CHAD E. COOK, PT, PhD, FAPTA[1,2,3] • GERONIMO BEJARANO, BS[4] • JENNIFER RENEKER, MPT, PhD[5]
ANDREW D. VIGOTSKY, MS[6] • DANIEL L. RIDDLE, PT, PhD, FAPTA[7,8]

Responder Analyses: A Methodological Mess

Randomized controlled trials (RCTs) assess the average treatment effects of 1 or more interventions against 1 or more comparators/controls. These designs are regarded as the gold standard when determining an overall causal effect for a treatment. Although an RCT identifies the average effect of an intervention, it does not identify individuals who may have experienced "clinically meaningful" improvement *because* of that treatment.[8] To identify those who experience a meaningful improvement, researchers often conduct secondary "responder analyses" in RCTs.

Responder analyses identify participants who achieved a predefined level of improvement on at least 1 outcome. Responder analyses are endorsed in many quarters,[3] yet responder analyses have several methodological shortcomings, which in our view should preclude their use. In this Viewpoint, we explain our concerns that responder analyses lend their findings to be at best unusable and, at worst, misleading.

SYNOPSIS: Responder analyses are methods for analyzing randomized controlled trials, which purport to identify individuals or subgroups of study participants who experienced a "clinically meaningful" improvement from a treatment. Unfortunately, responder analyses have numerous methodological shortcomings, which preclude inferences concerning individual response to treatments and, thus, adoption into clinical practice. In this Viewpoint, we summarize 2 major limitations of responder analyses: (1) their thresholds of success involve arbitrary criteria and (2) responder analyses do not capture true individual treatment effects. *J Orthop Sports Phys Ther 2023;53(11):652-654. Epub 20 June 2023. doi:10.2519/jospt.2023.11853*

KEY WORDS: *minimal clinically important difference, responder analyses, randomized controlled trial*

Concern 1: Responder Analyses Are Often Based on Arbitrary Criteria

Responder analyses commonly classify participants into 1 of 2 groups (responders or nonresponders) based on a minimal clinically important difference (MCID) score.[4] There are different ways to determine MCID-based responder thresholds, including anchor-based methods, distribution-based methods, consensus-based methods, expert opinion, and composite-based methods. When calculating an MCID, researchers use 1 of 3 numeric thresholds: (1) a single threshold identified as meaningful (ie, ≤ 20/100 for the Oswestry scale), (2) a percentage change in the outcome of interest that is calculated by comparing later outcomes to baseline (ie, a 30% improvement in the Oswestry scale), or (3) an endorsed absolute change from baseline (ie, a 12-point change in the Oswestry from baseline). These methods do not converge on the same MCID for a given outcome measure, leading to discordant recommendations of responders.[2] This can also lead to confusion across different studies using disparate values and difficulty reconciling findings across studies.

There are also concerns with the measurement properties of the MCID.[2,5,6] These include (1) recall bias (participants are more heavily influenced by their current state than their state at baseline when they are asked to estimate how much they have improved or worsened); (2) many MCID estimates only consider subsets of participants rather than the entire sample; (3) distribution of the outcome data substantially influences MCID estimates; (4) the reliability of global rating scales used in MCID estimation is poor, ranging from 0.27 to 0.48; and (5) mapping the outcome scale onto another scale implies the latter scale

[1]Department of Orthopaedics, Division of Physical Therapy, Duke University, Durham, NC. [2]Department of Population Health Sciences, Duke University, Durham, NC. [3]Duke Clinical Research Institute, Duke University, Durham, NC. [4]Department of Epidemiology, University of Texas Health Science Center (UT Health), Austin, TX. [5]School of Population Health, Department of Population Health Science, University of Mississippi Medical Center, Jackson, MS. [6]Departments of Biomedical Engineering and Statistics, Northwestern University, Evanston, IL. [7]Department of Physical Therapy, Virginia Commonwealth University, Richmond VA. [8]Department of Orthopaedic Surgery and Rheumatology, Virginia Commonwealth University, Richmond, VA. ORCID: Vigotsky, 0000-0003-3166-0688. No external or internal funding was associated with this study. This paper is neither a systematic review nor a trial and, thus, was not registered. The authors certify that they have no affiliations with or financial involvement in any organization or entity with a direct financial interest in the subject matter or materials discussed in the article. Address correspondence to Chad E Cook, Department of Orthopaedics, Division of Physical Therapy, Duke University, 311 Trent Dr, Durham, NC 27110. E-mail: chad.cook@duke.edu

is of principal interest, questioning the need for the former scale.

Beyond our concerns about how an MCID is derived, we have concerns about interpreting the defined threshold values.[2] Dichotomizing continuous or ordinal data—the process used in most responder analyses—leads to a substantial loss of information and can reduce statistical power.[9] By discarding information other than whether points are above or below an arbitrary threshold, dichotomizing continuous variables obfuscates the observed relation between the treatment variable and the outcome, potentially leading to either inflated or artificially reduced effect sizes. Dichotomizing data may lead to miscategorizations (false positives and false negatives) if there is truly some latent dichotomous outcome, or misspecification if there is no latent dichotomy, meaning that the model does not accurately reflect the underlying data or data-generating process.[1] If a 30% improvement (and higher) is defined as the "clinically meaningful" value, a person who exhibits a 29% improvement would not be considered as having a "clinically meaningful" change. A 90% improvement would have the same "clinically meaningful" interpretation as a 30% change. By subscribing to binary MCID thresholds, people are either responders or nonresponders with no in-between or gradation.

Concern 2: Responder Analyses Do Not Capture True Treatment Effects for Individual Patients

The common assumption of responder analyses is that observed improvements in a participant's outcome are caused by the treatment provided because they occur after the treatment was implemented. However, nearly all methods of defining responders involve "within-subject" analysis. The analyses are influenced by natural history, measurement error, extratherapeutic effects from other treatments taken by the person, and other factors unrelated to the treatment such as contextual factors (ie, components of therapeutic encounter such as therapeutic alliance or social and physical environment that substantially influence clinical outcomes of a treatment intervention).[11]

Perhaps the most critical concern with using within-subject analysis is regression to the mean—a statistical phenomenon in which subsequent early extreme values will regress to a common mean when evaluated later. Researchers might end up drawing the wrong conclusion that a measured effect (difference between prescores and postscores) is due to treatment when it is actually due to chance.[7] Because of these additional potential contributors to participant outcomes, it is impossible to know if the change is related to the treatment or due to something else.[11] Statistician Stephen Senn[10] describes this methodological pitfall as "subsequence, not consequence." In other words, the observed change in the outcome may have nothing at all to do with the treatment provided. Importantly, this pitfall is a concern regardless whether an outcome is truly dichotomous or not.

Most responder analyses only quantify a proportion of individuals in a treatment group who met a "clinically meaningful" threshold of improvement following an intervention. In other words, they report the proportions of people in each group who "responded". This method fails to identify "who" may truly benefit from the treatment. Less often, responder analyses assess the unique baseline characteristics of individuals who improve in 1 group (through some form of regression analysis) versus those who improve in the comparator group. This method of responder analysis is more likely to identify characteristics of individuals who may benefit from 1 treatment over another. Further, responder analyses rarely use planned subgroup analyses to examine how average treatment effects differ based on an attribute of interest—this form of analysis may lend value as well.

Summary

The number of papers discussing responder analyses has exploded (over 33,000 papers per year and greater than 580,000 overall in a recent PubMed search). Groups such as the National Institute of Health; the National Cancer Institute; the Federal and Drug Administration; the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT); the Outcome Measures in Rheumatoid Arthritis Clinical Trials - Osteoarthritis Research Society International (OMERACT-OARSI) group; and the Institute of Medicine have widely endorsed responder analyses.[3] We believe that the rapid growth and almost universal endorsement should concern clinicians, researchers, and policymakers alike.

Responder analyses aim to identify who will preferentially benefit from the experimental treatment. We argue that responder analyses fail to identify who benefits from a specific treatment because the methodology is flawed and the interpretation is misleading. Current responder analyses do not correctly help the reader interpret the results of RCTs and do not identify the individuals who are likely to improve *because* of a treatment they have received. This and the other methods and measurement problems we discussed have led us and others to seriously question the validity of responder analyses.

Key Points

- Responder analyses are advocated by many groups and are purported to improve our understanding regarding who benefits from the treatment provided in a clinical trial.
- Two major methods and measurement problems render responder analyses as questionable tools for interpreting clinical trials.
- Responder analyses oversimplify patient outcomes by classifying people as responders or nonresponders, which can lead to false interpretations ◉

STUDY DETAILS

AUTHORS CONTRIBUTIONS: All authors were involved in the conception and design of the work. All authors were involved in the drafting of the article. All authors

were involved in the critical revision of the article. All authors approved the final version of the paper.

DATA SHARING: There was no data involved in the paper.

PATIENT INVOLVEMENT STATEMENT: There were no patients involved in the design, interpretation, or development of this work.

REFERENCES

1. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ.* 2006;332:1080. https://doi.org/10.1136/bmj.332.7549.1080
2. Cook CE. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. *J Man Manip Ther.* 2008;16:82E-83E. https://doi.org/10.1179/jmt.2008.16.4.82E
3. Cook CE, Bejarano G, Reneker J, Vigotsky A. Responder analyses in musculoskeletal research. In: Jull G, Moore A, Falla D, Lewis J, McCarthy C, Sterling M, eds. in *Gregory Grieve's Modern Musculoskeletal Medicine*. Elsevier; 2023.
4. Ferreira GE, McLachlan AJ, Lin CC, et al. Efficacy and safety of antidepressants for the treatment of back pain and osteoarthritis: systematic review and meta-analysis. *BMJ.* 2021;372:m4825. https://doi.org/10.1136/bmj.m4825
5. Griffiths P, Terluin B, Trigg A, Schuller W, Bjorner JB. A confirmatory factor analysis approach was found to accurately estimate the reliability of transition ratings. *J Clin Epidemiol.* 2022;141:36-45. https://doi.org/10.1016/j.jclinepi.2021.08.029
6. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA.* 2014;312:1342-1343. https://doi.org/10.1001/jama.2014.13128
7. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *BMJ.* 2003;326:1083-1084. https://doi.org/10.1136/bmj.326.7398.1083
8. Mulder R, Singh AB, Hamilton A, et al. The limitations of using randomised controlled trials as a basis for developing treatment guidelines. *Evid Based Ment Health.* 2018;21:4-6. https://doi.org/10.1136/eb-2017-102701
9. Rhon DI, Teyhen DS, Collins GS, Bullock GS. Predictive models for musculoskeletal injury risk: why statistical approach makes all the difference. *BMJ Open Sport Exerc Med.* 2022;8:e001388. https://doi.org/10.1136/bmjsem-2022-001388
10. Senn S. Statistical pitfalls of personalized medicine. *Nature*. 2018;563:619-621. https://doi.org/10.1038/d41586-018-07535-2
11. Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin.* 2009;25:929-942. https://doi.org/10.1185/03007990902774765

@ MORE INFORMATION
WWW.JOSPT.ORG