How Strong Should My Anchor Be For Estimating Group And Individual Level Meaningful Change? A Simulation Study Assessing Anchor Correlation Strength And The Impact of Sample Size, Distribution of Change Scores And Methodology On Establishing A True Meaningful Change Threshold.

Philip Griffiths
IQVIA Reading: IQVIA Ltd
Joel Sims (joel.sims@adelphivalues.com)
Adelphi Values https://orcid.org/0000-0003-0474-2186
Abi Williams
Adelphi Values
Nicola Williamson
Adelphi Values
David Cella
Northwestern University Feinberg School of Medicine
Elaine Brohan
Adelphi Values
Kim Cocks
Adelphi Values

Research Article

Keywords: minimal important change, anchor correlation, meaningful change threshold, anchor relationship

Posted Date: July 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-384605/v1

Abstract

Purpose: Treatment benefit as assessed using clinical outcome assessments (COAs), is a key endpoint in many clinical trials at both the individual and group level. Anchor-based methods can aid interpretation of COA change scores beyond statistical significance, and help derive a meaningful change threshold (MCT). However, evidence-based guidance on the selection of appropriately related anchors is lacking.

Methods: A simulation was conducted which varied sample size, change score variability and anchor correlation strength to assess the impact of these variables on recovering the true simulated MCT at both the individual and group-level. At the individual-level, Receiver Operating Characteristic (ROC) curves and Predictive Modelling (PM) anchor analyses were conducted. At the group-level, group means of the 'not-improved' and 'improved' groups were compared.

Results: Sample sizes, change score variability and magnitude of anchor correlation affected accuracy of the estimated MCT. At the individual-level, ROC curves were less accurate than PM methods at recovering the true MCT. For both methods, smaller samples led to higher variability in the returned MCT, but higher variability still using ROC. Anchors with weaker correlations with COA change scores had increased variability in the estimated MCT. An anchor correlation of 0.50-0.60 identified a true MCT cut-point under certain conditions using ROC. However, anchor correlations as low as 0.30 were appropriate when using PM under certain conditions. At the group-level, the MCT was consistently underestimated regardless of the anchor correlation.

Conclusion: Findings show that the chosen method, sample size and variability in change scores influence the necessary anchor correlation strength when identifying a true individual-level MCT. Often, this needs to be higher than the commonly accepted threshold of 0.30. Stronger correlations than 0.30 are required at the group-level, but a specific recommendation is not provided. Results can be used to assist researchers selecting and assessing the quality of anchors.

Introduction

Statistical significance alone does not necessarily reflect a treatment benefit that is meaningful from the patient perspective. Interpretation of meaningful change from the patient perspective is therefore important when assessing concepts measured by clinical outcome assessments (COAs) [1, 2]. Such thresholds, though differently derived at the within-individual or group comparison level, can be used for assessing the treatment benefit experienced. This can occur either in a healthcare setting, in discussions with physicians or in a clinical trial to define responders or assess mean differences in treatment and comparator arms for assessing the potential treatment efficacy of a new product.

Anchor-based methods are the preferred approach employed to aid the interpretation of meaningful within-individual changes in COA scores [1, 2]. Anchors provide an external indicator to classify patients into groups representing the degree of change in their overall disease condition or the specific concept of interest. The Food and Drug Administration (FDA) recommend that anchors should be simply worded,

easy to understand and assess a specific concept [3]. Multiple anchors can be used to aid interpretation and selection of an appropriate meaningful change threshold at an individual-level or group-level depending on the methods employed [4]. When conducting anchor-based analyses, the FDA Patient Focused Drug Development (PFDD) guidance recommend use of both empirical cumulative distribution function (eCDF) and probability density function (PDF) curves to aid identification of an appropriate meaningful change threshold [5]. Importantly, for anchor-based analyses to be possible, external anchors should be sufficiently correlated with the COA of interest [4].

The strength of the correlation between the external anchor(s) and target COA, however, remains a topic of uncertainty and debate. Cohen's effect size (*d*) has frequently been used as a reference for choosing a suitable correlation coefficient (*r*) [2]. A large effect size for Cohen's *d* (i.e., the difference in size of two means) is 0.80 with the corresponding value of *r* being 0.371 (i.e., effect size indicating the degree of relationship between two sets of scores) [6]. Based on this, published literature recommends a correlation of 0.30–0.40 as an appropriate threshold [2, 7], while others have chosen higher values such as 0.50 despite this being acknowledged as an arbitrary cut-point [8–10]. Coon and Cappelleri (2015) have recommended a correlation of 0.40–0.70 as preferred [1]. Amidst the mixed recommendations for correlation thresholds, empirical reports continue to suggest a correlation of 0.30–0.40 as an acceptable lower-limit of the level of association between an external anchor and the target COA [11]. The FDA does not explicitly recommend a correlation threshold, with little guidance on how researchers should assess the strength of anchor correlations [3]. Therefore, the responsibility lies with researchers to evaluate the suitability of anchors prior to conduct of anchor-based analyses.

The strength of the relationship may affect the level of accuracy and reliability of results, alongside the meaningful change threshold that is ultimately selected. A poor correlation between an external anchor and the target COA can increase error in the derived threshold estimates [1]. Any value of *r* below 1.0 leads to attenuation of the meaningful change threshold, as demonstrated previously at the group-level [12]. Although limitations such as correlation strength and sample size have been noted [11], how thresholds may be affected by varying anchor correlations, sample size and distribution of change scores, remains empirically unexplored. Therefore, there is an unmet need to provide clearer guidelines on the suitability of anchors for anchor-based analyses. This has significant implications for studies aiming to aid the interpretation of COA meaningful change thresholds, both at the individual and group-level. To address this gap in the literature, we conducted a simulation incorporating different samples of varying change score distributions and anchor correlation strengths to assess the impact of these factors on estimating a meaningful change threshold (MCT) at both the individual and group-levels.

Methods

The methods for this simulation study were developed in accordance with best-practice guidelines for conducting simulation studies to evaluate statistical methods [13].

3.1 Simulation of data (Data Generating Mechanisms)

Data were simulated separately for the individual and group level meaningful change analyses. For each analysis, several conditions were created which varied the distribution of change scores and the sample size. Different conditions tested a range of between 100 and 2,000 patients, with differing variability (Standard Deviation scaled in relation to the simulated mean) and strength of correlations between the simulated anchor and target score (from 0.30 upwards). Separate simulations were conducted for the individual and group level scenarios. The simulations were different in order to keep an underlying meaningful change threshold of 15 points both at the individual level (where this represents the responder definition that specifies if a patient has improved) and group level (where this represents the mean difference between groups of not-improved and improved patients).

3.1.1 Individual level

For the individual level meaningful change data, sample sizes of either 100, 250, 500 or 2000 patient records were created. For each patient, a COA change score was developed by drawing a random number from a normal distribution with a mean of 15 and, depending on the condition, a standard deviation (SD) of 3.0, 5.0 or 7.5 to represent different levels of deviation. For each patient, anchor variables were created which correlated with the COA change score at 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and ~1.0. This was done by drawing a random number from a normal distribution with a mean equal to the sample mean and a SD equal to that of the sample SD and multiplying it by:

$$rho \times COA\ change\ score + \sqrt{1 - rho^2}$$

Where *rho* represents the simulated level of correlation between the COA change score and anchor variable.

This led to a series of variables which were correlated at the respective levels but had an inflated mean. A correction was applied to each sample to return the sample mean of the anchor variables to the scale of the original COA change score. This correction used the difference between the mean COA change score and the mean of each simulated anchor variable to return the simulated anchor variable to the original scale.

For each patient in each sample, a variable was created to represent patient improvement status. This variable was created based on the simulated anchor variables. Due to how the data was simulated, the simulated COA change score had a mean of 15, this value was used as the “true” cut point to determine change. Where anchor scores fell above 15, patients were defined as “responders” and where anchor scores fell below this value, patients were defined as “non-responders”.

3.1.2 Group level

Simulation of group level data followed the same methodology to individual level data (i.e., same sample sizes and standard deviation conditions) but with some small differences. A cohort of patients were simulated from a single normal distribution, with a mean of 7.5[1] and a SD of 3.0, 5.0 or 7.5. Simulated anchors were derived in the same manner as described above for individual-level change. Patients were then re-classified into “improved” and “not improved” groups using the simulated anchors. Given that the overall mean for the COA change score in each sample was 7.5, this value was used to classify patients on the anchors. When a patient’s simulated anchor score was ≤7.5 they were defined as “not improved”; where a patient’s simulated anchor was >7.5, they were defined as “improved”. A threshold of 7.5 is arbitrary, but allowed for relatively equal group sizes. The “true” difference in mean change between these groups was determined as the difference seen when patients were grouped by a perfectly correlated anchor, and this threshold was use for comparison of all other anchors with a correlation < 1.

3.2 Recovering the meaningful change threshold (MCT)

3.2.1 Individual level

To recover the MCT (that is, to output the estimated MCT for each simulated condition) at the individual level, a series of Receiver Operating Characteristic (ROC) curve-based and Predictive Modelling-based (PM) analyses were conducted. Briefly, using the simulated anchor groups (“non-responders” and “responders”), the ROC method identifies the number of “non-responder” and “responder” patients who are correctly and incorrectly classified by each possible threshold. For each sample, the COA change score which most accurately represents the ROC curve point closest to the top-left hand corner of the plot is used as the estimate for the responder definition. This is calculated for each COA change score using the following formula, with the smallest resulting value being selected as the responder definition:

$$\sqrt{(1 - se_{cutpoint})^2 + (1 - sp_{cutpoint})^2}$$

Where *se* represents sensitivity of the cut point and *sp* represents specificity.

The PM approach, on the other hand, uses logistic regression to predict the COA change score at which patients are most likely to be accurately classified into two groups [14]. This logistic regression model uses the improvement status on the anchor (“responder” vs “non-responder”) as the dependant variable and the COA change score is the independent variable. The PM approach then uses the regression output in the following formula, which takes into account the prevalence of the number of responders:

$$(\log\left(\frac{p}{1-p}\right) - C)/\beta$$

Where p is the proportion of patients in the improved group (responders) expressed as a decimal, C is the intercept from the logistic regression and β is the regression coefficient.

Once both of these analyses had been completed for each of the 1000 samples in each of the 12 conditions, the recovered cut points were displayed graphically using a histogram with the recovered threshold along the x-axis and number of samples the threshold was recovered in on the y-axis. This was done in order to assess accuracy and variation around the “true” simulated threshold.

3.2.2Group Level

To recover the MCT at the group level, using the groups based on the simulated anchor, the difference between the group means of the “not-improved” patients and “improved” patients was calculated for each sample in each condition (correlation with simulated anchor, standard deviation of change score and sample size). Recovered MCT frequencies for each of the 12 conditions (sample size*change score variability) were graphed using probability density functions, with the difference in change score on the x-axis and the density on the y-axis. All simulated anchor conditions were displayed on the same plot, with the *r*=1 condition shown for comparison of the other anchors against a “true” change. This was included so that anchors with a correlation <1 could be compared to what may be expected if a perfect anchor existed.

[1] Initially, two groups were simulated: a “not-improved” group from a normal distribution with a mean of 0 and an “improved” group from a normal distribution with a mean of 15. However, this les to a distribution with an overall mean of 7.5, but a bimodal distribution. We feared that this may have influenced the results so re-simulated this based on a single normal distribution with a mean 7.5. The results were the same in both scenarios. The approach presented here makes it easier for the reader to follow technical discussion of the results presented in the discussion.

Results

4.1 Individual level

Results showed that each of the variables controlled for in this simulation study (i.e., magnitude of correlation coefficient, sample size and variability of change score) played a role in the accuracy of the recovered MCT. As can be seen in Fig. 1 - Fig. 4, ROC curve-based methods were less accurate than PM-based methods at recovering the true MCT, as shown by the spread of the recovered MCT around the expected value of 15.

Further to this, smaller sample sizes led to higher variability in the returned MCT for both methods, but higher variability still for the ROC methods. Anchors which had a low correlation coefficient with the COA change score also led to increased variability in the returned MCT. Interestingly, although larger COA change score SDs led to a less accurate estimate of the MCT (particularly for ROC curve-based assessments), at the highest level of variability (SD = 7.5), for sample sizes less than 2000, there was some small variability in the returned MCT even when the anchor had a *near perfect* correlation with the

COA instrument (r = ~ 1). This is due to the variability inherent in performing such analyses on small samples with a large variance to estimate ratio.

4.2 Group level

When patients were split into groups based on a correlated anchor, results showed that the group level MCT was consistently underestimated. For example, an anchor correlated at 0.30 typically returned a value of 2.0–3.0 and an anchor correlated at 0.60 typically returned a value of 4.5–5.5 (Fig. 5). This compares to a "true" MCT of just under 8.0. Although the variability around these estimates changed with sample size and COA change score variability, the pattern remained the same (Fig. 6). Of note, the "true" MCT also changed as the SD of the COA change score increases. This is expected and will be discussed.

Discussion

This study aimed to provide some answers to the unmet need for clearer guidelines about what level of correlation between external anchor(s) and COAs can be considered sufficient to conduct anchor-based analyses and produce an appropriate MCT that reflects a true meaningful change. While prior studies have explored the effect of anchor correlations and sample sizes on MCTs at the group-level this has not been assessed at the individual-level. The overall objective of this study was to conduct a simulation to explore the degree to which various anchor, sample and change score conditions may influence selection of a true MCT at both the individual-level (responder definition; RD) and group-level. Different anchor correlations were assessed while evaluating different sample sizes and distributions of change scores to explore the effect on MCTs.

Overall, the results show that an ideal anchor correlation of 0.50–0.60 may be appropriate to identify a true MCT cut point at the individual-level when using ROC analyses. Notably, this range is dependent on the sample size and variability of the change score under assessment. Smaller samples and larger variability in the COA change score reduces the reliability of the results produced here using ROC analyses, regardless of anchor correlation. Although not assessed in this study, Terluin et al. (2020) described how the MCT derived from ROC curve analyses may be biased when anchor groups are unequal in size [15]. However, here we show that PM methods (as described elsewhere) [14] were better able to accurately represent the true MCT at the individual-level, even with weaker anchor correlations. Even correlations at the lower end of the literature recommended level (r = 0.30) provided PM results more or less equivalent to what was obtained with ROC analysis at a correlation of r = 1.0. In addition, the PM approach provided robust MCT estimates despite small sample sizes and high variability in COA change scores. These factors appear to reduce accuracy of ROC estimates more significantly in comparison to PM estimates.

At the group-level, any level of anchor correlation lower than *near perfect* (i.e., r = ~ 1.0) led to underestimation of the true MCT. This is theoretically unsurprising, given that an anchor which is used to divide patients into two groups will introduce noise and error into this division at a rate that is inversely

related to the magnitude of its correlation. Therefore, a weakly correlated anchor will be worse at correctly classifying patients into groups of those who have, and have not, improved. The result of this is not only a widening of the distribution of change scores in each of the “not-improved” and “improved” groups, but crucially, a shift in the group means. As more patients are misclassified by the anchor, their associated COA change score used to derive the MCT is also incorrectly classified. This directly leads to a shift in the mean of each group, whereby the erroneous inclusion of improved participants in the “not-improved” group increases the not-improved group’s mean and vice versa for the “improved” group. Differences in group means are therefore increasingly diminished as misclassification increases leading to an underestimated MCT. Relative to the true MCT, the returned MCT for any given correlation appears proportional to the magnitude of the correlation. These results appear similar to the attenuation of group-level MCTs as described elsewhere when linear regression techniques are used [12]. In the case of linear regression, r is directly incorporated into the calculation of the MCT such that any value of r below 1.0 leads to an underestimate of the true MCT, a method argued against by Fayers and Hays (2014) [12]. In addition, this analysis was shown to be heavily influenced by the SD of the COA change scores. Even with a perfect anchor, a wider distribution of change scores leads to a larger difference between groups of “improved” and “not-improved” patients. This is because the mean of “improved” and “not-improved” groups also diverges as the change score SD increases. Although this, in itself, is not an issue, it is then compounded with the strength of the anchor correlation. A poorly correlated anchor can underestimate a between-group MCT much more substantially when the “true” MCT is higher by virtue of a large change score SD.

Although the results presented here were stark and may require researchers to reconsider the MCT practices they employ, they were confirmed through the use of an alternate procedure for calculating the anchor correlations (not presented here). The alternate method used a Cholesky decomposition rather than the algebraic formula to specify the intended correlation between the COA change score and the anchor score.

The FDA PFDD guidance recommends use of both eCDF and PDF curves to aid identification of an appropriate MCT [5]. While this recommendation is clear, the PFDD guidance lacks specificity on what anchor-based analyses can be considered acceptable. While the PFDD guidance does make the distinction between individual-level between-patient change and between-group mean differences, in both the PFDD and PRO guidance documents there is lack of specific individual-level analyses which can be used to determine meaningful change, focusing only on assessment of group-level change and applying this level of change to define individual-level RDs [1]. As illustrated in this study, a threshold at the group-level is likely to be very different to a threshold considered appropriate at the individual-level and as such individual and group-level thresholds cannot, and should not, be considered interchangeable. Results from this study indicate that PM anchor analyses were more reliable and less susceptible to small sample sizes and variability in COA change scores. However, if sample sizes are not prohibitive, somewhat reliable results can also be produced using ROC analyses. Selection of an appropriate anchor-based analyses should be based on these factors.

Regardless of the analysis selected however, a sufficient correlation should ideally be demonstrated between the external anchor(s) and target COA. As shown here 'sufficient' will depend on whether meaningful change will be assessed at the individual or group-level. However, it is acknowledged that in reality achieving a 'sufficient' correlation between the external anchor(s) and target COA may be challenging and may result in some level of circularity whereby only self-report external anchors that are assessing the same concept of interest as the COA correlate at a sufficiently high enough level, ruling out the possibility of using more clinical assessments as external anchors. In practice, using multiple anchors that may include a combination of clinical assessments and self-reports and accounting for the level of correlation between the external anchor(s) and target COA when triangulating across multiple anchors, is recommended.

It is important to note limitations of the study when interpreting the results. The simulations presented here only included one "not-improved" and one "improved" group (representing the "minimally improved" patients) as it is these two groups which typically define meaningful change. One problem with this is that there was no opportunity for an imperfectly correlated anchor to "misclassify" patients outside of this range. It is likely that in real studies, imperfectly correlated anchors misclassify patients who have worsened as "not-improved" and misclassify patients with a moderate improvement as "minimally improved". This could have some effect on controlling the shift in the means observed in this study. However, further simulations conducted alongside this study (not reported here) have assessed the impact of this more realistic situation and have not led to a solution to the issues observed here. These additional simulations have involved groups of patients outside of the minimally improved and not-improved groups (those who have worsened or improved to a greater degree). As such, although it is necessary to complete and share work examining these extreme group-level results presented here and how typical they are of real-life studies, it is unlikely that future comprehensive simulations would offer any renewed faith in this method. Importantly, selection of an anchor should be based on one that is simple, easy to understand and representative of the concept the researcher is aiming to classify. Only in this way can misclassification error be reduced and some faith in the group-level anchor-based MCT be assured.

Another limitation is that, under some conditions, it was found that even a near-perfectly correlated anchor ($r = \sim 1.0$) had some variability in the returned MCT cut point. This could be a result of the anchor approaching $r = 1.0$ rather than being exactly $r = 1.0$. Equally, this could be due to the wide distribution of change scores influencing the ability of these anchor-based methods to determine the true MCT. Given that this effect also varies by sample size, it is likely that this is due to between-sample fluctuation where the true mean of the sample varies randomly in line with the procedures used to generate the data. Sample size alone effects the variability of frequencies for returned MCTs. While many studies are unavoidably limited by sample size, such as in rare diseases, longitudinal study designs can be used to increase the number of data points over time for a limited sample size, increasing the reliability of estimates and the likelihood of selecting a true MCT. Collection of data from multiple early timepoints and later timepoints may therefore serve to mitigate the limitations arising from small sample sizes.

Findings from this study support use of anchor correlations above 0.30 to identify an appropriate individual-level MCT when using PM based methods, while stronger correlations are needed for ROC based methods (perhaps around 0.50–0.60). At the individual-level, 0.50–0.60 may demonstrate an ideal threshold, however correlations in the 0.30–0.50 range remain a viable outcome in practice. In such cases, consideration of the PM method as a primary analysis would be beneficial, and close attention to the correlation when triangulating across multiple anchors is recommended. At the group-level, there will always be a bias in the MCT derived from a less than perfect correlation, but researchers should assess what they think is acceptable as an anchor correlation in their own work based on the results here, and perhaps err on the side of a more conservative estimate given the apparent under-estimation present in this method. No recommendation for the group level is offered here, as any relationship less than $r = 1.0$ leads to an attenuation of the true threshold. However, given this knowledge, it may be possible in future to develop an anchor correlation-based adjustment for group level MCTs which will help account for the bias observed. This adjustment would also need to account for the SD of the COA change score, but perhaps not the sample size. Further work is needed to support development of guidance for the conduct of appropriate anchor-based analyses.

Declarations

Funding

No funding was received to assist with the preparation of this manuscript.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Availability of data and material

Data available on request.

Code availability

Custom code in R, available on request.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Philip Griffiths and Abi Williams. The first draft of the manuscript was written by Philip Griffiths, Abi Williams, Joel Sims and Nicola Williamson and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

References

1. Coon, C. D., & Cappelleri, J. C. (2016). Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci, 50*(1), 22–29.
2. Revicki, D., Cella, H. R., & Sloan, D. (2008). J., Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Clin Epidemiol., 61*(2), 102–109.
3. Food, & Administration, D. *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims*. December 2009.
4. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res, 27*(1), 33–40.
5. Food, & Administration, D. *Patient-Focused Drug Development: Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments. Draft discussion document*. October 2018.
6. Cohen, J., *Statistical power analysis for the behavioral sciences*. 2013: Academic press.
7. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD, 2*(1), 63–67.
8. Guyatt, G. H., et al. (2002). A critical look at transition ratings. *Journal of clinical epidemiology, 55*(9), 900–908.
9. Escobar, A., et al. (2013). Total knee replacement; minimal clinically important differences and responders. *Osteoarthritis Cartilage, 21*(12), 2006–2012.
10. Devji, T., et al. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ, 369*, m1714.
11. Ousmen, A., et al. (2018). Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. *Health Qual Life Outcomes, 16*(1), 228.
12. Fayers, P. M., & Hays, R. D. (2014). Don't middle your MIDs: regression to the mean shrinks estimates of minimally important differences. *Qual Life Res, 23*(1), 1–4.
13. Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102.

14. Terluin, B., et al. (2015). Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *Journal of Clinical Epidemiology, 68*(12), 1388–1396.
15. Terluin, B., et al. (2020). Unlike ROC analysis, a new IRT method identified clinical thresholds unbiased by disease prevalence. *Journal of Clinical Epidemiology, 124*, 118–125.

Figures

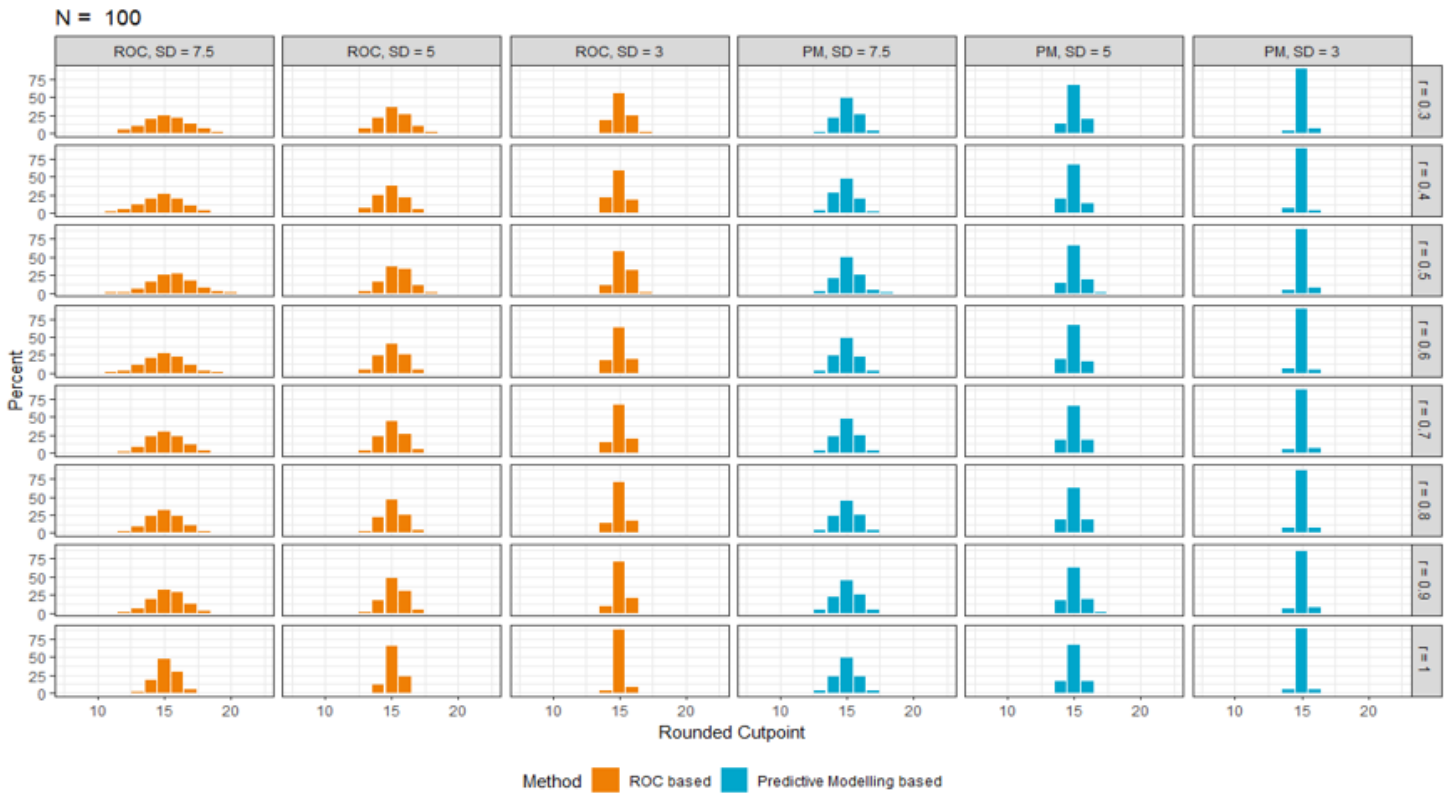

Figure 1

Simulation of individual level distributions with different anchor correlations and change scores (n=100)

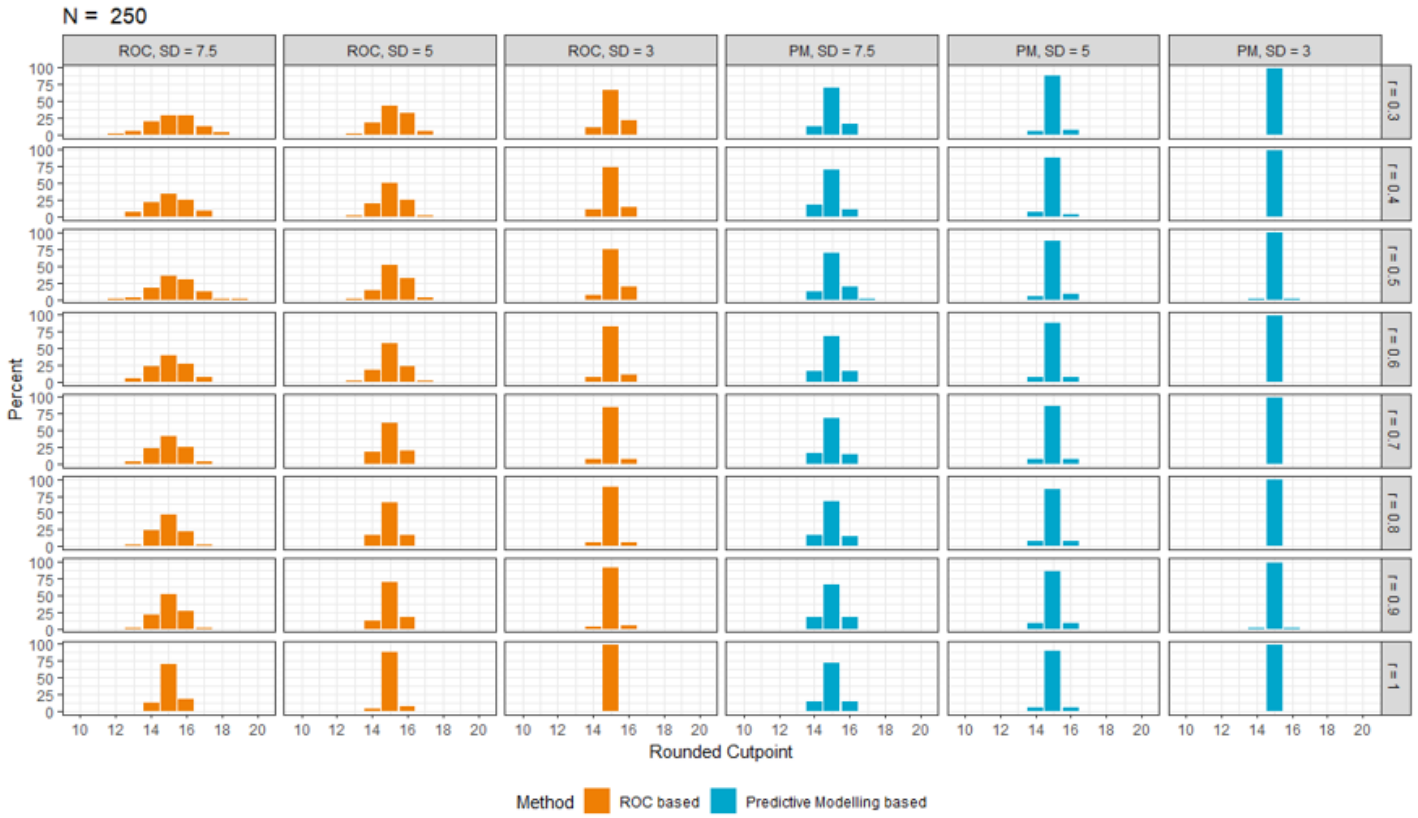

Figure 2

Simulation of individual level distributions with different anchor correlations and change scores (n=250)

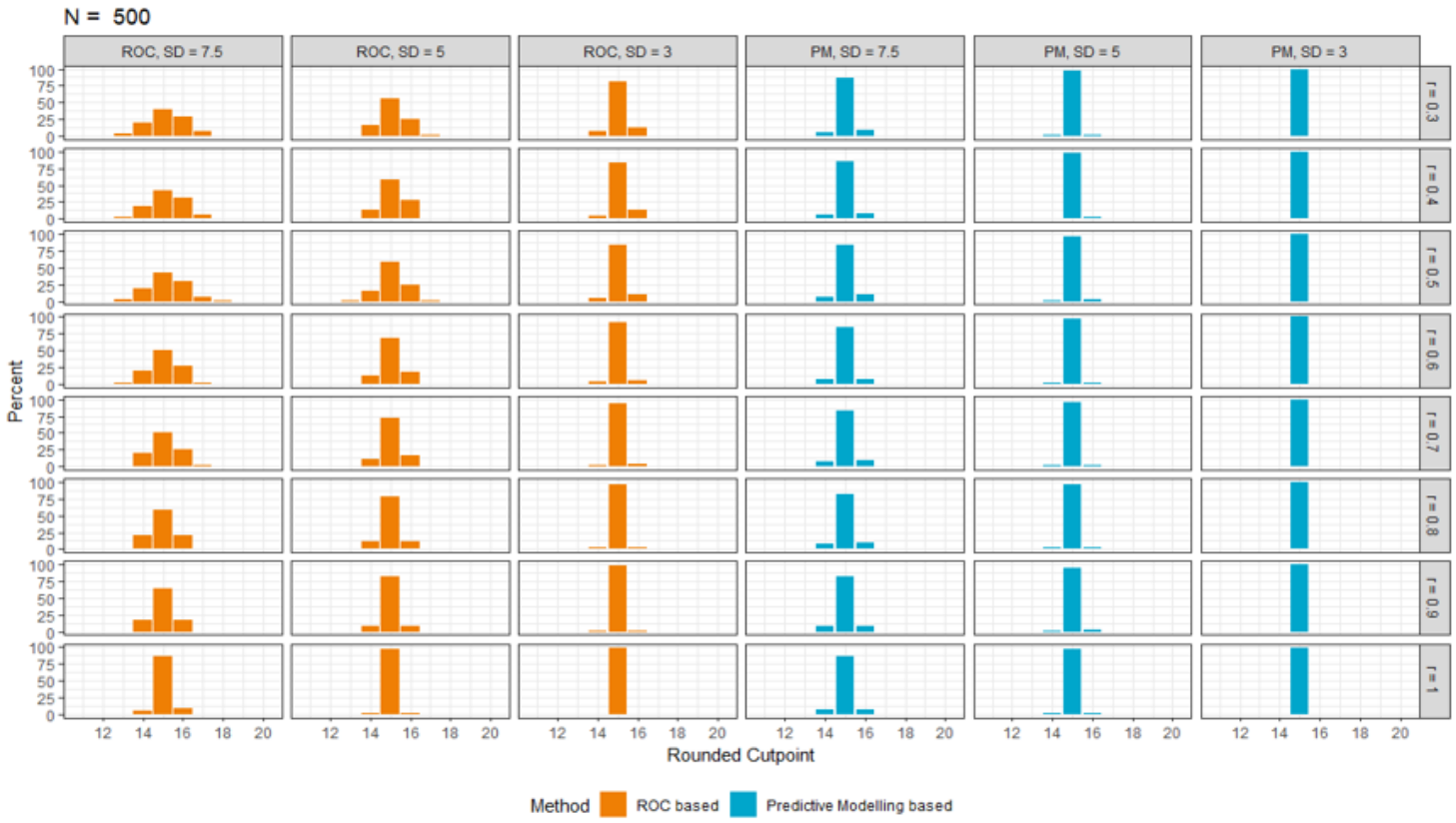

Figure 3

Simulation of individual level distributions with different anchor correlations and change scores (n=500)

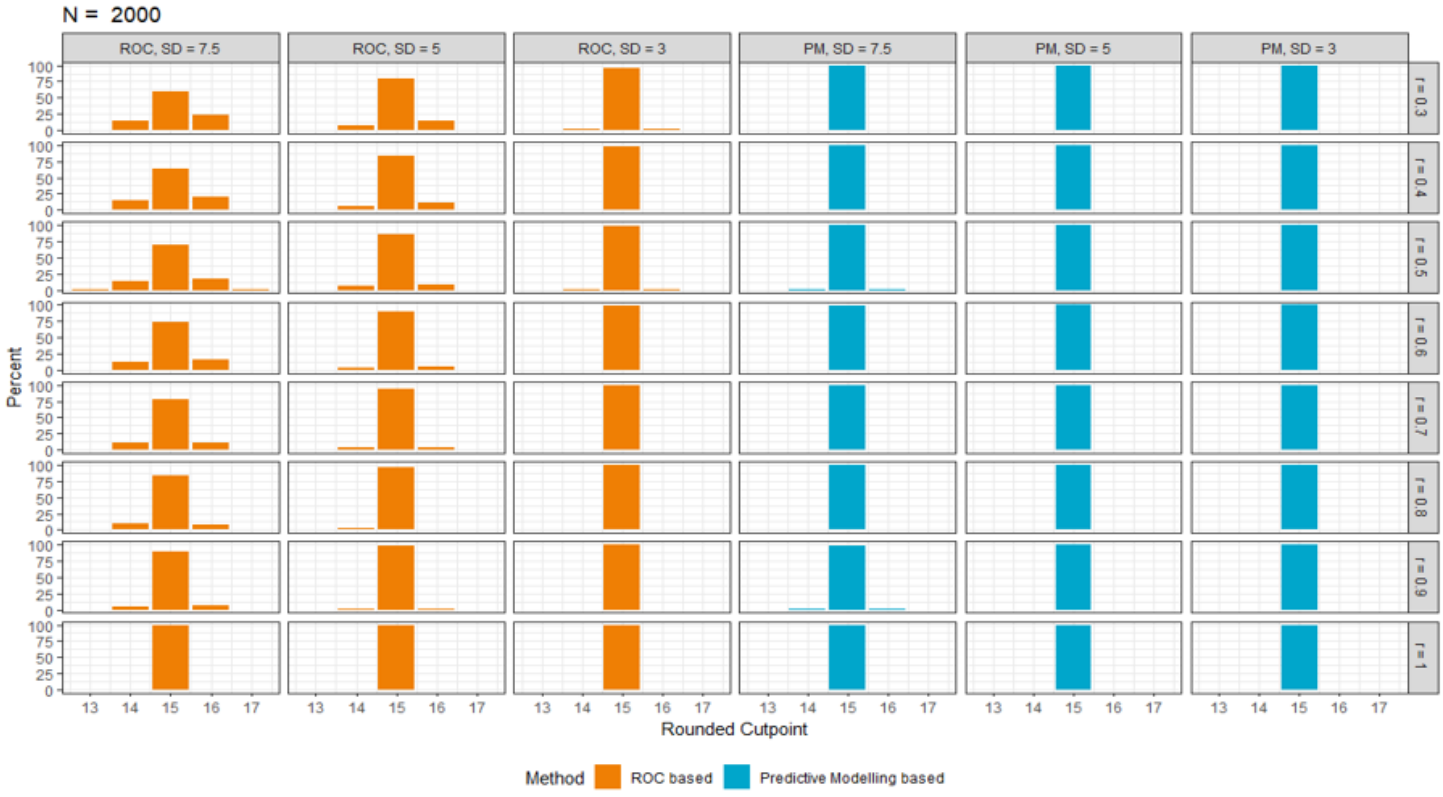

Figure 4

Simulation of individual level distributions with different anchor correlations and change scores (n=2000)

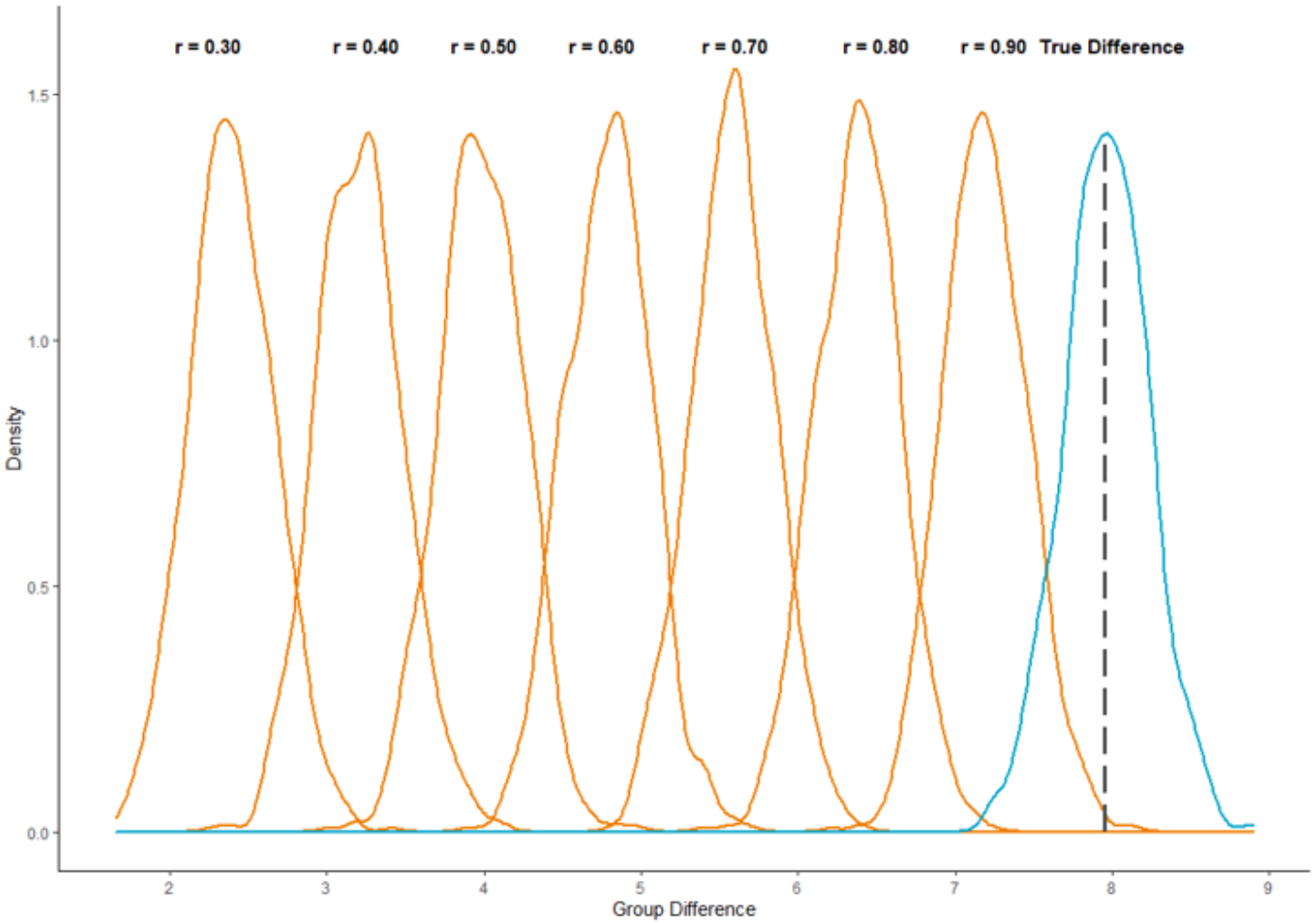

Figure 5

Simulation of group level distributions with different anchor correlations

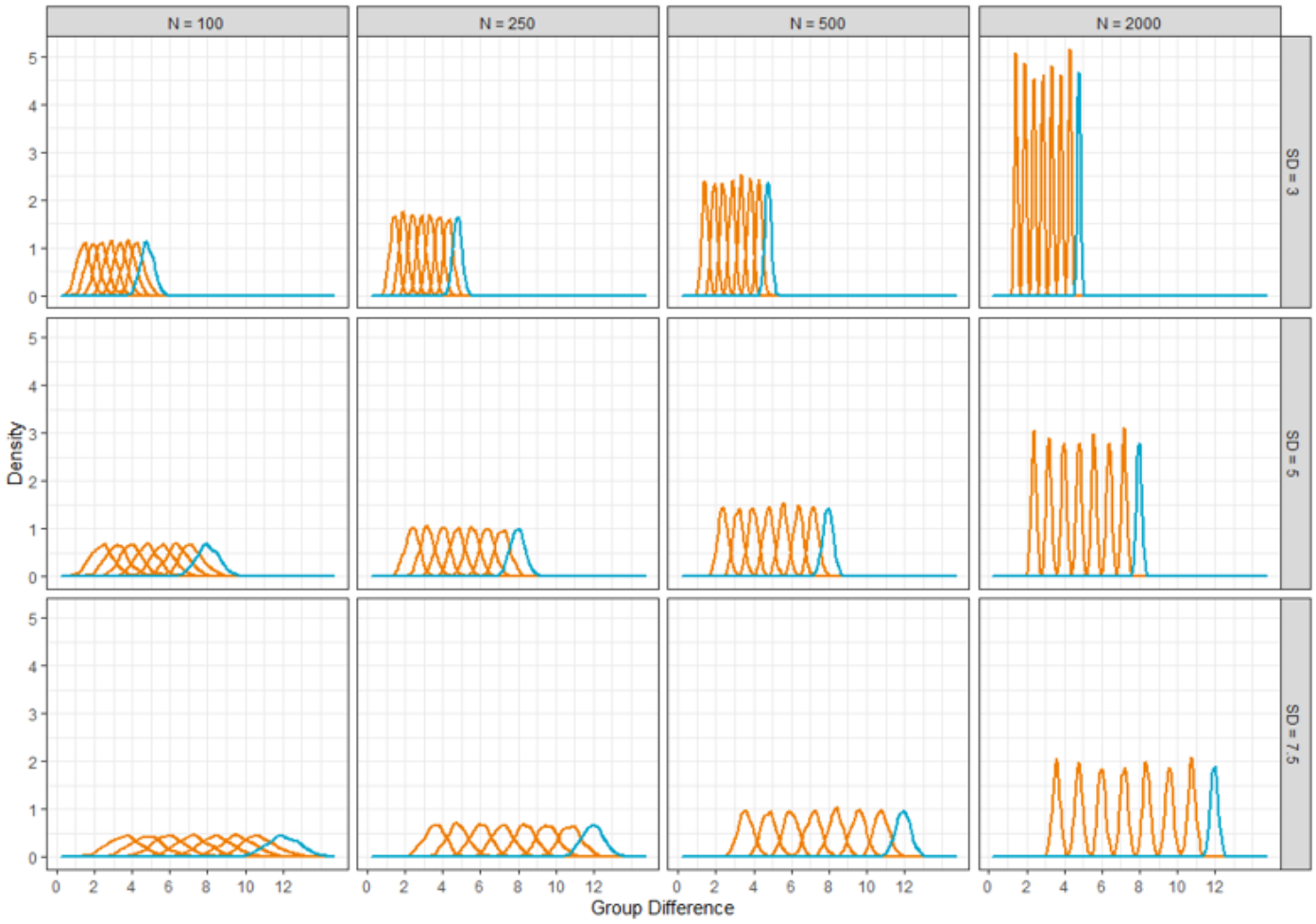

Figure 6

Simulation of group level distributions with different anchor correlations for all conditions