In the psychological literature, this effect is usually referred to as the demand characteristics of an experiment [5]. Finally, Fisher and colleagues [6] draw an important distinction between the placebo response and the placebo effect. The placebo response is the total response in the placebo condition, whereas the placebo effect is that portion of the effect that is independent of natural history, regression to the mean, and response bias.

As Kaptchuk and colleagues [7,8] have shown experimentally, the placebo effect can be further sub-divided into effects due to the placebo treatment itself (e.g., conditioning and/or expectancy) and effects due to the treatment relationship (e.g., a warm, friendly, and empathic physician). The various components of the treatment response are summarized in Box 1. So long as the components of the treatment response combine additively, an RCT can effectively isolate the specific effect of the active treatment so as to determine efficacy; in addition RCTs can provide an estimate of the magnitude of the specific effect.

Box 1

Components of the treatment response.

1. Treatment effect (specific effect)
2. Placebo response
    - **a.** Placebo effect (non-specific effect)
        - **i.** Effect due to placebo medication or device
        - **ii.** Effect due to treatment relationship
    - **b.** Natural history
    - **c.** Regression to the mean
    - **d.** Response bias

3. Identifying treatment responders: group versus individual effects

Senn [9] notes an important inferential limit to RCTs. If 50% of patients in an RCT respond to treatment, it is commonly inferred that this means that half the patients are responders and the other half are non-responders. However, an equally valid inference is that *all* of the patients in the trial were potential responders, but with only a 50% probability of responding. Of course, neither of these boundary conditions is especially likely (in spite of the fact that the former inference is so commonly made). Instead, individuals are likely to vary substantially in their response probabilities, with some responding virtually always, others responding virtually never, and many responding at various intermediate levels. Indeed, the response probability for any single individual is also likely to fluctuate over time, depending on the severity of the current episode of illness and a number of other varying factors such as environmental stress, social support, diet, exercise, and adjunctive therapy.

A high profile example of this problem can be found in one of the reports on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial published in the *New England Journal of Medicine* [10]. In this trial, patients who could not tolerate or did not respond to citalopram were randomly switched to one of three alternative treatments: (a) buproprion SR; (b) sertraline; or (c) venlafexine. There were no significant differences between the three alternatives in outcomes or tolerability, and approximately 25% of patients in each group achieved remission. In their concluding paragraph, the authors noted that the "three drugs had similar efficacy" (p. 1241). However, although one can say with reasonable confidence that the three treatments did not differ in efficacy, without an adequate control group, one cannot say with confidence that any of the treatments were effective. How would patients have done had they been switched to placebo, or continued on citalopram? In fairness to the authors, they did mention the lack of a placebo control group as a limitation in their discussion, however the take home message from the trial as noted in the last sentence of the abstract is “any one of the medications in the study provided a reasonable second-step choice for patients with depression” (p. 1231).

Fig. 1 is adapted from our recently completed trial of sham acupuncture in irritable bowel syndrome [7,8]. In this RCT, 262 patients were randomized to one of three groups: (1) *Waitlist*; patients were assessed at regular intervals by study staff, but treatment was not offered until after the trial had ended; (2) *Limited*: sham acupuncture was delivered in a neutral, business-like manner; and (3) *Augmented*: sham acupuncture was delivered in the context of a warm and empathic treatment relationship. The waitlist group controlled for natural history and regression to the mean, as well as any effects associated with participation in the trial, such as attention from the study staff who conducted assessments. The limited group received sham acupuncture but had only minimal interaction with the practitioner. Thus, improvement in the limited group in excess of that shown by the waitlist group is attributable to the sham acupuncture treatment. The augmented group also received sham acupuncture, but the treatments were delivered by a warm and empathic practitioner. Thus, improvement in the augmented group in excess of that shown by the limited group would show the effect attributable to the treatment relationship. We hypothesized that symptom improvement would be ordered as follows: Augmented>Limited>Waitlist.

As can be seen on the left side of Fig. 1, the results at the level of the group were consistent with our hypothesis. Our study demonstrated that sham acupuncture for irritable bowel syndrome produces a placebo response that exceeds regression to the mean, natural history, response bias, and any effects associated with participation in a treatment trial. In addition, the study showed that a warm and empathic patient-practitioner relationship (as contrasted with a neutral, business-like relationship) can more than double the effect associated with sham acupuncture treatment. Thus, at the level of the group, our trial detected a statistically significant placebo effect associated with sham acupuncture, as well as an additional effect associated with the treatment relationship. Moreover, these effects were not only statistically significant; they were also of substantial magnitude. Using Cohen’s method [11], the standardized effect sizes for the differences between the groups on the combined outcome were: limited versus waitlist, d=0.51; augmented versus limited, d=0.46; and augmented versus waitlist, d=0.99. Conventionally, the first two effect sizes are considered medium and the last effect size is large.

As can be seen on the right side of the Fig. 1, the results at the level of the individual are remarkably variable. This variability suggests that it would be extremely difficult to determine which patients are treatment responders and which are non-responders. For example, nine individuals in the waitlist condition showed improvements that exceeded the mean improvement for the augmented group, and an additional thirteen individuals showed improvement that exceeded the mean improvement in the limited condition. Thus, a total of 22 waitlisted patients (23% of the group) showed improvement that was in excess of the mean improvement of one or both of the other conditions.

These facts suggest that some individuals in the limited and augmented groups who improved substantially might have shown improvement not because of a response to sham acupuncture or the treatment relationship, but rather because of factors such as natural history, regression to the mean, or response bias. A similar argument can be made for the contrast between the limited and augmented groups. In particular, 33 patients who received the limited treatment (34%) had responses that exceeded the mean response in the augmented group. Thus, some augmented patients who showed improvement above the group mean might not be responders to the treatment relationship, but rather were responders to the sham acupuncture. In short, although one can say with confidence that the augmented and limited conditions produced statistically significant treatment effects at the level of the group, it is extremely difficult to conclude that any individual patient – even those who improved *above* their own group mean – was actually a responder.

Moreover, and even more disturbing, one can easily imagine situations in which the direction of symptomatic change for an individual patient is in the opposite direction to the specific effect of the treatment on that patient. For example, as illustrated in Fig. 2, an individual can show improvements in symptoms despite the fact that for this patient the treatment actually has a negative effect. In spite of treatment that is harmful, the treated patient still showed improvement due to natural history, regression to the mean, response bias, or a placebo effect. The treatment does not work, but the patient is better. Conversely, as illustrated in Fig. 3, an individual may show a worsening of symptoms despite the fact that the treatment is effective for that individual. Without treatment, this patient would still have shown an increase in symptoms; with treatment, however, the worsening in symptoms is attenuated. The treatment works, but the patient is worse.

4. Implications for clinical practice

Physicians are often confronted with patients whose disorders do not respond to treatments that RCTs have demonstrated to be effective. When this occurs, the patient is presumed to be a non-responder, and the clinician moves on to "trials" of other treatments (note how the language of clinical practice now mirrors that of RCTs). For many disorders, there are several treatment options from which to choose, and it is not uncommon for a series of trials to be conducted before an "effective" medication is finally found for the patient. Moreover, patients sometimes relapse, and it is not uncommon for the physician to conclude that the medication no longer works, perhaps due to a shift in the patient's physiology (e.g., down-regulation of neurotransmitter receptors in the monoamine treatment of psychiatric disorders). As explicated in the discussion above, however, these inferences about treatment response are much weaker than many physicians may realize. Indeed, some of the discarded treatments might be equally or even more efficacious than the one that is finally identified as most helpful to the patient.

Given the forgoing, how should physicians alter their behavior? We would argue that recognition of the weakness of inferences about the effectiveness of medication for individual patients should provoke physicians to formulate their hypotheses about treatment efficacy more tentatively. Moreover, alternate explanations for improvement or worsening of symptoms that go beyond pharmacology, such as ancillary treatment, changes in social support, or changes in life stressors should be investigated. Unfortunately, the current set of financial incentives imposed by third party payers does not reward such careful elucidation of additional factors influencing apparent treatment response. Depending on the circumstances, physicians should even be willing to consider another trial of a medication that previously appeared to have been ruled out as ineffective.

The n-of-1 trial is a particularly valuable strategy for testing empirically whether a treatment is effective for an individual patient [12]. These single patient trials utilize a double-blind, multiple cross-over design to tease apart the treatment effect from the placebo response, with several empirical studies supporting their utility in improving clinical practice [13,14]. We believe that broader adoption of n-of-1 trials in clinical practice would substantially reduce the likelihood of the inferential errors described in this paper.

We wish to emphasize that we are *not* suggesting that physicians ignore past evidence of pharmacological successes or failures. Nor are we arguing against the use of evidence-based treatment algorithms developed by governmental agencies and public health authorities that are based on the best available data from RCTs. Instead, we are advocating for a more nuanced clinical approach that takes into account the unavoidable uncertainty that attends the clinical treatment of individual patients. A better appreciation of this uncertainty would help physicians improve patient care.

5. Implications for research

In recent years, success in sequencing the human genotype and the advent of genetic testing has raised the promise of personalized medicine [15-17]. If the genetic, biochemical, and physiological characteristics of individual patients can be identified, and if the biochemical action of pharmaceuticals can be precisely understood, it is possible to envision a future in which medications and other therapies are chosen depending on the particular genetic and physiological make-up of the individual.

A common strategy for linking genetic or physiological characteristics to treatment response is to first identify treatment responders and non-responders and then search for markers that can distinguish between these two groups. Attempts to achieve this goal require that enormous time, effort, and expense be devoted to the application of sophisticated neuroimaging, genetic, and physiological testing. In addition, great care is taken to make an accurate assessment of baseline and post-treatment signs and symptoms of illness. Nevertheless, the arguments presented in this paper suggest that misspecification of treatment responders and non-responders is likely to be an under-appreciated but quite serious impediment to this scientific endeavor, and much greater attention should be paid to this problem and to its solution. In particular, researchers should consider more sophisticated methods for defining treatment response and non-response. Where possible, it would be useful to conduct multiple trials of the treatment for each individual (similar to the n-of-1 trials mentioned above), and then select only those patients who prove to be consistent responders and/or consistent non-responders for more in-depth genetic and physiological studies.

6. Conclusion

RCT methodology has been extraordinarily successful in improving patient care by transforming medicine from an anecdotal art into an empirical science. However, the inferences that can reliably be drawn from RCT methodology are limited to group effects. Analogous inferences made about individuals, whether they be research participants or clinical patients, should be much more tentative. In other words, researchers and clinicians should have much less confidence when they designate some patients as treatment responders and others as non-responders. We argue that a wider acknowledgement of these inferential limits and efforts to overcome them will yield improvements in both clinical research and patient care.

Summary Box

- The logic of randomized controlled trials (RCTs) applies to groups, not individuals.
- The reliable identification of individual treatment responders and non-responders is substantially more difficult than many researchers and clinicians realize.
- This problem has serious negative consequences for research and clinical practice.

- Increased awareness of and attention to this problem will result in improved research designs and better clinical outcomes.

Acknowledgments

This study was supported by National Institutes of Health (NIH) grant no. 1R01 AT001414-01 from the National Center for Complementary and Alternative Medicine (NCCAM) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); and grant nos. 1R21 AT002564 and 1K24 AT004095 from NCCAM. The contents of this article do not necessarily represent the official views of NCCAM, NIDDK, or the NIH.

References

1. Kent DM. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. J Am Med Assoc 2007;298:1209–12.
2. Rosser WW. Application of evidence from randomised controlled trials to general practice. Lancet 1999;353:661–4. [PubMed: 10030347]
3. van Weel C, Knottnerus JA. Evidence-based interventions and comprehensive treatment. Lancet 1999;353:916–8. [PubMed: 10093999]
4. Mant D. Can randomised trials inform clinical decisions about individual patients? Lancet 1999;353:743–6. [PubMed: 10073531]
5. Orne, MT. Demand characteristics and the concept of quasi-controls. In: Rosenthal, RR.; Rosnow, RL., editors. Artifacts in Behavioral Research. New York: Academic Press; 1969. p. 143-79.
6. Fisher S, Lipman RS, Uhlenhuth EH, Rickels KP. L C Drug effects and initial severity of symptomatology. Psychopharmacologia 1965;7:57–60. [PubMed: 5318924]
7. Kaptchuk TJ, Kelley JM, Conboy LA, Davis RB, Kerr CE, Jacobson EE, et al. Components of placebo effect: randomized controlled trial in patients with irritable bowel syndrome. Br Med J 2008;336:998–1003.
8. Kelley JM, Lembo AJ, Ablon JS, Villanueva JJ, Conboy LA, Levy R, et al. Patient and practitioner influences on the placebo effect in irritable bowel syndrome. Psychosom Med 2009;71:789–97. [PubMed: 19661195]
9. Senn S. Individual response to treatment: is it a valid assumption? Br Med J 2004;329:966–8. [PubMed: 15499115]
10. Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME, et al. Bupropion-SR, Sertraline, or Venlafaxine-XR after failure of SSRIs for depression. N Engl J Med 2006;354:1231–42. [PubMed: 16554525]
11. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
12. Guyatt GH, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy: randomized trials in individual patients. N Engl J Med 1986;314:889–92. [PubMed: 2936958]
13. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience. Ann Intern Med 1990;112:293–9. [PubMed: 2297206]
14. Larson EB, Ellsworth AJ, Oas J. Randomized clinical trials in single patients during a 2-year period. J Am Med Assoc 1993;270:2708–12.
15. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol 2001;19:491–6. [PubMed: 11711191]
16. Piquette-Miller M, Grant DM. The art and science of personalized medicine. Clin Pharmacol Ther 2007;81:311–5. [PubMed: 17339856]
17. Woodcock J. The prospects for “personalized medicine” in drug development and drug therapy. Clin Pharmacol Ther 2007;81:164–9. [PubMed: 17259943]

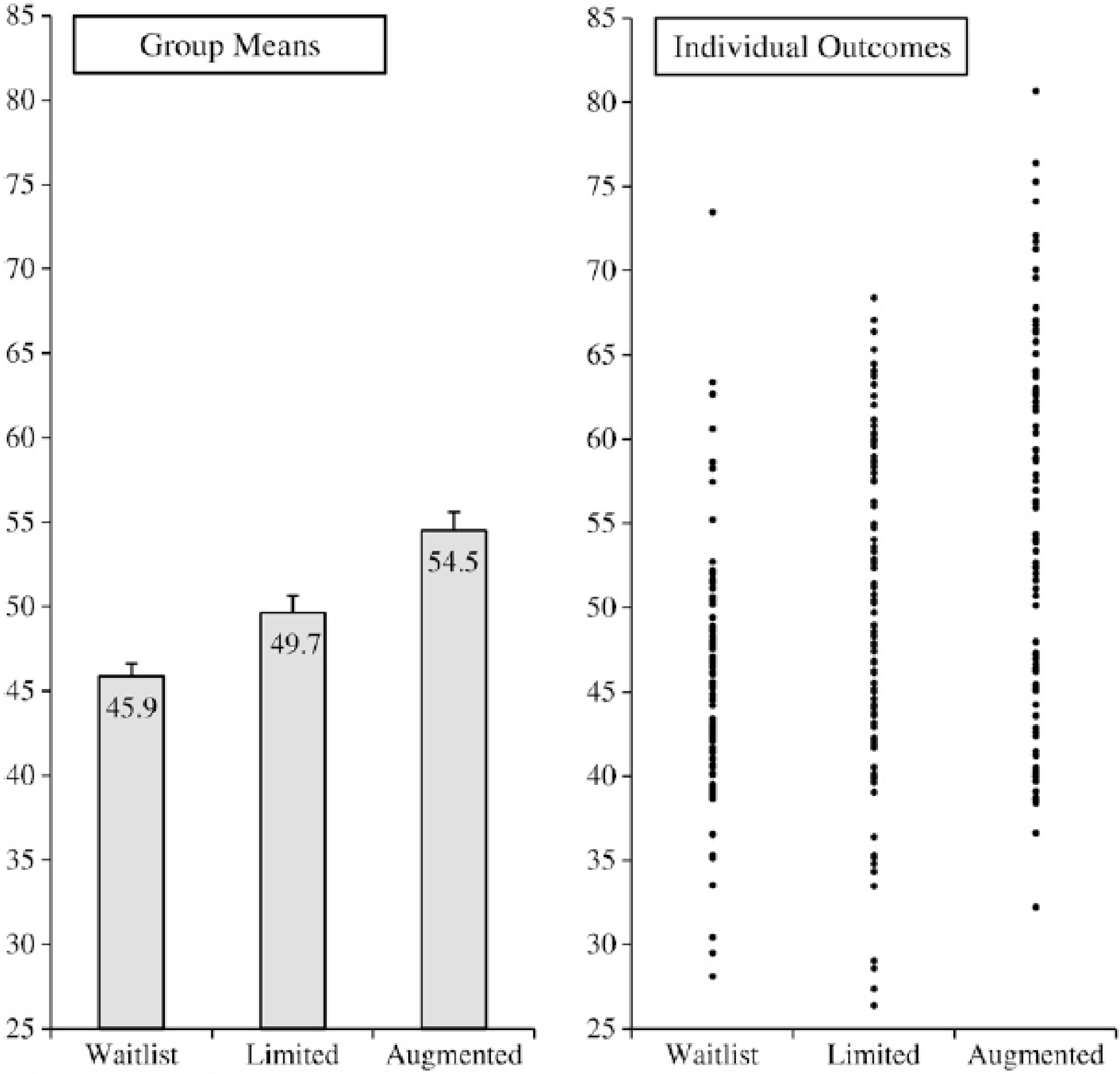

Figure 1.

Group means and individual outcomes by treatment group.

Note: Symptom improvement levels are presented as *t*-scores, with an overall mean of 50 and a standard deviation of 10.

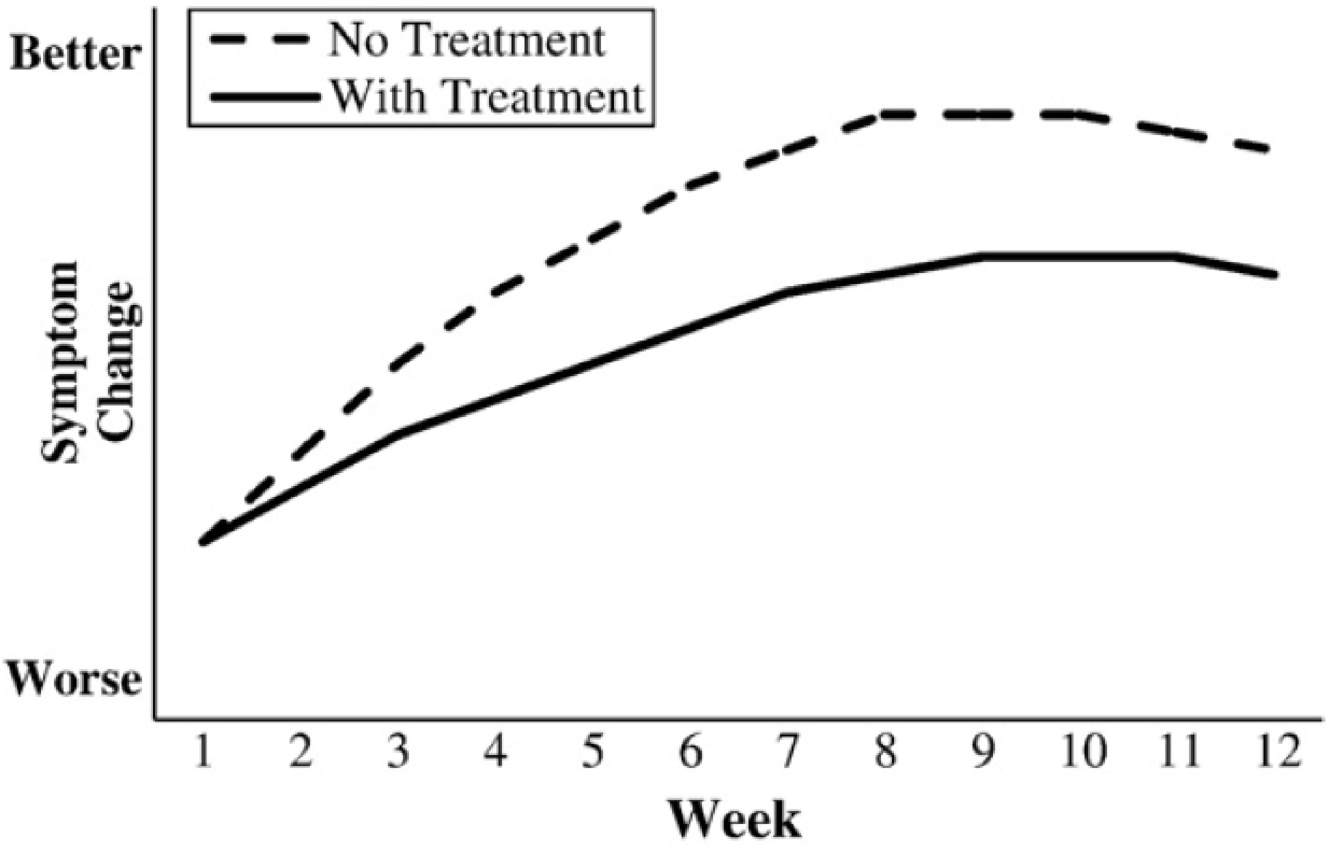

Figure 2.

Hypothetical disease courses for an individual patient treated with an ineffective medication.

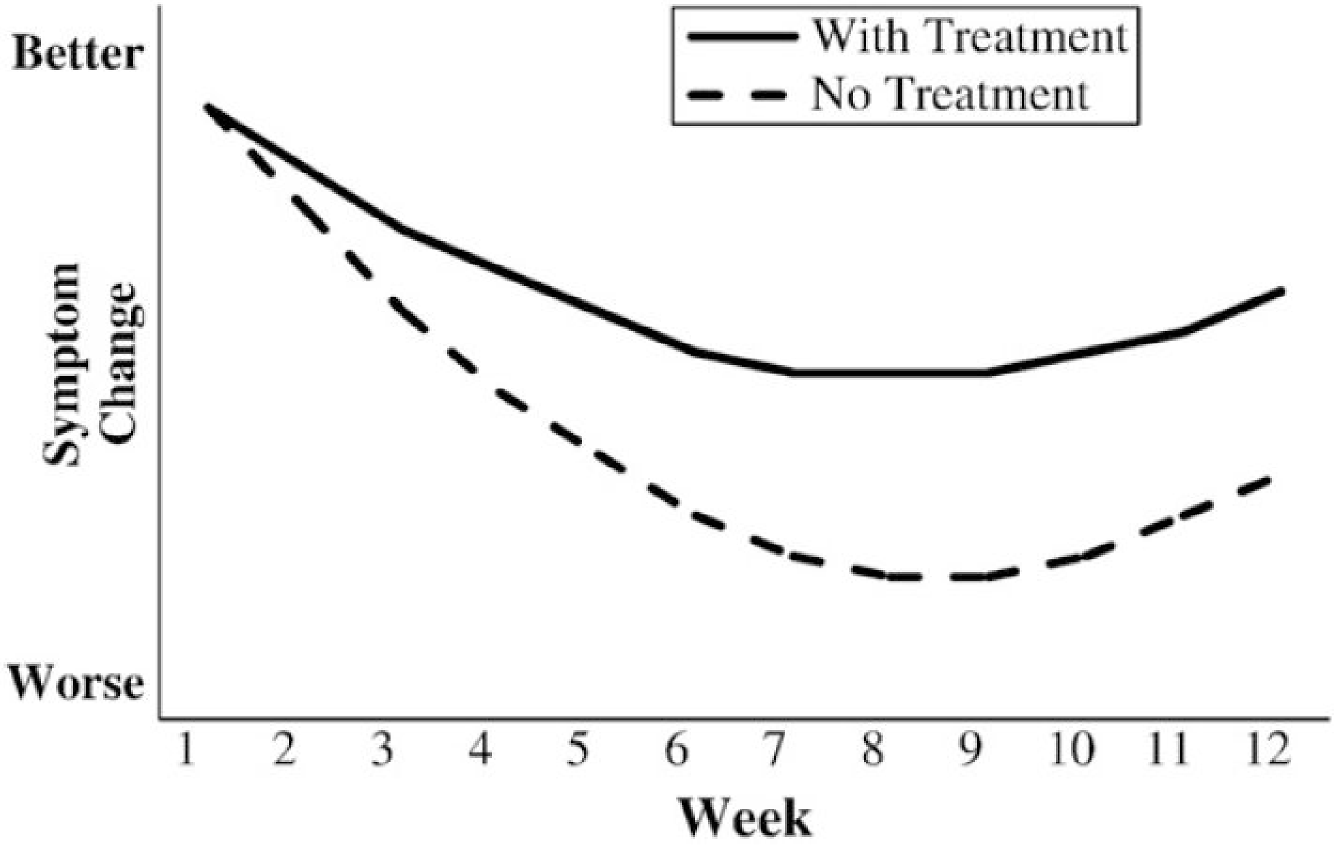

Figure 3.
Hypothetical disease courses for an individual patient treated with an effective medication.