Can psychological features predict antidepressant response to rTMS? A Discovery–Replication approach

Noralie Krepel1,2, A. John Rush3,4,5, Tabitha A. Iseger2,6, Alexander T. Sack1,* and Martijn Arns2,6,7,*

1Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands; 2Research Institute Brainclinics, Nijmegen, The Netherlands; 3Duke-National University of Singapore, Singapore; 4Duke Medical School, Durham, NC, USA; 5Texas Tech University Health Sciences Center, Permian Basin, TX, USA; 6Department of Experimental Psychology, Utrecht University, Utrecht, The Netherlands and 7neuroCare Group Netherlands, Nijmegen, The Netherlands

Abstract

Background. Few studies focused on the relationship between psychological measures, major depressive disorder (MDD) and repetitive transcranial magnetic stimulation (rTMS) response. This study investigated several psychological measures as potential predictors for rTMS treatment response. Additionally, this study employed two approaches to evaluate the robustness of our findings by implementing immediate replication and full-sample exploration with strict p-thresholding.

Methods. This study is an open-label, multi-site study with a total of 196 MDD patients. The sample was subdivided in a Discovery (60% of total sample, n = 119) and Replication sample (40% of total sample, n = 77). Patients were treated with right low frequency (1 Hz) or left high frequency (10 Hz) rTMS at the dorsolateral prefrontal cortex. Clinical variables [Beck Depression Inventory (BDI), Neuroticism, Extraversion, Openness Five-Factor Inventory, and Depression, Anxiety, and Stress Scale, and BDI subscales] were obtained at baseline, post-treatment, and at follow-up. Predictors were analyzed in terms of statistical association, robustness (independent replication), as well as for their clinical relevance [positive predictive value (PPV) and negative predictive value (NPV)].

Results. Univariate analyses revealed that non-responders had higher baseline anhedonia scores. Anhedonia scores at baseline correlated negatively with total BDI percentage change over time. This finding was replicated. However, anhedonia scores showed to be marginally predictive of rTMS response, and neither PPV nor NPV reached the levels of clinical relevance.

Conclusions. This study suggests that non-responders to rTMS treatment have higher baseline anhedonia scores. However, anhedonia was only marginally predictive of rTMS response. Since all other psychological measures did not show predictive value, it is concluded that psychological measures cannot be used as clinically relevant predictors to rTMS response in MDD.

Introduction

Major depressive disorder (MDD) is a chronic mental disease with a remitting and relapsing course. Repetitive transcranial magnetic stimulation (rTMS) as a treatment method for MDD has been studied thoroughly over the past few years. High frequency (HF, 10 Hz) rTMS applied to the left dorsolateral prefrontal cortex (DLPFC) (Schutter, 2009) as well as low frequency (LF, 1 Hz) rTMS applied to the right DLPFC (Schutter, 2010) showed to have antidepressant effects. Additionally, left and right DLPFC stimulation seem to have similar clinical effects (Fitzgerald et al., 2003, 2009). rTMS also showed to be effective in treatment-resistant depression (Gaynes et al., 2014). However, even though rTMS is well accepted as a treatment option for MDD, response rates remain relatively low, ranging from 29.3% (Berlim et al., 2014) to 58% (Carpenter et al., 2012) in HF-rTMS, which is similar to a study that employed LF as well as HF rTMS [response rates 45% and 44%, respectively (Fitzgerald et al., 2009)]. Finding psychological measures that are associated with treatment response may help to identify those patients who have a greater chance of achieving response. Additionally, finding predictors that can enhance treatment allocation accuracy might increase response rates by immediately indicating the optimal treatment to a given patient, thereby saving time and money.

Substantial research has been done investigating the relationship between behavior, depression, and treatment response. Frequently studied domains include the ‘Big Five Personality Traits’ [neuroticism, extraversion, openness, conscientiousness, and agreeableness (Goldberg, 1990)], anhedonia, depression severity, stress, and anxiety. A robust finding seems to be the relationship between the personality traits neuroticism and extraversion, wherein neuroticism...
seems to be positively associated to MDD (Jylhä and Isometsä, 2006; Griffith et al., 2010; Kotov et al., 2010; Rosellini and Brown, 2011; Hayward et al., 2013), and extraversion seems to be inversely associated with MDD (Jylhä and Isometsä, 2006; Kotov et al., 2010; Rosellini and Brown, 2011; Hayward et al., 2013). Other studies have also elaborated on the association between personality traits and MDD by including treatment response. For example, Bagby and colleagues found that MDD patients with higher scores on neuroticism are more likely to respond to pharmacotherapy, rather than to cognitive behavioral therapy (Bagby et al., 2008). Similarly, Quilty and colleagues found that neuroticism was indicative of a lower probability of response, whereas conscientiousness was predictive of a higher probability of response to combined pharmacotherapy and psychotherapy (Quilty et al., 2008). A review by Mulder evaluated antidepressant response in a variety of treatments, including psychotherapy, pharmacotherapy, electroconvulsive therapy (ECT), or a combination thereof, and reported that higher neuroticism generally predicts worse treatment outcome, especially in the long-term (Mulder, 2002). However, not many studies focused on rTMS as a treatment. Berlim and colleagues demonstrated that neuroticism was found to decrease during rTMS treatment, yet lacked predictive value, whereas baseline extraversion levels predicted greater treatment response (Berlim et al., 2013). In a study that focused on deep TMS (dTMS), higher agreeableness and higher conscientiousness were observed in patients who achieved remission (McGirr et al., 2014).

Another psychological dimension that gained interest as a predictor of treatment outcome in MDD is anhedonia. Anhedonia is a core MDD symptom. Recently, the importance of considering the role of anhedonia in MDD has been highlighted (Treadway and Deroche-Gamonet, 2008). It has been argued that anhedonia is a difficult symptom to treat (Treadway and Zald, 2011) and multiple studies have shown that higher levels of anhedonia are predictive of poorer treatment outcome (Spijker et al., 2001; McMakin et al., 2012). Likewise, improvements in anhedonia levels predicted increased psychosocial functioning in patients with MDD, which is in turn an important feature of treatment response and remission (Vinckier et al., 2017). In an rTMS study, Downar and colleagues found that their groups of MDD non-responders to rTMS treatment were marked by more anhedonic symptoms (Downar et al., 2014). Likewise, a recent study by Rostami and colleagues found that loss of interest (an anhedonia-related symptom) predicted rTMS treatment response (Rostami et al., 2017).

Next to such individual symptoms, overall depression severity is also considered an important treatment response predictor, with higher pretreatment depression severity being associated with lower response rates (Croughan et al., 1988; Trivedi et al., 2006), however, the interaction between depression severity and treatment response is unclear. For example, in a placebo-controlled study Fournier and colleagues found that at mild to moderate levels of MDD, beneficial effects of antidepressants were minimal to none, however, at very severe levels of depression severity, there was a substantial benefit of the usage of medications (Fournier et al., 2010). A similar trend of response was observed in those with high depression severity when assigned to the treatment or placebo group. That is, those with high depression severity that were assigned to the treatment group were more likely to have a greater response (i.e. the higher the severity, the greater the response), whereas those with high severity that were assigned to the placebo group were more likely to have a smaller response (i.e. the higher the severity, the smaller the response) (Khan et al., 2002). For rTMS, it has been reported that younger patients with a lower baseline depression severity had a modestly better treatment outcome (Carpenter et al., 2012). Likewise, Fitzgerald and colleagues found that rTMS responders had lower baseline depression severity, however, it did not sufficiently influence response rates to base treatment decisions on (Fitzgerald et al., 2016).

Finally, anxiety and stress are also associated with MDD and treatment response. Higher levels of anxiety have been associated with greater depression severity (Fava et al., 2004; Uher et al., 2011) and lower response to pharmacological treatment (Fava et al., 2008). It has even been proposed that depression with increased levels of anxiety could be a distinguished subtype of MDD (Fava et al., 2004). However, the association between higher anxiety levels and decreased response to treatment is modest (Joffe et al., 1993) and inconsistently replicated, including reports of patients with anxious depression responding better to ketamine treatment (Ioanescu et al., 2014), or reports in which anxious depression is not found to be predictive of worse pharmacological treatment response (Uher et al., 2011). In rTMS studies, it has been found that rTMS non-responders had higher baseline anxiety than responders (Brakemeier et al., 2007). For stress, it has been suggested that chronic stress is predictive of depression, even more so than acute stressors (McConagle and Kessler, 1990; Hammern, 2005). Some studies support this, see for example Deng et al. (2018), in which greater perceived stress in remitted older (age ≥ 60) people predicted recurrence of depression.

Given the above richness of available psychological factors that might be predictive of antidepressant treatment response, yet the lack of such research for rTMS, makes the purpose of this study to test various psychological factors that can predict rTMS treatment non-response in MDD. Finding such predictors in psychological measures is the most cost-effective way to optimize treatment allocation. To explore all possible psychological factors, this study used the total Beck Depression Inventory (BDI)-II-NL score, as well as subscales thereof, the Neuroticism, Extraversion, Openness Five-Factor Inventory (NEO-FFI), and the Depression, Anxiety, and Stress Scale (DASS) to investigate potential predictors of rTMS treatment outcome. The BDI subscales were literature-based. Next to this, the clinical relevance of these predictors was explored. That is, next to sensitivity and specificity, the positive predictive value (PPV) and negative predictive value (NPV) were examined. This method has been applied elsewhere [e.g. Kuk et al. (2010); Li et al. (2012)] and attempts to identify constructs that can accurately and reliably inform the therapist on treatment response and therefore treatment (dis)continuation (Li et al., 2012), while incorporating the false positives and false negatives. Additionally, the predictors were tested for their robustness by immediate replication in an independent sample. More specifically, given the recently highlighted interest for the replication of studies (Open Science Collaboration, 2012; Pashler and Wagenmakers, 2012; Patil et al., 2016), and our own recent non-replication of previous work (Krepel et al., 2018), it was decided to a priori split the complete dataset into a Discovery and Replication set. Hereby, it was possible to confirm or deny any findings in the Discovery set by executing the same analyses in the Replication set.

**Methods and materials**

**Design**

This study was a multi-site, open-label study. Data were collected at three sites (Brainclinics Treatment/neuroCare Nijmegen and The...
The NEO-FFI was used to examine personality analyses were performed in the Discovery set. Only when a significant difference between the two groups, the random division was frozen and the two groups were designated as Discovery set and Replication set.

Differences in age and BDI at intake were tested for using one-variables were assessed using the following instruments:

- The BDI-II-NL was used to assess depression severity. Of the BDI, several subscales were taken. These included:
  - the Anhedonia scale (items 4, 12, and 21) and the Non-Anhedonia scale (items 1–3, 5–11, and 13–20) (Leventhal et al., 2006)
  - the Cognitive-Affective scale (items 1–13) and the Somatic and Performance scale (items 14–21) (Trentini et al., 2005)
  - the Cognitive scale (items 2, 3, 5–9, and 14) and the Non-Cognitive scale (items 1, 4, 10–13, and 15–21) (Kumar et al., 2002)

These subscales were computed by adding the indicated items into one variable (e.g. the anhedonia scale was computed by adding item 4, item 12, and item 21). Scores for these scales were also calculated at outtake, as well as in change over time (in absolute numbers and percentages).

- The NEO-FFI was used to examine ‘the big five’ personality traits. The NEO-FFI is a 60-item, self-report instrument that measures five personality traits, being Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. These domains have shown good internal consistency (Cronbach’s $\alpha$ range 0.87–0.92) (Costa and McCrae, 1992).
- To measure anxiety and stress, the DASS (Henry and Crawford, 2005) was used.

**Response prediction**

Differences between responders (R) and non-responders (NR) were analyzed based on clinical baseline variables. The clinical variables were assessed using the following instruments:

- The specific treatment procedures and clinical outcomes have recently been published elsewhere (Donse et al., 2018). In short, baseline clinical variable measurements consisted of the BDI-II-NL, Depression, Anxiety and Stress Scale (DASS), and NEO-FFI. To track the course of the rTMS treatment, the BDI was assessed every fifth rTMS session.

Next, the sensitivity, specificity, PPV, and NPV for predicting non-response of all the above described variables were examined, using a discriminant analysis. The reason for this extra dimension of statistical evaluation was that a psychological measure may not be significantly different between R and NR, as tested by a GLM univariate analysis, yet it may prove to be highly predictive of non-response in a smaller specific subset of NR. Typically, PPV and NPV are set relatively high, i.e. $\geq 0.75$ or $75\%$ (Li et al., 2012). Thus, if it were the case that NPV or PPV exceeded 75% and replicated, this psychological measure was still examined in the following analyses.

Replicated psychological measures (or those showing a PPV or NPV higher than 75%) were also tested for dimensional association by partially correlating these measures with BDI percentage change (BDI% change) from intake to outtake, while controlling for age.

Next, a discriminant analysis on non-response was performed, using the predictors as indicated by GLM univariates of PPV and NPV. More specifically, the replicated baseline psychological measures that showed to be significantly different between R and NR, or that had a replicated NPV or PPV of 75% or higher, were used as independent variables in the discriminant analysis. In addition to these items, age was also used in this analysis. From these variables, a receiver operator curve (ROC) was derived to establish the predictive value of the discriminant analysis on treatment response.

It was also attempted to establish a useful construct based on the replicated predictors by examining the severity levels of these predictors. That is, a cut-off score of these predictors was established using sensitivity, specificity, and Youden’s $J$. This cut-off score provided a useful construct on which a therapist can act.

A further confirmatory approach that also compares a Discovery–Replication approach to a more traditional approach applied to the full sample (albeit using strict Bonferroni corrected $p$-values) is reported in the online Supplementary material. In previous treatment prediction studies in MDD we have reported many sex-specific predictors (van Dinteren et al., 2015; Arns et al., 2016; Iseger et al., 2017), yet statistically testing such interactions requires relatively large sample sizes (Leon and Heo, 2009). Therefore, in this full sample a post-hoc analysis on the complete dataset was conducted and statistically examined the
A total of 196 MDD patients were enrolled in this study (average age: 43.62, range 18–78 years; 99 females and 97 males). Dividing the complete dataset (n = 196) in a 60–40 distribution (60% Discovery, 40% Replication) resulted in a Discovery set of 119 MDD patients (average age: 43.60, range 18–73; 62 females and 57 males) and a Replication set of 77 MDD patients (average age: 43.64, range 19–78 years; 37 females and 40 males). The clinical outcome measures of the Discovery and Replication sample are summarized in Table 1. One subject was excluded based on missing data. Note that, given that most the MDD patients were from Nijmegen, the following analyses were also performed in the Nijmegen-only cohort. These analyses for data collected in Nijmegen only yielded similar statistical outcomes and did not result in different conclusions.

### Discovery set

Primary-analyses only yielded an effect of response for the Anhedonia scale (p = 0.072; F = 3.298; df = 1). For the secondary analyses only Openness differed between R and NR (p = 0.037; F = 8.516; df = 1). No significant effect was found for the Anhedonia scale (p = 0.072; F = 3.298; df = 1). For the secondary analyses only Openness differed between R and NR (p = 0.037; F = 8.516; df = 1). All other variables were not significantly different between R and NR (see Table 2). In Table 2, all the effect sizes (Cohen’s d) for each baseline psychological measure, subdivided for R and NR, are reported as well.

### Replication set

A GLM univariate test was performed on the Replication set, with the exact same parameters as in the Discovery set. When focusing on only the predicted psychological measures that were significantly related to response in the Discovery set (i.e. the Anhedonia scale and the NEO-FFI Openness scale), a significant effect was found for the Anhedonia scale (p = 0.014; F = 5.016; df = 1). No significant effect was found for Openness (p = 0.227; F = 1.490; df = 1). The results are shown in Table 2. Since all other variables showed to be non-significant in the Discovery set, significant differences between R and NR that were confined to only the Replication set were ignored. However, for purposes of completeness, all other variables and their statistical values are shown as well.

### Sensitivity, specificity, NPV, and PPV

As described, baseline psychological measures may not be significantly different between R and NR, yet be of high predictive value for a (specific) (sub)group. Therefore, each baseline psychological measure was individually assessed for these features, using one psychological measure and age in a discriminant analysis. The preset value of both PPV and NPV was 75%. As can be observed...
from Table 3, none of the baseline variables reached or exceeded and replicated the preset value of PPV and NPV.

### Correlations between psychological scales and BDI percentage change

A two-tailed correlation while controlling for age showed a significant correlation between BDI% change and the Anhedonia scale at intake ($p = 0.023$; $r = -0.221$; $r^2 = 4.9%$). Additionally, BDI% change and Openness at baseline were shown to be significantly correlated ($p = 0.025$; $r = 0.214$; $r^2 = 4.6%$).

In the Replication set, there was a significant one-tailed partial correlation between BDI% change and the Anhedonia scale ($p = 0.025$; $r = -0.244$; $r^2 = 6.0%$), and between BDI% change and Openness ($p = 0.047$; $r = 0.206$; $r^2 = 4.2%$).

### Discriminant analysis

Using the Discovery set, a discriminant analysis was performed using the Anhedonia scale and age. The resulting model was shown to be significant ($p = 0.034$; Wilks’ $\lambda = 0.937$; $\chi^2 = 6.771$; $df = 2$), with a sensitivity of 62.5% and a specificity of 53.7%. The PPV and NPV were 44.6% and 70.6%, respectively. The area under the ROC curve, regressed on non-response, was 0.643 (Fig. 1). This analysis was repeated in the Replication set. The model also reached significance ($p = 0.018$; Wilks’ $\lambda = 0.880$; $\chi^2 = 8.024$; $df = 2$), with a sensitivity of 73.9%, and a specificity of 74.4%. The PPV and NPV were 60.7% and 80.0%, respectively. The area under the ROC curve was 0.726 (Fig. 2).

### Clinical relevance (sensitivity, specificity, NPV, PPV, Youden’s J of the Anhedonia scale)

For the current calculations, the Discovery and Replication set were merged, since in this section the clinical relevance of this metric was investigated, rather than its methodological value. To specify the clinical relevance of the Anhedonia scale, a cut-off score was established by calculating Youden’s index.

A complete overview of statistics can be found in the online supplementary material. In short, using an ROC the cut-off, from which predicting NR based on the Anhedonia scale) was highest at severity level 6 ($J = 0.235$). Also, the PPV and NPV do not reach the preset value of 75%, and therefore the model is, based on this method, considered to not be clinically relevant.
Our study suggests that none of the psychological measures are clinically meaningful predictors of rTMS treatment response in an MDD sample. This is in line with studies that found evidence for predictive utilities of psychological measures, but did not found it to be highly influential on response rates (Fitzgerald et al., 2016).

However, this study suggests that NR robustly show higher anhedonia scores at baseline, and that this score was – to some degree – related to clinical improvement. More specifically, a higher score on the Anhedonia scale (as taken from the BDI) marginally predicted non-response to rTMS. However, the ROC curves were only mildly predictive of non-response, and NPV was 70.6–80.0% and PPV was 44.6–60.7%, thereby the NPV

Table 3. This table depicts the sensitivity, specificity, PPV, and NPV per individual baseline psychological measure

<table>
<thead>
<tr>
<th>(Sub)scale</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>PPV (%)</th>
<th>NPV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discovery</td>
<td>Replication</td>
<td>Discovery</td>
<td>Replication</td>
</tr>
<tr>
<td>BDI total</td>
<td>63.4</td>
<td>54.2</td>
<td>57.1</td>
<td>64.2</td>
</tr>
<tr>
<td>Anhedonia</td>
<td>62.5</td>
<td>73.9</td>
<td>53.7</td>
<td>74.4</td>
</tr>
<tr>
<td>Non-Anhedonia</td>
<td>56.4</td>
<td>56.5</td>
<td>54.5</td>
<td>64.3</td>
</tr>
<tr>
<td>Cognitive-Affective</td>
<td>57.5</td>
<td>47.8</td>
<td>53.0</td>
<td>59.5</td>
</tr>
<tr>
<td>Somatic and Performance</td>
<td>59.0</td>
<td>56.5</td>
<td>59.7</td>
<td>62.8</td>
</tr>
<tr>
<td>Cognitive</td>
<td>56.4</td>
<td>47.8</td>
<td>54.5</td>
<td>50.0</td>
</tr>
<tr>
<td>Non-Cognitive</td>
<td>62.5</td>
<td>47.8</td>
<td>58.2</td>
<td>72.1</td>
</tr>
<tr>
<td>NEO-FFI Neuroticism</td>
<td>56.4</td>
<td>59.1</td>
<td>56.9</td>
<td>45.7</td>
</tr>
<tr>
<td>NEO-FFI Extraversion</td>
<td>56.4</td>
<td>63.6</td>
<td>56.9</td>
<td>58.7</td>
</tr>
<tr>
<td>NEO-FFI Openness</td>
<td>56.4</td>
<td>59.1</td>
<td>58.3</td>
<td>47.8</td>
</tr>
<tr>
<td>NEO-FFI Agreeableness</td>
<td>59.0</td>
<td>59.1</td>
<td>55.6</td>
<td>47.8</td>
</tr>
<tr>
<td>NEO-FFI Conscientiousness</td>
<td>51.3</td>
<td>59.1</td>
<td>55.6</td>
<td>63.0</td>
</tr>
<tr>
<td>DASS Depression</td>
<td>56.1</td>
<td>58.3</td>
<td>53.4</td>
<td>55.1</td>
</tr>
<tr>
<td>DASS Anxiety</td>
<td>56.1</td>
<td>54.2</td>
<td>54.8</td>
<td>62.5</td>
</tr>
<tr>
<td>DASS Stress</td>
<td>48.8</td>
<td>54.2</td>
<td>57.5</td>
<td>59.2</td>
</tr>
</tbody>
</table>

A discriminant analysis was performed using one psychological measure plus age as independent variables. Based on the absolute number of true positives, true negative, false positives, and false negatives (as predicted by the discriminant analysis), the above metrics were calculated. If PPV or NPV exceeded 75% and replicated, this psychological measure would still be examined in the following analyses, even though it showed not to be significantly different between R and NR in the previous univariate analyses. None of the baseline psychological measures showed a PPV or NPV of 75% or higher in both the Discovery and Replication sample.

Fig. 1. ROC of the discriminant analysis on non-response using the Anhedonia scale and age as independent variables in the Discovery sample, with an area under the curve of 0.643. The ROC shows the sensitivity (62.5%) and specificity (53.7%) for non-responders (dotted line) and responders (striped line). NPV and PPV were 44.6% and 70.6%, respectively.

Fig. 2. ROC of the discriminant analysis on non-response using the Anhedonia scale and age as independent variables in the Replication sample, with an area under the curve of 0.726. The ROC shows the sensitivity (73.9%) and specificity (74.4%) for non-responders (dotted line) and responders (striped line). The PPV and NPV were 60.7% and 80.0%, respectively.
just being short of the a priori preset threshold of 75%. Therefore, even though NR seem to have a higher anhedonia score at baseline, its predictive value remains relatively low. This is partly in line with Rostami and colleagues, who found that the symptom ‘loss of interest’ was strongly predictive in rTMS response (Rostami et al., 2017). Even though in our sample it is also found that baseline anhedonia is different between R and NR, the subset of symptoms could not predict treatment response such that is clinically relevant.

When attempting to find the optimal cut-off score on which a therapist can act, it was found that those with a baseline anhedonia score of 6 or higher are less likely to respond to rTMS treatment. However, given the overall clinical non-relevance of the model based on anhedonia, this cut-off should be taken with caution.

**Replication**

In this study we employed two approaches that are often advocated as tools to overcome the ‘Replication crisis’. Some people have advocated to rely more strongly on replication studies (Open Science Collaboration, 2012; Pashler and Wagenmakers, 2012; Roediger, 2012; Simons, 2014) [although caution should be taken when designing and interpreting a replication study, see Maxwell et al. (2015)], as well as self-replication (i.e. replicate your own work before publishing) (Roediger, 2012; Simons, 2014), which is what we performed in this main manuscript. Others advocate to use stricter p-values, for example using a p-value of 0.005 (Benjamin et al., 2018). In the online supplementary material this latter approach was implemented, using the full sample and a strict Bonferroni corrected p-value of 0.0033. Interestingly, both approaches converged on identifying anhedonia as a predictor for response. A difference was that in the latter approach, as elaborated on in the online supplementary material, another psychological scale of the BDI (the Non-Cognitive scale) was identified as being statistically different between R and NR. However, it seems that this difference is mainly driven by BDI item overlap with the Anhedonia scale (since the Anhedonia scale is also part of the Non-Cognitive scale) and thus a spurious correlation. Further elaboration on these analyses can be found in the online supplementary material.

Interestingly, Table 2 can also be inspected as an example of how many false positives one could obtain using a p < 0.05 approach without requiring replication (i.e. if these would be published as two separate papers). When looking at any p < 0.05 values in either the Discovery or Replication dataset 6 out of 17 measures (35%) are significantly different, whereas only one of those metrics (6%) actually replicates. When using stricter thresholding and a larger sample size (p < 0.0033; online supplementary material) only one false positive was found (the Non-Cognitive scale) and the same result for anhedonia was confirmed.

Therefore, these results further highlight that false positive findings can be easily obtained. Both approaches (replication vs. stricter p-thresholding) increased the robustness of results. Based on these results we cannot draw a definitive conclusion, but for future manuscripts where we will be looking at other predictors [e.g. electroencephalography (EEG)], we will use this exact same approach and hope to obtain robust and clinically relevant predictors of treatment response.

**Limitations**

This study did not employ a double blind placebo controlled design, hence we are unable to rule out that the obtained results can be (partly) explained by such effects. Similarly, this study was open-labeled, which poses another weakness, albeit results in ecologically valid results. Furthermore, as is described by Donse et al. (2018), this sample consisted of MDD patients who received simultaneous rTMS and psychotherapy, rather than rTMS only. Lastly, while the complete sample consists of 196 MDD patients, reducing it to a Discovery and Replication set and dividing it into a R/NR classification, narrows down the sample size, resulting in a smallest sample size (Replication set, NR) of 24.

**Conclusion**

Our study indicates that the psychological measure anhedonia, as measured by the Anhedonia scale of the BDI at baseline, is related to clinical improvement on MDD symptoms in response to rTMS treatment in an MDD sample. More specifically, lower baseline anhedonia scores were related to better clinical improvement. However, our study also suggests that anhedonia is only mildly predictive of treatment response and does not achieve predefined levels of clinical relevance. Therefore, it can be concluded that lower anhedonia scores are favorable in the treatment of MDD, however, these are unlikely to be of clinical usage and relevance due to the low PPV and NPV.

Furthermore, the current study also highlights that false positives are relatively easy obtained, when handling a 0.05 significance level, advocating for the Replication approach and the usage of stricter p-values, as was done in this main manuscript and online supplementary material.

**Supplementary material.** The supplementary material for this article can be found at https://doi.org/10.1017/S0033291718004191.

**Acknowledgements.** We would like to thank Lana Donse, Vera Kruiver, Rosalinde van Ruth, Marleen Stam, Maaike Moolenaar, Sanne Bongers, Myrthe van Eerdt, Dagmar Timmers, Inge Janssen-Bouwmeester and Nicole van Merode for support and collecting the data used in this study.

**Author contributions.** NK managed the literature search, performed the analyses, and wrote the first draft of the manuscript. All other authors contributed, reviewed, and approved the final manuscript.

**Conflict of interest.** MA reports options from Brain Resource (Sydney, Australia); he is director and owner of Research Institute Brainclincs, a minority shareholder in neuroCare Group (Munich, Germany), and a co-inventor on four patent applications (A61B5/0402; US2007/0299323; A1; WO2010/139361 A1) related to EEG, neuromodulation and psychophysiology, but does not own these nor receives any proceeds related to these patents; Research Institute Brainclincs received funding from Brain Resource (Sydney, Australia) and neuroCare Group (Munich, Germany), and equipment support from Deymed, neuroConn and Magventure, however data analyses and writing of this manuscript were unconstrained. A JR reports consulting fees from the American Psychiatric Association, Brain Resource Ltd., Compass Inc., Curbstone Consultant LLC, Eli Lilly, Emmes Corp., Liva-Nova, Lundbeck A/S, National Institute of Drug Abuse, Taj Medical, Santium Inc., Sunovion, Taj Medical, Takeda USA; speaking fees from Live Nova; royalties from Guilford Publications and the University of Texas Southwestern Medical Center. The other authors report no disclosures or conflicts of interest.

**References**

with different antidepressant medications in the randomized iSPOT-D study. Clinical Neurophysiology 127, 509–519.


Leon AC and Heo M (2009) Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. Computational Statistics & Data Analysis 53, 603–608.


