Clinical trials methodology: Randomization, intent-to-treat, and random-effects regression

David C. Atkins
University of Washington

Published as (© Wiley):

Note: This article may not exactly replicate the final version published in the Wiley journal. It is not the copy of record.

Author Note:
David C. Atkins, Department of Psychiatry and Behavioral Science, University of Washington.
Correspondence concerning this article should be addressed to David C. Atkins, 1100 NE 45th St, Suite 300, Seattle, WA, 98105. datkins@u.washington.edu.
Clinical practice guidelines for psychiatric treatment, such as those by the American Psychiatric Association\textsuperscript{[1]} or Britain’s National Institute for Clinical Excellence,\textsuperscript{[2]} are defining the standard of care in terms of research evidence, broadly referred to as evidence-based practice (EPB). For the practicing clinician, this presents a challenge: What is the best way to keep up to date with the research base that informs EPB? Although there are a variety of ways to keep abreast with the empirical literature on psychiatric illness and treatment (e.g., continuing medical education, workshops, Cochrane reviews), clinicians who are serious about EBP will also read primary research sources. Here is the challenge. Research methods and statistics, as they evolve, steadily become more complex and farther away from introductory methods that clinicians may have learned in their training. This can lead to a chasm between cutting edge research and the practicing clinician. For clinicians reading primary research sources, how best to bridge this gap?

The present article provides an overview of several common features of randomized trials that clinicians may understand in part, or not at all. Specifically, I will focus on randomization and its connection to intent-to-treat analyses,\textsuperscript{[3]} as well as random-effects regression,\textsuperscript{[4]} which has become the standard statistical tool in analyzing clinical trial data. My goal is to provide a non-technical introduction to these topics while still presenting methodological nuances that regularly arise in clinical research. Additional resources are cited throughout for readers desiring more information. Hopefully, this brief review will help clinicians to be better consumers of clinical research.

\textit{Randomization and Intent-to-Treat}

What changes people? As psychiatrists and clinical psychologists, our goal is to help people change maladaptive behaviors and emotions in order to lead more functional, fulfilling
lives, and this is at the heart of clinical research. The randomized controlled trial (RCT) is widely considered the gold-standard in terms of research evidence. Understanding the motivating factors behind an RCT design will clarify the importance and limitations of the intent-to-treat principle. In reality, RCTs are not the optimal design for understanding the causal effects of a treatment. The ideal study would be to take a group of patients, randomly assign a treatment to them, observe their response to that treatment, then provide a second treatment, and observe their response under this second treatment. The difference in the patient’s response to the two treatments would be a strong, causal measure of the treatment effect.

Why can’t we do this study? The problem is that treatments change people. The experiment just described requires that we have an untreated, or naïve, patient prior to each treatment. In effective psychotherapy, learning and behavior change takes place, and this carries into a subsequent treatment, contaminating our ability to isolate the treatment effect. Even with pharmacological treatments, people’s expectations have been influenced by previous experiences with medications, and these expectations can influence treatment response (and likely lie at the heart of the placebo response). Furthermore, medications can sometimes produce longer lasting changes that persist for a time even after they are discontinued.

For these reasons RCTs are the primary design for testing new treatments, though they have an important limitation relative to the previous, cross-over design (in which patients randomly receive one treatment first, then stop and begin a second treatment). The beauty of having the same person receive both treatments is that all of the factors of the individual that could affect the treatment response are the same for both treatments, because the same person receives both treatments. For any given treatment, genetic factors, learning history, or contextual factors such as social support or socioeconomic status may affect treatment response. When the
same individual receives both treatments, the design effectively controls for these influences regardless of number or strength. With an RCT we make a difficult bargain.

The bargain is that we would like to have the same person take our treatment twice. Failing that, the next best alternative is to have identical people in each treatment group. For any pair of individuals, this is challenging to achieve. However, on average and in the long run, parity between treatment groups can be achieved using random assignment to treatment condition. That is, by randomly assigning patients to different treatment groups, over time the groups will, on average, be similar on all background characteristics. The bargain is really a tradeoff: Where we must make strong assumptions about lack of learning effects in cross-over designs, RCTs are only as good as their randomizations ability to yield similar groups. Some techniques adjust for differences in baseline characteristics (i.e., covariates) using either propensity analysis or, before the fact, stratifying the sample by key characteristics (e.g., presence of comorbidities) prior to randomization. (Some interesting, recent work on causal effects in treatment studies casts the causal effect as a problem in missing data (i.e., because each individual receives only one treatment and thus are “missing” on their response to the other treatment).\[5\])

This is the context in which the intent-to-treat (ITT) principle arose, and we can now consider its place and threats. Analyzing data on an ITT basis means that researchers analyze their data as randomized, regardless of what treatment was actually received. On the face of it, this seems sensible and uncontroversial. In fact, ITT has generated significant debate (reviewed [6]). Imagine a treatment for depression study in which a new drug is tested against fluoxetine (Prozac). Patients are randomized to receive one or the other treatment. Will every patient take the assigned med and continue with it for the duration of the study? Almost certainly not. Some
patients may experience side-effects or lack of improvement and discontinue their use of the assigned medication. Others may have been hoping to receive the experimental medication and drop-out when assigned to fluoxetine, which they can get at low-cost without all the hassles of a research study. Perhaps the most frustrating scenario is that some patients receiving the experimental medication might stop and begin a trial of fluoxetine. Thus, these patients are not receiving their randomly assigned treatment, but instead receiving the comparison treatment. Such scenarios are unfortunately quite common in clinical research.

Should data such as these still be analyzed using ITT? Almost certainly yes, as a first and primary analysis. There are arguments in favor of ITT based on actual data, mathematical proofs, and simulated data. However, it may be clearest to understand the preference for ITT analyses by considering an analogy to public policy. When we test a treatment, the question we are addressing is whether this medicine is safe enough and effective enough to use in routine practice – a policy statement. If the new medication is introduced in routine practice, its effect on the health of the population will be muted by most of the scenarios described above – some patients will have side-effects, some will prefer an alternative, some will discontinue due to lack of (or great) improvement. An ITT analysis essentially tests this policy perspective – will the medication overall have an effect, even considering lack of perfect compliance? For this reason, which also lines up with more detailed statistical reasons, ITT analyses are preferred and protect the overall causal statements available from the randomization process in RCTs.

Before leaving ITT it is worth considering what the impact of a treatment is for those who take it as they should. This, too, is a policy statement: What would the optimal effectiveness be of a new treatment? Statisticians have been actively engaged in precisely this
question, usually framed as estimating treatment effects in the face of noncompliance. This will work will likely make its way into clinical trials practice over the upcoming decade.\textsuperscript{[7,8]}

**Random-effects Regression**

The statistical methods for analyzing RCTs and other treatment designs have changed notably over the last 20 years. Historically, statistical analyses of treatment data focused on treatment differences at the end of therapy, or treatment differences in the change from beginning to end of therapy. These simple designs could be analyzed using a t-test for the treatment difference at the end of therapy or analysis of covariance, which controls for pre-therapy.\textsuperscript{[9]} These designs and their related statistical methods have a number of shortcomings. For example, they tell us nothing about the nature of change during therapy, and recent treatment studies reveal that different therapies and therapy combinations have different treatment trajectories.\textsuperscript{[10]} Multiple assessments during treatment are required to assess the trajectory of change, which quickly moves us beyond the aforementioned statistics.

Another aspect of treatment data with important ramifications for analysis concerns balance – balance in terms of timing of assessments and the total number of assessments. Imagine in your clinical practice that you begin a medication with a new patient. You ask the patient to return in two weeks to assess how their symptoms might be changing and whether they are having any difficulties with the medication. Will they return in two weeks? Some percentage of patients will return in two weeks time, but some will come later, a few earlier, and others not at all. This will happen for all of the natural reasons of life (e.g., vacations, business obligations) as well as reasons related to their psychiatric problems (e.g., depression makes it challenging for them to travel to your office). The point is that even though the assessment was scheduled for two weeks time, it could occur at one week or three weeks or not at all. This difference between
planned and actual assessments also occurs in RCTs, even with the well-meaning reminders of research staff.

The data just described are considered imbalanced in timing and number. That is, our research design may call for assessments at pre-therapy (i.e., week 0) and every two weeks thereafter until the trial ends at week 16, but our patients actual assessments will occur at somewhat – hopefully not radically – different times. This causes problems for classical statistical methods such as repeated measures analysis of variance (ANOVA). Conceptually, the issue rests on the fact that ANOVA compares means at different times points. If the time points vary, how do we calculate the means? Thus, an assessment scheduled for week two that actually occurred at three and a half weeks would need to be treated as though it happened at week two, so that it could be averaged with other week two values. Does this matter? Let’s assume that our patient meets criteria for a major depressive episode and fluoxetine was begun, which is known to effectively ameliorate depressive symptoms (on average). Moreover, depressive symptoms tend to change rapidly, early in treatment for many patients.\[11\] Thus, an assessment at week three and a half will be a biased estimate of depressive symptoms at week two. The imbalance in timing combined with repeated measures ANOVA would lead to a biased analysis.

What is more problematic is missing data – data points that are not collected at all. It is simple enough to take averages at different time points, even if the total number of data points varies across time points. However, repeated measures ANOVA runs into problems with its underlying mathematics when there are varying numbers of data points at different time points.\[12\] Hence, the standard course of action is that any patient with any missing data is removed from the analysis. Clearly, this is a suboptimal solution.
For these reasons and others, classical statistical methods (i.e., t-tests, analysis of covariance, and repeated measures ANOVA) have largely been replaced by random-effects regression models (RRM) in medical RCTs\textsuperscript{[12,13]} (also referred to as mixed-effects models or multilevel models\textsuperscript{[14]}). There are a number of advantages of RRM over classical statistics, but for our purposes, a critical difference is that RRMs focus on trajectories, not means. Whereas the individual gets averaged over in the means of repeated measures ANOVA, RRMs construct estimates of how each individual changes across time. These trajectories are often straight lines with an intercept and slope, though can take more complex, nonlinear functions.

Importantly, the RRM approach to trajectories does not require balanced data. To fit a straight line requires a minimum of two data points (and preferably more), but the timing of the assessments does not fundamentally cause problems. Thus, a week two assessment that occurs at three and a half weeks can be included at the actual timing of the assessment. Now, what about missing data? RRMs are able to use all available data (i.e., a straight line trajectory can be based on two or five or seven data points), and thus are not impaired by missing data in the same way that classical statistics are. In fact, simple statements in research studies routinely note that missing data were not a problem due to the use of RRMs. (It is worth noting in passing that trajectories from RRMs include the reliability of each individual’s data, and individuals with more data have more reliable trajectory estimates. These estimates are referred to as empirical Bayes estimates\textsuperscript{[15]})

Nonetheless, RRMs are not a panacea for all missing data, as we shall briefly consider. Earlier, when discussing ITT analyses, it was mentioned that patients drop out of treatment for a variety of reasons, including improvement or deterioration. Statisticians have proposed certain assumptions about why data are missing, with technical specifications (reviewed in [12,13,14]),
but we can understand the concepts through three simple examples. One scenario is that missing data are driven by the natural vicissitudes of life (e.g., vacations and business obligations), and these are equally likely for any patient in the study. In this scenario, some patients’ trajectories from RRM may be based on one or two fewer data points, but the trajectories themselves are likely accurate estimates of what they would have been without any missing data. A second example occurs when a patient drops out of an RCT due to lack of improvement (or sometimes large improvement). This is notably different than a vacation, as a drop-out implies missing data from a particular point in time forward into the future. However, even here, RRM can often provide reasonable estimates. If we have a series of measurements prior to drop-out, the patient’s trajectory (worsening or improving) may be reasonably approximated by the available data. The final missing data scenario is more pernicious but not wholly unusual. Imagine a patient who is improving in therapy. However, something occurs that triggers a major depressive relapse (e.g., an abusive ex-spouse reappears, lost job, etc.), and they subsequently drop out of the study. Here, we now have a problem because the trajectory specified by the RRM does not reflect the true trajectory, had the patient continued to provide data. In this scenario RRM will provide biased results.

The three missing data scenarios broadly map on to the statistical definitions of missing completely at random, missing at random, and missing not at random, where the last category is often called non-ignorable, as we are not able to ignore the missing data process in the data analyses. Thus, RRM can effectively handle missing data in the first two categories (i.e., “ignorable” missing data) but will provide biased estimates in the final category. Considerable statistical work is currently being devoted to this final category, where pattern-mixture models condition the analysis on patterns of missing data and shared parameter models simultaneously.
model the trajectory and drop-out processes.\textsuperscript{[12,14,16]} However, each of these new methods makes assumptions that can be tenuous.\textsuperscript{[16]} Needless to say, RCTs that use RRM$s need to consider carefully which scenario best represents their missing data.

\textit{Conclusion}

In this brief overview, I have covered some of the basic design characteristics of RCTs involving randomization, analytic principles (i.e., ITT), and specific analytic methods (i.e., RRM$s). I hope that presenting both the reasoning behind these RCT facets and some of the nuance of how patients and data behave in the real world will assist clinicians in being more educated consumers of the research literature.
References

1. American Psychiatric Association (2009). Practice guidelines. Available at: 
   http://www.psych.org/psych_pract/treatg/pg/prac_guide.cfm

   http://www.nice.org.uk/Guidance/CG/Published


10. Vitiello, B. Treatment of adolescent depression: What we have come to know. 2009;Depression and Anxiety, 26: 393-395.


