Journal of Statistical Software

November 2024, Volume 111, Issue 1. doi: 10.18637/jss.v111.i01

clinicalsignificance: Clinical Significance Analyses of Intervention Studies in R

Benedikt B. Claus Witten/Herdecke University

Julia Wager Witten/Herdecke University

Udo Bonnet University of Duisburg-Essen

Abstract

The analysis of clinical significance is helpful to decide if an intervention leads to practically relevant or meaningful changes for individual patients which is clearly different from the analysis of statistical significance. However, the framework is used rarely and inconsistently. We introduce the R package **clinicalsignificance** to harness the use of clinical significance analysis of intervention trials in clinical research. This package provides all relevant methods to calculate and present analyses of clinical significance in a consistent form and easy to use implementation. Despite its shortcomings, clinical significance analyses are a valuable tool to gain more insight into intended and potential unintended intervention effects and they may improve the interpretation and comparability of intervention trial results. Lastly, analyses of clinical significance may guide researchers and policy makers in determining which interventions are clinically effective.

Keywords: clinical significance, psychotherapy, clinical studies, intervention studies, treatment effect, intervention effect, R.

1. Introduction

Most researchers developing new interventions for specific disorders want to ascertain if that said intervention actually helps the individual patient. Most intervention studies, however, rely on tests of "statistical significance" (Cohen 2011; Gao 2020; Wasserstein, Schirm, and Lazar 2019). If for instance, a researcher wants to determine the effectiveness of a new antidepressant for the treatment of major depression, he/she measures the depressive syndrome with a psychometric instrument in a patient sample before and during taking the new antidepressant for a given period of time and then calculates the difference in these instrument scores. This is usually done in one of two frameworks: frequentist null hypothesis significance testing (NHST) or Bayesian significance testing. The result in the frequentist framework may be a p value which gains insight into how improbable the resulting or a more extreme

test statistic would be, if there was in fact no intervention effect (Goodman 2008). In the Bayesian framework, a Bayes Factor (BF) may be calculated to update prior beliefs about which of two hypotheses is more likely, e.g., null vs. alternative hypothesis, given the observed data (Wagenmakers *et al.* 2018). In both cases, the resulting parameter, p or BF, is based on groups. But "Statistical inferences drawn from groups of individuals may not accurately describe the individuals themselves" (Grice *et al.* 2020, p. 444). These parameters fail to provide information for a specific patient and the size of effects has long been considered as a more suitable measure of the effectiveness of an intervention. One way of determining the intervention effect's size are effect sizes (Cohen 1988; Lakens 2013). But again, these are calculated based on group-wise summary statistics and may be hard to interpret as well (Funder and Ozer 2019). Thus, with the methods described above, no statement on the *practical relevance* of intervention effects for *individual* patients can be made, and following that, a researcher cannot decide, if the intervention actually helps these patients. Jacobson, Follette, and Revenstorf (1984), as well as Ogles, Lunnen, and Bonesteel (2001) give an excellent overview of these issues and recommend to additionally analyze the clinical significance of intervention studies. So, let's "bring the person back into scientific psychology [and other fields], this time forever" (Molenaar 2004).

Jacobson *et al.* (1984) were among the first to introduce the framework of clinical significance for which the question regarding the practical relevance or meaningfulness of intervention effects for patients is the core idea. The goal of the clinical significance framework is not to probabilistically distinguish an intervention effect from no effect at all, but to differentiate, if an observed change is practically relevant or meaningful for a given patient. There does exist a plethora of statistical procedures that may be used to determine if a patient is clinically significant. Lavigne (2016) and Crosby, Kolotkin, and Williams (2003) give an overview of the most commonly used approaches and we will follow their nomenclature in this article. We will introduce the most relevant methods in Section 2 and will exemplify their use with **clinicalsignificance** in Section 3.

Currently, there is only one R package (R Core Team 2024) that addresses analyses of clinical significance, namely **clinsig** (Lemon 2016), which solely uses the method proposed by Jacobson and Truax (1991). This package also requires data to be extracted from a dataframe to a vector which can be cumbersome for some, especially new R users. Currently, this procedure is not implemented in other popular statistical software solutions such as IBM SPSS Statistics (IBM Corp. 2021) or **JASP** (Love *et al.* 2019; **JASP** Team 2022). To facilitate the use of clinical significance analyses in clinical trials and foster research on the topic itself, we developed the R package **clinicalsignificance** which employs multiple methods by various authors, is easy to use, and yields publication ready results in a clear and consistent form. Package **clinicalsignificance** (Claus 2024) is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=clinicalsignificance`.

We want to highlight that this article is not aimed at recommending any method over another; we merely provide a toolbox with different utensils to perform clinical significance analyses for the most common approaches in intervention studies. The decision to use any illustrated approach and method is ultimately made by the researcher and should always be based on and substantiated by the relevant literature. Nonetheless, the contemporary literature seems to favor the individual level anchor-based and combined approaches as outlined below.

2. Clinical significance methods

2.1. Anchor-based approaches

Anchor-based approaches are based on the minimal important difference[1] (MID; see Engel, Beaton, and Touma 2018; King 2011 for methodological details) of an instrument, which is the smallest change on an instrument that is regarded as practical or meaningful by patients. This is usually established by comparing or linking the change in the respective instrument with a patient reported global rating of that change (e.g., "I feel slightly better."; see Lakens 2013; Leucht, Fennema, Engel, Kaspers-Janssen, Lepping, and Szegedi 2013, 2017 for examples). For instance, if the MID for an instrument measuring well-being is 5 points, then a patient change during intervention of ≥ 5 points is deemed clinically significant. Furthermore, there are two levels at which anchor-based approaches can be used to examine the clinical significance of an intervention study, i.e., the individual and the group level, which will be outlined below.

Individual level

For the individual level, the individual patient change Δ from a pre intervention measurement x_1 to post intervention measurement x_2 is calculated as $\Delta = x_2 - x_1$ and is then compared against the MID. If the change exceeds the MID in size, then this change is clinically significant and can thusly imply an improvement, deterioration or no significant change if it does not exceed the MID. One also has to take into account the direction of the used instruments. Positive instruments measure a construct of which higher values are desirable, for instance, well-being. Negative instruments are ones that measure a construct for which lower values are desirable, e.g., symptom severity. Consequently, a clinical significant (meaningful) improvement for the individual is believed to have occurred if

$$\kappa\Delta \geq \text{MID}$$

and a clinical significant (meaningful) deterioration can be assumed if

$$\kappa\Delta \leq -\text{MID}$$

with κ indicating the desired direction of instrument. If a positive instrument is used, $\kappa = 1$ and if a negative instrument is used, then $\kappa = -1$. A change that is $-\text{MID} < \Delta < \text{MID}$ would be categorized as "unchanged" or, respectively, not clinically significant (for an overview, see Table 1).

For instance, imagine a psychotherapy patient filling out a negative instrument that measures depressive symptoms like the (open access) *Mind over Mood Depression Inventory* (Greenberger, Padesky, and Beck 2016). Suppose that this patient has changed from $x_1 = 32$ to $x_2 = 15$ during a cognitive behavior therapy and that the MID for this instrument is MID $= 9$ points. Then, $\kappa = -1$ and $\Delta = 15 - 32 \Leftrightarrow \Delta = -17$. Because $\kappa\Delta = (-1)(-17) \Leftrightarrow \kappa\Delta = 17$ and $17 \geq 9$, this patient is categorized as "improved". If however, a patient changed from $x_1 = 7$ to $x_2 = 16$ (still with $\kappa = -1$ because the instrument is the same), then

[1] Although the term "minimal important change" (MIC, see De Vet and Terwee 2010) fits better with the longitudinal nature of the underlying data, MID seems to be used more often, so we adapted to this naming convention as well.

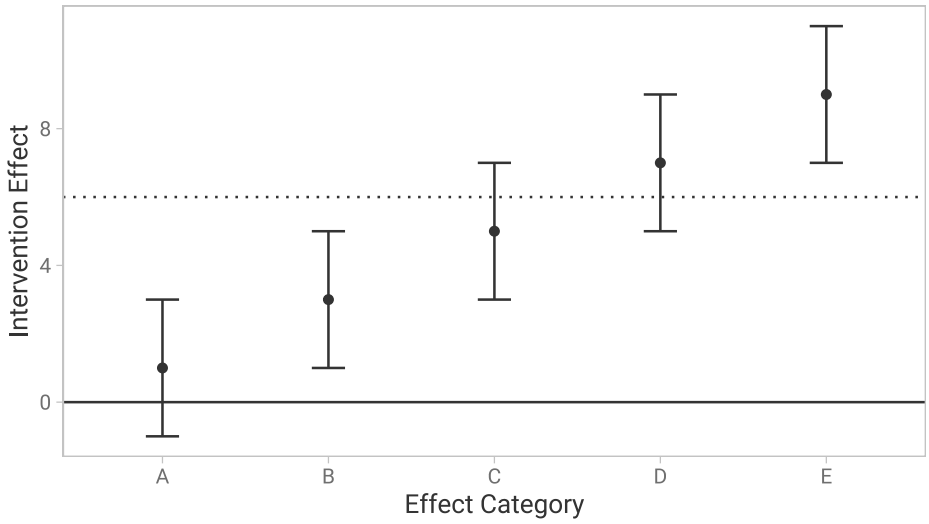

Figure 1: Clinical significance categories for the group anchor-based approach (for a positive outcome). The solid line represents a null-effect, the dotted line the instrument's MID. Points represent the average intervention effect, surrounded by the uncertainty interval with error bars.

$\kappa\Delta = (-1)9 \Leftrightarrow \kappa\Delta = -9$ and $-9 \leq -9$, the patient would therefore be classified as having meaningfully deteriorated during intervention.

It is further possible to obtain separate MIDs for an improvement and a deterioration. In this case, one could define MID_i as the MID that needs to be reached for an improvement and MID_d as the threshold that signifies a clinically significant deterioration. If only the MID for either an improvement or a deterioration is known, then $\text{MID}_\text{i} = \text{MID}_\text{d}$ is implicitly assumed.

Group level

For the group level, a shifted null hypothesis significance test can be performed to test the hypothesis that the whole group demonstrated a clinically significant change (Kieser and Hauschke 2005). For this, the mean change Δ_mean of a group is calculated, along with the associated confidence interval (CI), typically a 95 % CI (which would result in a significance level of 2.5 %). Based on the location of the mean difference and its confidence interval, a group-wise change can be categorized one of several groups (see Figure 1).

The effect can be not statistically significant (see effect category "A" in Figure 1) or statistically significant but both, the mean intervention effect and its confidence intervals are beneath the predefined MID (see effect category "B" in Figure 1). In this case, the effect is not clinically relevant. Or the effect can be not significantly less then the predefined MID if the MID falls into the range of the CI but the mean effect is still less than it (see effect category "C" in Figure 1). The effect may also be probably clinically significant, in which case the mean effect is greater than the MID but it still is in the range of the CI (see effect category "D" in Figure 1). In the last case both, the mean intervention effect and both CI limits are greater than the MID (see effect category "E" in Figure 1). Then, the mean treatment effect is regarded as a large clinically significant effect (see Table 2).

Approach	Category		
	Improved	Unchanged	Deteriorated
Anchor-based	$\kappa\Delta \geq \text{MID}$	$-\text{MID} < \Delta < \text{MID}$	$\kappa\Delta \leq -\text{MID}$
Percent-change	$\kappa\Delta_{\text{rel}} \geq \text{PCC}$	$-\text{PCC} < \Delta_{\text{rel}} < \text{PCC}$	$\kappa\Delta_{\text{rel}} \leq -\text{PCC}$
Distribution-based (Jacobson *et al.* 1984)	$\kappa\text{RCI} > z_{1-\frac{\alpha}{2}}$	$-z_{1-\frac{\alpha}{2}} \leq \text{RCI} \leq z_{1-\frac{\alpha}{2}}$	$\kappa\text{RCI} < -z_{1-\frac{\alpha}{2}}$
Distribution-based (Hageman and Arrindell 1999b)	$\kappa\text{RCI} > z_{1-\phi_{\max}}$	$-z_{1-\phi_{\max}} \leq \text{RCI} \leq z_{1-\phi_{\max}}$	$\kappa\text{RCI} < -z_{1-\phi_{\max}}$
Statistical (Jacobson *et al.* 1984)	$\kappa x_1 < \kappa C \wedge \kappa x_2 > \kappa C$	$x > C \vee x < C$ $x = (x_1, x_2)$	$\kappa x_1 > \kappa C \wedge \kappa x_2 < \kappa C$
Statistical (Hageman and Arrindell 1999b)	$\kappa x_1 < \kappa C \wedge \kappa\text{CS}_{\text{INDIV}} > z_{1-\phi_{\max}}$	$-z_{1-\phi_{\max}} < \text{CS}_{\text{INDIV}} < z_{1-\phi_{\max}}$ $\vee x > C \vee x < C$ $x = (x_1, x_2)$	$\kappa x_1 > \kappa C \wedge \kappa\text{CS}_{\text{INDIV}} < -z_{1-\phi_{\max}}$

Table 1: Clinical significance categories for the (individual) anchor-based, percentage-change, distribution-based, and statistical approach. α = significance level, C = cutoff value, κ = instrument direction, $\phi_{\max}$ = maximal risk of misclassification, MID = minimal important difference, PCC = percent-change cutoff, z = quantile function of the standard normal distribution.

	Point estimate		
Category	κCI_c	$\kappa\Delta_\text{mean}$	κCI_l
Statistically significant but not clinically relevant	$< \kappa\text{MID}$	$< \kappa\text{MID}$	$< \kappa\text{MID}$
Not significantly less than MID	$< \kappa\text{MID}$	$< \kappa\text{MID}$	$\geq \kappa\text{MID}$
Probably clinically significant effect	$< \kappa\text{MID}$	$\geq \kappa\text{MID}$	$\geq \kappa\text{MID}$
Large clinically significant effect	$\geq \kappa\text{MID}$	$\geq \kappa\text{MID}$	$\geq \kappa\text{MID}$

Table 2: Clinical significance categories for the group anchor-based approach. CI_c and CI_l refer to the conservative and liberal limits of the confidence interval for the mean effect. The conservative confidence limit is the one that is closer to 0 (the lower limit of the CI for "positive" outcomes as in Figure 1) and the liberal confidence limit the one that is further away from 0, so $|\text{CI}_\text{c}| < |\text{CI}_\text{l}|$. In all cases, $\kappa \times$ point estimate > 0 is assumed (indicating a statistically significant change).

Above, the frequentist framework for obtaining group-wise clinical significance estimates is described. However, the same reasoning can be adopted in a Bayesian framework (for an introduction see Etz and Vandekerckhove 2018; Rouder, Haaf, and Vandekerckhove 2018; Wagenmakers *et al.* 2018). In this case, the average group change is the median of the posterior distribution and the uncertainty interval is the credible interval, i.e., the highest density interval (HDI) of the posterior distribution (Kruschke 2015). The Bayesian framework offers unique advantages in the interpretation of the average treatment effect and the associated uncertainty interval. In the frequentist framework, the CI limits are merely points and – strictly speaking – offer no insight into how the mean change is distributed. Given the sampling procedure, the sample size, and imaginary infinite repetitions of the sampling, the true mean group change would lie within the confidence limits 95 % of the time. This is the usual long-run frequency interpretation of the frequentist approach.

In the Bayesian framework however, in which inference is based on a full posterior distribution given the observed data, the credible interval contains the true group change with 95 % probability (see Hespanhol, Vallio, Costa, and Saragiotto 2019 and Morey, Hoekstra, Rouder, Lee, and Wagenmakers 2016 for a thorough discussion on this topic). Thus, the use of the Bayesian framework may be more intuitive, offers several advantages over the frequentist CI, and is the default method for the group level anchor-based approach in our package **clinicalsignificance**.

In intervention studies, random variations or systematic time trends may be controlled by the use of an inactive comparator group, such as a placebo in medication studies or a sham treatment in psychotherapy studies, for instance in randomized controlled trials (RCTs). In these studies, the researcher is usually not directly interested in the change of both groups but the change of the target group (i.e., medication, psychotherapy, new intervention, etc.) in relation to the control group. In this case, the average between-group difference and its associated confidence (or credible) interval can be calculated. This difference and the associated uncertainty interval can then be categorized as the within group change as described above and in Table 2.

Although this group level approach to anchor-based methods for establishing clinical significance is very easy to use and naturally extends the common statistical procedures taught in most statistical courses (i.e., the t test), the very idea of clinical significance analyses is to

focus on the *individual* and whether individuals experienced a meaningful change during the intervention (as described in the introduction and Ogles *et al.* 2001 as well as Molenaar 2004). One may easily lead the group level analyses *ad absurdum*, if we assume a large dataset with two distinct but unknown groups. During a fictitious intervention one group can be described with a pre intervention mean $M_1 = 5$ and the post intervention mean $M_2 = 10$. The other group can be described with $M_1 = 10$ and $M_2 = 5$. We further assume the standard deviations to be equal in all instances. Here, the group level mean change would be $\Delta_{\text{mean}} = 0$ but a lot has happened on the individual level, i.e. many individuals improved but approximately the same number of individuals deteriorated as well. This is an extreme (and unlikely to be actually observed) example for an intervention study but these distinctions may be overlooked by generally focusing on group level analyses. Because of that, regarding the clinical significance of intervention studies, we do recommend always conducting an individual level analysis of intervention studies and other desired approaches *in addition*.

2.2. Percentage-change approach

The percentage-change approach is similar to the individual anchor-based approach (Section 2.1) but differing in the way not to evaluate the raw change Δ against the MID, but the *relative* change Δ_{rel} to the predefined percentage-change cutoff (PCC) that is believed to indicate a clinical significant change. Here, Δ_{rel} is defined as

$$\Delta_{\text{rel}} = \frac{\Delta}{x_1} \tag{1}$$

A practically meaningful, or clinically significant, improvement is believed to have occurred if the relative change is larger or equal to a predefined percentage change, so

$$\kappa \Delta_{\text{rel}} \geq \text{PCC}$$

The classification of clinical significance categories based on Δ_{rel} relative to the PCC can more generally be categorized according to Table 1. One exception is the case when $x_1 = 0$, for which Equation 1 is not defined. In this special case, a change in the beneficial direction, irrespective of its size, is classified as an improvement and a change in the harmful direction as a deterioration.

For example, reconsider the patient example from Section 2.1 that changed from $x_1 = 32$ to $x_2 = 15$ during an intervention as measured by a negative instrument. Suppose that it was previously shown by others that a 30 % reduction (PCC = 0.3) in depression scores may be considered clinically relevant. Again, $\kappa = -1$ and thusly, $\Delta_{\text{rel}} = \frac{-17}{32} \Leftrightarrow \Delta_{\text{rel}} \approx -0.53$. Because $\kappa \Delta_{\text{rel}} \geq \text{PCC} \Leftrightarrow 0.53 \geq 0.3$, this patient may be categorized as improved.

Similar to the MID, there may be different estimates for a PCC that represent an improvement (PCC_{i}) or a deterioration (PCC_{d}). If only one of those estimates is known, it will be implicitly assumed that $\text{PCC}_{\text{i}} = \text{PCC}_{\text{d}}$.

2.3. Distribution-based approach

The distribution-based approach takes into account the distribution of the observed instrument scores, as well as the instrument's reliability. This information is used to determine if an individuals's change is beyond a level that could be attributed to measurement error

and is consequently reliable. All psychometric instruments measure the desired outcome with imprecision which needs to be accounted for. This is usually referred to as the minimal detectable change (MDI). Following contemporary research, the distribution-based approach should not be used as the sole method to infer clinical significant change (Turner *et al.* 2010; De Vet and Terwee 2010). Nevertheless, several methods have been developed to be used in this approach, which are outlined below.

Jacobson and Truax

Jacobson and Truax (1991) proposed the reliable change index (RCI, Jacobson *et al.* 1984), incorporating small computational changes advocated by Christensen and Mendoza (1986), which can be calculated for each patient as follows.

Let s_i be the sample standard deviations when index i denotes the measurement before ($i = 1$) and after ($i = 2$) an intervention. Then, the RCI is calculated as

$$\text{RCI} = \frac{x_2 - x_1}{S_{\text{diff}}}$$

with

$$S_{\text{diff}} = \sqrt{2\left(S_{\text{E}}\right)^2}$$
$$S_{\text{E}} = s_1\sqrt{1 - r_{xx}}$$

Here, x_1 and x_2 are the pre- and post intervention scores for a given patient, s_1 is the pre intervention standard deviation of all scores, and r_{xx} is the instruments reliability. Both, Tingey, Lambert, Burlingame, and Hansen (1996) and Martinovich, Saunders, and Howard (1996) recommend using a measure of internal consistency, e.g., McDonald's ω (Flora 2020) as the instruments reliability although originally, r_{xx} refers to the test-retest reliability. S_{diff} is the standard error of difference, and S_{E} is the standard error of measurement. The standard error of difference "describes the spread of the distribution of change scores that would be expected if no actual change had occurred" (Jacobson and Truax 1991, p. 14).

The resulting RCI is then compared against a critical value to determine, whether an observed change is unlikely, if there was in fact no change, as is outlined in Table 1. Given a significance level α, the critical value can be obtained as the quantile of the standard normal distribution $z_{1-\frac{\alpha}{2}}$. For $\alpha = .05$, the critical value is $z_{.975} \approx 1.96$. For instance, if a patient's RCI > 1.96 or RCI < -1.96 then the observed change would be surprising ($p < .05$), if there was in fact no change. If $-1.96 \leq \text{RCI} \leq 1.96$, the change is assumed to fall in the range of scores which would be expected if in fact no change had occurred. Thus, an RCI in this range would categorize a patient as "unchanged". The Jacobson and Truax (1991) method is the default distribution-based method in **clinicalsignificance** and α can be changed if desired.

Edwards and Nunnally

Speer (1992) incorporated other statistical approaches (Edwards, Yarvis, Mueller, Zingale, and Wagman 1978; Nunnally 1967, 1975) to account for regression to the mean, which is the tendency of test scores to become less extreme over time. Consequently, the pre intervention score x_1 should be adjusted in the direction of the mean of all pre intervention scores M_1. Thus, patients with more extreme initial instrument scores must show a greater change from

before to after the intervention to show a reliable change. Hence, he recommended calculating an interval around the adjusted ("true") pre intervention score which signifies an area of no change with the upper bound I_u and lower bound I_l. If the post intervention score falls outside this interval, reliable change can be assumed. Speer (1992) advocated the interval

$$[I_\mathrm{u}, I_\mathrm{l}] = [r_{xx}\,(x_1 - M_1) + M_1] \pm 2S_E$$

which can be reformulated to

$$\mathrm{RCI} = \frac{x_2 - [r_{xx}\,(x_1 - M_1) + M_1]}{S_E}$$

The RCI can then be interpreted and classified according as described above (see Table 1).

Gulliksen, Lord, and Novick

Hsu (1989, 1995) refined the Edwars and Nunnally method further, incorporating work from Gulliksen (1950) and Lord and Novick (1968) by not only adjusting the pre intervention score x_1 but also the post intervention score x_2 and substituting the standard error of measurement S_E with the standard deviation of errors of prediction. The result is

$$\mathrm{RCI} = \frac{(x_2 - M_1) - r_{xx}\,(x_1 - M_1)}{s_1\sqrt{1 - r_{xx}^2}} \tag{2}$$

Hsu (1999) raises the caveat that the mean and standard deviation in Equation 2 "might be the pretest mean of a 'relevant' group" (p. 595) although he initially stated that these estimates might be from an entirely different sample constituting the clinical population as well. Our package **clinicalsignificance** estimates the required M and s from the pre intervention sample. Hence the index 1 in Equation 2. The interpretation and classification of the RCI is the same as described above (see Table 1).

Hsu, Linn, and Lord

Hsu (1989) additionally advocated a method that adjusts the post intervention scores x_2 with the mean of all post intervention scores M_2 instead of M_1

$$\mathrm{RCI} = \frac{(x_2 - M_2) - r_{xx}\,(x_1 - M_1)}{s_1\sqrt{1 - r_{xx}^2}}$$

Again, the resulting RCI is categorized according to Jacobson and Truax (1991) (see Table 1).

Nunnally and Kotsch

Nunnally and Kotsch (1983) explicitly recommended a measure of internal consistency r_{IC} in their approach to calculate the RCI and advocate that the internal consistency of the used instrument should be incorporated for both, the pre and the post intervention measurement. Their RCI estimate is

$$\mathrm{RCI} = \frac{x_2 - \left[r_{IC(1)}\,(x_1 - M_1) + M_1\right]}{\sqrt{\left[r_{IC(1)}^2 s_1^2\left(1 - r_{IC(1)}\right)\right] + \left[s_1^2\left(1 - r_{IC(2)}\right)\right]}}$$

and is also interpreted as outlined above (see Table 1).

Hageman and Arrindell

Hageman and Arrindell (1999b) presented the most thorough modification of the original RCI by Jacobson and Truax (1991). They center their approach around work of Cronbach and Gleser (1959) and propose an individual index for reliable change, namely the

$$\text{RCI}_{\text{INDIV}} = \frac{(x_1 - x_2)\, r_{DD} + (M_2 - M_1)\,(1 - r_{DD})}{\sqrt{r_{DD}}\sqrt{2S_E^2}}$$

with

$$r_{DD} = \frac{s_1^2 r_{xx(1)} + s_2^2 r_{xx(2)} - 2 s_1 s_2 r_{12}}{s_1^2 + s_2^2 - 2 s_1 s_2 r_{12}}$$

and

$$r_{xx(1)} = \frac{s_1^2 - S_E^2}{s_1^2} \tag{3}$$

$$r_{xx(2)} = \frac{s_2^2 - S_E^2}{s_2^2} \tag{4}$$

r_{12} is the Pearson correlation coefficient between pre and post intervention scores.

$\text{RCI}_{\text{INDIV}}$ can be interpreted like the Jacobson and Truax (1991) RCI but the critical value is defined as the quantile of the standard normal distribution $z_{1-\phi_{\max}}$, given the maximal risk of misclassification $\phi_{\max}$ (Cronbach and Gleser 1959), which is usually set to $\phi_{\max} = .05$ by default and then yields the critical value $z_{.95} \approx 1.64$ (see Table 1). If desired, $\phi_{\max}$ can be changed in our package.

Hageman and Arrindell (1999b) furthermore proposed a "straightforward" method to estimate the proportion of patients whose true difference score is below zero (for a "negative" outcome) and whose true score is below the cutoff at the group level. They advocate

$$z = \frac{0 - M_{\text{diff}}}{s_{\text{diff}}\sqrt{r_{DD}}}$$

where $\text{Proportion}_{\text{changed}} = F(z)$ and with M_{diff} and s_{diff} being the mean and standard deviation of the pre-post difference scores and F being the cumulative probability of the standard normal distribution. $\text{Proportion}_{\text{changed}}$ can be interpreted as the percentage of patients that have improved irrespective of the size of change (Hageman and Arrindell 1999a,b). Note that the given formula is for "negative" instruments. For "positive" instruments, $\text{Proportion}_{\text{changed}} = 1 - F(z)$. Additionally, they advocate

$$z = \frac{\text{TRC} - M_2}{s_2\sqrt{r_{xx(2)}}}$$

where $\text{Proportion}_{\text{beyond cutoff}} = F(z)$. Again, this is the formula for "negative" instruments. For "positive" instruments, $\text{Proportion}_{\text{beyond cutoff}} = 1 - F(z)$

De Vries and Morey

So far, all RCI methods can be seen as belonging to the frequentist framework. However, De Vries and Morey (2013) as well as De Vries, Meijer, Van Bruggen, and Morey (2016) describe methods to calculate Bayes Factors (Wagenmakers *et al.* 2018) quantifying the evidence for the hypothesis of no (i.e., zero) change against the hypothesis of a (non-zero) change. This approach requires at least three data points per patient (De Vries *et al.* 2016). As far as we know, this method was rarely applied but may be a viable alternative in the Bayesian framework. It is (currently) not implemented in **clinicalsignificance**.

Hierarchical linear modeling

All methods (except De Vries and Morey) concerning the detection of reliable change presented above do encounter one issue: they can only take into account two data points per patient. Researchers may, however, assess the desired outcomes more frequently than before and after an intervention. Hierarchical linear models (HLM) may be used to incorporate all available patient data. We can not give any detailed information on this topic here, but the interested reader may refer to Field (2018) for a light and Finch, Bolin, and Kelley (2019) for a thorough introduction. In this approach, all measurements can be seen as nested in the individual patients. In this case, a regression line can be estimated for each patient, with own intercept and slope, which incorporates information from the individual patient and the whole sample. If a patient-specific slope is "steep" enough (as compared to the slopes standard deviation), then a patient may be classified as reliably changed based on this ratio of slope against its variance. More formally, let y_{ij} be the instrument score of a patient j for measurement i, then

$$y_{ij} = \beta_{0j} + \beta_{1j} \cdot \text{time} + \varepsilon_{ij}$$

with β_{0j} denoting the individual intercept, β_{1j} denoting the individual impact of time (i.e., slope), and ε_{ij} as the error of measurement of that individual at measurement i. The regression coefficients can then be expressed as

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + U_{0j} \\ \beta_{1j} &= \gamma_{10} + U_{1j} \end{aligned}$$

with γ_{00} and γ_{10} being the overall intercept and slope. U_{0j} and U_{1j} are the individual deviations of patient j from the overall coefficients. Of interest in this analysis is a variant of the individual slope β_{1j}, which incorporates not only information from the patient but also from the sample. This estimate is the empirical Bayes estimate β^*_{1j} (see Liu, Kuppens, and Bringmann 2021; Raudenbush and Bryk 2002). Given the variance of this estimate, V^*_j, the RCI can then be calculated as

$$\text{RCI} = \frac{\beta^*_{1j}}{\sqrt{V^*_j}}$$

This RCI can, again, be interpreted according to Jacobson and Truax (1991) (see Table 1). Bear in mind that the estimation of hierarchical linear models for unbiased estimates requires further assumptions, that need to be considered (e.g., linearity or homoscedasticity, see Finch

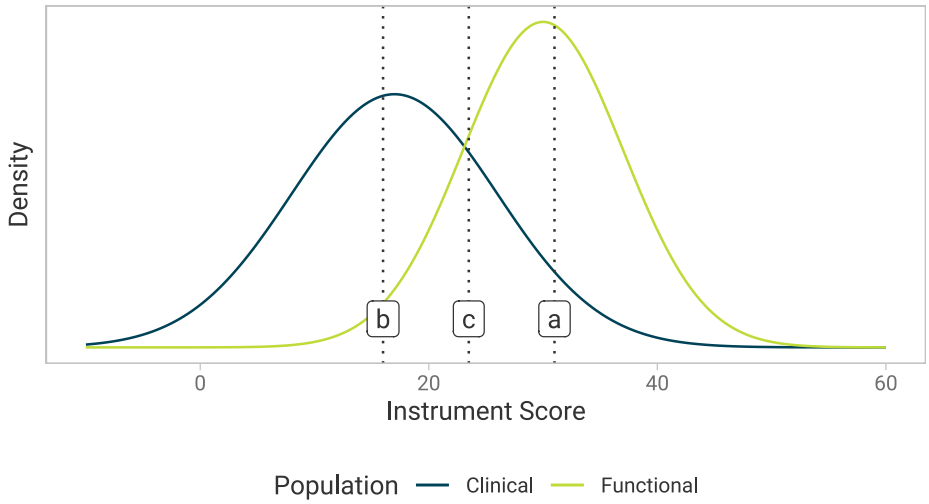

Figure 2: Jacobson *et al.* (1984) proposed three cutoffs to determine the border between two populations. These three cutoffs are shown as vertical lines over the two population distributions. Cutoff c incorporates information from both, the clinical and functional population and is thus the recommended cutoff choice. In this fictional example, a positive instrument is used.

et al. 2019). We included a function into our package to retrieve the fitted model, making it easy to check these assumptions.

2.4. Statistical approach

If it can be assumed that the outcome of interest (e.g., depressive symptom severity) can be described on a continuum and a clinical population (e.g., patients with major depression) as well as a functional population (e.g., the general population) form two distinct distributions on that continuum, then Jacobson *et al.* (1984) proposed the use of the statistical approach to clinical significance. They reason that if a patient belongs to the clinical population pre intervention and is likely to belong to the functional population post intervention, then this change should be meaningful for the given individual and thus clinically significant. At the same time, one might need to distinguish between symptom severity and general functioning, which do not necessarily need to covary. In such instances, two separate analysis or the development of a composite score may be a viable option to portray the whole picture.

With the statistical approach, a cutoff distinguishing the two populations is defined, which must be crossed during intervention. Jacobson *et al.* (1984) advocated three cutoffs that may be used, based on the available data.

Jacobson, Follette, and Revenstorf

Jacobson *et al.* (1984), as well as Jacobson and Truax (1991), proposed the three cutoffs a, b, and c that may be used to distinguish the clinical and functional population and are illustrated in Figure 2.

One reasonable cutoff would be the midpoint between the two populations. If M_i and s_i are

the mean and standard deviation of both populations with $i = 0$ denoting the functional and $i = 1$ denoting the clinical population, then this cutoff can be calculated as

$$c = \frac{s_0 M_1 + s_1 M_0}{s_0 + s_1}$$

As is evident, this cutoff incorporates information of both populations and can be regarded as the most objective one. Unfortunately, in some cases, the required summary statistics of the functional population are lacking. In this case, Jacobson *et al.* (1984) proposed two additional cutoffs, namely

$$a = M_1 + \kappa 2 s_1$$
$$b = M_0 - \kappa 2 s_1$$

For a patient to belong to the functional population, its test score must exceed the mean of the clinical population and two times the clinical standard deviation in the beneficial direction after the intervention (compare with Figure 2) to cross cutoff a. For b, the patient must cross the point that is the mean of the functional population and two times the clinical standard deviation in the harmful direction. Note that in both a and b, the clinical sample's standard deviation s_1 is used to establish the cutoff. Based on the individual's pre- and post intervention scores, it's change may be categorized according to the rules in Table 1.

For instance, consider the psychotherapy patient from Section 2.1 again. This patient changed from $x_1 = 32$ to $x_2 = 15$ on a negative outcome instrument, so $\kappa = -1$. Suppose, the population cutoff is $c = 16$. Because $\kappa x_1 < \kappa c \Leftrightarrow -32 < -16$ and $\kappa x_2 > \kappa c \Leftrightarrow -15 > -16$, this patient may be considered as having meaningfully improved.

As can be seen in Figure 2, if the populations overlap considerably, a may be too conservative and b to liberal to classify a patient belonging to the functional population. In this case c seems to be the best option. Yet, if the populations are very distant from each other, c may even become too conservative and a might be a viable option. Hence, we recommend checking the cutoff in each case but generally recommend c as it incorporates information of both populations. If there are no summary statistics for the normal population, we encourage researchers to estimate these first.

Hageman and Arrindell

Jacobson and Truax (1991) noted that the usage of the aforementioned cutoffs may be problematic because of the instrument's inherent measurement error. Consequently, Hageman and Arrindell (1999b) developed an individual cutoff, $\mathrm{CS}_{\mathrm{INDIV}}$, which establishes a band around the cutoff in which the risk that a patient is classified as "changed population", although it actually did not change the population, is $> 5\,\%$. Their cutoffs can be calculated as

$$\mathrm{CS}_{\mathrm{INDIV}} = \frac{M_2 + (x_2 - M_2) r_{xx(2)} - \mathrm{TRC}}{\sqrt{r_{xx(2)}} S_E}$$

In this case, the index refers to the measurement, so M_2 and x_2 denote the mean of instrument scores and individual patient score at post intervention measurement.

$$S_E = s_{\text{sample}} \sqrt{1 - r_{\text{sample}}}$$

Hageman and Arrindell (1999a) argue in response to McGlinchey and Jacobson (1999) that the calculation of the standard error of measurement S_E is crucial. This means that S_E is fixed and all reliability estimates are calculated based on it. Our package **clinicalsignificance** estimates s_{sample} from the pre intervention sample, thus, $s_{\text{sample}} = s_1$. r_{sample} refers to the samples' reliability coefficient which is estimated under optimal conditions.

TRC is the "true" cutoff score and an equivalent to the proposed cutoffs a, b, and c for which every standard deviation s_i is multiplied with the square root of the relevant reliability. So

$$
\begin{aligned}
c_{\text{true}} &= \frac{s_0\sqrt{r_{xx(0)}}M_1 + s_1\sqrt{r_{xx(1)}}M_0}{s_0\sqrt{r_{xx(0)}} + s_1\sqrt{r_{xx(1)}}} \\
a_{\text{true}} &= M_1 + \kappa 2 s_1\sqrt{r_{xx(1)}} \\
b_{\text{true}} &= M_0 - \kappa 2 s_1\sqrt{r_{xx(1)}}
\end{aligned}
$$

with index i denoting the functional ($i = 0$) and clinical ($i = 1$) population. Furthermore, $r_{xx(0)}$ may be calculated according to Equation 3 but using the respective standard deviation s_i. The CS_{INDIV} is the compared against a critical value to decide if a patient crossed the cutoff with a certain level of confidence as is shown in Table 1 and patients can, again, can be categorized as having "improved", "deteriorated" or are "unchanged".

Hsu

Hsu (1996) furthermore argues that the probability for a patient belonging to the functional or clinical population can only be estimated with Bayesian methods, although their use in practice may not be suitable because of unknown base rates for the examined clinical syndromes. Nonetheless, he provides the interested reader with the modified cutoffs as proposed by Jacobson *et al.* (1984) and more definite interpretations which can also be used to modify the cutoffs advocated by Hageman and Arrindell (1999b). This approach is (currently) not supported by **clinicalsignificance**.

2.5. Combined approaches

According to Lavigne (2016) and Crosby *et al.* (2003), it is possible to combine multiple approaches of the ones outlined above. Most common seems to be the combined approach by Jacobson *et al.* (1984) and Jacobson and Truax (1991) in which they recommend combining the statistical and distribution-based approach. They reason that a meaningful change occurs, if a patient changed from the clinical to the functional population *and* if that change is reliable. Based on the combination of fulfilled criteria for the two approaches, a patient may be categorized according to the following Table 3.

Note, that we introduce the new category "Harmed" to signify patients that changed reliably from the functional to the clinical population, which was not named in earlier works. This naming convention might be essential to also estimate the side effects or unintended harmful consequences of interventions (Margraf and Scholten 2018).

We propose another combined approach in which the statistical and anchor-based approach are combined instead of the statistical and distribution-based approach (see Table 3). In the approach above, the category "improved" refers to patients that did not change population but improved reliably in a *statistical* sense and showed a change greater than the MDC (and *vice versa* for a "deterioration"). In our proposed combined approach, "improved" refers to

	Combined approaches	
Category	Statistical	Distribution/anchor-based
Recovered	Improved	Improved
Improved	Unchanged	Improved
Unchanged	—	Unchanged
Deteriorated	Unchanged	Deteriorated
Harmed	Deteriorated	Deteriorated

Table 3: Clinical significance categories based on the combined approach by Jacobson *et al.* (1984) and our approach outlined above in which the statistical and distribution- or anchor-based approach are considered simultaneously.

patients that did not change populations but at least showed a clinically *meaningful* minimal improvement, i.e., a change that exceeds the MID. This changes the definition of the minimally required change. Like De Vet and Terwee (2010) argue, a meaningful change may also be smaller that the measurement error of an instrument (and hence smaller than the MDC), which is accounted for in this approach. "Recovered" then refers to patients that at least demonstrated a minimally meaningful change *and* can be seen as members of the functional population after treatment.

It is obvious that the combined approaches are stricter than any of the two approaches on their own but offer a further interpretation of results. The combined approaches also take into account a patient's pre intervention value and not just its change. Assuming an MID = 5 points, a patient with pre and post intervention values $x_1 = 20$ and $x_2 = 10$ would be classified in the same anchor-based approach clinical significance category as the patient that changed from $x_1 = 40$ to $x_2 = 5$, although the latter patient most certainly demonstrated a greater improvement. This information can be taken into account in combined approaches.

2.6. Visualizations

The categories proposed above may be easiest to interpret in visual form.

Clinical significance plot

Jacobson *et al.* (1984) proposed the default clinical significance plot. This plot can be constructed for each individual level approach and is thus recommended to be the primary visualization method for clinical significance analyses. Examples of such a plot can be seen in Figure 3A–C.

Plotted are uniform data to visualize the clinical significance categories. Suppose that in these figures, a negative instrument was used, so lower scores correspond to a beneficial outcome. The pre intervention scores are plotted on the x-axis, the post intervention scores on the y-axis. Each patient is plotted as a point. If points fall on the solid diagonal line, then the pre and post scores are identical. Points that are plotted above the solid diagonal represent patients with higher post intervention scores as compared to pre intervention and the points that fall below the diagonal represent patients with lower scores after the intervention. In Figure 3A, a shaded area is drawn around the solid diagonal, which indicates "the spread of the distribution of change scores that would be expected if no actual change had occurred" (Jacobson and

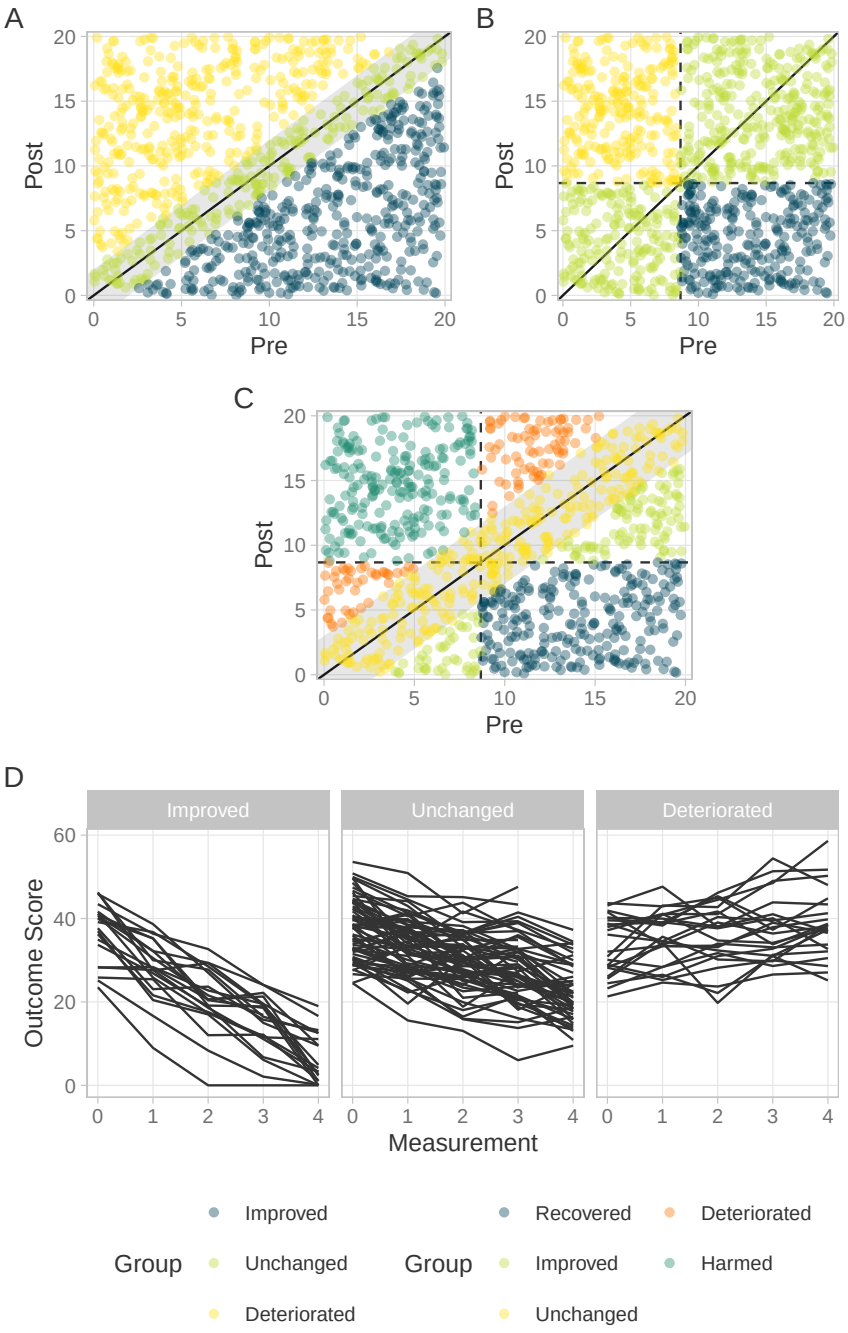

Figure 3: Clinical significance plots for (A) uniform data and the anchor-based, distribution-based or percentage-change approach, (B) the statistical approach, (C) a combined approach of statistical and distribution/anchor-based approach, and (D) trajectories of the distribution-based approach with the HLM method (based on a fictitious dataset).

Truax 1991, p. 14) and illustrates the patients that did not change reliably. Patients that fall in that region can be classified as "unchanged" (green points). Patients that are above can

be seen as "deteriorated" (yellow points) and those below (blue points) as "improved". The results of the individual anchor-based, percentage-change, and distribution-based approach can be visualized with a plot as depicted in Figure 3A.

In Figure 3B, the results of the statistical approach are presented. The plot contains dashed vertical and horizontal lines. These lines represent the cutoff value that distinguishes the populations (functional vs. clinical), or that a patient has to cross from pre to post intervention to demonstrate a change in populations. Patients that fall to the right of the vertical line can be seen as members of the clinical population before the intervention and those that fall to the left line can be seen as belonging to the functional population before the intervention (for a negative outcome instrument). Those patients that fall beneath the horizontal line can be seen as being part of the functional population after treatment and those above the horizontal line are in the clinical population after treatment. Because this is a negative outcome, we are especially interested in all patients that fall right to the vertical line and below the horizontal line. These patients demonstrated a clinical significant change and can be classified as "improved" (blue points). Patients that did not change populations are classified as "unchanged" (green points) and those that moved from the functional to the clinical population are categorized as "deteriorated" (yellow points).

Figure 3C depicts the result of the combined approach of distribution/anchor-based and statistical approach, originally described in Jacobson *et al.* (1984) and our approach proposed above. In addition to the statistical approach, a patient must demonstrate (in the distribution-based approach) a reliable change that is unlikely attributable to measurement error, or exceeds the MID (in the anchor-based approach). Thus, this plot also contains a shaded region that signifies patients as "unchanged" (yellow points). Contrary to the statistical approach alone, patients may be categorized as having improved or deteriorated, even if they still belong to the population in which they entered the study (categories "improved" and "deteriorated" as shown with light green and orange points, respectively). Patients can only be categorized as "recovered" or "harmed" is they fulfill both criteria, so additionally changed population and not only demonstrated a reliable change (as shown with blue and dark green points, respectively).

In all examples, the classification categories change, if a positive instrument is used. Then, the categories are mirrored along the solid diagonal, so for instance in Figure 3A, points that fall beneath the shaded area would be classified as "deteriorated" and points that are above the shaded are would be categorized as "improved".

Hierarchical linear modeling method

The distribution-based approach can be carried out with the use of a hierarchical linear model. In this case, because more than two data points are used, the reduction to only two data points (pre and post intervention) and consequently, the default clinical significance is unsuited for a visualization. In this case, we recommend showing individual trajectories per clinical significance category, as is shown in Figure 3D. In this figure, the course of each patient is shown as a solid line over time, while the x-axis represents the individual measurements and the y-axis the observed instrument score at any given measurement.

Group level clinical significance plot

A group level anchor-based clinical significance plot is shown in Figure 1. The average in-

tervention effect is plotted on the y-axis, surrounded by the uncertainty interval (confidence or credible interval) as error bars. An ineffective intervention would be expected to result in an average intervention effect of 0, which is depicted as a solid black line. The instrument's MID is shown as a dotted line and thus, the location of the average intervention effect and uncertainty interval in relation to a null-effect and the MID can be determined.

3. Practical examples

3.1. Harnessing the placebo effect

Claus, Scherbaum, and Bonnet (2020) conducted a study determined to enhance the placebo effect of antidepressants in the treatment of major depression because the placebo effect may be in great part responsible for their posited effectiveness (Hengartner and Plöderl 2018; Hengartner 2022; Jakobsen, Gluud, and Kirsch 2020; Kirsch and Sapirstein 1998; Kirsch, Deacon, Huedo-Medina, Scoboria, Moore, and Johnson 2008; Kirsch 2010; Moncrieff, Wessely, and Hardy 2004; Whitaker 2010). They sampled patients from an inpatient treatment site in Germany and randomized patients to receive either treatment as usual (TAU group) or an additional 30-minute intervention component that sought to amplify the placebo effect (PA group), mainly by boosting patients' realistic expectations regarding the effectiveness of antidepressants. The main outcome was the second edition of the Beck Depression Inventory (BDI-II; Beck, Steer, and Brown 1996) which is a self-report instrument for patients rating the severity and frequency of depressive symptoms. Lower values are beneficial because higher values indicate a more severe depressive syndrome. This outcome was measured four times for each patient: First, before the intervention, second and third during the intervention (weekly), and lastly at treatment termination. The original data are publicly available at https://osf.io/j439n/ and a reduced version is included in our package **clinicalsignificance**, which is readily available after loading the package.

```
R> library("clinicalsignificance")
R> set.seed(20230920)
R> claus_2020
```

```
 # A tibble: 172 × 9
       id   age sex    treatment  time   bdi shaps   who  hamd
   <dbl> <dbl> <fct>  <fct>     <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1    54 Male   TAU           1    33     9     0    25
 2     1    54 Male   TAU           2    28     6     3    17
 3     1    54 Male   TAU           3    28     9     7    13
 4     1    54 Male   TAU           4    27     8     3    13
 5     2    52 Female PA            1    26    11     2    15
 6     2    52 Female PA            2    26    10     0    16
 7     2    52 Female PA            3    25    10     0     7
 8     2    52 Female PA            4    19     9     3    11
 9     3    54 Male   PA            1    15     2     0    28
10     3    54 Male   PA            2    13     5     9    17
# ... with 162 more rows
```

As can be seen, this dataset is in long format, also known as tidy data (Wickham 2014), in which individual measurements of a single patient are represented in multiple rows. For instance, the first patient (for which variable `id` is 1) was measured, as expected, four times (each measurement corresponds to one value of variable `time`). The primary outcome – the BDI-II – is stored in variable `bdi`.

Our package **clinicalsignificance** offers the described approaches above via four main functions, `cs_anchor`, `cs_percentage`, `cs_distribution`, `cs_statistical`, and `cs_combined`. We will illustrate all methods using the data by Claus *et al.* (2020).

Anchor-based approach

To conduct a clinical significance analysis according to the anchor-based approach, the MID must be defined *a priori*. According to Hengartner and Plöderl (2022) and Button *et al.* (2015), the MID for the BDI-II can be estimated to be 6–7 points. Let's be conservative an consider MID = 7 for the BDI-II.

Given the dataset and the information on the MID, we can estimate the clinical significance of this very study by using the function `cs_anchor`.

```
R> anchor_individual <- cs_anchor(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, mid_improvement = 7)
```

First, we need to define the dataframe to use in the analysis with the `data` argument. In this case, the data is called `claus_2020`. Then, we need to supply the individual patient identifier column with the argument `id` (the column is also named `id` in the dataset), the column containing the individual measurement with the argument `time` (which is also called `time` in this dataset), and the column containing the outcome of interest with `outcome` (`bdi` in this case). Because the outcome was measured more than two times during the intervention, we also need to specify which measurements correspond to the pre and post intervention measurement with arguments `pre` and `post`, which are the first and fourth measurement in this case. Note that we would not need to specify `pre` and `post` if only two measurements were present in the data. Following that, we specify the MID with `mid_improvement`. One could also specify a different MID for a deterioration with `mid_deterioration`, but currently, we have no information on different MIDs for an improvement or a deterioration and consequently, `mid_deterioration` will be set to `mid_improvement` internally. Furthermore, the direction of a desired intervention effect can be set with the function argument `better_is`, which defaults to `better_is = "lower"` and is the correct decision for this negative instrument.

The results is an S3 object of class '`cs_analysis`' which is named `anchor_individual` in this case. When called alone, it prints the following table. From this output, we can verify the employed clinical significance approach (anchor-based in this case), the predefined MID and how many patients improved, deteriorated or did not change in relation to the MID.

```
R> anchor_individual

— Clinical Significance Results —

Individual anchor-based approach with a 7 point decrease in
instrument scores indicating a clinical significant improvement.
```

```
Category     |  n | Percent
---------------------------
Improved     | 25 |  62.50%
Unchanged    | 11 |  27.50%
Deteriorated |  4 |  10.00%
```

From this output, it can be determined, that 25 patients (62 % of the sample) showed a change that is at least as big as the MID in the beneficial direction, so a reduction in this case. 11 patients (28 %) changed lass than the MID and 4 patients (10 %) actually demonstrated an increase in BDI-II scores that exceeded the MID.

For further information, we can call the generic `summary` function with the '`cs_analysis`' object as the only argument.

```
R> summary(anchor_individual)

— Clinical Significance Results —

Individual anchor-based analysis of clinical significance with a 7
point decrease in instrument scores (bdi) indicating a clinical
significant improvement.

There were 43 participants in the whole dataset of which 40 (93%)
could be included in the analysis.


— Individual Level Results

Category     |  n | Percent
---------------------------
Improved     | 25 |  62.50%
Unchanged    | 11 |  27.50%
Deteriorated |  4 |  10.00%
```

This way, we can retrieve additional information, like the outcome defined by the user as well as the number of participants that could be used for the analysis, i.e., those with complete data at x_1 and x_2.

Furthermore, it is possible to plot the results of each analysis by passing the '`cs_analysis`' object to the `plot` generic.

```
R> plot(anchor_individual)
```

The resulting plot is shown in Figure 4A. Following the interpretation guidelines from Section 2.6, we can infer that the majority of patients showed a clinically significant change (improvement). Most patients had lower BDI-II scores after intervention as compared to before (the majority of points fall beneath the solid diagonal) and the 25 patients that are classified as "improved" are represented by the points beneath the shaded area. 11 patients

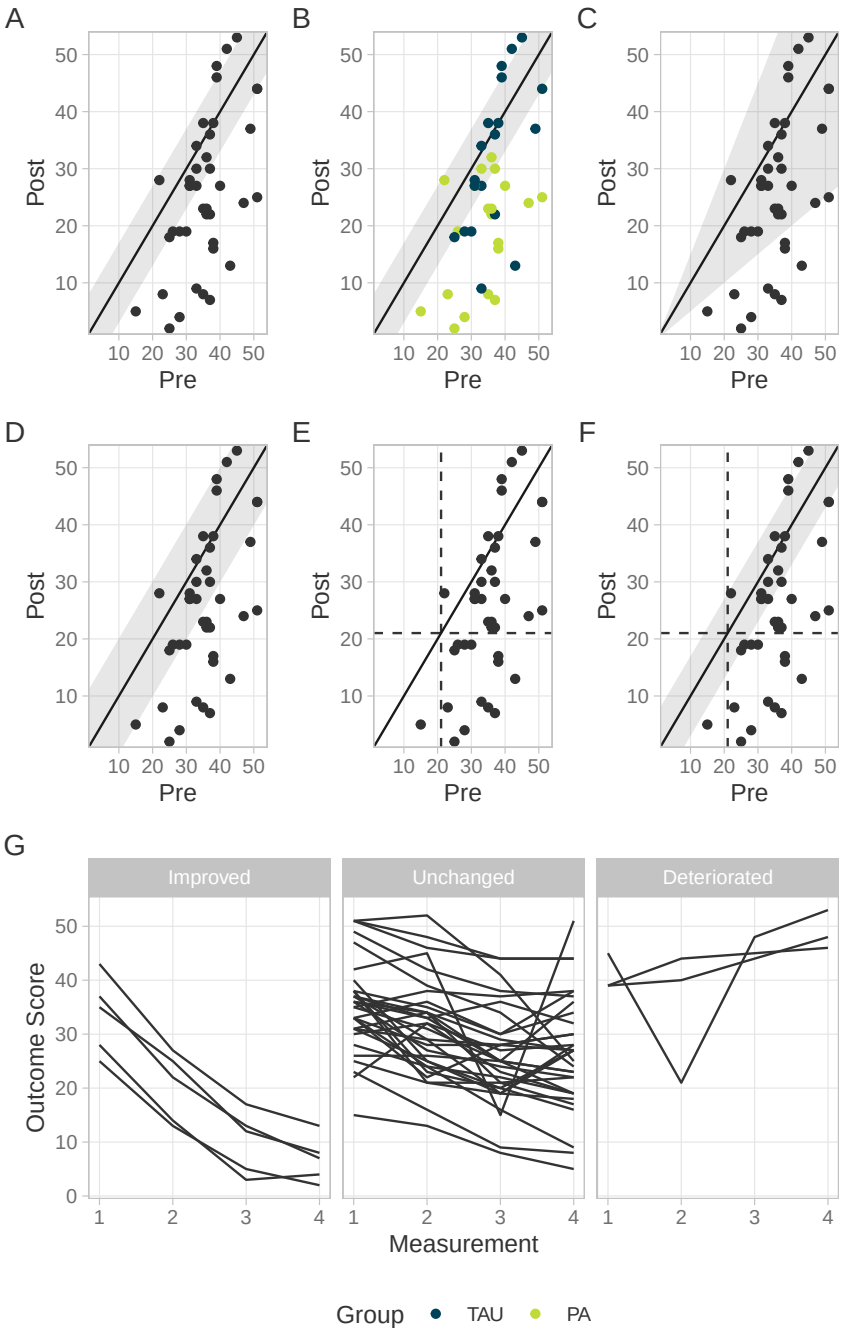

Figure 4: Clinical significance plots for different approaches and the same study by Claus *et al.* (2020), depicting (A) a basic clinical significance plot for the individual level anchor-based approach, (B) the same plot as in (A) but with a provided grouping variable, (C) the percentage-change approach, (D) the distribution-based approach, (E) the statistical approach, (F) the combined approach proposed by the authors, and (G) the HLM method for the distribution-based approach.

did not change meaningfully (these points are still inside the shaded area). Also, four patients had meaningfully worse scores after intervention (points above the shaded area).

A plot with colors indicating each participant's clinical significance category as is shown in Figure 3A can also be created by modifying the argument `show`. The resulting plot is not shown here to avoid redundancy but is created in the supplemental R script.

```
R> plot(anchor_individual, show = category)
```

Because the method of clinical significance gains information for every specific patient, one can extract this information by calling `cs_get_augmented_data` with the '`cs_analysis`' object being its only argument.

```
R> cs_get_augmented_data(anchor_individual)
```

```
 # A tibble: 40 × 8
 id   pre  post change improved deteriorated unchanged category
 <dbl> <dbl> <dbl>  <dbl> <lgl>    <lgl>        <lgl>     <fct>
 1     1    33    27     -6 FALSE    FALSE        TRUE      Unchanged
 2     2    26    19     -7 TRUE     FALSE        FALSE     Improved
 3     3    15     5    -10 TRUE     FALSE        FALSE     Improved
 4     5    39    46      7 FALSE    TRUE         FALSE     Deteriora...
 5     6    22    28      6 FALSE    FALSE        TRUE      Unchanged
 6     7    25    18     -7 TRUE     FALSE        FALSE     Improved
 7     8    33    30     -3 FALSE    FALSE        TRUE      Unchanged
 8     9    23     8    -15 TRUE     FALSE        FALSE     Improved
 9    10    47    24    -23 TRUE     FALSE        FALSE     Improved
10    11    43    13    -30 TRUE     FALSE        FALSE     Improved
```

From the resulting output, one can determine the individual incorporated pre and post intervention scores, the individual change, and clinical significance category. Thus, it is very simple to, for instance, filter out specific patients and review their change or identify unusual response patterns. We implemented various additional extractor functions to obtain necessary information from a '`cs_analysis`' object, which all are of the form `cs_get_*`.

Of particular interest is the fact, that Claus *et al.* (2020) examined two groups: the one undergoing treatment as usual (TAU) and the one receiving the placebo amplification treatment (PA). With every main function of our package we can specify a variable containing the grouping information and obtain group-specific results. The function call is identical to the one above, except that the `group` argument is defined.

```
R> anchor_individual_groups <- cs_anchor(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, mid_improvement = 7, group = treatment)
R> anchor_individual_groups

— Clinical Significance Results —

Individual anchor-based approach with a 7 point decrease in
instrument scores indicating a clinical significant improvement.
```

```
Group |     Category |  n | Percent
-----------------------------------
TAU   |     Improved |  8 |  20.00%
TAU   |    Unchanged |  7 |  17.50%
TAU   | Deteriorated |  4 |  10.00%
PA    |     Improved | 17 |  42.50%
PA    |    Unchanged |  4 |  10.00%
PA    | Deteriorated |  0 |   0.00%
```

More patients demonstrated a clinically meaningful improvement in the PA (42 %, placebo amplification) condition as compared to the TAU (20 %, treatment as usual) condition. Less patients in the PA condition can be classified as "unchanged" (10 % vs. 17 %) and all patients that showed a meaningful deterioration are members of the TAU condition (10 % of all incorporated patients). This result becomes even more apparent in the accompanying plot (see Figure 4B), which was created with

```
R> plot(anchor_individual_groups)
```

Another possibility would be to examine clinical significance with the anchor-based approach on the group level. This can be done by setting the function argument to `target = "group"`.

```
R> anchor_whole_group <- cs_anchor(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, mid_improvement = 7, target = "group")
R> anchor_whole_group

— Clinical Significance Results —

Groupwise anchor-based approach (within groups) with a 7 point
decrease in instrument scores indicating a clinical significant
improvement.

Median Difference | [Lower | Upper] | CI-Level | n  | Category
-------------------------------------------------------------------------------
-9.36             | -12.84 | -5.74  | 0.95     | 40 | Probably clinically si
gnificant effect
```

Note that by default, the Bayesian approach to significance testing is used (see Section 2.1). The median difference signifies that the sample as a whole demonstrated an average reduction in BDI-II scores of 9.36 points, with the 95 % CI ranging from −12.84 to −5.74. Because the median intervention effect exceeds the MID but the credible interval still contains it, the effect is probably clinically significant (category "D" in Figure 1). A figure depicting the results as Figure 1 can be created with

```
R> plot(anchor_whole_group)
```

but is not shown here.

Because Claus *et al.* (2020) examined two groups, it is wise to evaluate the intervention effect in both groups separately. This can be done by providing the `group` argument in `cs_anchor`.

```
R> anchor_group <- cs_anchor(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, mid_improvement = 7,
+    target = "group", group = treatment)
R> anchor_group
```

```
— Clinical Significance Results —

Groupwise anchor-based approach (within groups) with a 7 point
decrease in instrument scores indicating a clinical significant
improvement.

Group | Median Difference | [Lower | Upper] | CI-Level | n  | Category
------------------------------------------------------------------------
TAU   | -4.27             | -9.22  | 0.50   | 0.95     | 19 | Statistically
not significant
PA    | -13.73            | -18.23 | -9.18  | 0.95     | 21 | Large clinical
ly significant effect
```

From this we can see that the results for the whole group are biased because of the inherent grouping structure of the data. For the TAU group, not even a statistically significant change could be observed (because the credible interval does contain 0, see category "A" in Figure 1). The PA group change, however, demonstrated a large clinically significant effect (category "E" in Figure 1).

Additionally, one might not be interested in the effectiveness of both treatment groups as compared to a null effect but in the intervention effect of the active treatment condition in relation to the inactive one, i.e., PA vs. TAU. In this case, a between-groups analysis is needed, which can be performed by changing the `effect` argument from `effect = "within"` (the default) to `effect = "between"`. Note that the argument `post` must be set in this case to indicate at which measurement the groups should be compared. It is also possible to define a reference group to which (in the case of multiple groups) all subsequent groups are compared with the argument `reference_group` but it is not needed in this case because **clinicalsignificance** automatically picks the first factor level of an ordinary factor as the reference level, which is the TAU group in this dataset.

```
R> set.seed(20230920)
R> anchor_group_between <- cs_anchor(data = claus_2020, id = id,
+    time = time, outcome = bdi, post = 4, mid_improvement = 7,
+    target = "group", group = treatment, effect = "between")
R> anchor_group_between
```

The resulting output is not shown here for purposes of the layout but the median difference (and credible interval) is -10.31 $(-18.05, -2.62)$ which indicates a probable clinically significant effect (category "D" in Figure 1) in favor of the PA group. The function `set.seed(20230920)` is used to make the results reproducible, because the Bayesian approach requires random sampling and may vary slightly for each run.

Percentage-change approach

To conduct a percentage-change analysis of clinical significance, a PCC must be defined *a*

priori. In some medical field, a symptom reduction of 50 % is usually defined as a "response". Such an analysis could be conducted with `cs_percentage`. All arguments are identical to `cs_anchor` but instead of an MID, one defines the `pct_improvement` argument, i.e., the percentage-change to infer a meaningful change. As with `cs_anchor`, the PCC may be set for a deterioration separately and is assumed to be equal if only `pct_improvement` is defined. Also, a grouping variable may be provided to obtain group-wise results. If PCC = 0.5, then

```
R> percentage <- cs_percentage(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, pct_improvement = 0.5)
R> percentage

— Clinical Significance Results —

Percentage-change approach with a 50% decrease in instrument scores
indicating a clinical significant improvement.

Category     |  n | Percent
---------------------------
Improved     | 11 |  27.50%
Unchanged    | 29 |  72.50%
Deteriorated |  0 |   0.00%
```

From the output we can see, that the vast majority of patients, i.e., 72 % did not change in a meaningful way when a meaningful change is defined as a 50 % reduction of the initial instrument score. The BDI-II scores decreased more than 50 % for only 11 patients (28 %). A clinical significance plot can be created as well with

```
R> plot(percentage)
```

and the resulting output is shown in Figure 4C.

Distribution-based approach

A distribution-based oriented analysis of clinical significance can be done with the main function `cs_distribution`. For all RCI methods (except HLM), a reliability estimate must be provided by the user[2]. In the study by Claus *et al.* (2020), the reliability of the BDI-II was calculated from the data at hand and was estimated to be McDonald's $\omega = 0.801$. The resulting function call would then be

```
R> distribution_jt <- cs_distribution(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, reliability = 0.801)
R> distribution_jt
R> plot(distribution_jt)

— Clinical Significance Results —
```

[2]For the Nunnally and Kotsch (1983) method, two separate reliability estimates for the pre and post intervention measurement must be supplied.

```
Distribution-based approach using the JT method.

Category     |  n | Percent
---------------------------
Improved     | 18 |  45.00%
Unchanged    | 22 |  55.00%
Deteriorated |  0 |   0.00%
```

By default, the Jacobson and Truax (1991) (JT) method is used to calculate the RCI. From the output, we can see that, according to the distribution-based approach, that the majority of patients (55 %) did not change statistically reliably, but 28 patients (45 %) exhibited a change that was greater than the measurement error of the BDI-II . The results are shown in Figure 4D. To change the RCI method, the `rci_method` argument must be changed. If, for instance, the Hageman and Arrindell (1999b) (HA) method is desired, the function call would be

```
R> distribution_ha <- cs_distribution(data = claus_2020, id = id,
+    time = time, outcome = bdi, pre = 1, post = 4, reliability = 0.801,
+    rci_method = "HA")
R> distribution_ha

— Clinical Significance Results —

Distribution-based approach using the HA method.

Category     |  n | Percent
---------------------------
Improved     | 25 |  62.50%
Unchanged    | 15 |  37.50%
Deteriorated |  0 |   0.00%
```

One can easily see that the Hageman and Arrindell (1999b) method is more liberal regarding the RCI change criterion.

Another RCI is the HLM method, which can be used to incorporate all available data points per patient, if it's the case that each patient was measured more than twice. Again, the `rci_method` argument can be changed to achieve the following call. Notice that the definition of pre and post arguments are not needed here because all available data points will be considered for the analysis. Additionally, the reliability will be estimated from the data itself, so not reliability must be provided.

```
R> distribution_hlm <- cs_distribution(data = claus_2020, id = id,
+    time = time, outcome = bdi, rci_method = "HLM")
R> distribution_hlm
R> plot(distribution_hlm)

— Clinical Significance Results —
```

```
Distribution-based approach using the HLM method.

Category     |  n | Percent
---------------------------
Improved     |  5 |  12.50%
Unchanged    | 32 |  80.00%
Deteriorated |  3 |   7.50%
```

The accompanying plot is shown in Figure 4G. Of course, all distribution-based methods can be grouped by a variable present in the data as well.

Statistical approach

The statistical approach can be used to determine, which patients changed from a clinical to a functional population during the intervention. This can be achieved by using the main function `cs_statistical`. For the calculation of an optimal cutoff, summary statistics of our instrument for a functional population must be provided. Kühner, Bürger, Keller, and Hautzinger (2007) estimated that an adult German non-clinical sample yielded a mean value of $M = 7.69$ with a standard deviation of $s = 7.52$ on the BDI-II. Because we have this information, we can calculate cutoff of choice c (see Section 2.4). The appropriate function call would be

```
R> statistical_jt <- cs_statistical(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, m_functional = 7.69,
+    sd_functional = 7.53, cutoff_type = "c")
R> statistical_jt
R> plot(statistical_jt)

— Clinical Significance Results —

Statistical approach using the JT method.

Category     |  n | Percent
---------------------------
Improved     | 13 |  32.50%
Unchanged    | 27 |  67.50%
Deteriorated |  0 |   0.00%
```

with the clinical significance plot shown in Figure 4E. From the output, we can see that 68 % of patients did not change populations but 32 % did. Zero patient changed from the functional to the clinical population, i.e., no patient deteriorated according to the statistical approach. For the statistical method described by Hageman and Arrindell (1999b), the instrument's reliability must also be provided by the user with the argument `reliability`.

Combined approaches

Our package **clinicalsignificance** offers two combinations of the approaches outlined in Section 2. Firstly, the classic approach advocated by Jacobson *et al.* (1984) and Jacobson and

Truax (1991), i.e., the combination of the statistical with the distribution-based approach, as well as our proposed approach, so the combination of the statistical and individual anchor-based approach. Both approaches are implemented in the last main function, `cs_combined`.

To perform a clinical significance analysis for our approach, one would need an MID and (optionally, depending on the desired cutoff) summary statistics for the given instrument of a functional population. We take both information from above. Then the analysis can be carried out with

```
R> combined_cwb <- cs_combined(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, m_functional = 7.69,
+    sd_functional = 7.53, cutoff_type = "c", mid_improvement = 7)
R> combined_cwb
R> plot(combined_cwb)

— Clinical Significance Results —

Combined approach using the anchor-based and statistical approach.

Category     |  n | Percent
---------------------------
Recovered    | 13 |  32.50%
Improved     | 12 |  30.00%
Unchanged    | 11 |  27.50%
Deteriorated |  4 |  10.00%
Harmed       |  0 |   0.00%
```

with the resulting plot shown in Figure 4F. For 12 patients, change was at least as big as the MID but they did not change populations during intervention, hence they can be categorized as "improved". 13 patients additionally fulfilled the population change criterion and can be classified as "recovered" after the intervention. A minority of patients changed less than the MID (11 "unchanged" patients) and 4 patients exhibited a change equal to or greater than die MID but in the harmful direction, i.e., at least a 7 point increase in BDI-II scores. Fortunately, no patient changed from the functional to the clinical population during the intervention.

The classic clinical significance approach proposed by Jacobson and Truax (1991) can be run with the following command. As soon as the `mid_improvement` argument is not provided, the classic approach will be computed.

```
R> combined_jt <- cs_combined(data = claus_2020, id = id, time = time,
+    outcome = bdi, pre = 1, post = 4, m_functional = 7.69,
+    sd_functional = 7.53, cutoff_type = "c", reliability = 0.801)
R> combined_jt

— Clinical Significance Results —

Combined approach using the JT and statistical approach.
```

```
Category     |  n | Percent
-------------------------
Recovered    | 10 |  25.00%
Improved     |  8 |  20.00%
Unchanged    | 22 |  55.00%
Deteriorated |  0 |   0.00%
Harmed       |  0 |   0.00%
```

Because the RCI is greater than the MID in this specific example, more patients are classified as "unchanged" as opposed to our combined approach (22 vs. 11 patients). 8 patients belong to the "improved" group, i.e., their change was greater than the error of the measurement but they did not change populations during the intervention. 10 patients showed a change greater than the RCI and changed populations, thereby fulfilling both criteria.

From this comparison between the two approaches, the interpretation advantages of our approach become clear: The "improved" group in our approach offers a meaningful interpretation, because these patients demonstrated a change that, at least, can be believed to be practically relevant. In the classic Jacobson and Truax (1991) approach, this group only showed a reliable change which is not indicative of a clinically meaningful change (De Vet and Terwee 2010).

3.2. Treating children and adolescents with chronic pain

Chronic pain in children and adolescents is a highly prevalent and debilitating condition (Kashikar-Zuck *et al.* 2010; King 2011; Könning, Rosenthal, Brown, Stahlschmidt, and Wager 2021; Wilson and Palermo 2012; Zernikow *et al.* 2012). Hechler *et al.* (2014) report a clinical trial to estimate the effectiveness and efficiency of an intensive interdisciplinary pain treatment for the group of patients that have to undergo this kind of treatment because of their highly impacting pain. Claus *et al.* (2022) have recently shown through a systematic review and meta-analysis that this kind of treatment results in great improvements in pain-related disability at 6-month follow-up (FU, $g_{\text{rm}} = -1.77$). This statement was based on group-wise effect sizes, but are those changes in pain-related disability also meaningful for individuals?

The accompanying data related to this question is provided by the package again. It is called `hechler_2014` and can be loaded by

```
R> hechler_2014

# A tibble: 208 × 3
   patient measurement disability
     <dbl>       <dbl>      <dbl>
 1       2           1         46
 2       2           2         NA
 3       3           1         41
 4       3           2         12
 5       6           1         29
 6       6           2         13
 7       9           1         30
```

```
 8       9           2         14
 9      13           1         41
10      13           2         12
# ... with 198 more rows
```

Hechler *et al.* (2014) measured pain-related disability with the PPDI (Hübner *et al.* 2009). Because an MID for this instrument is currently not known, the anchor-based approach is no option. The distribution-based approach may be feasible because it only incorporates information from the data at hand. On the other hand, the consideration of information from outside of a given study is also needed to examine the meaningfulness of results. Thus, the classic approach proposed by Jacobson and Truax (1991) may be a suitable alternative to investigating the clinical significance in this study. Consequently, as in the previous example, we need descriptive data for pain-related disability from a functional population to carry out the clinical significance analysis. Lambert and Ogles (2009) as well as Ogles *et al.* (2001) rightfully question what the functional population should be in the case of a chronic health condition like chronic pain. In one approach, Tingey *et al.* (1996) propose to employ the dimensional approach of clinical significance again and define a population that is *more* functional than the clinical population, which may still experience the disorder in question but is more functional in other areas. For instance, Könning *et al.* (2021) determined the severity of chronic pain in German adolescents that have chronic pain but still go to school and did not see a health professional for their pain in the last three months. These criteria were fulfilled by $n = 595$ adolescents, who showed a mean pain-related disability of $M = 26.7$ ($s = 9.14$) as measured with the PPDI (these summary statistics were calculated by the present authors, based on previously unpublished data). Higher values indicate a higher pain-related disability, so lower values are desirable. According to Hübner *et al.* (2009), we can assume coefficient $\alpha = 0.865$ as the measure of reliability.

The clinical significance analysis can be done with the following function call.

```
R> hechler_results <- cs_combined(data = hechler_2014, id = patient,
+    time = measurement, outcome = disability, m_functional = 26.7,
+    sd_functional = 9.14, cutoff_type = "c", reliability = 0.865)
R> summary(hechler_results)

— Clinical Significance Results —

Combined analysis of clinical significance using the JT and
statistical approach method for calculating the RCI and population
cutoffs.

There were 104 participants in the whole dataset of which 92 (88.5%)
could be included in the analysis.

The outcome was disability and the reliability was set to 0.865.

The cutoff type was c with a value of 31.17 based on the following
summary statistics:
```

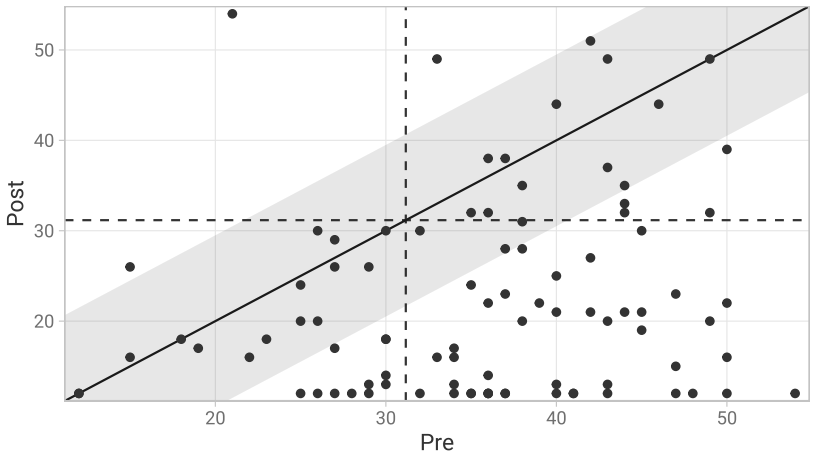

Figure 5: Clinical significance analysis of the trial reported by Hechler *et al.* (2014) for the 6-month FU results. Most children and adolescents can be either categorized as improved or recovered. One patient deteriorated and one patient can be classified as harmed.

```
— Population Characteristics
M Clinical | SD Clinical | M Functional | SD Functional
-----------------------------------------------------------
35.73      |        9.32 |        26.70 |          9.14

— Individual Level Results
Category     |  n | Percent
---------------------------
Recovered    | 44 |  47.83%
Improved     | 15 |  16.30%
Unchanged    | 30 |  32.61%
Deteriorated |  2 |   2.17%
Harmed       |  1 |   1.09%
```

From this summary, we can see that, in fact, 48 % of examined patients can be seen to have changed clinical significantly, relative to a sample of adolescents who go to school but have chronic pain. One can also observe that two patients were classified to have deteriorated and one even harmed (i.e., showed a reliable change and changed from the functional to the clinical population at 6-month FU). This becomes also apparent by examining the clinical significance plot, which can be seen in Figure 5.

```
R> plot(hechler_results)
```

4. Reporting recommendations

We recommend researchers to provide the following parameters when reporting a clinical significance analysis to ensure reproducibility and replicability of research findings:

- Means and standard deviations of pre and post intervention scores.
- Effect size estimates for this difference.
- The employed clinical significance approach.
- Clinical significance analysis results formatted as a table.
- A clinical significance plot.
- All defined arguments besides those needed for data wrangling. State why you chose specific values (e.g., for the MID) with corresponding, relevant references.
- The software used to calculate the results (including version numbers).
- If possible, publish your anonymized data and analysis scripts in scientific online repositories.

5. Summary and discussion

We introduced the new R package **clinicalsignificance**, in which the most relevant methods to conduct clinical significance analyses of intervention trials are implemented, with results displayed in a consistent and publication ready form. In doing so, we hope to foster research on the *clinical*, as opposed to statistical, significance of interventions and research on the very method itself. As can be seen from Section 3, the results of such an analysis can differ based on which clinical significance method is used. Fortunately, this package can certainly be used to simplify research on clinical significance which may yield even more informed recommendations as to which method should be used.

The validity of the clinical significance approach has been shown, for instance, by Ronk, Hooke, and Page (2016) who compared the results of clinical significance analyses regarding depression against a measure of life satisfaction and enjoyment and found a good correspondence and, hence, validity of this approach. See Lambert and Ogles (2009) for further details and other examples.

Nonetheless, clinical significance analyses have shortcomings as well. First and foremost, even if researchers examine the clinical significance and state the employed analysis methods, indices and results can vary considerably between studies (Lambert and Ogles 2009). This issue may be solved by using this package in routine research practice to ensure that all indices are calculated in a consistent way. Lavigne (2016), Crosby *et al.* (2003), and De Vet and Terwee (2010) list various advantages and disadvantages for the clinical significance approaches outlined here and implemented in **clinicalsignificance**. We encourage future researchers to bear those in mind and compare as well as potentially improve these approaches as well.

To further facilitate the replicable use of this method, we offer to host a database of MIDs and functional population descriptive statistics on the package website at `https://pedscience.github.io/clinicalsignificance/` which then can be chosen when using the package, or when conducting a clinical significance analysis without using the package. This prevents researchers from relying on different population statistics which may improve consistency greatly and is urgently needed to interpret the results across studies. We encourage researchers

to get in touch with us to include MIDs, or descriptive statistics of functional populations which are based on big representative samples in a variety of cultures and languages.

Computational details

The estimation of hierarchical linear models (HLM) relies on the **lme4** package (Bates, Mächler, Bolker, and Walker 2015), Bayesian t tests are calculated using the R package **BayesFactor** (Morey and Rouder 2023), and the plots are created with **ggplot2** (Wickham 2016) internally.

References

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using **lme4**." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Beck AT, Steer RA, Brown GK (1996). *Manual for the BDI-II*. 1st edition. The Psychological Corporation, San Antonio.

Button KS, Kounali D, Thomas L, Wiles NJ, Peters TJ, Welton NJ, Ades AE, Lewis G (2015). "Minimal Clinically Important Difference on the Beck Depression Inventory-II According to the Patient's Perspective." *Psychological Medicine*, **45**(15), 3269–3279. doi:10.1017/s0033291715001270.

Christensen L, Mendoza JL (1986). "A Method of Assessing Change in a Single Subject: An Alteration of the RC Index." *Behavior Therapy*, **17**(3), 305–308. doi:10.1016/s0005-7894(86)80060-0.

Claus B (2024). ***clinicalsignificance**: A Toolbox for Clinical Significance Analyses in Intervention Studies*. doi:10.32614/CRAN.package.clinicalsignificance. R package version 2.1.0.

Claus BB, Scherbaum N, Bonnet U (2020). "Effectiveness of an Adjunctive Psychotherapeutic Intervention Developed for Enhancing the Placebo Effect of Antidepressants Used within an Inpatient-Treatment Program of Major Depression: A Pragmatic Parallel-Group, Randomized Controlled Trial." *Psychotherapy and Psychosomatics*, **89**(4), 258–260. doi:10.1159/000505855.

Claus BB, Stahlschmidt L, Dunford E, Major J, Harbeck-Weber C, Bhandari RP, Baerveldt A, Neß V, Grochowska K, Hübner-Möhler B, Zernikow B, Wager J (2022). "Intensive Interdisciplinary Pain Treatment for Children and Adolescents with Chronic Non-Cancer Pain: A Preregistered Systematic Review and Individual Patient Data Meta-Analysis." *Pain*. doi:10.1097/j.pain.0000000000002636.

Cohen HW (2011). "P Values: Use and Misuse in Medical Literature." *American Journal of Hypertension*, **24**(1), 18–23. doi:10.1038/ajh.2010.205.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Lawrence Erlbaum Associates.

Cronbach LJ, Gleser GC (1959). "Interpretation of Reliability and Validity Coefficients: Remarks on a Paper by Lord." *Journal of Educational Psychology*, **50**(5), 230–237. doi:10.1037/h0042848.

Crosby RD, Kolotkin RL, Williams GR (2003). "Defining Clinically Meaningful Change in Health-Related Quality of Life." *Journal of Clinical Epidemiology*, **56**(5), 395–407. doi:10.1016/s0895-4356(03)00044-1.

De Vet HCW, Terwee CB (2010). "The Minimal Detectable Change Should Not Replace the Minimal Important Difference." *Journal of Clinical Epidemiology*, **63**(7), 804–805. doi:10.1016/j.jclinepi.2009.12.015.

De Vries RM, Meijer RR, Van Bruggen V, Morey RD (2016). "Improving the Analysis of Routine Outcome Measurement Data: What a Bayesian Approach Can Do for You." *International Journal of Methods in Psychiatric Research*, **25**(3), 155–167. doi:10.1002/mpr.1496.

De Vries RM, Morey RD (2013). "Bayesian Hypothesis Testing for Single-Subject Designs." *Psychological Methods*, **18**(2), 165–185. doi:10.1037/a0031037.

Edwards DW, Yarvis RM, Mueller DP, Zingale HC, Wagman WJ (1978). "Test-Taking and the Stability of Adjustment Scales." *Evaluation Quarterly*, **2**(2), 275–291. doi:10.1177/0193841x7800200206.

Engel L, Beaton DE, Touma Z (2018). "Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation." *Rheumatic Diseases Clinics of North America*, **44**(2), 177–188. doi:10.1016/j.rdc.2018.01.011.

Etz A, Vandekerckhove J (2018). "Introduction to Bayesian Inference for Psychology." *Psychonomic Bulletin & Review*, **25**(1), 5–34. doi:10.3758/s13423-017-1262-3.

Field A (2018). *Discovering Statistics Using IBM SPSS Statistics*. 5th edition. Sage Publications, Los Angeles.

Finch WH, Bolin JE, Kelley K (2019). *Multilevel Modeling Using R*. 2nd edition. Chapman & Hall/CRC, Boca Raton. doi:10.1201/9781351062268.

Flora DB (2020). "Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates." *Advances in Methods and Practices in Psychological Science*, **3**(4), 484–501. doi:10.1177/2515245920951747.

Funder DC, Ozer DJ (2019). "Evaluating Effect Size in Psychological Research: Sense and Nonsense." *Advances in Methods and Practices in Psychological Science*, **2**(2), 156–168. doi:10.1177/2515245919847202.

Gao J (2020). "*P*-Values – A Chronic Conundrum." *BMC Medical Research Methodology*, **20**(1), 167. doi:10.1186/s12874-020-01051-6.

Goodman S (2008). "A Dirty Dozen: Twelve *p*-Value Misconceptions." *Seminars in Hematology*, **45**(3), 135–140. doi:10.1053/j.seminhematol.2008.04.003.

Greenberger D, Padesky CA, Beck AT (2016). *Mind over Mood: Change How You Feel by Changing the Way You Think*. 2nd edition. Guilford Publications, New York.

Grice JW, Medellin E, Jones I, Horvath S, McDaniel H, O'lansen C, Baker M (2020). "Persons as Effect Sizes." *Advances in Methods and Practices in Psychological Science*, **3**(4), 443–455. doi:10.1177/2515245920922982.

Gulliksen H (1950). *Theory of Mental Tests*. 1st edition. John Wiley & Sons, New York.

Hageman WJ, Arrindell WA (1999a). "Clinically Significant and Practical! Enhancing Precision Does Make a Difference. Reply to McGlinchey and Jacobson, Hsu, and Speer." *Behaviour Research and Therapy*, **37**(12), 1219–1233. doi:10.1016/s0005-7967(99)00036-4.

Hageman WJ, Arrindell WA (1999b). "Establishing Clinically Significant Change: Increment of Precision and the Distinction between Individual and Group Level Analysis." *Behaviour Research and Therapy*, **37**(12), 1169–1193. doi:10.1016/s0005-7967(99)00032-7.

Hechler T, Ruhe AK, Schmidt P, Hirsch J, Wager J, Dobe M, Krummenauer F, Zernikow B (2014). "Inpatient Based Intensive Interdisciplinary Pain Treatment for Highly Impaired Children with Severe Chronic Pain: Randomized Controlled Trial of Efficacy and Economic Effects." *Pain*, **155**(1), 118–128. doi:10.1016/j.pain.2013.09.015.

Hengartner MP (2022). *Evidence-Biased Antidepressant Prescription: Overmedicalisation, Flawed Research, and Conflicts of Interest*. 1st edition. Springer-Verlag, Cham. doi:10.1007/978-3-030-82587-4.

Hengartner MP, Plöderl M (2018). "Statistically Significant Antidepressant-Placebo Differences on Subjective Symptom-Rating Scales Do Not Prove That the Drugs Work: Effect Size and Method Bias Matter!" *Frontiers in Psychiatry*, **9**, 517. doi:10.3389/fpsyt.2018.00517.

Hengartner MP, Plöderl M (2022). "Estimates of the Minimal Important Difference to Evaluate the Clinical Significance of Antidepressants in the Acute Treatment of Moderate-to-Severe Depression." *BMJ Evidence-Based Medicine*, **27**(2), 69–73. doi:10.1136/bmjebm-2020-111600.

Hespanhol L, Vallio CS, Costa LM, Saragiotto BT (2019). "Understanding and Interpreting Confidence and Credible Intervals around Effect Estimates." *Brazilian Journal of Physical Therapy*, **23**(4), 290–301. doi:10.1016/j.bjpt.2018.12.006.

Hsu LM (1989). "Reliable Changes in Psychotherapy: Taking into Account Regression Toward the Mean." *Behavioral Assessment*, **11**(4), 459–467. doi:10.1037/h0085462.

Hsu LM (1995). "Regression Toward the Mean Associated with Measurement Error and the Identification of Improvement and Deterioration in Psychotherapy." *Journal of Consulting and Clinical Psychology*, **63**(1), 141–144. doi:10.1037/0022-006x.63.1.141.

Hsu LM (1996). "On the Identification of Clinically Significant Client Changes: Reinterpretation of Jacobson's Cut Scores." *Journal of Psychopathology and Behavioral Assessment*, **18**(4), 371–385. doi:10.1007/bf02229141.

Hsu LM (1999). "Caveats Concerning Comparisons of Change Rates Obtained with Five Methods of Identifying Significant Client Changes: Comment on Speer and Greenbaum (1995)." *Journal of Consulting and Clinical Psychology*, **67**(4), 594–598. doi:10.1037/0022-006x.67.4.594.

Hübner B, Hechler T, Dobe M, Damschen U, Kosfelder J, Denecke H, Schroeder S, Zernikow B (2009). "Schmerzbezogene Beeinträchtigung bei Jugendlichen mit Chronischen Schmerzen. Erste Überprüfung des Pediatric Pain Disability Index (P-PDI)." *Schmerz*, **23**(1), 20–32. doi:10.1007/s00482-008-0730-0.

IBM Corp (2021). "IBM SPSS Statistics for Windows Version 28.0."

Jacobson NS, Follette WC, Revenstorf D (1984). "Psychotherapy Outcome Research: Methods for Reporting Variability and Evaluating Clinical Significance." *Behavior Therapy*, **15**(4), 336–352. doi:10.1016/s0005-7894(84)80002-7.

Jacobson NS, Truax P (1991). "Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research." *Journal of Consulting and Clinical Psychology*, **59**(1), 12–19. doi:10.1037/0022-006x.59.1.12.

Jakobsen JC, Gluud C, Kirsch I (2020). "Should Antidepressants Be Used for Major Depressive Disorder?" *BMJ Evidence-Based Medicine*, **25**(4), 130. doi:10.1136/bmjebm-2019-111238.

JASP Team (2022). "**JASP** Version 0.16.4." URL https://jasp-stats.org/.

Kashikar-Zuck S, Johnston M, Ting TV, Graham BT, Lynch-Jordan AM, Verkamp E, Passo M, Schikler KN, Hashkes PJ, Spalding S, Banez GA, Richards MM, Powers SW, Arnold LM, Lovell D (2010). "Relationship between School Absenteeism and Depressive Symptoms among Adolescents with Juvenile Fibromyalgia." *Journal of Pediatric Psychology*, **35**(9), 996–1004. doi:10.1093/jpepsy/jsq020.

Kieser M, Hauschke D (2005). "Assessment of Clinical Relevance by Considering Point Estimates and Associated Confidence Intervals." *Pharmaceutical Statistics*, **4**(2), 101–107. doi:10.1002/pst.161.

King MT (2011). "A Point of Minimal Important Difference (MID): A Critique of Terminology and Methods." *Expert Review of Pharmacoeconomics & Outcomes Research*, **11**(2), 171–184. doi:10.1586/erp.11.9.

Kirsch I (2010). *The Emperor's New Drugs: Exploding the Antidepressant Myth*. 1st edition. Basic Books.

Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT (2008). "Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration." *PLoS Medicine*, **5**(2), e45. doi:10.1371/journal.pmed.0050045.

Kirsch I, Sapirstein G (1998). "Listening to Prozac but Hearing Placebo: A Meta-Analysis of Antidepressant Medication." *Prevention & Treatment*, **1**(2). doi:10.1037/1522-3736.1.1.12a.

Könning A, Rosenthal N, Brown D, Stahlschmidt L, Wager J (2021). "Severity of Chronic Pain in German Adolescent School Students: A Cross-Sectional Study." *The Clinical Journal of Pain*, **37**(2), 118–125. doi:10.1097/ajp.0000000000000898.

Kruschke JK (2015). *Doing Bayesian Data Analysis: A Tutorial Introduction with R, JAGS, and Stan*. 2nd edition. Elsevier Academic Press, Amsterdam.

Kühner C, Bürger C, Keller F, Hautzinger M (2007). "Reliabilität Und Validität Des Revidierten Beck-Depressionsinventars (BDI-II). Befunde Aus Deutschsprachigen Stichproben." *Der Nervenarzt*, **78**(6), 651–656. doi:10.1007/s00115-006-2098-7.

Lakens D (2013). "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science a Practical Primer for *t*-Tests and ANOVAs." *Frontiers in Psychology*, **4**, 863. doi:10.3389/fpsyg.2013.00863.

Lambert MJ, Ogles BM (2009). "Using Clinical Significance in Psychotherapy Outcome Research: The Need for a Common Procedure and Validity Data." *Psychotherapy Research*, **19**(4-5), 493–501. doi:10.1080/10503300902849483.

Lavigne JV (2016). "Systematic Review: Issues in Measuring Clinically Meaningful Change in Self-Reported Chronic Pediatric Pain Intensity." *Journal of Pediatric Psychology*, **41**(7), 715–734. doi:10.1093/jpepsy/jsv161.

Lemon J (2016). "**clinsig**: Clinical Significance Functions." doi:10.32614/CRAN.package.clinsig. R package version 1.2.

Leucht S, Fennema H, Engel R, Kaspers-Janssen M, Lepping P, Szegedi A (2013). "What Does the HAMD Mean?" *Journal of Affective Disorders*, **148**(2-3), 243–248. doi:10.1016/j.jad.2012.12.001.

Leucht S, Fennema H, Engel RR, Kaspers-Janssen M, Lepping P, Szegedi A (2017). "What Does the MADRS Mean? Equipercentile Linking with the CGI Using a Company Database of Mirtazapine Studies." *Journal of Affective Disorders*, **210**, 287–293. doi:10.1016/j.jad.2016.12.041.

Liu S, Kuppens P, Bringmann L (2021). "On the Use of Empirical Bayes Estimates as Measures of Individual Traits." *Assessment*, **28**(3), 845–857. doi:10.1177/1073191119885019.

Lord FM, Novick MR (1968). *Statistical Theories of Mental Test Scores*. 1st edition. Addison-Wesley, Reading.

Love J, Selker R, Marsman M, Jamil T, Dropmann D, Verhagen J, Ly A, Gronau QF, Šmíra M, Epskamp S, Matzke D, Wild A, Knight P, Rouder JN, Morey RD, Wagenmakers EJ (2019). "**JASP**: Graphical Statistical Software for Common Statistical Designs." *Journal of Statistical Software*, **88**(2), 1–17. doi:10.18637/jss.v088.i02.

Margraf J, Scholten S (2018). "Risiken Und Nebenwirkungen." In J Margraf, S Schneider (eds.), *Lehrbuch Der Verhaltenstherapie, Band 1*, pp. 213–228. Springer-Verlag, Berlin. doi:10.1007/978-3-662-54911-7_14.

Martinovich Z, Saunders S, Howard K (1996). "Some Comments on 'Assessing Clinical Significance'." *Psychotherapy Research*, **6**(2), 124–132. doi:10.1080/10503309612331331648.

McGlinchey JB, Jacobson NS (1999). "Clinically Significant but Impractical? A Response to Hageman and Arrindell." *Behaviour Research and Therapy*, **37**(12), 1211–1217. doi:10.1016/s0005-7967(99)00035-2.

Molenaar PCM (2004). "A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever." *Measurement: Interdisciplinary Research & Perspective*, **2**(4), 201–218. doi:10.1207/s15366359mea0204_1.

Moncrieff J, Wessely S, Hardy R (2004). "Active Placebos versus Antidepressants for Depression." *The Cochrane Database of Systematic Reviews*, **2004**(1), CD003012. doi:10.1002/14651858.cd003012.pub2.

Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ (2016). "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin & Review*, **23**(1), 103–123. doi:10.3758/s13423-015-0947-8.

Morey RD, Rouder JN (2023). "**BayesFactor**: Computation of Bayes Factors for Common Designs." doi:10.32614/CRAN.package.BayesFactor. R package version 0.9.12-4.7.

Nunnally JC (1967). *Psychometric Therory*. 1st edition. McGraw-Hill, New York.

Nunnally JC (1975). "The Study of Change in Evaluation Research: Principles Concerning Measurement Experimental Design and Analysis." In EL Streuning, M Guttentag (eds.), *Handbook of Evaluation Research*. Sage Publications, Beverly Hills.

Nunnally JC, Kotsch WE (1983). "Studies of Individual Subjects: Logic and Methods of Analysis." *British Journal of Clinical Psychology*, **22**(2), 83–93. doi:10.1111/j.2044-8260.1983.tb00582.x.

Ogles BM, Lunnen KM, Bonesteel K (2001). "Clinical Significance: History, Application, and Current Practice." *Clinical Psychology Review*, **21**(3), 421–446. doi:10.1016/s0272-7358(99)00058-6.

Raudenbush SW, Bryk AS (2002). *Hierarchical Linear Models – Applications and Data Analysis Methods*. 2nd edition. Sage Publications, Thousand Oaks.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ronk FR, Hooke GR, Page AC (2016). "Validity of Clinically Significant Change Classifications Yielded by Jacobson-Truax and Hageman-Arrindell Methods." *BMC Psychiatry*, **16**, 187. doi:10.1186/s12888-016-0895-5.

Rouder JN, Haaf JM, Vandekerckhove J (2018). "Bayesian Inference for Psychology, Part IV: Parameter Estimation and Bayes Factors." *Psychonomic Bulletin & Review*, **25**(1), 102–113. doi:10.3758/s13423-017-1420-7.

Speer DC (1992). "Clinically Significant Change: Jacobson and Truax (1991) Revisited." *Journal of Consulting and Clinical Psychology*, **60**(3), 402–408. doi:10.1037/0022-006x.60.3.402.

Tingey R, Lambert M, Burlingame G, Hansen N (1996). "Assessing Clinical Significance: Proposed Extensions to Method." *Psychotherapy Research*, **6**(2), 109–123. doi:10.1080/10503309612331331638.

Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, Guyatt GH (2010). "The Minimal Detectable Change Cannot Reliably Replace the Minimal Important Difference." *Journal of Clinical Epidemiology*, **63**(1), 28–36. doi:10.1016/j.jclinepi.2009.01.024.

Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, Selker R, Gronau QF, Šmíra M, Epskamp S, Matzke D, Rouder JN, Morey RD (2018). "Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications." *Psychonomic Bulletin & Review*, **25**(1), 35–57. doi:10.3758/s13423-017-1343-3.

Wasserstein RL, Schirm AL, Lazar NA (2019). "Moving to a World Beyond '$p < 0.05$'." *The American Statistician*, **73**(sup1), 1–19. doi:10.1080/00031305.2019.1583913.

Whitaker R (2010). *Anatomy of an Epidemic: Magic Bullets, Psychiatric Drugs, and the Astonishing Rise of Mental Illness in America*. Broadway Books, New York.

Wickham H (2014). "Tidy Data." *Journal of Statistical Software*, **59**(10), 1–23. doi:10.18637/jss.v059.i10.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. 1st edition. Springer-Verlag. doi:10.1007/978-0-387-98141-3.

Wilson AC, Palermo TM (2012). "Physical Activity and Function in Adolescents with Chronic Pain: A Controlled Study Using Actigraphy." *The Journal of Pain*, **13**(2), 121–130. doi:10.1016/j.jpain.2011.08.008.

Zernikow B, Wager J, Hechler T, Hasan C, Rohr U, Dobe M, Meyer A, Hübner-Möhler B, Wamsler C, Blankenburg M (2012). "Characteristics of Highly Impaired Children with Severe Chronic Pain: A 5-Year Retrospective Study on 2249 Pediatric Pain Patients." *BMC Pediatrics*, **12**, 54. doi:10.1186/1471-2431-12-54.

Affiliation:

Benedikt B. Claus
German Paediatric Pain Centre
Department of Children's Pain Therapy and Paediatric Palliative Care
Witten/Herdecke University
Dr.-Friedrich-Steiner-Str. 5, 45711 Datteln, Germany
E-mail: b.claus@pedscience.de

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
November 2024, Volume 111, Issue 1 *Submitted:* 2022-12-20
doi:10.18637/jss.v111.i01 *Accepted:* 2023-11-17