Two steps forward, one step back

TWO STEPS FORWARD, ONE STEP BACK

REPLICABILITY AND CLINICAL RELEVANCE OF POTENTIAL BIOMARKERS IN DEPRESSION AND ADHD

NORALIE KREPEL

Two Steps Forward, One Step Back

Replicability and Clinical Relevance of
Potential Biomarkers in
Depression and ADHD

Noralie Carlijn Krepel

The work described in this doctoral dissertation was carried out at Stichting Brainclinics Foundation, in collaboration with Maastricht University.
The support from Total Brain and neuroCare Group is acknowledged.

Krepel, Noralie Carlijn
Two Steps Forward, One Step Back
Replicability and Clinical Relevance of Potential Biomarkers in Depression and ADHD

Published by Brainclinics Insights,
Nijmegen, The Netherlands

An electronic version of this dissertation is available at:
brainclinics.com

TWO STEPS FORWARD, ONE STEP BACK

REPLICABILITY AND CLINICAL RELEVANCE OF POTENTIAL BIOMARKERS IN DEPRESSION AND ADHD

TWEE STAPPEN VOORUIT, ÉÉN STAP TERUG

REPLICEERBAARHEID EN KLINISCHE RELEVANTIE VAN POTENTIËLE BIOMARKERS IN DEPRESSIE EN ADHD

DISSERTATION

to obtain
the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus,
Prof. Dr. Rianne M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public on
Friday, the 18th of June 2021, at 14:00 hrs.

by

Noralie Carlijn Krepel
born on the 4th of November 1993
in Dordrecht, The Netherlands

Supervisor
Prof. dr. A.T. Sack

Co-supervisors
Dr. M. Arns,
Maastricht University/University of Amsterdam/
Brainclinics Foundation, Nijmegen
Dr. H. van Dijk,
University of Amsterdam/
Brainclinics Foundation, Nijmegen

Assessment Committee
Prof. dr. B.M. Jansma (Chair)
Prof. dr. F.P.M.L. Peeters
Prof. dr. L.L. Carpenter,
Warren Alpert Medical School at
Brown University, USA
Prof. dr. K.R.J. Schruers
Prof. dr. J.K. Buitelaar,
Radboud University Nijmegen

Cover and layout: Brainclinics Insights, Nijmegen
Printed by GVO, Ede, The Netherlands
ISBN-978-90-830013-2-6

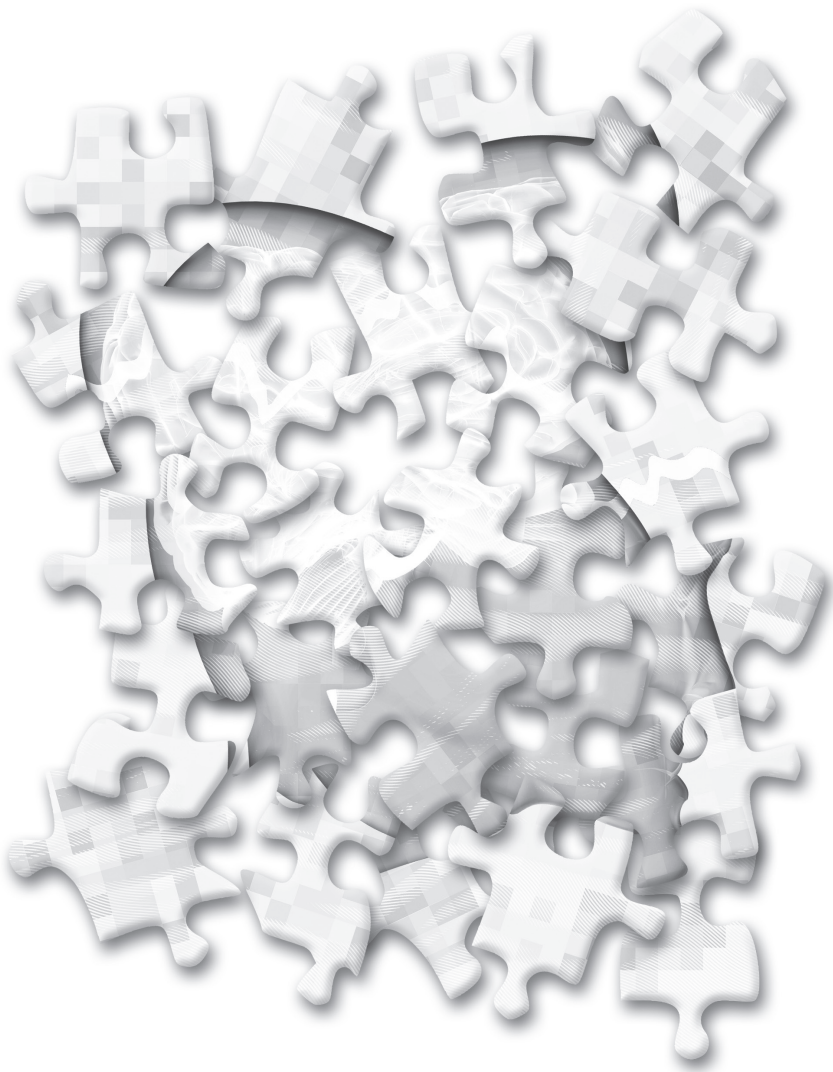
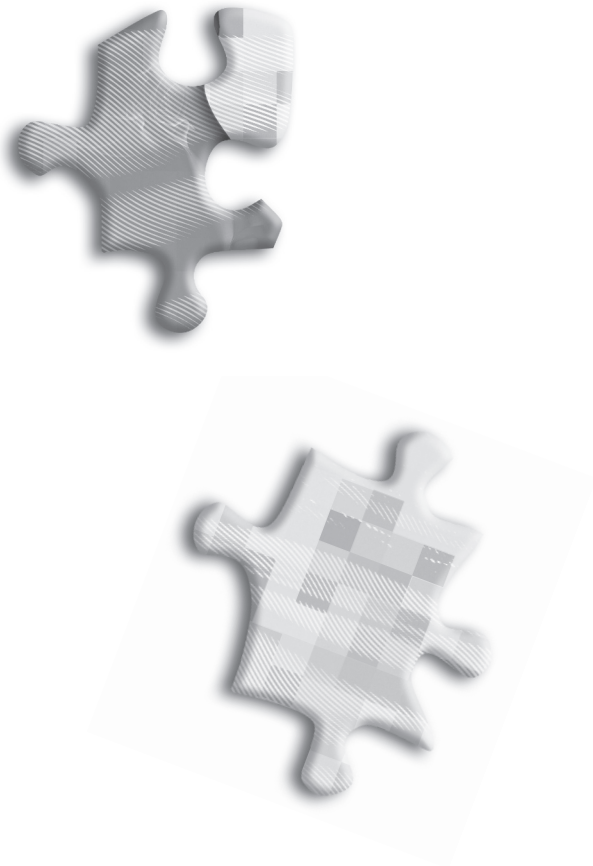


TABLE OF CONTENTS



1

INTRODUCTION

Parts of this chapter are adapted from the following book chapter: Arns, M., Gunkelman, J., Olbrich, S., Krepel, N., Vollebregt, M., Sander, C., Hegerl, U. (under review). Assessment of quantitative EEG (QEEG) and brain arousal: Relevance for diagnosis, prognosis and treatment in ADHD and Depression. Editors: Coben, R. & Evans, J.: In *Neurofeedback and Neuromodulation Techniques and Applications.*

THE ROAD TO PERSONALIZED MEDICINE

THE STARTING POINT: ONE-SIZE-FITS-ALL

The current approach to pharmacological treatment in psychiatry is illustrated by a quote from Lombard and Doraiswamy (2013): *"Currently, the selection of medications for a given patient in psychiatry is primarily based on a trial-and-error process with some consideration for trying to use a drug's known side effects to help a symptom (e.g., use of a sedating antidepressant for a depressed patient with insomnia)."* (Lombard & Doraiswamy, 2013, p. 1). This is reflective of a matched care approach, in which therapy choice is based on patient's characteristics and preferences. An alternative approach is the 'stepped care' approach, in which patients start with a low-intensity evidence-based treatment and, in cases of insufficient clinical improvement, proceed to other, more intensive treatments (Bower & Gilbody, 2005; van Straten, Hill, Richards, & Cuijpers, 2015). Both matched and stepped care reflect one-size-fits-all approaches, in which treatment assignment is relatively arbitrary. One-size-fits-all approaches are, by themselves, not problematic, as long as symptom alleviation is sufficient. However, the treatment efficacy of psychiatric disorders using one-size-fits-all approaches seems to be falling short

in a relatively large group of patients. In depression, response rates to antidepressant treatment have been reported at 60.5% for escitalopram, 66.3% for sertraline, 59.7% for venlafaxine-XR (Saveanu et al., 2015), 53% for Cognitive Behavioral Therapy (CBT), 46% for Interpersonal Psychotherapy (IPT), and 44% for other psychotherapies (Cuijpers et al., 2014). Remission (i.e., minimal symptoms) rates have been reported at 48.1% for escitalopram, 46.3% for sertraline, 41.6% for venlafaxine-XR (Saveanu et al., 2015), 49% for CBT, 38% for IPT, and 36% for other psychotherapies (Cuijpers et al., 2014). Newer treatment options such as repetitive Transcranial Magnetic Stimulation (rTMS) also have limited response rates, as reflected by a recent non-inferiority trial comparing two types of rTMS, reporting response rates of 47% and 49% and remission rates of 27% and 32% (Blumberger et al., 2018). Importantly, attempting new types of treatment does not guarantee treatment success, as the multi-center Sequenced Treatment Alternatives to Relieve Depression (STAR*D) showed that after having attempted up to four different, carefully considered, sequenced treatment steps, 67% of depressed individuals went into remission, yet 33% of patients were still experiencing symptoms.

For attention-deficit/hyperactivity disorder (ADHD) a slightly different picture emerges. It has been shown that methylphenidate (MPH) effectively reduces symptoms of ADHD in adults (Faraone, Spencer, Aleardi, Pagano, & Biederman, 2004; Spencer et al., 2005), children and adolescents (Cortese et al., 2018), and preschoolers (Greenhill et al., 2006). The Multimodal Treatment Study of Children With ADHD (MTA) showed that treatment with stimulant medication or treatment combining stimulant medication with behavioral treatment was more effective in reducing symptoms of ADHD than community care or behavioral treatment alone, although the dosage used in the combined treatment was significantly lower than in treatment solely consisting of stimulant medication (The MTA Cooperative Group, 1999). Maia et al. (2017) also found long-term efficacy of MPH usage in children. Yet, even though medication treatment seems efficacious on the group level, response rates show that treatment is not as efficacious for every individual. Steele, Jensen, and Quinn (2006) evaluated multiple randomized controlled trials and found a response rate varying between 65 – 75% for various pharmacological treatments.

Additionally, it should be taken into consideration that response in ADHD is generally defined as a 25 – 30% (Steele et al., 2006) or 40% (Newcorn et al., 2008; Prasad et al., 2007) symptom reduction. When the clinical outcome is defined as remission, rates are estimated at 40 – 60% in randomized clinical trials (Steele et al., 2006). Likewise, rates of excellent response (similar to remission) in the MTA trial have been reported at 68% for combined treatment, 56% for stimulant medication, 34% for behavioral treatment, and 25% for community care (Swanson et al., 2001). Also, for a subgroup of children an increase in MPH dosages is required after a longer period of treatment to maintain optimal control of ADHD symptoms (Vitiello et al., 2001; Wilens et al., 2005).

A potential reason for the limited efficacy of psychiatric treatment is that treatment may not be perfectly targeting the symptoms to be alleviated. Wong, Yocca, Smith, and Lee (2010) describe that pathophysiological heterogeneity underlying patterns of pathology are unified into one disorder, which may compromise the effectiveness of psychiatric medication. This may also complicate scientific research (Wong et al., 2010). Cuthbert and Insel (2013) also describe how progress in mental health research has stagnated – reflected by a lack of clinical tests for diagnosis, delayed detection of disorders, and a lack of preventive interventions. One potential reason for this may be that the diagnostic categories from the Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Statistical Classification of Diseases and Related Health Problems (ICD) may not map well onto emerging findings from genetics, systems neuroscience, and behavioral science (Cuthbert & Insel, 2013). Relating to this is the 'assumption of diagnostic discrimination' introduced by Tabb (2015). Tabb (2015) explains this as an optimistic view in which current diagnostic tests (here referring to the diagnostic criteria of the DSM, or diagnostic screening instruments derived from these) can discriminate between patient groups that allow for relevant facts to be discovered for that given disorder. The term relevant facts, here, specifically means *"... those about the underlying mechanisms causing the signs and symptoms with which patients present that count as a significant discovery within the experimental context. They are the sorts of validators that biomedical scientists hope to find: genetic signatures, neurological or cognitive dysfunctions, focal brain*

lesions, and so forth." (Tabb, 2015, p. 1049). The DSM may not be the perfect instrument for the discovery of these relevant facts because, historically, the DSM has not been designed to discriminate between disorders on a biomedical level. Likewise, diagnoses derived from the DSM are relatively heterogeneous (Tabb, 2015). Specifically, since the third edition of the DSM, a patient needs to meet a certain number of symptoms out of a list of symptoms for a given disorder. The result is that patients who are diagnosed similarly may present different symptom profiles. Hence, heterogeneity within a given diagnosis is large. For example, Park et al. (2017) identified 119 different symptom combinations that all may lead to the diagnosis of depression. Hyman (2010) has also identified several issues with the DSM, including the reliance on categorical diagnoses, comorbidity between disorders, and a limited fit of family and genetic data to DSM-IV disorder boundaries. Also, a relatively narrow, specific set of criteria and wording on which diagnoses are based may result in a low concordance between the ICD and DSM and a relatively large proportion of diagnoses based on Not Otherwise Specified (NOS) (Hyman, 2010).

As such, in the current system, in which treatment assigned is relatively arbitrary, treatment success is limited. One potential explanation may be that these one-size-fits-all approaches target symptoms associated with one particular disorder, yet heterogeneity within a disorder is large. Thus, patients receive similar treatments while the disorder presentation may be different. Treatment systems may be improved by individualizing treatment allocation, taking individual differences into account.

FROM ONE-SIZE-FITS-ALL TO PERSONALIZED MEDICINE

An alternative to the one-size-fits-all approach is personalized medicine (otherwise termed as 'precision medicine' by the National Research Council). This approach may improve psychiatric treatment by fitting treatment to the individual. One way through which this, partly, can be achieved is through the use of biomarkers. A biomarker is defined as *"A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention."* (Biomarkers

Definitions Working Group, 2001, p. 91) and can be used diagnostically (i.e., to diagnose disorders) or prognostically (i.e., to predict treatment outcome). In the current thesis, biomarkers in relation to personalized medicine are investigated to study how well they predict and optimize treatment outcome. Chapter 2 (Krepel et al., 2018), Chapter 3.1 (Bailey et al., 2021), Chapter 5 (Krepel, Rush, Iseger, Sack, & Arns, 2019), and Chapter 6 (Krepel et al., 2020) investigate whether features obtained at baseline can predict rTMS treatment response. Chapter 3 (Roelofs et al., 2021) approaches a biomarker slightly different, namely whether a relation exists between a baseline parameter (specific to the patient) and one of the treatment parameters (specific to rTMS). Chapter 4 (Krepel et al., in press) and Chapter 7 (Krepel et al., under review) again take a different approach to biomarkers, specifically looking at associations between baseline neurophysiological features and baseline behavioral constructs, without taking into account the complete symptom profile. Each of these approaches is aimed at making clinical decision-making more informed and at improving clinical effectiveness. By identifying biomarkers at baseline that are related to treatment outcome it may be possible to allocate treatment according to biomarkers, increasing the likelihood that a patient responds to treatment.

It has been suggested that a neuroimaging biomarker should be developed in multiple phases, starting with the identification of a clinically relevant question. Then, a biomarker should show internal validity by undeniably measuring the entity associated with a given disorder, followed by a demonstrable external validity expressed as having a high predictive value. Then, lastly, biomarkers should show to be clinically useful (Abi-Dargham & Horga, 2016). Biomarkers also should have a favorable cost-benefit ratio and provide clinical risk information. They also should be easy to measure, accurate, reproducible, and internationally standardized (Wium-Andersen, Vinberg, Kessing, & McIntyre, 2017). Biomarkers may follow from a variety of research approaches, one of which is the Research Domain Criteria (RDoC) approach. This was initiated by the National Institute of Mental Health (NIMH) (Insel, 2014; Insel et al., 2010; Insel & Wang, 2010) and aims to build a framework for research on pathophysiology in psychiatric disorders, with a focus on genomics and neuroscience (Insel et al., 2010).

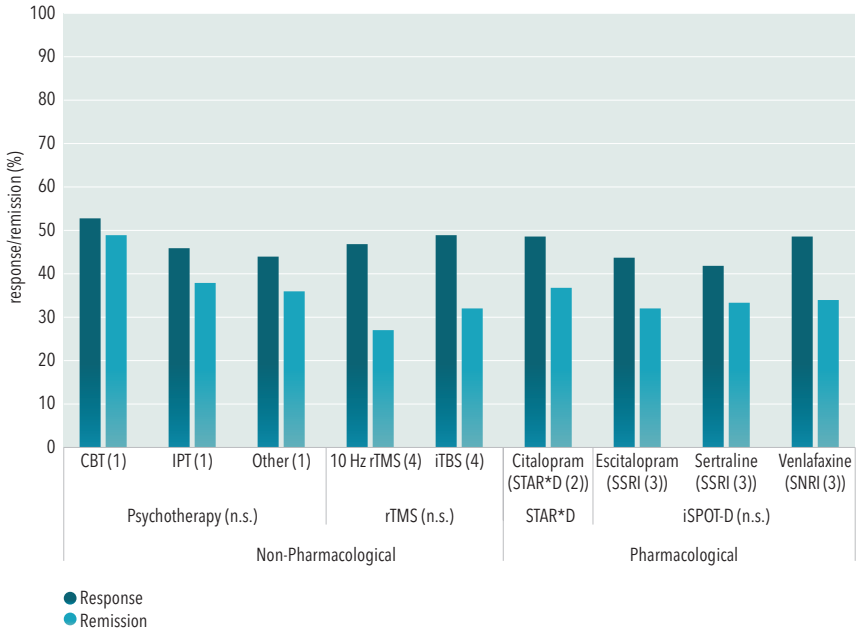
Although personalized medicine based on biomarkers may be the treatment allocation strategy of the future, it is still in its infancy at this point. For example, Thase (2014) showed that, although some biomarkers are promising, no biomarker has emerged as a definitive predictor of response or non-response to depression treatment (Thase, 2014). Likewise, although some characteristics of a useful biomarker have been proposed (Cook, 2008), no fixed set of requirements for a useful biomarker has been decided on. Some critiques towards biomarkers in psychiatry have also been expressed. For example, a lack of gold standard in diagnostics complicates the development of biomarkers, as the current diagnostic system is not designed to disentangle biologically distinct markers. Yet, a system that can identify such markers is still to be developed, resulting in a chain of circularity (Venkatasubramanian & Keshavan, 2016). Likewise, methodological issues such as underpowered studies and unconvincing (non-)replications may play a part in the difficulty of establishing reliable clinical tests (Kapur, Phillips, & Insel, 2012). Even in cases when biological findings survive the test of time and replication, the question remains how clinically useful these findings are. As pointed out by Kapur et al. (2012), findings that are based on a comparison between textbook patients and perfectly healthy controls may not show to be clinically useful, since clinical decision making rarely comes down to distinguishing a patient from a healthy individual. Rather, the difficulty lies in distinguishing between patients that show a similar symptom profile but whose treatments and outcomes differ. Another issue is that personalized medicine may imply 'to treat' or 'not to treat' and this is ethically questionable; when is a biomarker considered to be trustworthy enough to withhold an evidence-based treatment (van der Vinne, 2020)? Some critiques have also been expressed towards RDoC, the framework through which biomarker development can occur. Frances (2014) describes that, although RDoC is necessary and already widely celebrated, it will probably need a good amount of time before this approach can be implemented. Frances (2014) highlights that *"Lost in the bombast of the NIMH press release was that RDoC has absolutely nothing to offer in the present except an untested research tool."* (Frances, 2014, p. 48). Thus, although personalized medicine might show to be of great importance in the future, there are still some issues that need to be overcome.

AN IN-BETWEEN STEP: FROM STRATIFIED PSYCHIATRY TO PERSONALIZED MEDICINE

An alternative to personalized medicine, that may serve as an interim step, is stratified psychiatry. Stratified psychiatry aims to identify biomarkers that predict treatment outcome for a subgroup of people. The most important differences between precision medicine and stratified psychiatry are that stratified psychiatry does not require a complete understanding of the etiology of the illness, nor does it aim to personalize treatment for each individual. Rather, stratified psychiatry aims to identify subgroups of patients that can be assigned to different evidence-based treatment options that, based on research, may respond well to that specific (type of) treatment (Wium-Andersen et al., 2017). Also, stratified psychiatry can be applied in harmony with the current diagnostic system. After diagnoses are made, subgroups within a diagnostic category can be stratified to a specific treatment based on scientific research (Kapur et al., 2012; Wium-Andersen et al., 2017). As such, clinical decision making can be more informed than it is now while leaving the current diagnostic system intact. van der Vinne et al. (2021) describe such an approach. In this study, biomarkers were used to stratify depressed individuals to a particular type of antidepressant medication (escitalopram, sertraline, or venlafaxine) based on individual characteristics of the EEG. A group receiving treatment as usual (TAU) was used as a comparative group. The researchers found that individuals assigned to the EEG-informed treatment group had a significantly greater symptom reduction (36.8% compared to 23.9% in TAU) and the remission rate almost doubled (29% compared to 17% in TAU) (van der Vinne et al., 2021). A similar approach is taken in Chapter 6 (Krepel et al., 2020), in which the neurofeedback protocol was informed by EEG characteristics.

The ethical implications of stratified psychiatry are less problematic compared to personalized medicine, as no treatment is withheld and the efficacy of treatments is not expected to be lower than in the current treatment system. This is illustrated in Figure 1 on the following page. In this figure, the response and remission rates from various large effectiveness trials as well as meta-analyses of antidepressant treatments are presented. Figure 1 shows that the response

and remission rates for all treatments are relatively similar – and, importantly, these rates do not significantly differ within treatment modalities. In other words, the likelihood of achieving clinical improvement in TAU is similar within each treatment modality. Assigning patients to any of these treatments without using biomarkers is not expected to do any harm. As such, assigning treatment according to a biomarker that shows potential stratification value is not expected to do any harm either but may show to improve clinical effectiveness relative to TAU. Thus, stratified psychiatry may improve response and remission rates by making treatment decisions more informed.



CBT	cognitive behavior therapy
IPT	interpersonal therapy
rTMS	repetitive transcranial magnetic stimulation
iTBS	intermittent theta burst stimulation
SSRI	selective serotonin reuptake inhibitor
SNRI	selective serotonin-norepinephrine reuptake inhibitor

Figure 1: *Group-level response and remission rates derived from the largest non-industry sponsored effectiveness trials or meta-analyses for various antidepressant treatments. These results show no significant within-modality differences (n.s. = not significant). This is demonstrated by Cuijpers et al. (2014) for various psychotherapies, Blumberger et al. (2018) for two different forms of rTMS, Saveanu et al. (2015) for the effectiveness of three widely prescribed antidepressants assessed in iSPOT-D, and Rush et al. (2006) for the first step in the STAR*D trial.*

SUMMARY

Currently, treatment in psychiatry primarily rests on a one-size-fits-all principle. This would not be a problem if this approach would result in treatment success for each individual that reports psychiatric complaints. However, the overall efficacy of treatments in psychiatry is limited. Personalized medicine, in which treatment is fitted to the individual based on research, may show to be the treatment model of the future, but a lot of work needs to be done before personalized medicine can be implemented. A treatment approach that may function as an interim step towards personalized medicine is stratified psychiatry. This approach attempts to stratify patients into different treatment options, and the current treatment and diagnostic systems are less impacted relative to the personalized medicine approach. Personalized medicine as well as stratified psychiatry should be updated with recent scientific advances (Cuthbert & Insel, 2013) and may be complemented with different treatment options (e.g., electroconvulsive therapy (ECT), mindfulness-based treatment, or a combination of different treatments, and so forth), research techniques, and so on. A one-size-fits-all approach, stratified psychiatry, and personalized medicine are illustrated in Figure 2 on page 20-21.

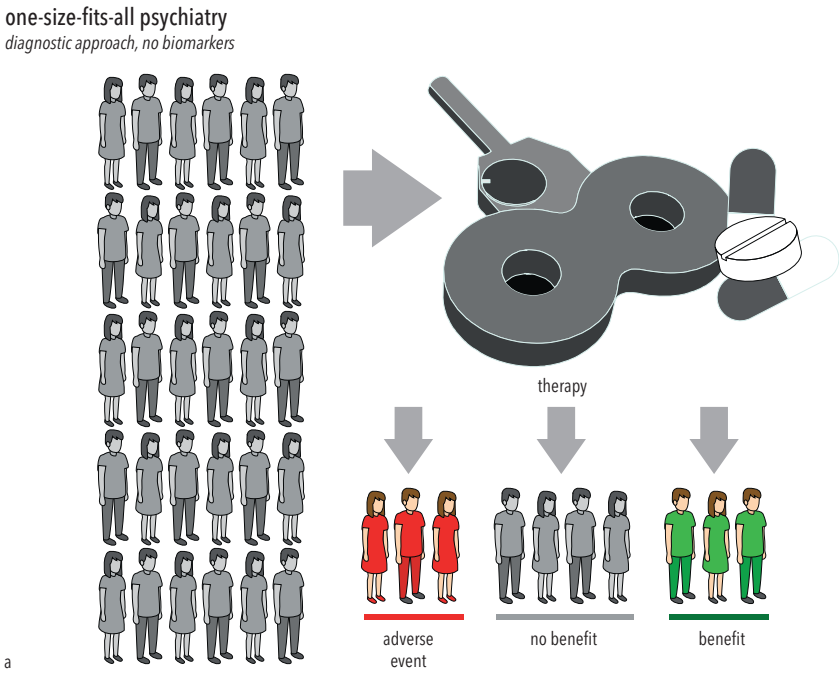
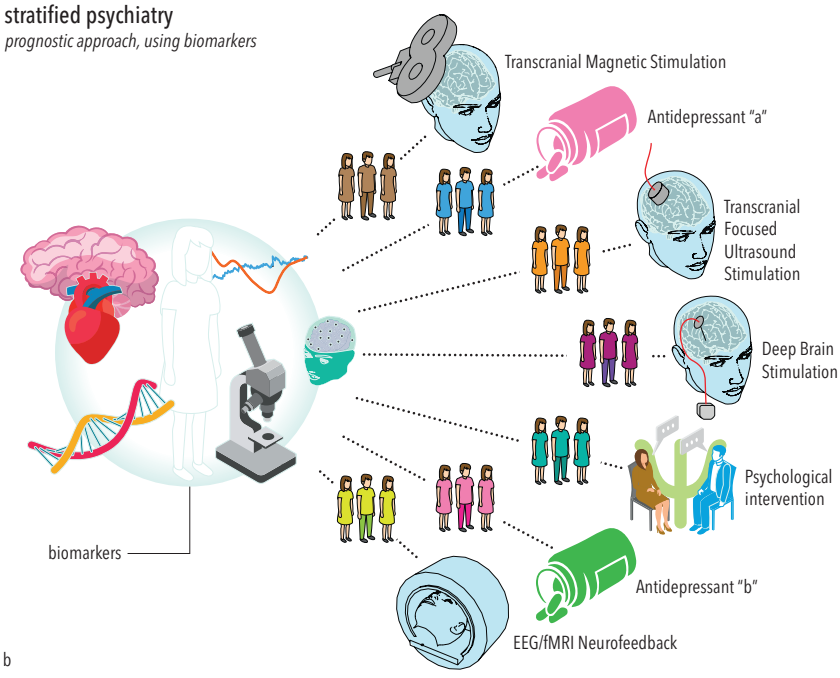


Figure 2: *A schematic overview of the different treatment approaches discussed. Currently, treatment in psychiatry rests on a one-size-fits-all approach (a). Treatments are implemented without discrimination between groups, resulting in a group that benefits from the treatment, a group that has no benefit from the treatment, and a group that experiences an adverse event from the treatment. Using biomarkers treatment can be tailored to the individual, resulting in personalized medicine (c). However, at this moment in time, personalized medicine cannot be implemented. An alternative approach, that can serve as an interim step towards personalized medicine, is stratified psychiatry (b). Using biomarkers, groups of patients can be stratified to a particular treatment.*

stratified psychiatry

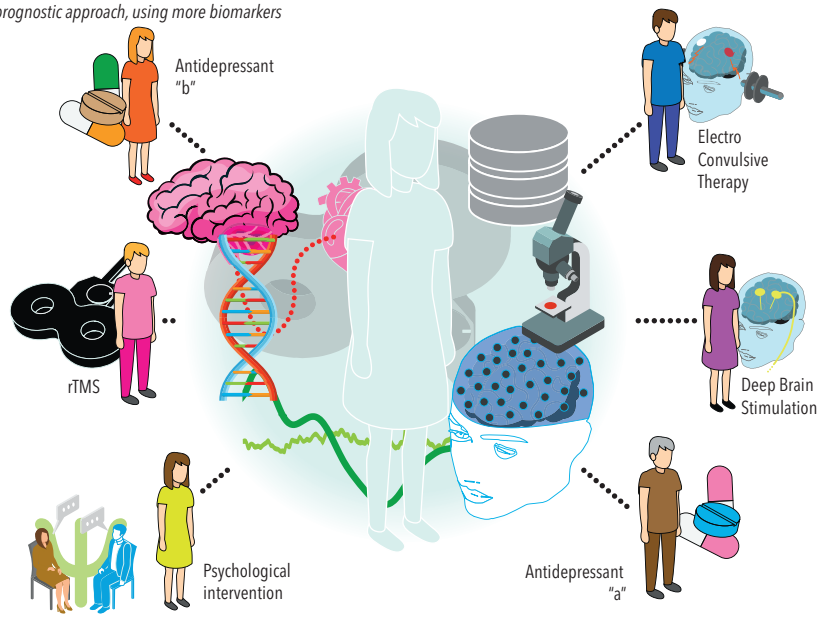
prognostic approach, using biomarkers



b

personalized medicine

prognostic approach, using more biomarkers



c

disorder independent individualized therapy or prevention

ELECTROENCEPHALOGRAPHY (EEG)

One focus of this thesis is the use of EEG in biomarker research. The EEG measures electrical activity in the brain. Communication between neurons happens through action potentials (AP), which, in turn, elicit inhibitory postsynaptic potentials (IPSP's) or excitatory postsynaptic potentials (EPSP's) in the next neuron. Near the surface of the brain, pyramidal neurons are spatially aligned and the summation of the post-synaptic potentials of many pyramidal neurons makes it possible to measure this activity at the surface of the brain using EEG electrodes (Speckmann, Elger, & Gorji, 2011).

Field potentials fluctuate, depending on the discharge pattern of the given group of neurons (Speckmann et al., 2011). Periodical fluctuations of the EEG are widely referred to as neuronal oscillations and different patterns of oscillations can be distinguished. These include delta (1 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 13 Hz), beta (13 – 35 Hz), and gamma (> 35 Hz) by convention (Miller, 2007).

The EEG is celebrated for its temporal resolution, low cost, and relatively easy implementation (Farnsworth, 2019, July 12th), which makes it a good candidate for biomarker research. However, the spatial resolution of EEG is limited.

BIOMARKERS IN DEPRESSION AND ADHD

Research in biomarkers involves many areas of research including, but not limited to, genetics, neuroscience, mRNA, blood, proteins, and the cardiovascular system. The following section gives a selective review of potential biomarkers in depression and ADHD and thus does not entail all biomarkers that have been identified.

BIOMARKERS IN DEPRESSION

Frontal alpha asymmetry (FAA)

One biomarker that is commonly discussed in depression the frontal alpha asymmetry (FAA), which is a hemispheric asymmetrical distribution of alpha activity in the frontal regions. Early studies reported

an alpha power asymmetry in depression (d'Elia & Perris, 1973; Schaffer, Davidson, & Saron, 1983), and some studies showed that FAA could discriminate between depressed and non-depressed individuals (e.g., (Baehr, Rosenfeld, Baehr, & Earnest, 1998; Cantisani et al., 2016; Gollan et al., 2014)). However, a recent meta-analysis showed a non-significant effect of FAA as a diagnostic tool in depression, potentially explained by studies that employed small samples reporting large effect sizes (*ES*) (van der Vinne, Vollebregt, van Putten, & Arns, 2017). This is illustrated in Figure 3 (below), which shows that the *ES* of the FAA nears zero as the sample size increases.

Yet, some studies suggest that it is possible to use FAA as a prognostic tool (Arns et al., 2016; Arns, Etkin, et al., 2015; Bruder et al., 2001). It has been reported that responders to fluoxetine had greater right over left alpha power in the occipital regions, whereas responders showed an opposite asymmetry pattern (Bruder et al., 2008). Likewise, Arns et al. (2016) found that female responders to escitalopram and sertraline also showed greater right over left alpha power, and this was successfully used as a stratification tool in a prospective replication (van der Vinne et al., 2021).

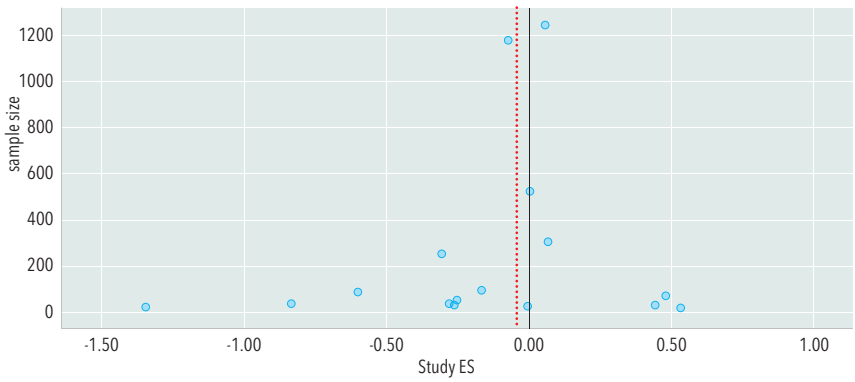


Figure 3: *Summarizing the results of the van der Vinne et al. (2017) systematic review and meta-analysis on FAA in depression, where no significant difference was found between depressed and non-depressed patients for FAA (red dotted line, Grand mean* ES = -0.007*). This figure demonstrates that most of the positive findings were driven by small sample sizes (Y-axis), whereas for larger studies an* ES *closer to 0 is found, further demonstrating the lack of a significant difference in FAA for depressed patients as a group.*

Individual alpha frequency (IAF)

The IAF, the peak frequency at which alpha oscillations occur in an individual, has also been investigated in depression. The results do not seem to converge towards a similar outcome. Recent studies found no association between IAF at baseline and clinical outcome following 5 Hz rTMS in post-traumatic stress disorder (PTSD) and comorbid major depressive disorder (MDD) (Petrosino, Zandvakili, Carpenter, & Philip, 2018), nor was baseline IAF found to be predictive of antidepressant medication response (although the authors note that the analyses were limited due to missing treatment days and variable EEG recording quality) (Widge, Avery, & Zarkowski, 2013). This last study was commented on by Arns and Olbrich (2014) for confining analyses to Fp1 and Fp2, determining alpha in anterior regions while alpha is most prominent in posterior regions, and determining alpha from an eyes open EEG, while alpha is most prominent during eyes closed. It has also been shown that a low IAF was associated with a favorable response to sertraline, but not to escitalopram or venlafaxine (Arns, Gordon, & Boutros, 2015) and a low IAF has also been associated with non-response to rTMS treatment in depression (Arns, Drinkenburg, Fitzgerald, et al., 2012). These latter results were attempted to be replicated (Krepel et al., 2018; Chapter 2) and Chapter 3 (Roelofs et al., 2021) is a follow-up and refinement of the association of IAF and rTMS response.

Other predictors of treatment (non-)response in depression

Various clinical, neuroscientific, genetic, demographic characteristics (Kemp, Gordon, Rush, & Williams, 2008; Kozel et al., 2008; Saveanu et al., 2015; Trivedi et al., 2006), and psychological features (Krepel et al., 2019; Chapter 5) have been associated to antidepressant treatment outcome. However, at this moment no predictors have clinical use in predicting treatment outcome to various antidepressant treatments (Bagby, Ryder, & Cristi, 2002; Krepel et al., 2019; Chapter 5; Simon & Perlis, 2010). In Chapter 5 (Krepel et al., 2019), which investigates the potential value of psychological features as predictors of rTMS non-response, this will be further elaborated on. Gamma oscillations have also been proposed as a potential biomarker of depression (Fitzgerald & Watson, 2018). The use of gamma as a biomarker in depression will also be discussed in Chapter 4 (Krepel et al., in press), in

which a replication study considering the role of beta/gamma activity in suicidal ideation is reported.

BIOMARKERS IN ADHD

Excess theta and theta/beta ratio (TBR)

Lubar introduced the concept of TBR as a diagnostic measure for ADHD (Lubar, 1991) with the clearest replication from Monastra et al. (1999) who demonstrated that TBR could discriminate ADHD individuals from healthy controls with an accuracy of 88%. However, a meta-analysis could not confirm that TBR is a reliable diagnostic marker for ADHD (Arns, Conners, & Kraemer, 2013), and van Dijk et al. (2020) showed that different methods for EEG signal processing can result in significantly different TBRs (albeit this could not explain the lack of a difference in TBR between ADHD and non-ADHD in recent studies). Yet, TBR may show to be useful as a prognostic tool (Arns & Gordon, 2014). A substantial proportion (26 – 38%) of ADHD subjects present with a high TBR and excess theta activity and these subgroups have been found to respond to stimulant medication (Arns, Gunkelman, Breteler, & Spronk, 2008; Clarke, Barry, McCarthy, & Selikowitz, 2002; Suffin & Emory, 1995), although this was not replicated a recent study (Arns et al., 2018). It has also been reported that baseline excess theta is associated with a favorable response to Theta/Beta neurofeedback (Arns, Drinkenburg, & Kenemans, 2012; Gevensleben et al., 2009; Janssen et al., 2016; Monastra, Monastra, & George, 2002), which suggests that for the subgroup with increased TBR, Theta/Beta neurofeedback is a preferred treatment option. This will be further elaborated on in Chapter 6 (Krepel et al., 2020).

Individual alpha frequency (IAF)

Several studies have demonstrated that within the previously described excess theta group a group exists that is primarily characterized by a slowed IAF rather than excess theta (Arns et al., 2008; Lansbergen, Arns, van Dongen-Boomsma, Spronk, & Buitelaar, 2011; Vollebregt, van Dongen-Boomsma, Slaats-Willemse, Buitelaar, & Oostenveld, 2015). A low IAF in male adolescents has been found to be associated with non-response to MPH (Arns et al., 2008) which was replicated in the multicenter Study to Predict Optimized Treat-

ment for ADHD (iSPOT-A) for male adolescents (Arns et al., 2018), suggesting a slow IAF as a predictor for non-response to psychostimulant medication. However, IAF seems to be differently related to neurofeedback remission. This will be further elaborated on in Chapter 6 (Krepel et al., 2020).

SUMMARY

This section describes multiple biomarkers that have been investigated in depression and ADHD. For depression, FAA, IAF, and a variety of other markers such as psychological/behavioral concepts may provide baseline information that can be evaluated as potential biomarkers. For ADHD, the TBR and IAF each show some predictive value in ADHD treatment and may potentially be used as stratification markers.

A lot of research has been performed in the search for biomarkers in mental healthcare. However, at this moment there is no consensus about what biomarkers are unquestionably true and which ones require more investigation. For example, a promising study by Drysdale et al. (2017) showed that four different biotypes based on fMRI connectivity could be distinguished in depression, each of which showed a different clinical profile. More so, these biotypes were related to distinct response patterns to DMPFC rTMS treatment. However, in a response to Drysdale et al. (2017), Dinga et al. (2019) demonstrated that this finding could not be replicated. The non-replication could be explained by overfitting and the assumption that clusters existed within the fMRI data (while this did not need to be the case) in the original study (Dinga et al., 2019). Conflicting results within different modalities of psychiatric research is a reappearing theme and one potential explanation for this is a lack of replication studies. The replicability of findings is important to consider since unstable findings will not be useful in clinical practice. Therefore, next to the development of biomarkers in mental health, a secondary focus of this thesis is the replicability of scientific findings.

THE IMPORTANCE OF REPLICATION

Psychological and social science have been argued to be in a Replication Crisis (for a more in-depth discussion, see, for example, Diener & Biswas-Diener, 2015; Shrout & Rodgers, 2018; Yong, 2018, November 19), yet replicability is considered an essential trait of scientific findings by philosophers (Popper, 1959/2002), methodologists (Fisher, 1926), and early psychologists (Dunlap, 1926). A broad distinction can be made between 'direct replication' and 'conceptual replication' (Schmidt, 2009), either of which serves a similar purpose but do so differently. That is, direct replications are useful in reducing false positives and closely follow the experimental design (taking into account current knowledge of the to-be-replicated study). This does not mean that direct replications take into account every aspect of the original study, rather, only the critical elements that are needed to replicate the original study are considered (Zwaan, Etz, Lucas, & Donnellan, 2017). Conceptual replications attempt to replicate the previously tested hypothesis but may do so using different methods (Schmidt, 2009), and changes to the original procedures might be as such that the effect size changes as a result thereof (Zwaan et al., 2017). Zwaan et al. (2017) also argue that conceptual replications test the generalizability of the results, act as an extension on the original research, and extrapolate the results to other contexts. It is thus important for theory building.

Attempts have been made to increase the replication rate (Open Science Collaboration, 2012; Pavlov et al., 2020, November 27), but initial analyses showed that a mere 35 out of 97 studies replicated (Open Science Collaboration, 2015). However, other reports show that it is possible to achieve high successful replication rates. Protzko et al. (2020, September 10) performed a prospective replication study and from this report, it followed that if the original research could answer to specific requirements (including preregistering the study (i.e., to submit the design and analysis plan, including materials, protocols, and procedures for data cleaning, exclusion, and analysis) to self-confirmatory tests, reporting the results irrespective of the outcome, using large sample sizes, and sharing specialized materials for the experimental design), successful replication rates increased. Protzko et al. (2020, September 10) reported a replication rate of 86% - in stark

contrast with earlier reported findings where rates generally vary between 36 – 61.9% (Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015). In accordance with this is 'The Replication Recipe', which states that convincing replications should adhere to five conditions, being: 1) a careful definition of the to-be-replicated methods and effects, 2) exactly following the methods of the original study, 3) high statistical power, 4) being transparent to other experts by making the details of the replication attempt available, and 5) evaluating and comparing the replication results to the original results (Brandt et al., 2014). Also, Simons (2014) highlights that direct replication of one's own results is important, but that it preferably should be done by other, independent labs.

Even when replications are performed and the methods are sound, the results of replications are not always published. For example, Rosenthal (1979) introduced the 'file drawer problem', in which non-significant results are not considered for publication and literally get tucked away in file drawers (although this effect does not solely apply to replication studies). A recent study supports this, as the results indicated that 50.0% of 2,155 surveyed academic psychologists admitted to selectively report studies that "worked" (John, Loewenstein, & Prelec, 2012). Another possibility is that replications do not get published because of a publication bias against replication studies and Neuliep and Crandall (1990) found support for this notion. Likewise, a study that reviewed the 'instructions to authors' and 'aims and scope' of 1151 psychology journals, specifically looking for sections stating the acceptance of replications, found that only 33 (3%) stated to accept replications. Of the remaining journals considered, 728 (63%) journals did not state to accept replications but did not discourage replications either, 379 (33%) journals implicitly discouraged replications, and 12 (1%) actively discouraged replications (Martin & Clarke, 2017). Overall, although the number of replication studies since 2000 was 1.84 times higher than it was from 1950 – 1999, only 1.07% of all publications are replication studies (Makel, Plucker, & Hegarty, 2012). Importantly, not reporting negative or null-findings can influence the validity of scientific research. This is illustrated by a recent meta-analysis by Widge et al. (2019), who investigated the predictive power of biomarkers based on the Quantitative EEG (QEEG)

in depression. It was found that QEEG biomarkers show some predictive power, but this result was primarily driven by small-sample studies reporting large effect sizes. Negative or small effect size studies were missing. Thus, it was concluded that the predictive power of QEEG biomarkers is not well supported and that the QEEG is not ready for widespread use (Widge et al., 2019). These findings suggest that non-replications and null-findings are published relatively little and this possibly undermines the reliability of findings.

As such, (non-)replications are needed to establish the reliability of (QEEG) biomarkers and to draw more accurate conclusions about the applicability thereof. Following our non-replication (Krepel et al., 2018; Chapter 2) and later the sobering conclusion of Widge and colleagues (2019), it was decided to direct more attention to the aspect of replication in our research.

SUMMARY

Replication is an important tool through which the robustness of scientific findings can be tested. Different forms of replication exist, among which are direct and conceptual replication. Successful replications occur relatively little and to increase the rate of successful replications, the original, as well as the replication study, should adhere to certain conditions. Even when replication studies are performed, they are not always considered for publication.

AIMS AND OUTLINE

Within this thesis, biomarkers and personalization parameters, as well as the replicability and clinical relevance thereof will be investigated.

Chapter 2 (Krepel et al., 2018) describes a non-replication. The original study reported a slowed IAF, a larger P300 amplitude, and more frontal theta as predictors of rTMS non-response. In a newly acquired sample, none of these predictors survived replication. Following this, our future studies focused on the discovery as well as the replication of scientific findings. Additionally, a sharing of the

database was proposed to encourage other researchers to perform replication studies.

In Chapter 3 (Roelofs et al., 2021) a result of this data-sharing proposal is described. As an initial communication from the International Consortium On Neuromodulation – Biomarker Discovery (ICON-DB; initiated at the 3rd International Brain Stimulation Conference (BRST2019) in Vancouver, Canada), it was attempted to replicate the findings reported by Corlier et al. (2019). This study reported an association between IAF and distance to 10 Hz in a depressed sample treated with 10 Hz rTMS. Specifically, the closer the IAF to 10 Hz, the better the clinical outcome. In the replication study (Roelofs et al., 2021; Chapter 3) again the relation between 10 Hz and IAF proximity to 10 Hz was found in the group treated with 10 Hz rTMS. Post-hoc analyses revealed a quadratic association between IAF proximity and rTMS response, rather than a linear one (as was previously assumed in Chapter 2 (Krepel et al., 2018)). Chapter 3.1 (Bailey et al., 2021) shows another communication from ICON-DB, this time being a non-replication. The original study reported higher resting EEG theta connectivity and low alpha power in rTMS responders (Bailey et al., 2019). Using the dataset proposed for data-sharing (Krepel et al., 2018; Chapter 2), these effects could not be replicated.

Chapter 4 (Krepel et al., in press) shows another collaboration that was dedicated to replicating a study investigating a biomarker found in depressed females with suicidal ideation. The original study found that frontal beta/gamma hypoactivity was indicative of a higher suicide risk and this result was attempted to be replicated in the large iSPOT-D sample. Results showed that this biomarker could not be replicated.

Chapter 5 (Krepel et al., 2019) describes a study based on a Discovery-Replication approach. In a sample of 196 depressed individuals (treated with rTMS), psychological features were investigated for their predictive utility. The total sample was divided into a Discovery (n = 119) and a Replication (n = 77) sample. In the Discovery sample, psychological features at baseline were investigated. In the Replication sample, only the psychological features that showed to be different between responders and non-responders in the Discovery

sample were considered. If the psychological feature survived replication, then this variable would be used in a discriminant model. These results were also investigated using the total sample with a Bonferroni-corrected *p*-thresholding. The results indicated a significant effect of anhedonia, but the analyses also showed that out of 32 analyses, four initially positive findings did not replicate. Using anhedonia in a discriminant model did not result in a clinically relevant model.

Chapter 6 (Krepel et al., 2020) describes a replication study on the effectiveness of QEEG-*informed* neurofeedback, as first reported by Arns, Drinkenburg, and Kenemans (2012). Using a naturalistic replication approach, clients reporting symptoms of ADHD were assigned to a neurofeedback protocol in accordance with specific QEEG characteristics (in line with Arns, Drinkenburg, and Kenemans (2012)). The original study reported a response rate of 76% (based on a 50% reduction in ADHD symptom presentation) and this did not significantly differ from the response rates reported in this replication study. Additionally, a pooled remission rate of 57.4% was obtained. Post-hoc analyses identified several predictors of neurofeedback remission. Remitters showed to have lower hyperactivity levels at baseline compared to non-remitters and female remitters showed shorter P300 latencies. Based on earlier work (Arns et al., 2018), IAF analyses were performed in boys only. This showed that boys who remitted had a lower IAF compared to non-remitters.

In Chapter 7 (Krepel et al., under review) another replication study is presented. In 2015, Arns and colleagues showed, in a heterogeneous sample, that frontocentral spindling excessive beta (SEB) was related to more impulse control problems as well as sleep maintenance problems. In a new heterogeneous sample, analyses (confined to adults with no SEB or SEB) demonstrated that individuals with frontocentral SEB had more impulse control problems on a self-rated scale and more false positive responses on a Continuous Performance Task (CPT). No associations with sleep were found.

Chapter 8 provides the summary and general discussion of this thesis.

2

NON-REPLICATION OF NEUROPHYSIOLOGICAL PREDICTORS OF NON-RESPONSE TO RTMS IN DEPRESSION AND NEUROPHYSIOLOGICAL DATA-SHARING PROPOSAL

Published as: Krepel, N., Sack, A. T., Kenemans, J. L., Fitzgerald, P. B., Drinkenburg, W. H., Arns, M. (2018). Non-replication of neurophysiological predictors of non-response to rTMS in depression and neurophysiological data-sharing proposal. *Brain Stimulation, 11*(3), 639-641. doi: 10.1016/j.brs.2018.01.032

DEAR EDITOR,

The application of repetitive transcranial magnetic stimulation (rTMS) as a treatment for major depressive disorder (MDD) has been shown to be effective when applied to either the right or left dorsolateral prefrontal cortex (DLPFC) in placebo controlled studies (Schutter, 2009, 2010) as well as open-label studies (Donse, Padberg, Sack, Rush, & Arns, 2018). Given the effectiveness of rTMS as a treatment for MDD, the interest for finding clinical or (neuro)physiological predictors has been increasing. In 2012, we (Arns, Drinkenburg, Fitzgerald, et al., 2012) reported neurophysiological predictors of non-response (NR) for rTMS treatment in MDD. These predictors included the EEG metrics: increased frontocentral theta, a low individual alpha frequency (IAF), and a large P300 amplitude at site location Pz in a sample of 90 MDD patients, however these biomarkers still require replication. The aim of the current study is to investigate the replicability of these findings in a newly collected sample, and also to make our EEG and ERP data available to scientific use for replication analyses that have specific formulated hypotheses, and thus facilitating future replication studies.

METHODS AND MATERIALS

Design

This study was an open-label study (details published recently in this journal (Donse et al., 2018). In summary, data for this replication cohort were collected in two clinics (Brainclinics Treatment/neuroCare Nijmegen and The Hague, The Netherlands) between November 2009 and March 2016. Only data from patients with 1) a primary diagnosis of Depression or Dysthymic disorder according to the MINI (MINI Plus Dutch version 5.0.0) and 2) a Becks Depression Inventory (BDI) score of 14 or higher who were treated with left DLPFC HF rTMS (10 Hz) or right DLPFC LF rTMS (1 Hz) were included for this study. Exclusion criteria were: previously treated with ECT, epilepsy, wearing a cardiac pacemaker, metal parts in the head and pregnancy. All patients signed an informed consent form before treatment was initiated. Response was defined by achieving response (≥ 50% decrease on BDI) or remission (BDI ≤ 12), like in the earlier study. EEG and ERP acquisition and analysis were identical to the methods used in the earlier 2012 study.

Analysis

Given the confirmatory nature of this data analysis where we specifically aimed to replicate earlier reported measures, we initially only ran One-Way ANOVAs to test differences between responders and non-responders in IAF, P300 amplitude, and frontocentral theta (for exact processing details see Arns, Drinkenburg, Fitzgerald, et al. (2012)).

Results

A total of 106 patients were included in this study (average age: 43.92 yrs, range 18 – 78 years; 50 females and 56 males; 63 responders). No differences between responders (R) and non-responders (NR) were found for age, gender, or rTMS protocol (each $p > .236$). BDI baseline scores were significantly lower for responders than for non-responders ($p = .018$, $F = 5.761$, $df = 1$).

EEG biomarkers

No significant differences were found between R and NR for frontal theta (F7, F3, F4; Figure 1a on page 37); P300 amplitude at electrode site Fz and Pz (Figure 1b), nor for IAF (Figure 1c). The patterns of results were in the same direction as the original study for P300 and

IAF albeit not significant and with small effect sizes.

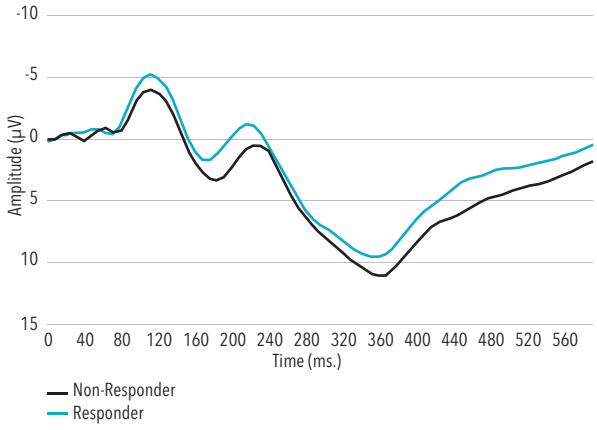
Exploratory analysis

Additional analyses were performed to test for subgroup interactions. One-Way ANOVAs for males and females separately also yielded no significant differences between responders and non-responders for frontocentral theta, P300 amplitude at Pz, and IAF. We also performed a univariate analysis with age as a covariate, but this too did not yield significant differences between responders and non-responders on the targeted variables.

1a. ANOVA outcomes for frontal theta, IAF and P300

	Responder	Non-responder	*F*	*p*	Cohen's *d*
Theta per electrode site					
F7	.7941 (.25923)	.7641 (.28731)	.305	.582	.110
F3	1.0795 (.27429)	1.0310 (.28521)	.753	.388	.173
F4	1.0749 (.27767)	1.0550 (.30055)	.120	.730	.069
P300 amplitude					
Frontal	7.414 (7.333)	5.106 (6.911)	2.315	.132	.324
Parietal	11.824 (7.964)	11.969 (7.087)	.008	.929	.019
IAF					
Frontal	9.048 (1.227)	8.891 (1.1048)	.449	.504	.138
Parietal	9.522 (1.337)	9.597 (.960)	.094	.760	.064

1b. P300 amplitude at Pz



1c. Power spectrum of IAF at F3, Fz, F4 and Pz

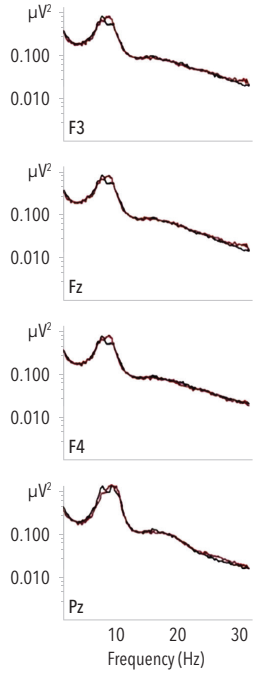


Figure 1a-c: *Figure 1a shows the ANOVA outcomes for the tested parameters demonstrating that there were no significant differences between R and NR for the tested parameters. Figure 1b visualizes the P300 amplitude at Pz, for which a difference between R and NR was found in Arns, Drinkenburg, Fitzgerald, and Kenemans (2012), lacking such a statistical significant difference between groups in the current sample. Figure 1c shows the visualization of the power spectral content for R (black) and NR (red), demonstrating no differences between R and NR.*

DISCUSSION

The aim of the current study was to replicate the findings from our earlier study (Arns, Drinkenburg, Fitzgerald, et al., 2012), however, we were unable to replicate the earlier obtained findings for frontocentral theta, P300 amplitude and IAF: Numerically the trends and direction of the results were the same for IAF and P300, however non-significant and with small effect sizes. The results for theta are in line with rather opposite findings throughout the literature where sometimes increased frontal midline theta with a putative generator in the rostral anterior cingulate cortex versus decreased frontocentral theta has been found to be related to antidepressant treatment response including rTMS (also see Pizzagalli (2011) and Arns, Etkin, et al. (2015) for reviews and data). In a previous paper combining our earlier sample and this new sample we were unable to find meaningful clinical predictors for treatment response to rTMS treatment in MDD, the predictors including depression severity (rated with BDI), comorbid depression, anxiety and stress (using DASS scales) (Donse et al., 2018) as well as personality traits (NEO-FFI; unpublished findings). These findings demonstrate that future treatment prediction studies should be adequately powered with sample sizes preferably larger than 100, and furthermore should aim to include replication analyses in order to more reliably report on biomarkers for treatment response. For our future EEG biomarker studies, we have implemented this by a priori dividing our current database into a discovery and replication dataset, which enables us to prospectively verify findings found in the discovery dataset. In addition, to reduce the likelihood of future non-replication, we hereby offer our full sample of EEG and ERP data (N = 196) for scientific use in replication analyses employing specifically formulated hypotheses.

3

INDIVIDUAL ALPHA FREQUENCY PROXIMITY ASSOCIATED WITH CLINICAL OUTCOME AFTER RTMS: A FIRST INDEPENDENT REPLICATION STUDY FROM THE ICON-DB CONSORTIUM

Published as: Roelofs, C. L., Krepel, N., Corlier, J., Carpenter, L. L., Fitzgerald, P. B., Daskalakis, Z. J., Tendolkar, I., Wilson, A., Downar, J., Bailey, N. W., Blumberger, D. M., Vila-Rodriguez, F., Leuchter, A. F., Arns, M. (2021). Individual alpha frequency proximity associated with clinical outcome after rTMS: A first independent replication study from the ICON-DB consortium. *Clinical Neurophysiology, 132*(2), 643-649. doi: 10.1016/j.clinph.2020.10.017

Author contributions: CR managed the literature search, performed the analyses, and wrote the first draft of the manuscript. NK supervised the first author and helped establish the first draft of the manuscript. All other authors contributed, reviewed, and approved the final manuscript.

ABSTRACT

Objective

The aim of the current study was to attempt to replicate the finding that the individual alpha frequency (IAF) as well as the absolute difference between IAF and 10 Hz stimulation frequency (IAF-prox) is related to treatment outcome.

Methods

Correlations were performed to investigate the relationship between IAF-prox and percentage symptom improvement in a sample of 153 patients with major depressive disorder treated with 10 Hz (n = 59) to the left dorsolateral prefrontal cortex (DLPFC) or 1 Hz (n = 94) to the right DLPFC repetitive Transcranial Magnetic Stimulation (rTMS).

Results

There was a significant negative correlation between IAF-prox and the percentage of symptom improvement only for the 10 Hz group. Curve fitting models revealed that there was a quadratic association between IAF and treatment response in the 10 Hz group, with a peak at 10 Hz IAF.

Conclusion

The main result of Corlier and colleagues was replicated, and the findings suggest that the distance between 10 Hz stimulation frequency and the IAF may influence clinical outcome in a non-linear manner.

INTRODUCTION

Currently, antidepressant drugs are often the first treatment option for patients with major depressive disorder (MDD). However, not all patients benefit from this treatment, usually achieving response rates of 30 – 40% to the first treatment course (Rush et al., 2006). Multiple studies have demonstrated that repetitive transcranial magnetic stimulation (rTMS) applied to the left or right dorsolateral prefrontal cortex (DLPFC) is an efficacious (Brunelin et al., 2014; George et al., 2010; O'Reardon et al., 2007) and effective treatment in MDD (Carpenter et al., 2012; Donse et al., 2018; Fitzgerald, Hoy, Anderson, & Daskalakis, 2016), offering a viable alternative for patients that do not respond to antidepressants. Increased mechanistic understanding of both MDD and how rTMS exactly works in the treatment of MDD could lead to optimized selection of treatment for patients as well as individualized stimulation parameters, thereby potentially improving treatment outcome.

The dominant activity in the resting state EEG is the alpha rhythm, with a frequency between 7 – 13 Hz. Research exploring the individual alpha frequency (IAF) has identified inter- and intra-individual differences (Haegens, Cousijn, Wallis, Harrison, & Nobre, 2014),

yet the IAF seems to be heritable and stable over time (Smit, Wright, Hansell, Geffen, & Martin, 2006). Ever since the first description that the alpha frequency can be entrained by photic stimulation by Adrian and Matthews (Adrian & Matthews, 1934), this phenomenon has been considered a possible mechanism of action of various neuromodulation treatments such as tACS and rTMS (Leuchter, Hunter, Krantz, & Cook, 2015). This neural oscillation entrainment also has the property that phase coupling can be more pronounced with increasing stimulation intensity as well as at stimulation frequencies closer to each participants' intrinsic frequency (Notbohm, Kurths, & Herrmann, 2016).

The primary purpose of this paper is to attempt to replicate the recent findings reported by Corlier and colleagues, where the proximity of IAF to the 10 Hz TMS stimulation frequency was associated with better treatment response (Corlier et al., 2019).

Oscillations in the alpha band are thought to represent a thalamocortical oscillation (Bollimunta, Mo, Schroeder, & Ding, 2011). Interference of this rhythm has been hypothesized to reset thalamocortical oscillators that are abnormal in depressed individuals and therefore may be related to the therapeutic mechanisms of rTMS (Leuchter, Cook, Jin, & Phillips, 2013). Applying the concept of entrainment to rTMS, stimulation at 10 Hz will hypothetically result in a stronger entrainment of alpha oscillations in an individual whose IAF is closer to 10 Hz, rather than in an individual whose IAF is further away from 10 Hz (Notbohm et al., 2016). This effect of rTMS has already been demonstrated in schizophrenia patients. Jin and colleagues delivered rTMS to schizophrenia patients at their IAF, which resulted in better clinical effects relative to stimulation at other frequencies, as well as increased alpha power from pre- to post-treatment ((Jin et al., 2005) which was replicated (Jin et al., 2012)).

The alpha frequency varies between individuals, and multiple studies have investigated the IAF and its association with treatment response to antidepressant treatments such as rTMS (Arns, Drinkenburg, Fitzgerald, et al., 2012; Corlier et al., 2019), antidepressant medication (Arns, Gordon, et al., 2015), as well as in other disorders (for review

see Olbrich, van Dinteren, and Arns (2015)). With respect to rTMS two approaches have been reported. Several studies have reported a linear association between IAF and response to rTMS, where overall a slower IAF is considered a predictor for non-response (Arns, Drinkenburg, Fitzgerald, et al., 2012; Arns, Spronk, & Fitzgerald, 2010) generally interpreted as reflecting abnormality (Arns, Gordon, et al., 2015). However, more recently the same group failed to replicate the linear association (Krepel et al., 2018; Chapter 2). On the other hand, Corlier and colleagues reported an association between the proximity of IAF to the 10 Hz TMS stimulation frequency and treatment response (Corlier et al., 2019). This latter 'proximity' approach implicates a quadratic association between IAF and treatment response (higher treatment responses in individuals with an IAF of 10 Hz, and lower response in individuals with an IAF of both higher and lower frequencies). In addition to the relationship between IAF-prox and treatment outcome, Corlier and colleagues also found that the absolute IAF was related to treatment outcome (Corlier et al., 2019). Of interest, Corlier and colleagues used an average reference montage, which is known to reflect more focal cortical activity, whereas Arns and colleagues used a linked-ears montage in their earlier studies. Since the primary purpose of this study was to replicate the study performed by Corlier and colleagues, the primary IAF employed was based on average reference. In addition, a linked-ears montage has also been tested as secondary analysis, and the amount of (dis)agreement between these EEG referencing montages. In this study, the alpha activity as measured below the coil is investigated. This is in line with replication, since Corlier and colleagues used the average reference and therefore focal frontal alpha was studied.

The aim of the current study was to investigate the replicability of the finding of Corlier et al. (2019) using the data from Krepel et al. (2018; Chapter 2). It was hypothesized that the proximity of an individual's alpha frequency to 10 Hz is associated with clinical improvement in MDD after 10 Hz rTMS applied to the left DLPFC. In order to provide a test of the proposed explanation that it is the proximity of IAF to the stimulation frequency that drives this association (rather than simply the proximity of IAF to 10 Hz regardless of stimulation), an additional analysis was performed. In this analysis, the same

comparison between proximity to 10 Hz and treatment effect was used, but in a group of participants treated with 1 Hz stimulation to the right DLPFC. It was hypothesized that the association would not be present in the group that had 1 Hz rTMS applied to the right DLPFC. Additionally, it was hypothesized that the absolute IAF is not correlated with treatment outcome, given the previously mentioned non-replication (Krepel et al., 2018; Chapter 2). Furthermore, the effect of two different EEG montages was investigated; a linked ears montage (as previously used in Arns, Drinkenburg, Fitzgerald, et al. (2012) and Arns et al. (2010) and an average reference montage (as used in Corlier et al. (2019)).

METHODS

Study design

An open-label study was conducted with data of patients who were diagnosed with major depressive disorder (MDD) or dysthymia and treated with rTMS in combination with psychotherapy. All subjects signed an informed consent form. Inclusion criteria were: 1) A diagnosis of MDD or dysthymia; 2) BDI-II-NL (Beck Depression Inventory) of 14 or higher at baseline; and 3) Treatment consisting of rTMS (left DLPFC rTMS at 10 Hz or right DLPFC at 1 Hz) combined with psychotherapy for at least 10 sessions. In contrast with Corlier et al. (2019), the choice of the applied rTMS protocol was not based on clinical criteria, but the first few years the standard protocol was 10 Hz, and when it was found that the clinical benefits for 10 Hz and 1 Hz were similar, the standard protocol became 1 Hz. Exclusion criteria were: 1) Previous ECT treatment; 2) Epilepsy; 3) Wearing a cardiac pacemaker; 4) Metal implants in the cranium; and 5) Pregnancy. Assessments including BDI, DASS, and PSQI were taken at baseline, every fifth session, and clinical endpoint (last rTMS session). Patients received an average of 20.8 (*SD* 7.3) rTMS treatment sessions, which did not differ between the 10 Hz and 1 Hz treatment groups ($t(173) = -0.260$, $p = .795$). Medication usage was not systematically tracked. Further details about treatment and clinical variables can be found in Donse et al. (2018) and Krepel et al. (2019; Chapter 5).

EEG procedure

Resting-state EEG recordings were performed using a standardized methodology (Brain Resource Ltd., Australia). EEG data were acquired from 26 channels (NuAmps; 10 – 20 electrode international system) and were recorded for two minutes with eyes closed (EC), and two minutes with eyes open (EO) with the participant asked to fixate on a red dot on the screen. Participants were instructed to remain relaxed for the duration of the recording. Vertical eye movements were recorded with electrodes above the middle of the left eyebrow and below the middle of the left bottom eyelid. Horizontal eye movements were recorded with electrodes placed lateral to the outer canthus of each eye. Skin impedance was < 10 kOhm for all electrodes (sampling rate = 500 Hz; Low-pass filter of 100 Hz with attenuation of 40 dB per decade and no high-pass filter).

EEG pre-processing and IAF

EEG data were analyzed in Brain Vision Analyzer 2.0 (Brain Products). Data were EOG-corrected using the regression-based Gratton technique (Gratton, Coles, & Donchin, 1983), re-referenced to the average reference, filtered (0.3 – 100 Hz and notch), segmented in 4-second epochs, and artifactual epochs were removed using an automated procedure, with a maximal allowed difference of 150 mV within an interval of 100 ms. An average of 29.5 segments (*SD* = 1.01) were included per subject, resulting in an average of 118 seconds of included data per subject. This implicated 98.3% of usable data. The IAF was extracted from eyes closed resting states and calculated for F3 and F4. In short, calculating the IAF consisted of the following steps: 1) A Fast Fourier Transform applied to EC using 4 sec. segments with 50% overlap to get a power spectrum for each site, with a Hamming window applied to each segment; 2) The IAF for each site was determined by identifying the maximum value within the 7 – 13 Hz alpha range. If the power of the alpha frequency peak was lower than 1.5 Z-score below the mean, the patient was considered not to have a dominant IAF rhythm and thus was not included in the analysis (these EEGs are also known as low voltage alpha EEG). This 1.5 Z-score cut-off was chosen because a bimodal frequency distribution was visible for IAF (see Supplementary Figure S1 on page 59), and this cut-off reflected the majority of the people and incorporated the bi-

modal distribution as well as possible. Secondly, visual inspection of the raw EEG data confirmed that no dominant alpha power was present, and no clear alpha peak could be distinguished from the background EEG. This criterion yielded a similar percentage of subjects having no dominant IAF as reported by Corlier and colleagues. Of the total sample of 175 patients, 12.6% was determined not to have a dominant IAF, which is consistent with the 15.1% reported by Corlier and colleagues (2019). IAF-prox was calculated as the absolute value of the distance from IAF to 10 Hz.

Statistics

Since the primary aim was to replicate the results from the (Corlier et al., 2019) study, an a priori defined analysis plan was drafted by the first author (CR) was shared with the ICON-DB consortium and amended/approved by all members. Data analysis was carried out exactly according to this analysis plan and the primary outcome was thus defined as a correlation between continuous symptom improvement (BDI change) and IAF at F3 (for 10 Hz rTMS) quantified using an average reference, covaried for age. Due to the a priori defined primary hypothesis and replication nature, a one-tailed partial correlation was conducted.

Descriptive statistics at pre-treatment, post-treatment and change scores can be found in Table 1 (opposite page). Since TMS protocol specific effects were expected, only data was included from patients that received one rTMS protocol (1 Hz or 10 Hz), hence 21 were excluded from the original sample (N = 196) resulting in a sample of 175. 12.6% had no dominant IAF and thus were excluded, resulting in a sample of 153 MDD patients included in the analyses. The dataset (N = 153) was divided into groups based on rTMS protocol: a group of patients treated with 10 Hz rTMS applied to the left DLPFC and a group of patients treated with 1 Hz rTMS applied to the right DLPFC. For the 10 Hz group the IAF was estimated at F3, while for the 1 Hz group the IAF was estimated at F4 to match the calculation of IAF to the stimulation site. To ensure sample comparability of the two groups, a Chi-square test was performed to compare gender and response ratios and an ANOVA to compare age, baseline depression severity, clinical improvement, and IAF between the two groups.

Due to the a priori defined, directional and replication nature of our primary hypothesis, a one-tailed Spearman correlation was conducted. One-tailed statistical tests are recommended when a result in the opposite direction to our previous research would provide the same rejection of our previous conclusion as no difference between groups (Ruxton & Neuhäuser, 2010). To further investigate whether the relation between IAF and BDI percentage change was linear or quadratic, post-hoc curve fitting was applied using GraphPad Prism (version 6.00 for Macintosh, GraphPad Software, La Jolla California USA, www.graphpad.com). A straight line was statistically compared to a quadratic function, where the quadratic function was constrained to a maximum at 10 Hz, in line with the hypothesis. Using Akaike's Information Criteria it was determined how well the data supports the straight line fit or the quadratic fit, and additionally an analysis of variance determined how much a line model improves by chance with a quadratic function.

As the study was designed to test the single primary hypothesis, that IAF-prox to 10 Hz is associated with treatment response in the 10 Hz group only (with other statistical tests conducted only to demonstrate the specificity of the replication result to variation in pre-processing steps and treatment parameters), no multiple comparison corrections were necessary.

	Total sample	10 Hz	1 Hz
	N = 153	*n* = 59	*n* = 94
Mean age (*SD*)	43.00 (12.95)	40.27 (11.89)	44.72 (13.34)
Sex (*n* male)	77 (50%)	30 (51%)	47 (50%)
BDI pre; mean (*SD*)	30.99 (9.60)	30.41 (8.90)	31.35 (10.04)
BDI post; mean (*SD*)	13.99 (12.16)	13.71 (11.56)	14.17 (12.59)
BDI mean % change (*SD*)	55.96 (34.24)	54.38 (35.22)	56.95 (33.77)
IAF; mean (*SD*)	9.47 (1.16)	9.42 (1.16)	9.50 (1.10)
IAF-prox; mean (*SD*)	0.97 (0.82)	0.96 (0.86)	0.98 (0.79)

BDI: Beck Depression Inventory
IAF: individual alpha frequency.

Table 1: *Clinical outcome measures of the total sample, the 1 Hz rTMS sample and the 10 Hz rTMS sample.*

RESULTS

153 MDD patients were included in the analyses. In Table 1 (previous page), the mean baseline values with standard deviations are reported. No significant differences were found between the groups for sex, age, baseline depression severity (BDI intake), BDI change, or IAF (all $p > 0.086$).

A one-tailed Spearman correlation test demonstrated a significant correlation ($r(59) = -0.250$; $r^2 = 0.063$; $p = .028$) between BDI percentage change and IAF-prox in the 10 Hz sample (see Figure 1). However, repeating this analysis in the 1 Hz group showed no significant correlation ($r(94) = -0.119$; $r^2 = 0.014$; $p = .126$). Across the total sample a significant correlation was observed ($r(153) = -0.162$; $r^2 = 0.026$; $p = .022$; one-tailed) (without any exclusion there is still a correlation trend: $r(174) = -0.125$; $r^2 = 0.016$; $p = .051$). No significant correlations were found between BDI percentage change and the absolute IAF for both the 10 Hz sample ($r(59) = 0.006$; $r^2 < 0.001$; $p = .483$), the 1 Hz sample ($r(94) = -0.024$; $r^2 = 0.001$; $p = .410$), as well as the total sample ($r(153) = -0.007$; $r^2 < 0.001$; $p = .466$). Additionally, oscillation strength (alpha peak amplitude) and treatment response were not correlated ($r(153) = 0.003$; $r^2 < 0.001$; $p = .487$).

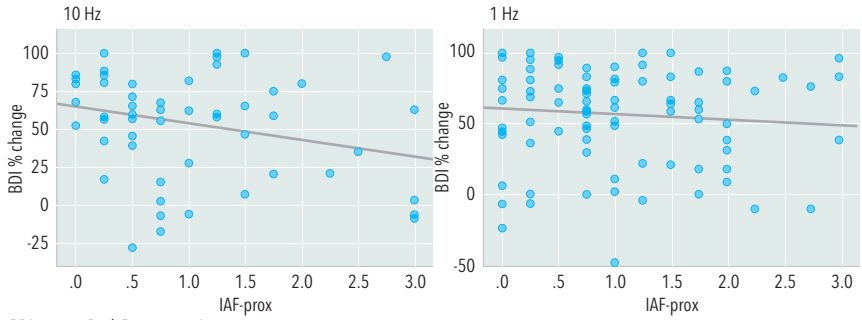


Figure 1: The association between IAF-prox and symptom improvement (BDI percentage change) for the 10 Hz rTMS sample (left) and the 1 Hz rTMS sample (right), where only a significant correlation was found for the 10 Hz rTMS sample (rTMS: repetitive Transcranial Magnetic Stimulation).

POST-HOC ANALYSES

In the study of Arns, Drinkenburg, Fitzgerald, et al. (2012) as well as Krepel et al. (2018; Chapter 2) a linked ears reference was used, whereas in this replication study an averaged reference EEG montage was used in line with Corlier et al. (2019). To examine a possible influence of referencing, the IAF was recalculated using the linked-ears montage, and the analyses were repeated. These analyses yielded no significant correlations between BDI percentage change and IAF-prox (all $p > .395$). A scatterplot of the comparison of these two IAFs can be found in Supplementary Figure S2 on page 59. Additional analysis demonstrated that the mean IAF calculated with a linked-ears montage (mean = 8.94; SD = 1.12) was significantly lower than the mean IAF measured with the average reference montage (mean = 9.48; SD = 1.22; $p < .000$).

Since the main hypothesis implies a quadratic association (Notbohm et al., 2016), a Loess fit was plotted for the 10 Hz group (Figure 2). This plot visualizes that indeed the data is best explained by a quadratic association, with a peak close to 10 Hz. To further test this statistically, curve fitting was applied. It was tested if a quadratic model, constrained to a maximum at 10 Hz, would fit the data better than a linear model. In the 10 Hz group, the quadratic fit was the correct model with 91.4% probability thus favoring the quadratic model over a linear model. Removal of the single outlier at 12.5 Hz did not change the results.

Figure 2: *The relationship between absolute individual alpha frequency (IAF) and BDI percentage change symptom improvement (Beck Depression Inventory) for the 10 Hz rTMS group only, fitted with a Loess fit (80% overlapping; left) and the quadratic fit with the peak constrained to 10 Hz which was a significantly better model for the data as compared to a line (right) (rTMS: repetitive Transcranial Magnetic Stimulation).*

DISCUSSION

The aim of this study was to replicate the association between clinical outcome and the proximity of the IAF to 10 Hz rTMS in an MDD sample as was recently published by Corlier et al. (2019). Our results replicated the earlier reported findings, where an association was observed between the proximity of the IAF to 10 Hz and clinical response to 10 Hz left DLPFC rTMS, which explained 6.3% of the variance. This association was not significant for the 1 Hz right DLPFC rTMS group. Additionally, in contrast with the results of Corlier et al. (2019), no association between absolute IAF and treatment response was found using both the average reference as well as linked ears references, thereby in line with the recent results from Krepel et al. (2018; Chapter 2). Curve fitting and a Loess fit – also see Figure 2 on the previous page – further confirmed that the association is quadratic, meaning that individuals with an IAF closer to 10 Hz showed most clinical improvement with 10 Hz rTMS. However, it is important to note that these results depend on the choice of EEG montage.

These results are theoretically in line with the properties of oscillation entrainment, which are conceptualized by the Arnold Tongue. The Arnold Tongue predicts that the degree of synchronization (entrainment) of an oscillator coupled to a rhythmic driving force depends on the amplitude of the driving force and the driving frequency (Fröhlich, 2015). With a driving frequency that approaches the intrinsic frequency, entrainment is more likely to occur (Notbohm et al., 2016). For example, tACS has a relatively weak stimulation strength and should therefore be aimed more accurately at the intrinsic stimulation frequency. On the other hand, rTMS has a relatively higher stimulation intensity and therefore requires less precise frequency matching (i.e., a relatively larger mismatch between stimulation frequency and intrinsic frequency is allowed). To summarize, the more closely an externally applied stimulation frequency matches an intrinsic frequency, the more likely it is that entrainment will occur. A higher degree of entrainment might be related to a stronger neuroplasticity effect, that eventually mediates long-term clinical and behavioral changes (Vosskuhl, Strüber, & Herrmann, 2018). It is unclear, however, how large a frequency difference can exist and still elicit optimal rTMS entrainment effects.

The current results help resolve the earlier contradictory results (Arns, Drinkenburg, Fitzgerald, et al., 2012; Corlier et al., 2019) where a linear association between IAF and rTMS response was reported, which was not replicated (Krepel et al., 2018; Chapter 2). Firstly, the use of an average reference montage in this study vs. a linked ears montage in these prior studies yielded different results. Secondly, a further inspection of the original data revealed that in the first sample from Arns, Drinkenburg, Fitzgerald, et al. (2012) there was a relative high proportion of IAFs below 10 Hz (Arns et al sample: 82% of subjects had IAF < 10 Hz vs. Krepel et al., 68.3% of subjects had IAF < 10 Hz), thereby explaining the earlier reported linear finding (i.e. if the majority of IAF are below 10 Hz, the quadratic association with a peak at 10 Hz will be modelled as a linear association). This could also explain why Corlier et al. (2019) observed a relationship between absolute IAF and clinical improvement (due to a high proportion of IAFs below 10 Hz), yet in the current study all data point to a quadratic association.

These findings should be interpreted in the context of resonance of brain circuitry. Each brain circuit has one or more preferred resonant frequencies at which its activity can be best modulated (Zaehle, Lenz, Ohl, & Herrmann, 2010). Studies using several different neuromodulation techniques have shown that cortical regions are particularly susceptible to the effects of stimulation at intrinsic peak frequencies in the delta (Marshall, Helgadóttir, Mölle, & Born, 2006; Schmidt, Iyengar, Foulser, Boyle, & Fröhlich, 2014), theta (Albouy, Weiss, Baillet, & Zatorre, 2017; Polanía, Nitsche, Korman, Batsikadze, & Paulus, 2012), alpha (Klimesch, Sauseng, & Gerloff, 2003; Thut et al., 2011), beta (Pogosyan, Gaynor, Eusebio, & Brown, 2009; Romei et al., 2016), or gamma (Helfrich et al., 2014) frequency ranges. Specific resonant frequencies vary among brain regions, as well as across individuals. Future rTMS studies should examine such endogenous resonant frequencies across the frequency spectrum, and the clinical outcomes of treatment in relation to resonant frequencies outside of the alpha band.

The findings of this study emphasize the need for replication, not only to confirm or refute previous results but also to sculpt and refine currently existing research. Interestingly, a recent report showed

that of 97 attempts to replicate previous research, only 35 were successful (Open Science Collaboration, 2015). The high rate of non-replication is of concern in the context of low rates of replication attempts; the rate of replications published in 100 journals has been studied and it was concluded that 1.07% of all publications were replications (Makel et al., 2012). Insufficient direct replication of previous EEG findings has also characterized the literature describing biomarkers for response prediction when treating depression (Widge et al., 2019). The gap between the low rate of replication attempts and the high rate of non-replication creates a false positive knowledge space, where studies provide evidence for conclusions that are not accurate or generalizable. The need for replication has been reported in multiple papers (Brandt et al., 2014; Makel et al., 2012; Simons, 2014), and the current study is an example of how studies may fit into this approach. This was also the primary reason for establishing the ICON-DB consortium at the 2019 Brain Stimulation conference in Vancouver (see acknowledgements for more details) of which this publication is the first result. Not only did the current study verify the results as obtained by Corlier and colleagues, but it also refined the result by considering the stimulation protocol. This builds on the currently existing body of knowledge, and aids in the development of a knowledge base which future research may extend upon and facilitates translation into clinical practice.

With regards to implications for rTMS treatment, the observed association between a patient's IAF and 10 Hz may imply that the specific frequency at which a patient is treated plays a role in clinical outcome in the treatment of MDD. The result might suggest that DLPFC stimulation with rTMS at the IAF could be more successful at entraining ongoing alpha oscillations in line with the Arnold Tongue model. This has already been demonstrated in schizophrenia patients (Jin et al., 2012; Jin et al., 2005), where individualized rTMS showed a significantly larger therapeutic effect than conditions with stimulation frequencies of 3 Hz or 20 Hz. However, an earlier smaller study where IAF + 1 Hz was applied did not find any advantages (Arns et al., 2010). They did, however, find a trend for reduced response to 9 Hz rTMS, which warrants caution and requires further research before such frequency individualization is implemented in clinical practice.

There were several limitations in this study. First, in this study psychotherapy was combined with rTMS making it difficult to distinguish whether the obtained relation of a marker to treatment outcome reflects a generic relationship for treatment improvement or a relationship for treatment improvement to either rTMS, psychotherapy, or the combination of both. Second, since the 1 Hz stimulation was applied on a different brain area than the 10 Hz stimulation, respectively right DLPFC and left DLPFC, further work is required to find out whether the finding generalizes to other brain areas or other stimulation patterns. Thirdly, most patients used antidepressant medication at the start and during the rTMS treatment (although all patients still met clinical criteria for MDD). Still, it is possible that the interaction between IAF, stimulation frequency, and clinical outcome was influenced by medication status or other uncontrolled factors. For example, benzodiazepines have the most marked effects on the EEG by slowing down the IAF (Sim & Tsoi, 1992). Third, even though the main result of Corlier et al. (2019) was replicated, it cannot be ruled out that somatosensory and auditory aspects of the rTMS mediated the effect, instead of the transcranial magnetic stimulation. Finally, although the curve fitting data confirmed that a model constrained to a 10 Hz peak was the best model, the Loess fit, in Figure 2 on page 51, suggests the optimal IAF is just below 10 Hz. Although an insufficient sample size prevents from drawing firm conclusions from this, this notion was confirmed in a recent model simulation by Li and colleagues (2019) where it was demonstrated that stimulation with a frequency slightly higher than the endogenous frequency results in optimal entrainment and enhancement. Future studies should investigate this in more detail, using larger samples.

With respect to the EEG pre-processing parameters, in the current study the common average reference was used, which is in line with Corlier et al. (2019). In earlier studies, a linked ears montage was used. To examine a possible influence of referencing, the IAF was recalculated using the average mastoid-reference, and the analyses were repeated. These analyses did not yield any significant relationship between IAF-prox and BDI percentage change. The IAF values for both montages are shown in Supplementary Figure S2 on page 59. In general, linked ears is used as a reference if the signal in central

areas or along the midline is of interest, as the mastoid electrodes are expected to pick up relatively little cortical activity from the top of the head. Therefore, a linked ears montage shows a more volume conducted alpha, where the average reference montage represents the more focal alpha activity. The linked ears montage is valid under the assumption that the average of the potentials recorded over two mastoids is close to zero or neutral. However, some argue that this is not the case (Hagemann, Naumann, & Thayer, 2001).

Alternatively, the average reference is the average electrical activity measured across all scalp channels. The average reference is useful to delineate focal activity. When using this reference, amplitudes will overall be reduced, but each channel will contribute equally to the new reference (Lei & Liao, 2017). Qin and colleagues reported that average reference is a better choice than linked-ears when applied to both stimulated and real resting-state EEG data (Qin, Xu, & Yao, 2010). Therefore, the fact that the association was only found for the average reference, which can be considered to be sensitive to more focal cortical activity, strengthens the main hypothesis that 10 Hz rTMS entrains endogenous EEG activity underneath the coil.

In conclusion, the main result of Corlier et al. (2019) was replicated, and the findings suggest that the distance between 10 Hz stimulation frequency and the IAF may influence clinical outcome, suggesting the most optimal rTMS frequency is the one identical to the frontal IAF. Further research should examine a broader range of stimulation frequencies to specifically examine the effect of the magnitude of difference between stimulation frequency and the IAF on clinical outcome, and additionally investigate what would be the optimal stimulation frequency for the 12.6% of patients that were classified as low voltage alpha EEG. Secondly, future studies should investigate changes in IAF over the course of treatment. If present, this would call for changes in rTMS frequency across treatment as well. Finally, in a future study it would be of interest to obtain two separate measures of eyes closed data with recordings separated by a few hours, to investigate the reliability of the IAF.

ACKNOWLEDGMENTS

This report forms the first communication of the 'International Consortium On Neuromodulation – Discovery of Biomarkers (ICON-DB)', which was established during the 3rd International Brain Stimulation Conference held in Vancouver in 2019. A group of EEG and TMS researchers decided to initiate this consortium in order to facilitate direct replication of EEG and TMS-EEG findings by facilitating immediate and independent cross-dataset replication in order to foster robustness of research findings and facilitate translation into clinical practice.

Requests for replication studies can be emailed to Andrew Leuchter (afl@ucla.edu) or Martijn Arns (martijn@brainclinics.com).

DISCLOSURES

MA is unpaid chairman of the Brainclinics Foundation, a minority shareholder in neuroCare Group (Munich, Germany), and a coinventor on 4 patent applications related to EEG, neuromodulation and psychophysiology, but receives no royalties related to these patents; Research Institute Brainclinics received research funding from Brain Resource (Sydney, Australia), Urgotech (France) and neuroCare Group (Munich, Germany), and equipment support from Deymed, neuroConn, Brainsway and Magventure.

FVR receives research support from Canadian Institutes of Health Research, Brian Canada, Michael Smith Foundation for Health Research, Vancouver Coastal Health Research Institute, and in-kind equipment support for investigator-initiated trial from MagVenture. He has participated in an advisory board for Janssen.

PBF is supported by a NHMRC Practitioner Fellowship (1078567). PBF has received equipment for research from MagVenture A/S, Medtronic Ltd, Neuronetics and Brainsway Ltd and funding for research from Neuronetics. He is on scientific advisory boards for Bionomics Ltd and LivaNova and is founder of TMS Clinics Australia.

AFL discloses that he has received research support from NIH, Neu-

ronetics, Breast Cancer Foundation, Department of Defense, CHDI Foundation, and Neurosigma. He has served as a consultant to Ionis Pharmaceuticals, CHDI Foundation, and NeoSync, Inc. He is Chief Scientific Officer of Brain Biomarker Analytics LLC (BBA). He has stock options in NeoSync, Inc. and equity interest in BBA.

DMB has received research support from the CIHR, NIH, Brain Canada and the Temerty Family Foundation through the CAMH Foundation and the Campbell Research Institute. He received research support and in-kind equipment support for an investigator-initiated study from Brainsway Ltd., and he is the principal site investigator for three sponsor-initiated studies for Brainsway Ltd. He received in-kind equipment support from Magventure for investigator-initiated research. He received medication supplies for an investigator-initiated trial from Indivior. He has participated in an advisory board for Janssen.

LLC discloses research support or in-kind equipment support from Neuronetics, Neosync, Nexstim, AffectNeuro, and Janssen. She has received consulting income for advisory board work from AffectNeuro, Janssen, Neurolief, Sage Therapeutics, Neuronetics, and Neuronix.

In the last 5 years, ZJD has received research and equipment in-kind support for an investigator-initiated study through Brainsway Inc and Magventure Inc. His work was supported by the Ontario Mental Health Foundation (OMHF), the Canadian Institutes of Health Research (CIHR), the National Institutes of Mental Health (NIMH) and the Temerty Family and Grant Family and through the Centre for Addiction and Mental Health (CAMH) Foundation and the Campbell Institute.

JC, NK, NWB, and CR have nothing to disclose.

SUPPLEMENTARY MATERIALS

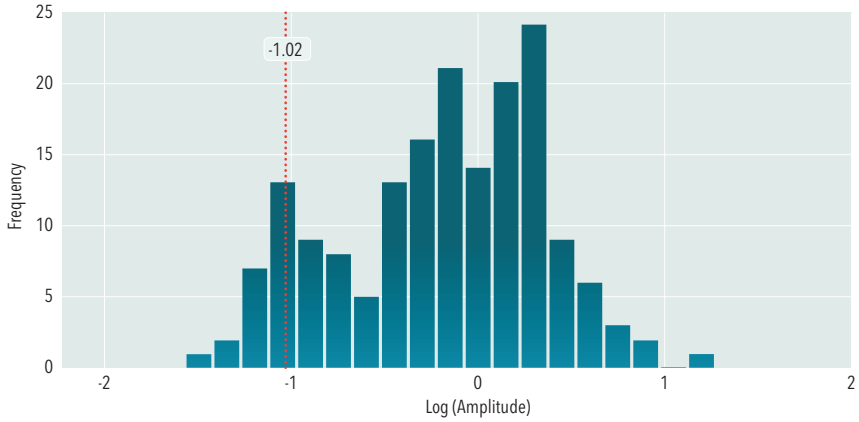


Figure S1: The frequency spectrum of the amplitude of the alpha peak for all subjects.

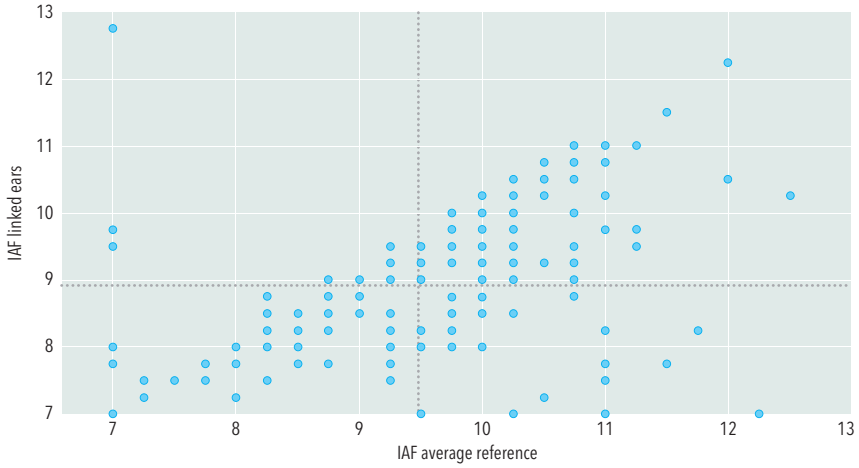


Figure S2: Individual alpha frequencies (IAF) for the different reference montages. The dashed lines represent the mean.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2020.10.017.

3.1

RESTING EEG THETA CONNECTIVITY AND ALPHA POWER TO PREDICT REPETITIVE TRANSCRANIAL MAGNETIC STIMULATION RESPONSE IN DEPRESSION: A NON-REPLICATION FROM THE ICON-DB CONSORTIUM

Published as: Bailey, N. W., Krepel, N., van Dijk, H., Leuchter, A. F., Vila-Rodriguez, F., Blumberger, D. M., Downar, J., Wilson, A., Daskalakis, Z. J., Carpenter, L. L., Corlier, J., Arns, M., Fitzgerald, P. B. (2021). Resting EEG theta connectivity and alpha power to predict repetitive transcranial magnetic stimulation response in depression: A non-replication from the ICON-DB consortium. *Clinical Neurophysiology, 132*(2), 650-659. doi: 10.1016/j.clinph.2020.10.018

ABSTRACT

Objective

Our previous research showed high predictive accuracy at differentiating responders from non-responders to repetitive transcranial magnetic stimulation (rTMS) for depression using resting electroencephalography (EEG) and clinical data from baseline and one-week following treatment onset using a machine learning algorithm. In particular, theta (4 – 8 Hz) connectivity and alpha power (8 – 13 Hz) significantly differed between responders and non-responders. Independent replication is a necessary step before the application of potential predictors in clinical practice. This study attempted to replicate the results in an independent dataset.

Methods

We submitted baseline resting EEG data from an independent sample of participants who underwent rTMS treatment for depression (N = 193, 128 responders) (Krepel et al., 2018; Chapter 2) to the same between group comparisons as our previous research (Bailey et al., 2019).

Results

Our previous results were not replicated, with no difference between responders and non-responders in theta connectivity (p = 0.250, Cohen's d = 0.1786) nor alpha power (p = 0.357, η_p^2 = 0.005).

Conclusions

These results suggest that baseline resting EEG theta connectivity or alpha power are unlikely to be generalizable predictors of response to rTMS treatment for depression.

Significance

These results highlight the importance of independent replication, data sharing and using large datasets in the prediction of response research.

INTRODUCTION

Recently we published a study demonstrating accurate prediction of response to repetitive transcranial magnetic stimulation (rTMS) treatment for depression using machine learning (84% sensitivity and 89% specificity) of a number of resting electroencephalography (EEG) measures in combination with measures of early change in mood (Bailey et al., 2019). Differences between the responder and non-responder groups in the EEG measures of theta connectivity and alpha power were consistent at both baseline and after one week of treatment, suggesting these measures reflected stable traits that were related to treatment outcome. However, the dataset was comprised of 42 participants, with only 12 responders. While cross-validation was used to ensure results were not due to over-fitting in a small sample, and permutation tests showed the machine learning results were significantly more accurate than chance, independent replication of previous results is necessary to ensure findings are valid and reliable. In particular, independent replication of the successful prediction of response to rTMS is required before the results could be generalized to the broader population of depressed patients undergoing rTMS treatment (Widge et al., 2019). Successful replication of treatment response prediction is of significant clini-

cal relevance, as rTMS results in distinct response or non-response outcomes, and rTMS treatments involve costly and time-consuming treatment regimens (Berlim, van den Eynde, Tovar-Perdomo, & Daskalakis, 2014; Fitzgerald et al., 2016; George & Post, 2011). Additionally, conducting a replication study also enables the testing of other possibly relevant variables that might influence the results. For example, previous results from Arns et al. (2016) indicated that frontal alpha asymmetry (FAA) was associated with response to selective serotonin reuptake inhibitors (SSRIs) in females only. The sample size of our original study was too small to enable interactions with sex to be tested (Bailey et al., 2019), but the results from Arns et al. (2016) demonstrate the importance of determining if interactions between response prediction variables and sex are present in order to enable maximum predictive accuracy.

To enable independent replications (as we aimed to perform) a large dataset (N = 193, with 128 responders) of baseline resting EEG data from an open-label trial of rTMS treatment of depression across two separate clinics was recently made available via a data sharing proposal (Krepel et al., 2018; Chapter 2). Although minor differences between our original study and this replication dataset were present in data collection and processing (different depression severity assessment tools were used, different electrode montages, recording equipment and settings, absence of week 1 recordings, and different EEG pre-processing procedures) predictive variables should be robust to minor parameter variation to be clinically useful. We therefore deemed the data similar enough to enable an independent replication of the previous results.

We hypothesized that responders in the replication dataset would show higher theta connectivity from within the same group of electrode pairs that differentiated responders from non-responders in our original research (a broad group of electrode pairs involving frontal, parietal and occipital connections). Additionally, following research showing that predictors of response can be sex specific (Arns et al., 2016), we had a non-directional hypothesis that the difference between responders and non-responders in theta connectivity would be influenced by sex. Following the results of our original research,

we also hypothesized that responders would show less alpha power in frontal and occipital electrodes than non-responders, and responders would show a smaller difference in alpha power between frontal and occipital regions than non-responders. If these measures showed replication of the results from our original dataset, we hypothesized that a machine learning algorithm would show accurate response prediction from this baseline data, with similar specificity and sensitivity to our original dataset.

METHODS

Participants

Participants with EEG recordings included 193 participants (95 male) with major depression aged 18 – 78 (Mean = 43.2, *SD* = 12.9, which can be compared to the original dataset, with a Mean = 45.86, *SD* = 13.95) treated with simultaneous psychotherapy and rTMS (Mean = 20.9 sessions, *SD* = 7.5). Participants were treated with either high frequency (10 Hz) left dorsolateral prefrontal cortex (DLPFC), low frequency (1 Hz) right DLPFC, or both sequentially (similar to our original research). Over 97% of the sample had at least one previous antidepressant treatment without response (in contrast to the original dataset, which only included participants who had tried at least two separate antidepressant treatments from different classes of antidepressants without response). Participants were separated into responders (n = 128) and non-responders (n = 65) defined by ≥ 50% reduction in Beck Depression Inventory II Dutch Language Version (BDI-II-NL) score between baseline measurement and the final visit. Data from these participants has been previously reported, and further details of participant and treatment characteristics can be found in Donse et al. (2018). Power calculation using the effect size for differences in connectivity from Bailey et al. (2019) (d = 1.097) suggested 52 participants were necessary to obtain 0.95 power with an alpha of 0.05, and post-hoc power analysis showed the number of participants used provided > 0.999 power. This demonstrated that the number of participants in the current study provided a more than large enough sample size to detect significant effects.

Electrophysiological recording and pre-processing
Two minutes of baseline resting EEG recordings with both eyes open (EO) and eyes closed (EC) were obtained using a 26 sintered Ag/AgCL electrode Quikcap (Neuroscan) and NuAmps amplifier (Compumedics, Neuroscan). Data was referenced online to averaged mastoids with a ground at FPz, impedances of < 5 kΩ were maintained, and EEG activity was sampled at 500 Hz with a DC high pass and 100 Hz low pass filter. Horizontal eye movements were recorded by electrodes placed 1.5 cm lateral to the outer canthus of each eye, and vertical eye movements were recorded with electrodes placed 3 mm above the middle of the left eyebrow and 1.5 cm below the middle of the left bottom eyelid.

Offline data was processed using a standardized methodology (Arns et al., 2016; Arns, Drinkenburg, Fitzgerald, et al., 2012). Data was bandpass filtered from 0.3 to 100 Hz with notch filters of 50 or 60 Hz (depending on which country data was recorded in) using zero phase Infinite Impulse Response filters. Eye movements were corrected using the Gratton and Coles method (Gratton et al., 1983). Data was epoched into two second windows across the recording period. Individual epochs per channel were automatically marked as artifact and rejected based on the following criteria: 1) a ratio of > 0.375 for 30 – 90 Hz gamma power relative to the rest of the signal ratio, 2) the presence of shifts in the voltage slope from 16 consecutive samples that exceeded 25x the epoch average, 3) kurtosis values > 8, 4) extreme frequency power in the epoch, with power values > 350 from the summation of power in the 1 – 5.25 Hz range and the 22 – 45 Hz range after a Fast Fourier Transform of each epoch, scaled for electrode location by a linear increase in the threshold from 350 at the most anterior electrodes to 525 at the most posterior electrodes (as power is usually higher at posterior electrodes), 5) the presence of residual eye blink detection based on cross correlation values of > 0.55 between eye electrodes and EEG electrodes, and 6) extreme voltage swing detection of > 200 mv across the epoch. See Arns et al. (2016) for more details. Data was re-referenced to the average reference prior to analysis. All participants provided 35 or more noise free epochs for analysis from both eyes open and eyes closed conditions (mean = 56.7, *SD* = 1.8).

This can be compared with Bailey et al. (2019), who obtained three

minutes of eyes open and eyes closed resting data for each participant using a 30 sintered Ag/AgCL electrode EasyCap (Easy-Cap, Woerthsee-Etterschlag, Germany) and recording EEG activity using a Synamps 2 amplifier (Compumedics, Neuroscan). CPz was used as the online reference and AFz as the ground, impedances of < 5 kΩ were maintained, and EEG activity was sampled at 10,000 Hz with a DC high pass and 2000 Hz low pass filter. Horizontal eye movements were recorded by electrodes placed 1.5 cm lateral to the outer canthus of each eye, and vertical eye movements were recorded with electrodes place 3 mm above the middle of the left eyebrow and 1.5 cm below the middle of the left bottom eyelid.

Data was processed using MATLAB (The Mathworks, Natick, MA) and EEGLAB (sccn.ucsd.edu/eeglab) (Delorme & Makeig, 2004). Data were initially downsampled to 1000 Hz, then second order butterworth filtering was applied with a bandpass from 1 to 80 Hz and a band stop filter from 47 to 53 Hz. Data was epoched into two second windows across the recording period with a 500 ms overlap. Single electrodes containing artifacts in more than 3% of epochs were rejected (indicated by variations in voltage that were larger than 250 mv, kurtosis values > 5, or values exceeding 100 or 30 dB in the 25 – 45 Hz range). Epochs containing artifacts were also rejected (indicated by kurtosis values > 3 for all electrodes or > 5 for single electrodes, and more than 100 to 30 dB in the 25 – 45 Hz range). Artifact rejections were manually verified by visual inspection by an experienced EEG researcher (NWB), then Fast independent component analysis (FastICA) (Hyvärinen, 1999; Hyvärinen & Oja, 2000) was used to manually select and remove eye blinks and movements and remaining muscle activity artifacts. The 'symmetric approach' and the 'tanh' contrast function were used for the algorithm. Recordings were re-referenced offline to an averaged reference. All participants provided 63 or more noise free epochs for analysis from both eyes open and eyes closed conditions.

To summarize the differences in the EEG recording and processing of the two datasets: 1) The replication dataset used a Quickcap and NuAmps amplifier while the original dataset used an EasyCap and Neuroscan Synamps 2 amplifier (as far as we are aware, these amplifiers are highly compatible). 2) The data were recorded at a different sam-

pling rate (the replication dataset recorded at 500 Hz while the original dataset recorded at 10,000 Hz and down-sampled to 1000 Hz for analysis). 3) The data were processed using different methods. In the replication dataset (Krepel et al., 2018; Chapter 2), epochs were rejected through a six step automated process that excluded epochs showing excessive power, kurtosis and voltage shift values, and eye blinks and movements were corrected using the Gratton and Coles method (Gratton et al., 1983), which corrects for this activity in the EEG trace without rejecting epochs contaminated by eye blink and movement. In the original study (Bailey et al., 2019), epoch rejection was based on kurtosis values of > 5 in single electrodes or > 3 for all electrodes, or power exceeding -100 to 30 dB in the 25 – 45 Hz range, and FastICA (Hyvärinen, 1999; Hyvärinen & Oja, 2000) was used to manually select and remove components containing eye blinks, movements, and remaining muscle activity artifacts, which corrects for this activity in the EEG trace without rejecting epochs contaminated by eye blink and movement. 4) After extracting the significant group of electrode pairs from our previous research and applying that group of electrode pairs to the replication dataset (Krepel et al., 2018; Chapter 2), we were left with only 14 electrode pairs in the proposed group of electrode pairs, compared to 66 electrode pairs in the network that differentiated responders and non-responders in the original research. However, when we re-analyzed the original dataset using theta connectivity values from only the restricted montage and compared this group of electrode pairs between responders and non-responders, differences were still highly significant in that original dataset. This suggests that the restricted montage does not explain the lack of differences in the replication dataset.

Alpha power and theta connectivity computation

In order to determine whether our previous results replicated in this independent dataset, alpha power and theta connectivity values were computed by an independent team (NK and HvD) in the same manner as our previous research (Bailey et al., 2019). For the connectivity computation, EEG data was submitted to a single Hanning taper time-frequency transform, determining instantaneous phase values for the complex Fourier-spectra from 1 to 45 Hz with a 0.5 Hz resolution across sliding time windows corresponding to 4 cycles in length. These

values were slightly higher than the original study, which used 1 Hz resolution across sliding time windows corresponding to 3 oscillation cycles in length. As such, the replication connectivity measures were assumed to be more robust than the original measures. The weighted phase lagged index (wPLI) was then calculated between each electrode to measure phase synchronization between electrodes (Vinck, Oostenveld, van Wingerden, Battaglia, & Pennartz, 2011). Following this, wPLI values in the theta frequency (4 – 8 Hz) were averaged across these frequencies, and across epochs in preparation for statistical analysis. Total average theta wPLI was also computed across all available electrode pairs in the group of electrode pairs that differentiated responders and non-responders in our original study. Not all electrodes from the original significant group of electrode pairs were present in the replication dataset, so the original dataset was tested on this reduced group of electrode pairs to confirm those results were not altered by reducing the number of electrode pairs included in the analysis (reported below). Electrode pairs that were both significantly different between responders and non-responders in the original study, and present in the replication dataset were Fz-FC4, FC3-FC4, F3-P3, FC3-P3, P3-P4, F3-O1, FC3-O1, P3-O1, P4-O1, F3-O2, FC3-O2, P3-O2, P4-O2, and O1-O2.

Alpha power was computed using a multi-taper fast Fourier frequency transformation with a Hanning taper to calculate power in the alpha range (8 – 13 Hz). Alpha power was calculated across each epoch, then averaged across the frequency window, across all epochs, and across both eyes open and eyes closed recordings, in exact replication of the procedure from Bailey et al. (2019). As per our previous research, F3, F4, O1 and O2 electrodes were selected for analysis.

Statistical analysis

Traditional frequentist statistical comparisons were conducted using SPSS version 23. Bayesian comparisons were conducted using JASP version 0.11.1 (Love et al., 2019) to provide an indicator of the strength of evidence for null results. Where our previous results suggested directional finding that we would expect for the results in the current study, one-tailed Bayesian comparisons were used, as a result in the opposite direction would provide the same rejection of our previous conclusion as no difference between groups (Ruxton & Neuhäuser,

2010). For analyses involving more than a single factor, comparisons were made between Bayesian models containing a hypothesized effect to equivalent models stripped of the effect. Comparisons of connectivity values across all electrodes were performed using the network based statistics (NBS) (Zalesky, Fornito, & Bullmore, 2010). In order to confirm the comparisons of theta connectivity made using the reduced electrode montage available in the replication data was still a valid test of our initial result, we conducted an independent samples t-test of averaged wPLI values from electrode pairs that were both within the group of electrode pairs that significantly differentiated responders and non-responders shown in our initial study, and present in the electrode montage from the replication study. Next, to test our primary hypothesis of replication of increased theta connectivity in responders within the same group of electrode pairs as the original research, we performed an independent samples t-test comparing responders and non-responders in averaged theta wPLI values across the group of electrode pairs including electrode pairs common to both studies and averaged across EO and EC conditions. In order to test whether the original result might also be specific to a stimulation type, we performed a sub-analysis with the same t-test but restricted to only participants who underwent 10 Hz left side treatment. To test our hypothesis that sex would influence these results, we also performed a repeated measures ANOVA on averaged theta wPLI values from this group of electrode pairs using group (responder and non-responder) and sex (females and males) as between subject factors and condition (EO and EC) as the within-subject factor, with age as a covariate. Thirdly, in order to assess connectivity across all electrodes (in case a different group of electrode pairs separated responders and non-responders in the replication sample) we submitted the replication dataset to a t-test comparison of responders and non-responders using the NBS cluster analysis of connectivity values across all pairs of electrodes available in the replication dataset (Zalesky et al., 2010). Finally, the last comparison of theta connectivity administered this same NBS test separately for each sex. Note that Bayesian statistics are not currently able to replicate the analyses performed by the NBS, so we were unable to test for the strength of our conclusion with regards to the analysis including all pairs of electrodes. In order to assess alpha power differences between responders and non-responders, we

conducted a repeated measures ANOVA including group (responder and non-responder) and sex (females and males) as between-subject factors, with hemisphere (right and left), and region (frontal [F3, F4] and occipital [O1, O2]) as within-subject factors, and age as a covariate. We also conducted this analysis restricted to participants who underwent 10 Hz left sided treatment. As reported below, our results were non-significant, so no machine learning algorithm was applied.

RESULTS

Clinical results from the dataset have been reported previously (Donse et al., 2018). When the averaged wPLI values from the original study were restricted to just the electrodes that overlapped between the two labs, comparisons between responders and non-responders were still significant $t(40) = 2.824$, $p = 0.015$, Cohen's $d = 1.0968$ (responder mean = 0.0901, $SD = 0.0667$, non-responder mean = 0.0338, $SD = 0.0286$). However, in the replication dataset, no significant difference was found in averaged connectivity from within the same group of electrode pairs as the original study between responders ($M = 0.02279$, $SD = 0.02240$) and non-responders ($M = 0.02825$, $SD = 0.02083$), $t(191) = 1.638$, $p = 0.103$, Cohen's $d = 0.25241$, BF0- = 15.132 (see Figure 1 on page 72). Additionally, even though the result was not significant and showed a small effect size, the effect was in the opposite direction to the original study. The sub-analysis focusing on only participants who underwent 10 Hz left side treatment also showed no differences between responders ($M = 0.02123$, $SD = 0.01790$) and non-responders ($M = 0.02577$, $SD = 0.01753$), $t(71) = 1.013$, $p = 0.314$, Cohen's $d = 0.255$, BF0- = 7.101. Furthermore, there was neither an interaction between response-group and sex, nor response- group, sex and eyes open or closed (all $p > 0.10$ and BFexcl > 3, see Table 1 on page 72 for detailed statistics). An interaction was observed between age and EO or EC, $F(1,189) = 5.141$, $p = 0.025$, $\eta_p^2 = 0.027$, such that age positively correlated with EC connectivity $r(193) = 0.237$, $p = 0.001$ but not EO connectivity $r(193) = 0.082$, $p = 0.255$. Using NBS to compare across all pairs of electrodes in the replication dataset revealed no differences between responders and non-responders, nor differences when data was split by sex (all $p > 0.05$), similar to the analyses performed in SPSS.

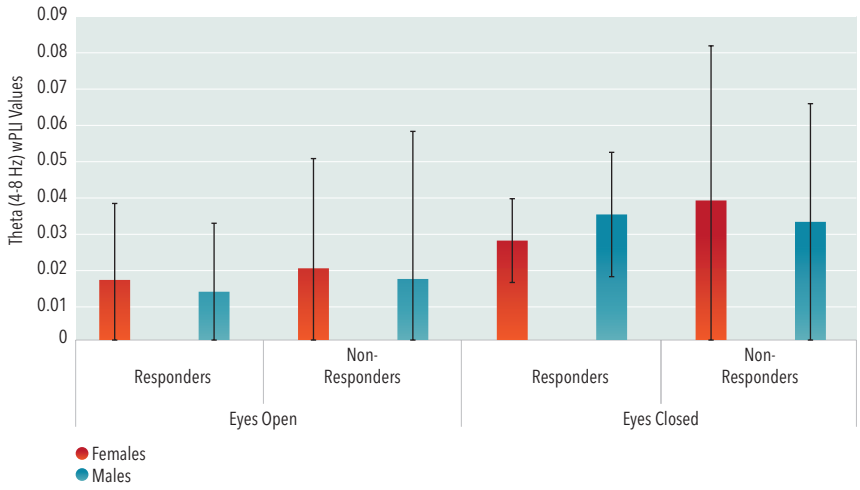


Figure 1: *Mean theta (4 – 8 Hz) weighted phase lag index (wPLI) connectivity values from responder and non-responder groups in the replication dataset (error bars reflect standard deviations). No significant differences were detected between responders and non-responders, in contrast to the original dataset and our hypotheses.*

Measurement	Responders		Non-Responders		Statistics
	Mean *(SD)* Females ($n = 69$)	Mean *(SD)* Males ($n = 59$)	Mean *(SD)* Females ($n = 29$)	Mean *(SD)* Males ($n = 36$)	Between group comparison: $t(191) = 1.638$, $p = 0.103$, Cohen's $d = 0.2524$, BF0- = 15.132
EO	.0165 (.0211)	.0132 (.0115)	.0197 (.0189)	.0168 (.0170)	
EC	.0273 (.0302)	.0345 (.0425)	.0384 (.0406)	.0384 (.0325)	Interaction between group and sex: $F(1,188) = 2.001$, $p = 0.159$, $\eta_p^2 < 0.011$, BFexcl = 4.062 Interaction between group, sex and eyes: $F(1,188) = 1.061$, $p = 0.304$, $\eta_p^2 < 0.006$, BFexcl = 3.588

EO Eyes Open
EC Eyes Closed
SD Standard Deviation

Table 1: *Mean values for resting wPLI theta connectivity, standard deviations and statistical comparisons between responders and non-responders in the replication data (values averaged across the group of electrode pairs that differentiated responders and non-responders in the original research, excluding electrode pairs that did not overlap between the two studies).*

Alpha power comparisons also showed no differences between responders and non-responders $F(1,189) = 0.851$, $p = 0.357$, $\eta_p^2 = 0.005$, BFexcl = 4.086 (see Figure 2 below, details in Table 2 on page 74). There was also no interaction between response-group and electrode region $F(1,189) = 0.578$, $p = 0.448$, $\eta_p^2 = 0.003$, BFexcl = 6.880. The main response-group effect was not influenced by sex $F(1,189) = 0.303$, $p = 0.582$, $\eta_p^2 = 0.002$, BFexcl = 3.529, nor was the interaction between response-group, sex, and electrode region ($F(1,189) = 0.037$, $p = 0.848$, $\eta_p^2 < 0.001$, BFexcl = 3.580). Lastly, there was no interaction between hemisphere, region, sex and response-group $F(1,189) = 0.008$, $p = 0.927$, $\eta_p^2 < 0.001$, BFexcl = 4.312. When performing the same comparisons restricted to participants who underwent 10 Hz left-sided rTMS, we likewise observed no differences (see Figure 3 on page 74, all $p < 0.2$, BFexcl > 2, details in Table 3 on page 75). All data met the assumption of equal variances (all $p > 0.2$).

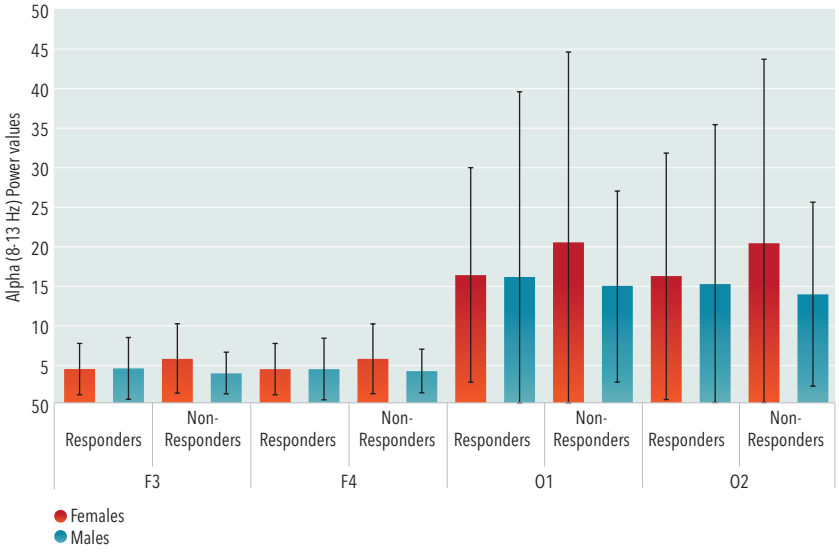


Figure 2: Mean alpha power values from responder and non-responder groups in the replication dataset (error bars reflect standard deviations). No significant differences were detected between responders and non-responders, nor interaction between region and response-group, in contrast to the original dataset and our hypotheses.

Electrode	Responders		Non-Responders		Statistics
	Mean *(SD)* Females ($n = 69$)	Mean *(SD)* Males ($n = 59$)	Mean *(SD)* Females ($n = 29$)	Mean *(SD)* Males ($n = 36$)	Between group comparison: $F(1,189) = 0.851$, $p = 0.357$, $\eta_p^2 = 0.005$, BFexcl = 4.086
F3	4.31 (3.30)	4.38 (3.90)	5.66 (4.40)	3.76 (2.65)	
F4	4.33 (3.27)	4.35 (3.90)	5.65 (4.42)	4.01 (2.78)	Interaction between group and region: $F(1,189) = 0.578$, $p = 0.448$, $\eta_p^2 = 0.003$, BFexcl = 6.880
O1	16.27 (13.66)	16.01 (23.62)	20.46 (24.13)	14.75 (12.15)	
O2	16.12 (15.69)	15.14 (20.24)	20.32 (23.35)	13.82 (11.68)	Interaction between group and sex: $F(1,189) = 0.303$, $p = 0.582$, $\eta_p^2 = 0.002$, BFexcl = 3.529
					Interaction between group, sex and region: $F(1,189) = 0.037$, $p = 0.848$, $\eta_p^2 < 0.001$, BFexcl = 3.580
					Interaction between group, sex, region and hemisphere: $F(1,189) = 0.008$, $p = 0.927$, $\eta_p^2 < 0.001$, BFexcl = 4.312

SD Standard Deviation

Table 2: Resting alpha power means, standard deviations and statistical comparisons between responders and non-responders in the replication data.

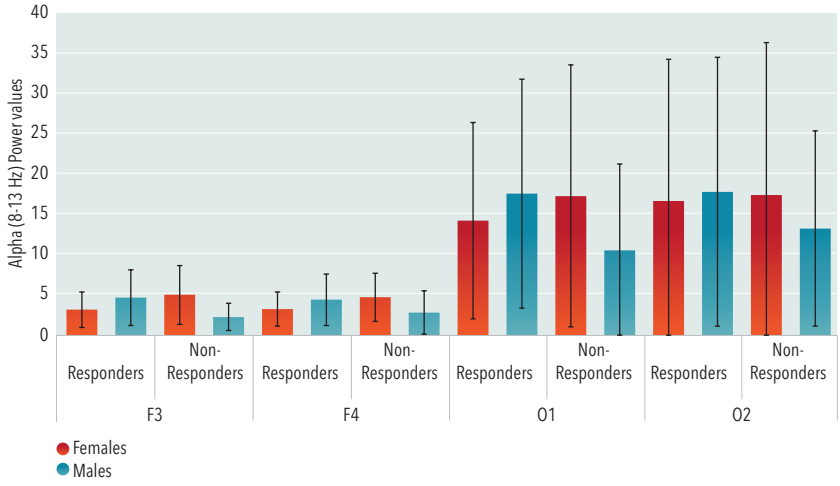


Figure 3: Mean alpha power values from responder and non-responder groups in the replication dataset for participants who underwent 10 Hz left hemisphere repetitive transcranial magnetic stimulation (rTMS) treatment only (error bars reflect standard deviations). As with the full dataset, no significant differences were detected between responders and non-responders.

Electrode	Responders		Non-Responders		Statistics
	Mean *(SD)* Females ($n = 25$)	Mean *(SD)* Males ($n = 25$)	Mean *(SD)* Females ($n = 11$)	Mean *(SD)* Males ($n = 12$)	Between group comparison: $F(1,68) = 0.026$, $p = 0.872$, $\eta_p^2 < 0.001$, BFexcl = 3.647
F3	3.23 (2.22)	4.73 (3.47)	5.07 (3.63)	2.34 (1.66)	
F4	3.29 (2.11)	4.48 (3.20)	4.76 (3.00)	2.89 (2.69)	Interaction between group and region: $F(1,68) = 0.048$, $p = 0.827$, $\eta_p^2 = 0.001$, BFexcl = 4.220
O1	14.26 (12.20)	17.58 (14.19)	17.32 (16.24)	12.49 (10.63)	
O2	16.73 (17.50)	17.80 (16.65)	17.49 (18.82)	13.31 (12.11)	Interaction between group and sex: $F(1,68) = 0.675$, $p = 0.414$, $\eta_p^2 = 0.010$, BFexcl = 2.090
					Interaction between group, sex and region: $F(1,68) = 0.017$, $p = 0.896$, $\eta_p^2 < 0.001$, BFexcl = 2.637
					Interaction between group, sex, region and hemisphere: $F(1,68) = 0.307$, $p = 0.581$, $\eta_p^2 < 0.004$, BFexcl = 3.245

SD Standard Deviation

Table 3: *Resting alpha power means, standard deviations and statistical comparisons between responders and non-responders to 10 Hz left sided treatment only in the replication data. SD = standard deviation.*

DISCUSSION

The aim of this study was to determine whether our previous research demonstrating that responders to rTMS treatment for depression showed higher resting EEG theta connectivity and lower alpha power than non-responders (Bailey et al., 2019) would replicate in a larger independent sample (Krepel et al., 2018; Chapter 2), indicating clinical relevance and applicability of these measures. The results of this study did not replicate our previous research as we did not observe similar differences between the responders and non-responders. Furthermore, although the selected measures did not differ between response-groups, the pattern for theta connectivity was reversed compared to our original results (with non-responders showing higher values), strongly suggesting our original finding does not generalize. Additionally, we aimed to assess whether our previous results would be modified by including sex in the analyses, as previous research has indicated that predictors of response may be modulated by sex (Arns et al., 2016). As with the main comparisons, no interaction between sex and theta connectivity or alpha power were found. These results were also consistent when analyses were restricted to participants receiving 10 Hz left sided treatment only. The results suggest that our previous findings do not generalize to independent samples, and as such the particular resting theta connectivity and resting alpha power measures examined in this study are unlikely to be clinically useful biomarkers for response to rTMS treatment for depression.

The non-replication was unexpected. In the original dataset (Bailey et al., 2019), the theta connectivity differences between responders and non-responders were present in comparisons across both baseline and week 1 time-points, suggesting a robust effect. Machine learning predictions including theta connectivity and alpha power were also accurate across learning and test samples. This consistency, comprised of test-retest replication across time and within sample replication across divisions of the same sample, suggested that differences were likely to reflect genuine findings. However, the results do not replicate, which prompts the question of why our results seemed to show consistency within the study, but not in data obtained and processed by other researchers.

There were a number of differences between the two datasets in response definition, treatment resistance definition, EEG recording, and EEG pre-processing steps (summarized in Table 4 on page 79). Perhaps most importantly, the specific electrode locations used differed between the two studies. However, when we reanalyzed the original dataset using theta connectivity values from only the restricted montage that overlapped with the replication dataset and compared this group of electrode pairs between responders and non-responders, differences were still highly significant in that original dataset. This suggests that the restricted montage does not explain the lack of differences in the replication dataset. Secondly, the two datasets were recorded using different (but highly similar) amplifiers, sampling rates, and were pre-processed using different artifact rejection/correction procedures. Thirdly, the inclusion criteria involving treatment resistance differed between the two studies. The original dataset (Bailey et al., 2019) used an inclusion criterion of at least two failed antidepressant treatments from different classes of antidepressants, while the replication dataset (Krepel et al., 2018; Chapter 2) did not have a formal inclusion criterion around treatment resistance (although 97% of participants had at least one failed antidepressant treatment). This point suggests that the replication dataset is likely to consist of a more heterogenous sample with a broader range of treatment resistance. If the measures that predicted response in the original dataset are influenced by the severity of treatment resistance, this difference between the datasets could offer an explanation for the inconsistency between the two studies. However, we think it is unlikely that the inclusion of less treatment resistant participants would have reversed the pattern of theta connectivity, as this would suggest that individuals who had only tried one unsuccessful antidepressant would show the opposite relationship between theta connectivity at baseline and treatment response compared to individuals who had tried two or more antidepressants.

Additionally, while participants in the original dataset (Bailey et al., 2019) were mostly taking antidepressants (and also mood stabilizers or antidepressants, or no medications in a small number of cases), the replication dataset (Krepel et al., 2018; Chapter 2) was a naturalistic sample so did not obtain verified data on medication use. However, recent research examining frontal alpha asymmetry has shown the

measure could accurately predict response both prior to and after SSRI treatment, suggesting that successful prediction of treatment response with EEG measures is likely to be robust to differences in medication status (van der Vinne, Vollebregt, van Putten, & Arns, 2019). Lastly, the two samples measured depression severity (and as such defined response to treatment) using different scales. The original research (Bailey et al., 2019) used the Hamilton Depression Rating Scale (HDRS), while the replication dataset (Krepel et al., 2018; Chapter 2) used the BDI-II-NL (with both studies defining response as a 50% reduction from baseline scores to endpoint). However, there is no indication that the different measures have a different relationship to the EEG measures, and the two depression severity measures have been shown to be highly correlated (Fitzgibbon, Cella, & Sweeney, 1988). As such, none of the differences in response definition, EEG recording, or EEG pre-processing steps suggest to us an obvious confound that would have led to systematic differences between responders and non-responders in theta connectivity or alpha power. Additionally, for our findings to be generalizable and clinically useful, they should be robust against minor variations in data collection or pre-processing, and generalizable to patients across a broad range of inclusion criteria.

In addition to the differences in data measurement between the two datasets, the participants in the replication dataset (Krepel et al., 2018; Chapter 2) underwent cognitive behavior psychotherapy concurrently with rTMS treatment, while participants in the original dataset (Bailey et al., 2019) did not. It is likely that psychotherapy treats depression through a mechanism that is different to rTMS, and it is possible that the mechanism underpinning psychotherapy is unrelated to theta connectivity or alpha power at baseline (or even related to theta connectivity/alpha power at baseline, but in the opposite direction to the direction shown with rTMS in our original study). While we were unable to find research directly addressing the relationship between theta connectivity or alpha power and response to psychotherapy, a review drawing evidence from parallel studies of psychotherapy and rTMS has suggested that the two therapies target different mechanisms within emotional processing networks (Thase, 2014), with intact executive function being suggested to predict response to psychotherapy (Harmer, 2014). Additionally, low levels

Parameter	Bailey et al. (2019)	Krepel et al. (2018)	Same/Different?	Explains Results?
Electrode montage	30 electrodes	26 electrodes	F3, Fz, F4, FC4, FCz, FC3, P3, Pz, P4, O1, O2 were present in both studies, other electrodes differed	No - limiting comparisons of original dataset to only include overlapping electrodes did not alter results
EEG Amplifier	Neuroscan Synamps 2	Neuroscan NuAmps	Same chipset	Very unlikely to explain results
EEG Sampling Rate	10,000 Hz downsampled to 1,000 Hz	500 Hz	Different	No reason to believe this would explain results
EEG Artifact Rejection	Epoch rejection via kurtosis values or excessive power in the 25-45 Hz range, FastICA to reject eye movements and other remaining artifacts	Epochs rejected following a 6-step automated process rejecting high kurtosis and power values, eye movements corrected using the Gratton and Coles method	Different, but both methods remove artifacts and use methods to correct for eye blinks and movements (rather than simply delete these artifacts)	No reason to believe this would explain results
Depression Severity	HDRS	BDI-II-NL	Different, but highly correlated	No reason to believe this would explain results
Treatment Resistance	Failure to respond to at least two antidepressants from two separate classes	97% of the sample showed failure to respond to at least one antidepressant	Different	Possible – but if it explains results then predictive potential of the original study is limited to a specific population
rTMS treatment	3 weeks of 10 Hz left DLPFC rTMS, then randomized to continue for 3 weeks, or to 1 Hz right DLPFC, or bilateral rTMS. 110% of RMT	10 Hz left DLPFC, 1 Hz right DLPFC, or bilateral, >10 sessions for inclusion, 110 – 120% of RMT	The minimum number of rTMS treatments for inclusion from Bailey et al. (2018) was higher, otherwise parameters were highly similar	No reason to believe this would explain results
Concurrent therapy	Antidepressants, and a minority of participants taking additional mood stabilizers, antipsychotics, or no medications	Cognitive Behavior Psychotherapy, medications not recorded	Different	Unclear
Sample size	42 (12 responders)	193 (128 responders)	Different	Possible – replication dataset is a more generalized sample, original dataset may contain sampling bias
Age	Mean 45.86, *SD* = 13.95	Mean 43.2, *SD* = 12.9	Very similar	Unlikely to explain results

DLFPC	Dorsolateral Pre-Frontal Cortex	BDI-II-NL	Beck Depression Inventory II - Dutch version
rTMS	repetitive Transcranial Magnetic Stimulation	*SD*	Standard Deviation
HDRS	Hamilton Depression Rating Scale		

Table 4: Summary of parameters that were the same or differed between the datasets.

of rostral and subgenual cingulate activity has been proposed by a theoretical perspective to be a predictor for psychotherapy response (DeRubeis, Siegle, & Hollon, 2008), whereas high levels of anterior cingulate activity and high resting fMRI anti-correlations between subgenual cingulate activity and the left prefrontal cortex have been shown to predict response to rTMS (Baeken et al., 2014; Fox, Buckner, White, Greicius, & Pascual-Leone, 2012 ; Langguth et al., 2007). If psychotherapy and rTMS do act via different mechanisms that are relevant to theta connectivity and alpha power, the effect of psychotherapy in the replication dataset may have diluted the statistical signal in the predictive relationship between theta connectivity, alpha power, and response, offering a potential explanation for the non-replication. However, there is no direct evidence to support the proposition that the presence of psychotherapy eliminated the statistical signal for theta connectivity and alpha power in the replication dataset. Given the fact that the pattern for theta connectivity was reversed in the replication dataset, we think the most parsimonious explanation is simply that the pattern from the original dataset (Bailey et al., 2019) was specific to that sample, but does not generalize to the broader population of individuals undergoing rTMS treatment for depression. Despite this conclusion based on parsimony, it may be valuable for future research to examine theta connectivity and alpha power measures as predictors of rTMS treatment without concurrent cognitive behavior psychotherapy.

A final difference between the two datasets is unrelated to the current results, but may be relevant for future research to consider. While the replication dataset only contained resting EEG, the original dataset additionally included working memory related EEG (which was reported in Bailey et al. (2018)). In addition to the higher resting theta connectivity, responders in the original dataset also showed higher working memory related theta connectivity (Bailey et al., 2018). While the replication sample demonstrated that resting theta connectivity differences between responders and non-responders did not generalize to an independent sample, they do not demonstrate the same for working memory related theta connectivity. However, given the responders in the original dataset showed higher theta connectivity across both baseline and week 1 measures that were consistent across the resting

EEG and the working memory EEG, we think it is likely the higher theta connectivity reflects a phenotype in those participants that is common across both resting and cognitive processes (note that the resting EEG recordings were performed prior to the working memory EEG, so our resting results were not influenced by a delayed cognition related increase in theta connectivity). As such, if our assumption that the higher theta connectivity was specific to that sub-set of responders is accurate, then we think it would be unlikely that higher theta connectivity related to working memory could predict responders in independent samples.

As such, because there are no measurement or treatment differences between the two datasets that offer obvious explanations for the non-replication, our interpretation of the two studies is that the 12 responders in our original research showed trait higher connectivity values on average merely by chance, and this may have been due to an unspecified sampling bias which was not apparent in the much larger replication dataset that included 128 responders. The higher theta connectivity as a trait in these individuals explains the consistency across both baseline and week 1 recordings in the original research (Bailey et al., 2019) as well as the working memory EEG measures (Bailey et al., 2018), but a potential unspecified sampling bias in the original dataset would explain the non-replication when examining the replication dataset (Krepel et al., 2018; Chapter 2). The proposition that depression is likely to be comprised of multiple underlying phenotypes (Insel & Wang, 2010; Widge et al., 2017) is one possible explanation. It may be that some of these phenotypes respond to rTMS while others do not, in which case knowing a patient's phenotype may lead to response prediction (Drysdale et al. (2017), however, also see Dinga et al. (2019)). As such, it may be that smaller sample sizes that may be less representative of the broad population contain more of certain depression phenotypes, leading to apparently high prediction accuracy which does not replicate when using a more representative sample containing the full spectrum of phenotypes. This suggests that research examining response prediction may be more complicated than simply finding a biomarker that can be used across all patients (Insel & Wang, 2010; Widge et al., 2017).

Although the non-replication means the particular theta connectivity and alpha power measures in this study may not have clinical utility, the result is valuable, as it narrows the search space for potential predictors of rTMS response by process of elimination. Non-replication studies are particularly important to publish, as the robustness and reliability of prediction studies is questionable, and publication bias has been demonstrated in the prediction of depression treatment response literature (Widge et al., 2019). Rigorous methodology and reporting as well as replication attempts have been proposed as the solution to this issue (Widge et al., 2019). Additionally, while resting theta connectivity and resting alpha power appear not to be generalizable predictors of response to rTMS, a number of measures have been replicated both by our original research and other labs. In particular, proximity of the alpha frequency to the 10 Hz rTMS stimulation frequency, as reported by Corlier et al. (2019) was successfully replicated by this ICON-DB consortium (Roelofs et al., 2021; Chapter 3) and early change in mood showed the largest effect size for differences between responders and non-responders of the measures in our original research (Bailey et al., 2019) and in other research (Donse et al., 2018). Early change in cognitive performance also seems to be a replicable predictor, particularly early change in working memory performance (Bailey et al., 2018; Hoy, Segrave, Daskalakis, & Fitzgerald, 2012). Also, fronto-midline theta during a working memory task was shown to differentiate responders and non-responders in the original sample (Bailey et al., 2019) even though fronto-midline theta during resting EEG did not differentiate the two groups. Fronto-midline theta has been suggested to be related to attention and cognitive control, to be generated by the anterior cingulate cortex, and to be negatively correlated with default mode network activity, all of which are implicated in depression (Nigbur, Ivanova, & Sturmer, 2011; Onton, Delorme, & Makeig, 2005; Pizzagalli et al., 2001; Sauseng, Hoppe, Klimesch, Gerloff, & Hummel, 2007; Scheeringa et al., 2008). Task related EEG has the added benefit of showing good test-retest reliability, increasing its potential utility as a predictor (Tenke et al., 2017). As such, we suggest that cognition related fronto-midline theta activity is also worth further exploration as a potential predictor (however, see Haller et al. (2018) for methods to ensure oscillation measurements are not confounded by non-oscillatory activity).

In addition to the recommendations for potential biomarkers in future research, we would also recommend that future research examining connectivity use multiple measures of connectivity, as recent reviews have suggested the use of a single connectivity measure may either fail to reveal true connectivity or falsely identify connectivity in the absence of connectivity differences (Bakhshayesh, Fitzgibbon, Janani, Grummett, & Pope, 2019). While the wPLI measure of connectivity we used is one of the measures least affected by the volume conduction of artifacts of all connectivity measures (Anastasiadou et al., 2019), and the consistency across time and within the sample in our original research suggests true connectivity differences in that sample which were detected using the wPLI method (but do not generalize external to the sample), the point still stands that future research will be able to more rigorously test connectivity as a potential predictor using multiple measures.

To conclude, the current study indicated that resting EEG measures of alpha power and theta connectivity which predicted response to rTMS treatment for depression did not replicate in a large independent sample. This suggests that other measures are more likely candidates for prediction of response and demonstrates the importance of replication research.

DECLARATION OF COMPETING INTEREST

MA is unpaid research director of the Brainclinics Foundation, a minority shareholder in neuroCare Group (Munich, Germany), and a co-inventor on 4 patent applications related to EEG, neuromodulation and psychophysiology, but receives no royalties related to these patents; Brainclinics Foundation received research funding from Brain Resource (Sydney, Australia), Urgotech (France) and neuroCare Group (Munich, Germany), and equipment support from Deymed, neuroConn, Brainsway and Magventure.

FVR receives research support from Canadian Institutes of Health Research, Brian Canada, Michael Smith Foundation for Health Research, Vancouver Coastal Health Research Institute, and in-kind equipment support for investigator-initiated trial from MagVenture.

He has participated in an advisory board for Janssen.

PBF is supported by a NHMRC Practitioner Fellowship (1078567). PBF has received equipment for research from MagVenture A/S, Medtronic Ltd, Neuronetics and Brainsway Ltd and funding for research from Neuronetics. He is on scientific advisory boards for Bionomics Ltd and LivaNova and is founder of TMS Clinics Australia.

AFL discloses that he has received research support from NIH, Neuronetics, Breast Cancer Foundation, Department of Defense, CHDI Foundation, and Neurosigma. He has served as a consultant to Ionis Pharmaceuticals, CHDI Foundation, and NeoSync, Inc. He is Chief Scientific Officer of Brain Biomarker Analytics LLC (BBA). He has stock options in NeoSync, Inc. and equity interest in BBA.

DMB has received research support from the CIHR, NIH, Brain Canada and the Temerty Family Foundation through the CAMH Foundation and the Campbell Research Institute. He received research support and in-kind equipment support for an investigator-initiated study from Brainsway Ltd., and he is the principal site investigator for three sponsor-initiated studies for Brainsway Ltd. He received in-kind equipment support from Magventure for investigator-initiated research. He received medication supplies for an investigator-initiated trial from Indivior. He has participated in an advisory board for Janssen.

LLC discloses that she has received research support from NIH, Neuronetics, Nexstim, Janssen, Neosync, and Affect Neuro. She has received consulting income from Janssen and Affect Neuro.

In the last 5 years, ZJD has received research and equipment in-kind support for an investigator-initiated study through Brainsway Inc and Magventure Inc. His work was supported by the Ontario Mental Health Foundation (OMHF), the Canadian Institutes of Health Research (CIHR), the National Institutes of Mental Health (NIMH) and the Temerty Family and Grant Family and through the Centre for Addiction and Mental Health (CAMH) Foundation and the Campbell Institute.

JC, NK, HvD, NWB, and CR have nothing to disclose.

ACKNOWLEDGEMENTS

This report forms the second communication of the 'International Consortium On Neuromodulation – Discovery of Biomarkers (ICON-DB)', which was established during the 3rd International Brain Stimulation Conference held in Vancouver in 2019. A group of EEG and TMS researchers decided to initiate this consortium in order to facilitate direct replication of EEG and TMS-EEG findings by facilitating immediate and independent cross-dataset replication in order to foster robustness of research findings and facilitate translation into clinical practice.

Requests for replication studies can be emailed to Andrew Leuchter (afl@ucla.edu) or Martijn Arns (martijn@brainclinics.com).

4

AN EEG SIGNATURE OF SUICIDAL BEHAVIOR IN FEMALE PATIENTS WITH MAJOR DEPRESSIVE DISORDER? A NON-REPLICATION

Submitted as: Krepel, N., Benschop, L., Baeken, C., Sack, A. T., and Arns, M. (in press). An EEG Signature of Suicidal Behavior in Female Patients with Major Depressive Disorder? A Non-Replication. *Biological Psychology*.

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.

ABSTRACT

Introduction

A recent study showed hypoactivity in the beta/gamma band in female suicide ideators and suicide attempters diagnosed with depression, relative to a low-risk group. The current study aimed to conceptually replicate these results.

Methods

In the iSPOT-D sub-sample ($N = 402$), suicide ideators and low-risk individuals were identified. Confining analyses to females only, differences between low-risk individuals and suicide ideators were tested for using the electroencephalogram (EEG) frequency bands SMR (Sensori-Motor-Rhythm; 12 – 15 Hz), beta (14.5 – 30 Hz), beta I (14.5 – 20 Hz), beta II (20 – 25 Hz), beta III (25 – 30 Hz), gamma I (31 – 49 Hz) using LORETA-software.

Results

None of the tested frequency bands showed to be significantly different between suicide ideators and low-risk individuals.

Conclusions

The current study could not conceptually replicate the earlier published results. Several reasons could explain this non-replication, among which possible electromyographic (EMG) contamination in the beta/gamma band in the original study.

Trial Registration

ClinicalTrials.gov identifier: NCT00693849
URL: http://clinicaltrials.gov/ct2/show/NCT00693849.

INTRODUCTION

Recently, a study examining the resting state EEG signature of depressed individuals at multiple levels of suicide risks was published by several of us (Benschop et al., 2019). Benschop and colleagues, limiting the analysis to females only, demonstrated that suicide ideators and suicide attempters showed low frontal beta and gamma activation as compared to low-risk individuals, with quite similar topography for both ideator and attempter groups. Furthermore, higher occipital alpha was observed in ideators. Interestingly, research into the electroencephalogram (EEG) correlates of suicidal behavior is scarce. A recent report by Arikan and colleagues showed that suicide ideators exhibited higher gamma (Arikan, Gunver, Tarhan, & Metin, 2019), contrasting the results by Benschop and colleagues. It has also been found that lower gamma power was related to an increased response to paroxetine treatment in depression (Arikan, Metin, & Tarhan, 2018), whereas Whitton et al. (2018) found higher 18.5 – 21 Hz activity within the default mode network as well as higher 12 – 18 Hz connectivity between the default mode network and the frontoparietal network in individuals with depression, as compared to individuals who remitted from depression. Additionally, higher between-network connectivity was related to more frequent

depressive episodes since the first depression onset (Whitton et al., 2018). Lee, Jang, and Chae (2017a) found that individuals exhibiting suicidal ideation showed higher frontal theta power. Another study, albeit confined to polysomnography, showed higher alpha and beta (albeit the result for beta was trend level significant) power in individuals experiencing higher levels of suicidal ideation (Dolsen et al., 2017). All in all, the EEG correlates of suicidal ideation and behavior are still unclear. Given the clinical relevance of early and accurate identification of individuals experiencing symptoms of suicidal ideation or behavior, more research is needed in this area.

The aim of the current study was to partially replicate the findings by Benschop and colleagues (2019), using data from the randomized, controlled, multicenter iSPOT-D study in 1008 MDD subjects (for additional information on this study and its parameters, please see: (Arns et al., 2016; Saveanu et al., 2015; Williams et al., 2011)). Since 'attempted suicide in the last 30 days' was an exclusion criterion for the iSPOT-D study (Williams et al., 2011), and thus no attempter group could be formed, the a priori defined hypothesis was that in female suicide ideators, relative to the low-risk group, beta/low gamma hypoactivation would be observed in the pre-frontal regions.

METHODS AND MATERIALS

The iSPOT-D study (registered at ClinicalTrials.gov (identifier: NCT00693849)) included 1008 patients diagnosed with major depressive disorder (MDD). For the purpose of this study and replication, only the suicide items administered using the Mini-International Neuropsychiatric Interview (MINI-plus) were considered and only females were included (in line with Benschop et al. (2019)). Also, individuals reporting to have attempted to commit suicide in a lifetime but not in the last 30 days, and who also did not report any suicidal ideation, were excluded. This was done to solely focus on acute suicide risk and is in line with Benschop et al. (2019).

STATISTICS

In the current study, the EEG frequency bands as reported by Benschop et al. (2019) were prospectively tested in LORETA.

LORETA analysis

1. Analyses were performed using LORETA v20200106 in Windows 10, standalone version. LORETA tests statistical group differences based on nonparametric permutation tests for functional neuroimaging. The original bands that were found to be significantly different between low-risk individuals and ideators were SMR (12 – 15 Hz), beta (14.5 – 30 Hz), beta I (14.5 – 20 Hz), beta II (20 – 25 Hz), beta III (25 – 30 Hz), and gamma I (31 – 49 Hz) (Benschop et al., 2019). Thus, in the current study these identical bands were investigated. Using LORETA, differences between suicide ideators and low-risk individuals in the iSPOT-D sample were examined. Significance level was set at $p < 0.01$, in line with other iSPOT-D studies (e.g., Arns et al. (2016)). Note: the original study also reports alpha hyperactivity in the occipital regions in suicide ideators. However, this finding was not found using LORETA software. To enhance comparability with the original study, it was decided to primarily focus on 14.5 – 48 Hz in the ROI analyses. In the first results section, alpha (9 – 13 Hz, in line with Benschop et al. (2019)) will only be shortly elaborated on.

2. After this initial replication, a second exploratory analysis was performed. Based on the results presented in Benschop et al. (2019), a single target frequency band of 14.5 – 48 Hz was established (consisting of the frequencies showing most prominent differences between suicide ideators and low-risk individuals). Then, using this frequency band ('suicide-specific band') a Region of Interest (ROI) was created in LORETA. Note that this ROI was created based on voxels being significantly different between low-risk individuals and ideators in the original study. Thus, this is not a pre-existing ROI. This ROI was used to extract beta/low gamma current source density (CSD) for all patients, after which statistical differences between high-risk and low-risk individuals were examined. In SPSS 26 for Mac, a GLM Univariate using beta/low gamma activation as a dependent variable, suicide

group as a between-subject variable, and age as a covariate was performed. This analysis was performed for both the original data as well as the iSPOT-D sample, in order to compare results and consistency between the samples. The *p*-value was set on 0.05 and 0.01 for the original and iSPOT-D sample, respectively, given the larger sample size and exploratory nature in iSPOT-D. This is in line with other iSPOT-D reports that employed a 0.01 threshold (e.g., Arns et al. (2016)).

RESULTS

Excluding individuals with no acute suicide risk (consisting of individuals with past suicide attempts, but no current suicidal ideation), females with missing data, and males, the resulting sample size was $N = 402$. This sample consisted of 188 suicide ideators (age range (in years): 18 – 65, $M = 37.2$, $SD = 13.3$) and 214 low-risk individuals (age range (in years): 19 – 85, $M = 37.7$, $SD = 12.5$).

1. No significant differences were found between low-risk individuals and ideators for SMR ($p = .383$), beta ($p = .291$), beta I ($p = .339$), beta II ($p = .344$), beta III ($p = .201$), and gamma I ($p = .255$), thereby failing to replicate the earlier result as reported by Benschop et al. (2019). Alpha also did not show to be significantly different between ideators and low-risk individuals ($p = .143$).

ROI analyses

2. A ROI was developed based on the original sample ($n_{\text{low-risk}} = 23$ (29.5%), $n_{\text{ideators}} = 36$ (46.2%), $n_{\text{attempters}} = 19$ (24.4%)) as used by Benschop et al. (2019), using the 14.5 – 48 Hz frequency range, consisting of the voxels showing to be significantly different between low-risk individuals and ideators at $p < .01$, one-tailed. In the original sample, a GLM Univariate analysis showed a significant main effect of suicide group ($F(2,74) = 3.505$, $p = .035$). Significant differences were observed when contrasting ideators with low-risk individuals ($F(1,56) = 5.904$, $p = .018$; $d = -.61$), but not when contrasting attempters with ideators ($F(1,52) = 2.079$, $p = .155$; $d = .28$), nor for attempters and low-risk individuals ($F(1,39) = 1.326$, $p = .257$; $d = -.44$), also see Figure 1 (opposite page). Importantly, the direction of the results comparing low-risk to

ideators and low-risk to attempters is identical (specifically, ideators and attempters show decreased activity compared to low-risk individuals) and the effect sizes are medium to large. This confirms that the established ROI and frequency band reflects the results as originally reported. Repeating this analysis on iSPOT-D, a GLM univariate showed a non-significant and opposite effect between ideators and low-risk individuals ($F(1,399) = 6.153$, $p = .014$; $d = .25$). This effect can be observed in Figure 1.

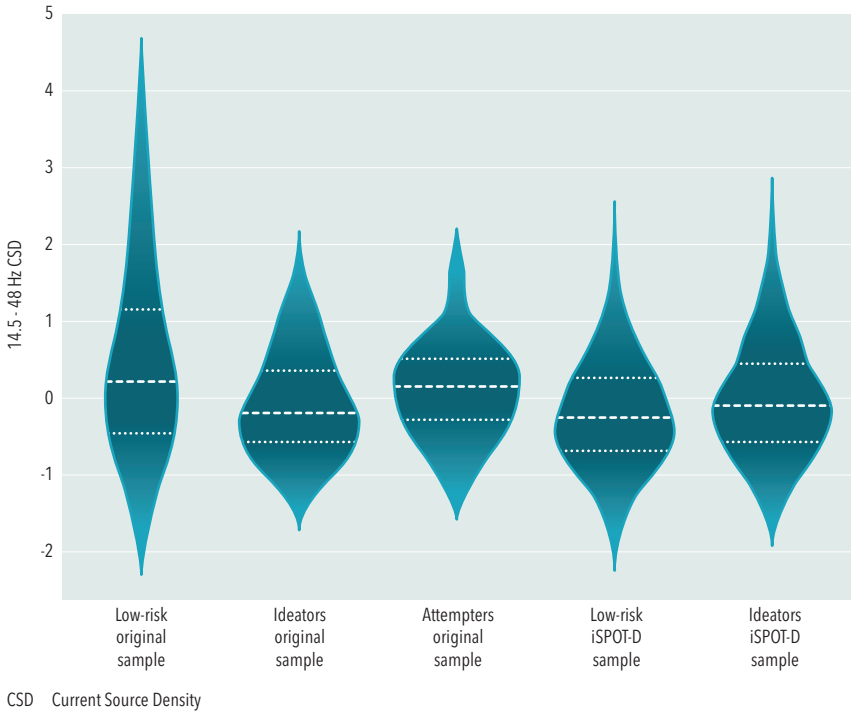


Figure 1: *This violin plot displays the CSD extracted from the ROI (14.5 – 48 Hz), separated by suicide group and sample group. The ROI is based on the original sample. Note that the 'Low-risk original sample' seems to have the largest spread.*

DISCUSSION

The current study aimed to partially replicate the results as reported by Benschop et al. (2019), which demonstrated a beta/gamma hypoactivation in female suicide ideators and attempters, as compared to

low-risk individuals. Unfortunately, the current results did not confirm the results of the original study, meaning that no difference in the beta/low gamma band was found between suicide ideators and low-risk individuals. The current results were more in line with a recent report by Arikan et al. (2019), in which high gamma power at electrodes F4, Fz, C4, Cz, O2, F8, T5, and T6 was observed in individuals reporting higher levels of suicidal ideation. Another possibility is that gamma is not specifically related to suicidal ideation, but to a more general presentation of symptoms of depression. A recent report by Fitzgerald and Watson (2018) suggested that various gamma activity patterns in different brain areas may be related to depression. It also was suggested that gamma may even be able to discriminate between bipolar and unipolar depression (Fitzgerald & Watson, 2018). Additionally, although suicidal ideation and depression are heavily intertwined, suicidal ideation is not a symptom solely present in depression. For example, suicidal ideation has been reported in obsessive-compulsive disorder (Pellegrini et al., 2020), post-traumatic stress disorder (Krysinska & Lester, 2010), schizophrenia (Chapman et al., 2015), and attention-deficit/hyperactivity disorder (Furczyk & Thome, 2014; Taylor, Boden, & Rucklidge, 2014). Future studies should explore EEG signatures of suicidal ideation across different disorders to investigate if an EEG correlate is specific to a particular disorder, or whether this relation behaves transdiagnostically and is apparent in multiple disorders.

It is unclear why the current study did not confirm the earlier obtained results. One reason could be a lack of high-risk individuals (i.e., 'suicide attempters'), since this was an exclusion criterion for iSPOT-D, thereby decreasing the signaling contrast between groups. Another reason could be medication effects. In the original study, the original sample was not tapered of medication at the time of baseline QEEG assessment, whereas subjects in iSPOT-D were off medication at baseline. It has been reported that barbiturates and benzodiazepines augment beta activity, specifically in the 15 – 25 Hz range (Blume, 2006), yet medication effects on gamma have not been widely studied. Some animal studies also report NMDAR antagonist-induced gamma oscillations (Hiyoshi, Kambe, Karasawa, & Chaki, 2014; Jones et al., 2012; Phillips et al., 2012), but NMDAR antagonists (ke-

tamine, among others) have not been reported in the original sample. A third reason for the current non-replication may be sample size. The original sample consisted of 78 females, and dividing that group into three subgroups resulted in relatively small samples of the subgroups. It is possible that individual contributions of the EEG within relatively small sample sizes introduced some noise which may have altered the signal-to-noise ratio in the samples. As can be observed in Figure 1 on page 93, it appears that the low-risk TMS sample has the widest spread (CSD for low-risk individuals: min = -1.01; max = 3.39; mean = .51, *SD* = 1.19), even relative to the large iSPOT-D groups (CSD for low-risk individuals: min = -1.75; max = 2.06; mean = -.21, *SD* = .69; CSD for ideators: min = -1.40; max = 2.35; mean = -.04, *SD* = .72). All EEG data were collected using identical procedures, amplifiers, and automated artifact processing (also see Arns et al., 2016 for full details), reducing the likelihood that the current non-replication can be explained by differences in hardware or signal pre-processing. Post-hoc visual inspection of individuals from the low-risk group with 14.5 – 48 Hz CSD > 2.0 showed intermediate levels of frontal muscle tension (visualized in Figure 2 on page 96). Possibly, EMG contaminated the gamma activity (Whitham et al., 2007), yielding a quite strong signal in a relatively small sample (specifically, in the low-risk group), thus explaining the lower 'gamma' in the risk groups. In the original study, muscle artifacts were controlled for using a pre-processing pipeline and additional machine learning ICA artifact detection (MARA), thus the possible influences of muscle artifacts on the dataset would be expected to be minimal to none. Yet, individual contributions to the signal possibly may have skewed the results. Concluding, in the current study, no consistent differences between female suicide ideators and females with low suicide risk could be found in the EEG beta and gamma bands.

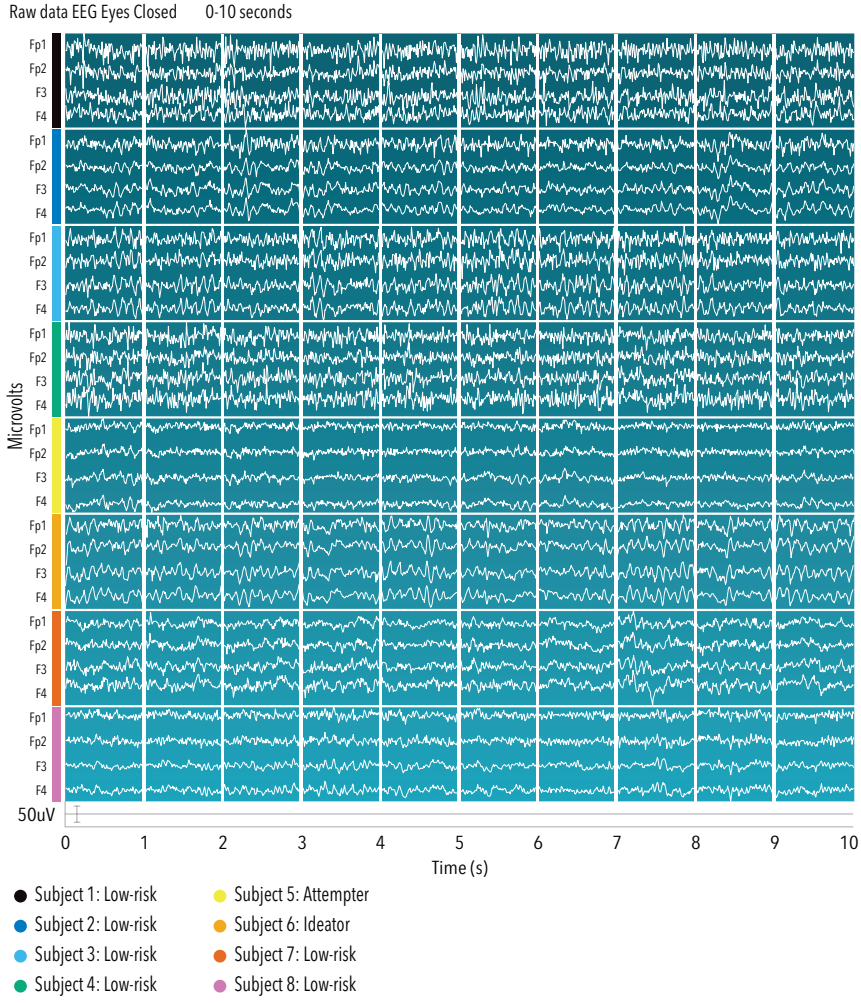


Figure 2: *Snapshot of the first ten seconds of the EC EEG for the eight people scoring highest on the ROI analyses, for the original sample only, ranked from scoring highest to lowest. Subject 1 scored highest and subject 8 scored lowest on the ROI analyses. Note that the four highest-scoring individuals all were part of the low-risk group.*

ACKNOWLEDGEMENTS

We would also like to thank Frederik Van De Steen for his support in establishing this manuscript. The iSPOT-D study is sponsored by Brain Resource Company Operations Pty Ltd. The iSPOT-D Investigators Group, the contributions of principal investigators at each site, and the monitoring support the central iSPOT-D management team are gratefully acknowledged.

POTENTIAL CONFLICTS OF INTEREST

The authors report no financial or other relationship relevant to the subject of this article.

FUNDING/SUPPORT

None.

DISCLOSURES

MA is unpaid research director of the Brainclinics Foundation, a minority shareholder in neuroCare Group (Munich, Germany), and a co-inventor on 4 patent applications (A61B5/0402; US2007/0299323, A1; WO2010/139361 A1) related to EEG, neuromodulation, and psychophysiology but does not own or receive any proceeds related to these patents. The other authors report no conflict of interest.

5

CAN PSYCHOLOGICAL FEATURES PREDICT ANTIDEPRESSANT RESPONSE TO RTMS? A DISCOVERY-REPLICATION APPROACH

Published as: Krepel, N., Rush, A. J., Iseger, T. A., Sack, A. T., Arns, M. (2019). Can psychological features predict antidepressant response to rTMS? A Discovery-Replication approach. *Psychological Medicine, 50*(2), 264-272. doi: 10.1017/S0033291718004191

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.

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; Fitzgerald, Hoy, Daskalakis, & Kulkarni, 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 (Griffith et al., 2010; Hayward, Taylor, Smoski, Steffens, & Payne, 2013; Jylhä & Isometsä, 2006; Kotov, Gamez, Schmidt, & Watson, 2010; Rosellini & Brown, 2011) and extraversion seems to be inversely associated with MDD (Hayward et al., 2013; Jylhä & Isometsä, 2006; Kotov et al., 2010; Rosellini & Brown, 2011). 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, McGirr, Beaulieu, Van den Eynde, & Turecki, 2013). In a study that focused on deep TMS (dTMS), higher agreeableness and higher conscientiousness were observed in patients who achieved remission (McGirr, Van den Eynde, Chachamovich, Fleck, & Berlim, 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 (Pizzagalli, 2014; Treadway & Zald, 2011). It has been argued that anhedonia is a difficult symptom to treat (Treadway & Zald, 2011) and multiple studies have shown that higher levels of anhedonia are predictive of poorer treatment outcome (McMakin et al., 2012; Spijker, Bijl, de Graaf, & Nolen, 2001). 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, Gourion, & Mouchabac, 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, Kazemi, Nitsche, Gholipour, & Salehinejad, 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, Leventhal, Khan, & Brown, 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 col-

leagues 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, Bagby, & Levitt, 1993) and inconsistently replicated, including reports of patients with anxious depression responding better to ketamine treatment (Ionescu 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, Luborzewski, Danker-Hopfe, Kathmann, & Bajbouj, 2007). For stress, it has been suggested that chronic stress is predictive of depression, even more so than acute stressors (Hammen, 2005; McGonagle & Kessler, 1990). 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 specific-

ity, the positive predictive value (PPV) and negative predictive value (NPV) were examined. This method has been applied elsewhere (e.g. (Kuk, Li, & Rush, 2010; Li, Kuk, & Rush, 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 & Wagenmakers, 2012; Patil, Peng, & Leek, 2016), and our own recent non-replication of previous work (Krepel et al., 2018; Chapter 2), 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 Hague, Psychologenpraktijk Timmers, Oosterhout, The Netherlands) between May 2007 and November 2016. Inclusion criteria were: (1) a primary diagnosis of major depressive or dysthymic disorder as confirmed by M.I.N.I. (M.I.N.I. Plus Dutch version 5.0.0), (2) a BDI-II of 14 or higher, and (3) a left DLPFC HF (10 Hz) rTMS or a right DLPFC LF (1 Hz) rTMS treatment combined with psychotherapy. Exclusion criteria included: (1) prior treatment with ECT, (2) epilepsy, (3) wearing a cardiac pacemaker, (4) wearing metal parts in the head, and (5) pregnancy. All participants signed an informed consent form before treatment was initiated.

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.

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 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, Chasson, Tapia, Miller, & Pettit, 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, Steer, Teitelman, & Villacis, 2002).

These subscales were computed by adding the indicated items into one variable (e.g., the anhedonia scale was computed by adding items 4, 12, and 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 α range 0.87 – 0.92) (Costa & McCrae, 1992).
- To measure anxiety and stress, the DASS (Henry & Crawford, 2005) was used.

DISCOVERY AND REPLICATION SET

To obtain a Discovery and Replication set, the complete, original dataset (N = 196) was randomly divided into a 60% Discovery sample and 40% Replication sample. Differences in gender, response, and remission were tested for using χ^2 statistics. Differences in age

and BDI at intake were tested for using one-way analysis of variance tests. Once none of these variables differed significantly between the two groups, the random division was frozen, and the two groups were designated as Discovery set and Replication set. This resulted in a Discovery set (n = 119; 62 females) and Replication set (n = 77; 37 females). All exploratory analyses were performed in the Discovery set. Only when a significant result was obtained in the Discovery set, the same statistical test was used in the Replication set to confirm or deny the prior obtained finding.

Statistics

SPSS version 24 was used for statistical analyses. Response was defined as a ≥ 50% decrease on the BDI score from intake to outtake (outtake BDI scores were taken around the last session of a patient, on average at session 21). All predictors were tested for statistical differences, robustness (independent replication), dimensional association, predictive value, and clinical relevance.

First, a generalized linear model (GLM) univariate analysis was performed to test for baseline psychological differences between R and NR, with response as a fixed factor, age as a covariate, and the BDI subscales (including total BDI, Anhedonia, Non-Anhedonia, Cognitive Affective, Somatic and Performance, Cognitive, and Non-Cognitive), NEO-FFI subscales (including Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness), and DASS scales (Depression, Anxiety, and Stress) as dependent variables. Effect sizes reported are Cohen's d.

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. ≥ 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 (Arns et al., 2016; Iseger et al., 2017; van Dinteren et al., 2015), yet statistically testing such interactions requires relatively large sample sizes (Leon & Heo, 2009). Therefore, in this full sample a post-hoc analysis on the complete dataset was conducted and statistically examined the data for potential sex × response interactions by including sex as a between-subject factor. In the online supplementary material, it is also investigated whether anhedonia could accurately predict long-term response.

Primary and secondary hypotheses
By dividing our complete sample into a Discovery and Replication set, the sample sizes naturally decreased in both samples. To prevent a type-II error from occurring due to reduced sample sizes, it was decided that the hypotheses were defined as primary or secondary hypotheses. A primary hypothesis was defined as a hypothesis that was predicted from the literature and the *p*-value was set at $p \leq 0.1$ in the Discovery set. In the Replication set this finding had to reach $p \leq 0.05$ to be considered a replication. A secondary hypothesis was defined as a hypothesis that was not predicted from the literature, i.e., an exploratory analysis, yet showed statistical significance in the dataset. This *p*-value was set at $p \leq 0.05$. Given the confirmatory nature of the Replication set, it was decided that the correlation would be one-tailed (under the condition that the direction of the two-tailed test was the same as in the Discovery set).

As primary hypotheses, the following were defined as predictors for non-response: high total BDI, high anhedonia, high neuroticism, low extraversion, high anxiety, and high stress.

All other psychological measures were marked as secondary hypotheses.

RESULTS

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 on page 110. 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.

Clinical response	Discovery sample	Replication sample
Responders	77 (65.3%)	53 (68.8%)
Total number of sessions	22.3 (*SD* 8.0)	19.0 (*SD* 6.0)
BDI intake	31.3 (*SD* 10.3)	31.2 (*SD* 9.7)
BDI outtake	13.6 (*SD* 11.6)	14.8 (*SD* 13.0)
BDI percentage change	57.1%	54.1%
Cohen's *d* intake-outtake	*ES* = 1.6	*ES* = 1.4
Trial center (Nijmegen, The Hague, Oosterhout)	109 (91.6%), 2 (1.7%), 8 (6.8%)	68 (88.3%), 2 (2.6%), 7 (9.1%)
rTMS protocol		
HF (10 Hz left)	45 (38.1%)	29 (37.7%)
LF (1 Hz right)	70 (58.8%)	45 (58.4%)
Both sequentially	4 (3.4%)	3 (3.9%)

Table 1: *Clinical outcome measures of the Discovery and Replication sample. This table depicts the number of responders, total number of sessions, total BDI at intake, total BDI at outtake, the BDI change from intake to outtake (in percentage), as well as the* ES (*Cohen's* d) *of total BDI change from intake to outtake, trial center, and rTMS protocol. BDI at intake, BDI at outtake, BDI percentage change, and number of responders did not differ significantly between groups* ($p \geq 0.502$).

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 ($p = 0.029$; $F = 4.889$; $df = 1$). All other variables were not significantly different between R and NR (see Table 2, opposite page). 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 ($p = 0.005$; $F = 8.516$; $df = 1$). 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-signif-

icant 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.

(sub)scale	Responder (*m*, (*sd*))		Non-Responder (*m*, (*sd*))		*p*		*ES* (Cohen's *d*)	
	Discovery	Replication	Discovery	Replication	Discovery	Replication	Discovery	Replication
Age	41.9 (10.8)	43.9 (14.3)	46.6 (14.5)	43.0 (12.6)	.049	.788	.37	-.07
BDI total	30.7 (9.5)	29.4 (8.4)	32.4 (11.8)	35.3 (11.1)	.198	.012	.16	.60
Anhedonia	**5.2 (2.1)**	**4.5 (1.8)**	**6.0 (2.3)**	**5.8 (1.6)**	**.072**	**.005**	**.36**	**.76**
Non-Anhedonia	25.9 (7.7)	24.5 (8.5)	26.8 (10.1)	29.0 (10.2)	.314	.066	.10	.48
Cognitive-Affective	18.4 (6.1)	16.9 (7.4)	19.1 (8.1)	20.7 (7.7)	.315	.064	.10	.50
Somatic & Performance	12.8 (4.0)	12.1 (3.1)	13.6 (4.8)	14.1 (4.2)	.259	.029	.18	.54
Cognitive	10.3 (4.5)	9.8 (5.5)	10.0 (5.7)	10.9 (5.7)	.821	.483	-.06	.20
Non-Cognitive	21.0 (6.2)	19.3 (4.9)	22.7 (7.8)	23.9 (6.4)	.133	.002	.24	.81
NEO-FFI...								
...Neuroticism	32.7 (6.3)	30.4 (7.8)	32.7 (7.2)	31.8 (6.8)	.844	.455	0	.19
...Extraversion	19.1 (6.8)	20.1 (7.1)	18.9 (8.6)	17.0 (8.0)	.919	.106	-.03	-.41
...Openness	**27.0 (6.2)**	**25.7 (7.3)**	**24.6 (7.4)**	**23.4 (6.8)**	**.029**	**.227**	**-.35**	**-.33**
...Agreeableness	32.1 (5.3)	31.8 (6.8)	32.0 (5.9)	32.9 (4.3)	.854	.476	-.02	.19
...Conscientiousness	26.3 (6.5)	28.6 (7.7)	27.3 (7.4)	26.7 (5.2)	.561	.294	.14	-.29
DASS...								
...Depression	29.0 (9.2)	26.4 (11.5)	29.1 (9.5)	31.0 (9.4)	.666	.088	.01	.44
...Anxiety	13.9 (8.7)	12.3 (8.4)	14.7 (9.3)	14.5 (8.8)	.801	.343	.09	.26
...Stress	24.8 (10.0)	18.8 (11.1)	21.8 (10.4)	23.8 (8.9)	.164	.053	-.29	.50

Table 2: *This table depicts the mean, standard deviation,* p*-value, and* ES *for each baseline psychological measure, showing values for Discovery as well as Replication sample, subdivided into R and NR. A univariate analysis, controlled for age, showed that the Anhedonia scale is significantly different between R and NR in both the Discovery* (p = *0.072;* d = *0.36) as well as for the Replication set* (p = *0.005;* d = *0.76). Openness shows to be significantly different between R and NR in the Discovery sample* (p = *0.029;* d = *–0.35), however, this result fails to be replicated in the Replication sample* (p = *0.227;* d = *–0.33). These results are indicated in bold.*

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 on page 112, none of the baseline variables reached or exceeded and replicated the preset value of PPV and NPV.

(sub)scale	Sensitivity (%)		Specificity (%)		PPV (%)		NPV (%)	
	Discovery	Replication	Discovery	Replication	Discovery	Replication	Discovery	Replication
BDI total	63.4	54.2	57.1	64.2	44.1	40.6	74.6	75.6
Anhedonia	62.5	73.9	53.7	74.4	44.6	60.7	70.6	84.2
Non-Anhedonia	56.4	56.5	54.5	64.3	42.3	46.4	67.9	73.0
Cognitive-Affective	57.5	47.8	53.0	59.5	42.6	39.3	67.3	67.6
Somatic & Performance	59.0	56.5	59.7	62.8	46.0	44.8	71.4	73.0
Cognitive	56.4	47.8	54.5	50.0	42.3	34.4	67.9	63.6
Non-Cognitive	62.5	47.8	58.2	72.1	47.2	47.8	72.2	72.1
NEO-FFI...								
...Neuroticism	56.4	59.1	56.9	45.7	41.5	32.5	70.7	67.9
...Extraversion	56.4	63.6	56.9	58.7	41.5	42.4	70.7	77.1
...Openness	56.4	59.1	58.3	47.8	42.3	35.1	71.2	71.0
...Agreeableness	59.0	59.1	55.6	47.8	41.8	35.1	71.4	71.0
...Conscientiousness	51.3	59.1	55.6	63.0	38.5	43.3	67.8	76.3
DASS...								
...Depression	56.1	58.3	53.4	55.1	40.4	38.9	68.4	73.0
...Anxiety	56.1	54.2	54.8	62.5	41.1	41.9	69.0	73.2
...Stress	48.8	54.2	57.5	59.2	39.2	39.4	66.67	72.5

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

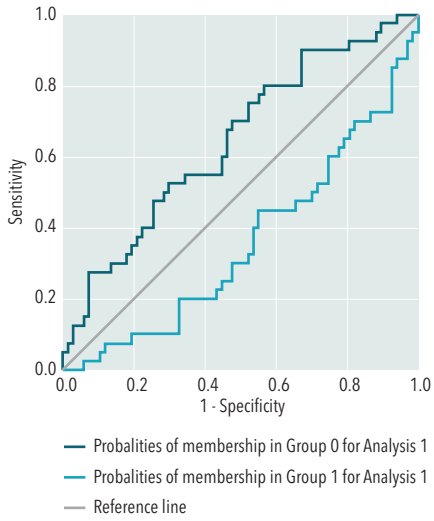
CORRELATIONS BETWEEN PSYCHOLOGICAL SCALES AND BDI PERCENTAGE CHANGE

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' λ = 0.937; χ^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 curve of the ROC, regressed on non-response, was 0.643 (Figure 1). This analysis was repeated in the Replication set. This model also reached significance (p = 0.018; Wilks' λ = 0.880; χ^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 (Figure 2).



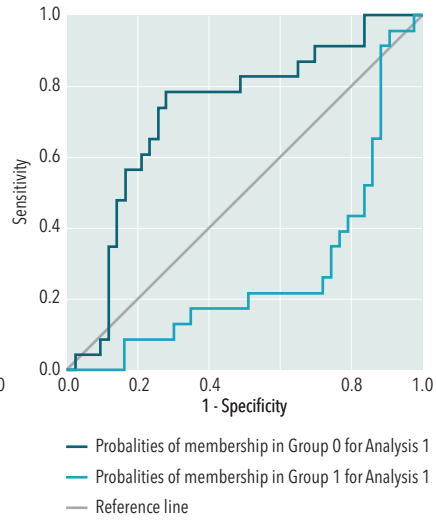


Figure 1 (left): ROC of the discriminant analysis on non-response using the Anhedonia scale and age as independent variables in the Discovery sample, with an area under the curve of 0.643. The ROC shows the sensitivity (62.5%) and specificity (53.7%) for non-responders (dark blue line) and responders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

Figure 2 (right): 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 (dark blue line) and responders (blue line). The PPV and NPV were 60.7% and 80.0%, respectively.

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 most accurate, was estimated. Youden's *J* 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 not to be clinically relevant.

DISCUSSION

This study focused on finding psychological measures and their potential ability to predict rTMS treatment response in an MDD sample. Additionally, attempting to overcome the issue of (non-)replication, this study tried to immediately replicate obtained findings by a priori dividing the complete dataset into a Discovery and Replication set. 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 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 it was 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 & Wagenmakers, 2012; Roediger III, 2012; Simons, 2014) (although caution should be taken when designing and interpreting a replication study, see Maxwell, Lau, and Howard (2015), as well as self-replication (i.e. replicate your own work before publishing) (Roediger III, 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 on page 111 can also be inspected as an example of how many false positives one could obtain using a $p < 0.05$ ap-

proach 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 v. stricter *p*-thresholding) increased the robustness of results. Based on these results we cannot draw a definite 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 anhedo-

nia 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.

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.

CONFLICTS OF INTEREST

MA reports options from Brain Resource (Sydney, Australia); he is director and owner of Research Institute Brainclinics, 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 Brainclinics 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.

AJR 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.

SUPPLEMENTARY MATERIALS

Given the piloting approach by applying a Discovery and Replication set, the precise parameters and validity of this approach still need to be considered.

Firstly, sex differences in symptom presentation (Marcus et al., 2005; Schuch, Roest, Nolen, Penninx, & de Jonge, 2014), developmental course of MDD (Essau, Lewinsohn, Seeley, & Sasagawa, 2010) and sex specific predictors of antidepressant treatment response (Arns et al., 2016; Iseger et al., 2017; van Dinteren et al., 2015) are well documented. This means that in the current sample it is possible that there are gender differences in how the psychological measures are structured and how rTMS NR is predicted.

Likewise, it is possible that with the Discovery-Replication method the current sample loses too much power (even though we were lenient with our primary and secondary outcomes), meaning that important differences between R and NR might have been missed that would have been visible in the total sample.

This supplement will expand on the main manuscript by examining the above methodological issues, performing analyses on Youden's *J* and clinical relevance, and examining follow-up data. In the first section, interaction effects between gender, response, and the psychological measurements will be explored. Then, the same statistical tests as in the manuscript (i.e. univariate analyses, dimensional approaches, and discriminant analyses) will be performed on the complete dataset, using a strict, Bonferroni-corrected *p*-value to prevent a type I error from occurring (thereby adhering to the proposed method of decreasing the *p*-value (Benjamin et al., 2018) in tackling the 'Replication crisis', as was elaborated on in the main manuscript).

In this section, the complete analyses for Youden's *J* (as described in the main manuscript) are also reported. Lastly, it will be investigated whether anhedonia can accurately predict long-term response using follow-up data.

Clinical response	Females	Males	Complete dataset
Responders	70 (70.7%)	60 (62.5%)	130 (66.7%)
Total number of sessions	21.5 (*SD* 7.2)	20.5 (*SD* 7.6)	21.0 (*SD* 7.4)
BDI Intake	33.2 (*SD* 10.6)	29.2 (*SD* 9.0)	31.2 (*SD* 10.0)
BDI Outtake	14.5 (*SD* 12.7)	13.7 (*SD* 11.6)	14.1 (*SD* 12.2)
BDI percentage change	58.5%	53.2%	55.9%
Cohen's *d* Intake – Outtake	*ES* = 1.6	*ES* = 1.5	*ES* = 1.5

Table S1: *Clinical outcome measures of females and males separately, as well as of the total sample. This table depicts the number of responders, total number of sessions, total BDI at intake, total BDI at outtake, the BDI change from intake to outtake (in percentage), as well as the* ES *(Cohen's* d*) of total BDI change from intake to outtake.*

SEX INTERACTIONS

The clinical features of the males (average age: 44.14, range 18 – 71 yrs) and females (average age: 43.11, range 19 – 78 yrs) separately and the complete dataset (average age: 43.62, range 18 – 78 yrs; 99 females and 97 males) can be found in Table S1.

First, a univariate analysis with response and gender as fixed factor, age as a covariate, and a psychological measure as dependent variable was performed. Since analyses of interactions require relatively large sample sizes (Leon & Heo, 2009), it was decided to merge the Discovery and Replication set to ensure that possible interaction effects are visible. Given the eliminatory nature of this analysis, gender * response interactions were looked at specifically. The results of these tests can be found in Table S2 on page 120. One subject was excluded based on missing data.

(sub)scale	Gender * Response
BDI total	$p = .252$
Anhedonia	$p = .731$
Non-Anhedonia	$p = .207$
Cognitive-Affective	$p = .343$
Somatic and Performance	$p = .226$
Cognitive	$p = .344$
Non-Cognitive	$p = .288$
NEO-FFI Neuroticism	$p = .844$
NEO-FFI Extraversion	$p = .179$
NEO-FFI Openness	$p = .176$
NEO-FFI Agreeableness	$p = .447$
NEO-FFI Conscientiousness	$p = .415$
DASS Depression	$p = .801$
DASS Anxiety	$p = .911$
DASS Stress	$p = .859$

Table S2: *This table depicts sex * response interactions. Each baseline psychological measure was examined using a univariate analysis with this measure as a dependent variable, age as a covariate, and sex and response as fixed factors. This table shows that none of the psychological variables show sex * response interactions.*

As can be seen in Table S2, no significant interactions between gender and response were found. Therefore, it can be concluded that there is no need to examine the psychological data separately for females and males in the analyses that were performed in the manuscript.

FULL-SAMPLE ANALYSES

In the next analysis, a more classic approach was taken by performing the statistics on the complete dataset ($N = 196$). To prevent a type I error from occurring, a Bonferroni-corrected p-value was used. The resulting Bonferroni-corrected was $\alpha = .05 / 15 = .00333$. A univariate analysis with response as a fixed factor, age as a covariate, and a psychological measure as dependent variable was performed. The results of these analyses can be found in Table S3 (opposite page).

Interestingly, the results from this analysis slightly differ from those in the manuscript. Handling a Bonferroni-corrected p-value of .00333, it can be concluded that the Anhedonia scale and the Non-Cognitive scale significantly differ at baseline between R and NR. Note, Openness does not reach statistical significance and is discarded immediately.

Next, a dimensional approach was applied. A correlation between BDI% change over the course of the treatment and the Anhedonia scale, while controlling for age, showed a significant correlation ($p = .006$; $r = -.210$; $r^2 = 4.4\%$). The correlation between BDI% change

and the Non-Cognitive scale did not reach significance (p = .058; r = -.145; r^2 = 2.1%).

(sub)scale	Responders	Non-responders	*p*	*ES* (Cohen's *d*)
Age	42.7 (12.3)	45.3 (13.8)	.195	.20
BDI total	30.1 (9.1)	33.5 (11.50)	.011	.33
Anhedonia	4.9 (2.0)	5.9 (2.1)	.003	.49
Non-Anhedonia	25.4 (8.0)	27.6 (10.1)	.045	.24
Cognitive-Affective	17.9 (6.7)	19.7 (8.0)	.049	.24
Somatic and Performance	12.5 (3.7)	13.8 (4.6)	.026	.31
Cognitive	10.1 (4.9)	10.3 (5.7)	.498	.04
Non-Cognitive	20.3 (5.7)	23.1 (7.3)	.003	.43
NEO-FFI Neuroticism	31.8 (7.0)	32.4 (7.0)	.479	.09
NEO-FFI Extraversion	19.5 (6.9)	18.2 (8.4)	.307	-.17
NEO-FFI Openness	26.5 (6.7)	24.2 (7.2)	.021	-.33
NEO-FFI Agreeableness	32.0 (5.9)	32.3 (5.4)	.869	.05
NEO-FFI Conscientiousness	27.2 (7.1)	27.1 (6.6)	.805	-.02
DASS Depression	28.0 (10.2)	29.8 (9.5)	.182	.18
DASS Anxiety	13.3 (8.6)	14.6 (9.1)	.341	.15
DASS Stress	22.4 (10.8)	22.5 (9.8)	.923	.01

Table S3: *This table depicts the mean, standard deviation,* p-*value, and* ES *for each baseline psychological measure, showing values for the complete dataset, subdivided into R and NR. A univariate analysis, controlled for age, using a strict Bonferroni-corrected* p-*value, showed that the Anhedonia scale is significantly different between R and NR* (p = *.003;* d = *.49). The Non-Cognitive scale also shows a significant difference between R and NR at baseline* (p = *.003;* d = *.43).*

A discriminant analysis using the Anhedonia scale, the Non-Cognitive scale, and age as independent variables, regressed on non-response, yielded a significant model (p = .007; Wilks' Lambda = .931; Chi-square = 12.054; df = 3). This resulted in an ROC (Figure S1, on page 122) with an area under the curve of .657.

An item to further explore is the fact that in the manuscript only the Anhedonia scale at baseline is indicated as being significantly different between NR and R, whereas in the current analysis the Non-Cognitive scale is also implicated. The Non-Cognitive scale is a psychological scale based on 13 items of the BDI, including item 1, 4, 10 – 13, and 15 – 21. Since the Anhedonia scale consists of items 4, 12, and 21, it could be the case that the currently obtained significant difference is mainly driven by item overlap with the Anhedonia scale, which is part of the Non-Cognitive scale. For this reason, a variable was created that consisted of Non-Cognitive items, minus the Anhedonia

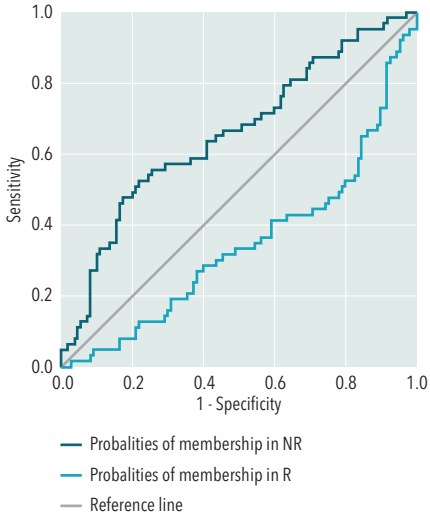


Figure S1: *Receiver operator curve (ROC) of the discriminant analysis on non-response using the Anhedonia scale, the Non-Cognitive scale, and age as independent variables in the complete sample, with an area under the curve of 0.657. The ROC shows the sensitivity (63.5%) and specificity (58.2%) for non-responders (dark blue line) and responders (blue line). The resulting PPV and NPV were 46.5% and 73.6%, respectively.*

items (i.e., item 1, 10, 11, 13, 15 – 20). A GLM univariate analysis was performed with response as fixed factor, age as a covariate, and the Non-Cognitive scale minus the Anhedonia items as a dependent variable. Using the Bonferroni-corrected *p*-value of .00333, this analysis yielded a non-significant result ($p = .011$; $F = 6.678$; $df = 1$). This implicates that the previously indicated result, namely that the Non-Cognitive scale at baseline is significantly different between R and NR, is mainly driven by the items that construct the Anhedonia scale.

CLINICAL RELEVANCE (SENSITIVITY, SPECIFICITY, PPV, NPV, AND YOUDEN'S *J*)

In the main manuscript, it was reported that a baseline anhedonia score of 6 or higher is a cut-off score on which a therapist can decide to discontinue rTMS treatment. In the current section, the analyses for this result will be elaborated on.

Statistics are based on BDI anhedonia scores at baseline. BDI anhedonia score was used as an independent variable in a ROC, regressed on non-response. The resulting sensitivities, specificities, and Youden's *J* can be found in Table S4 (opposite page). Based on Youden's *J*, it can be concluded that the cut-off of anhedonia severity of 6 or higher results in the optimal sensitivity and specificity of the model.

Note: this analysis was performed without including age as a supporting variable. This was done to ensure the clinical usefulness of such a cut-off. The results can be found in Table S4.

Cut-off point	Sensitivity (%)	Specificity (%)	Youden's *J*
≥0	100	0	0
≥1	100	0	0
≥2	96.8	6.4	.032
≥3	82.5	30.9	.134
≥4	69.8	46.4	.162
≥5	58.7	61.8	.205
≥6	44.4	79.1	.235
≥7	25.4	85.5	.109
≥8	12.7	93.6	.063
≥9	0.0	100	0

Table S4: *Sensitivity, specificity, and Youden's* J *calculated for each level of severity in the Anhedonia scale. For this analysis, a discriminant analysis using age and the Anhedonia scale was performed. Based on resulting receiver operator curve (ROC) of this analysis, the sensitivity, specificity, and Youden's* J (J = *Sensitivity + Specificity – 1) for each level of severity were calculated. Youden's* J *is highest at severity level of 6 or higher.*

As can be seen in Table 4, Youden's *J* is biggest at anhedonia severity level of 6 or higher (*J* = .235). This means that a score of 6 or higher is the indicated cut-off score on which a therapist can act.

LONG-TERM RESPONSE

Lastly, the option of long-term treatment response prediction based on anhedonia was explored. At this point, the Discovery and Replication set were merged. The reason for this was that only initial R (i.e. BDI decrease ≥ 50%) were included in the follow-up (FU) sample. This means that we reduced our sample size by 33.3% (based on the whole sample), next to missing values in the R. The final FU sample consisted of 65 subjects (age: 44.35 yrs, range 19 – 78 yrs; 35 females and 30 males). The clinical outcomes of this sample can be found in Table S5.

Clinical response	FU sample
Responders	42 (64.6%)
Total number of sessions	22.1 (*SD* 7.8)
BDI Intake	29.1 (*SD* 9.0)
BDI Outtake	7.3 (*SD* 6.4)
BDI FU	12.2 (*SD* 10.5)
BDI percentage change pre-post	75.4%
BDI percentage change pre-FU	55.5%
Cohen's *d* Intake – Outtake	*ES* = 2.79
Cohen's *d* Intake – FU	*ES* = 1.73

Table S5: *Clinical outcome measures of the follow-up (FU) sample. This table depicts the number of responders, total number of sessions, total BDI at intake, total BDI at outtake, total BDI at FU, the BDI change from intake to outtake (in percentage), the BDI change from intake to FU, the* ES (*Cohen's* d) *of total BDI change from intake to outtake, as well as the* ES (*Cohen's* d) *of total BDI change from intake to FU.*

A GLM univariate analysis using FU response as a fixed factor, age as a covariate, and the Anhedonia scale at baseline did not yield a significant result (p = .537; F = .387; df = 1). A discriminant analysis using age and the Anhedonia scale as independent variables, regressed on FU NR, resulted in a sensitivity of 71.4%, a specificity of 57.5%, a PPV of 46.9%, and an NPV of 79.3%. A dimensional approach in a two-tailed partial correlation using baseline-FU BDI percentage change and the Anhedonia scale at baseline, while controlling for age, yielded a non-signficant association (p = .604; r = .068; r^2 = 0.46%). A discriminant analysis using the Anhedonia scale and age as independent variables, regressed on non-response at FU, yielded a non-significant model (p = .151; Wilks' Lambda = .937; Chi-square = 3.777; df = 1), which resulted in an ROC curve with an area under the curve of .631 (Figure S2).

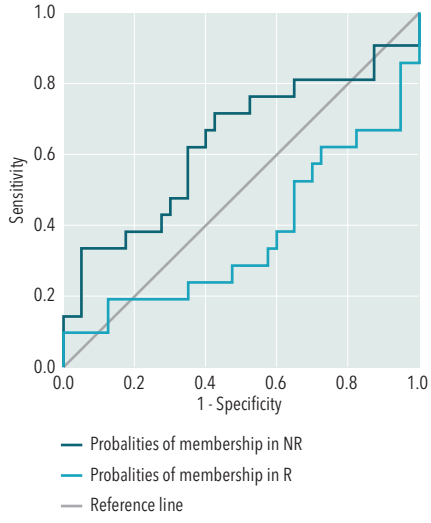


Figure S2: *Receiver operator curve (ROC) of the discriminant analysis on FU non-response using the Anhedonia scale and age as independent variables in the FU sample, with an area under the curve of 0.631. The ROC shows the sensitivity (71.4%) and specificity (57.5%) for non-responders (dark blue line) and responders (blue line). The resulting PPV and NPV were 46.9% and 79.3%, respectively.*

Given the above analyses, the following can be concluded. Based on the statistics in the manuscript, anhedonia is the only psychological measure that significantly and robustly differs at baseline between R and NR, but only marginally predicts treatment response. Based on the current statistics, anhedonia is still implicated as an important psychological feature that differs between R and NR at baseline. However, according to the full-sample, Bonferroni-corrected analysis, the Non-Cognitive scale (a scale based on 13 items of the BDI)

is also implicated as being significantly different between R and NR at baseline, while in the manuscript it is discarded. However, this result seems to be mainly driven by anhedonia. Regarding correlational analyses, the method used in the manuscript showed a higher explained variance for the correlation between BDI% change and Anhedonia at baseline ($r^2_{Discovery}$ = 4.9%, $r^2_{Replication}$ = 6.0%, vs. 4.4% in the current supplement). Lastly, anhedonia does not seem to accurately predict long-term rTMS response. However, even though univariate analyses and correlations did not show to be significant, the NPV of this analysis was higher than 75%, meaning that anhedonia might be able to predict long-term response in a subgroup of MDD patients. More research is needed to explore this possibility.

More important is the implication of the current usage of a different statistical method. While in the manuscript only anhedonia is being implicated as being significantly different between R and NR, in this supplement also the Non-Cognitive scale is indicated. Interestingly, the ES and the correlation for the Non-Cognitive scale are both weaker than for the Anhedonia scale. Importantly, when the items of the Anhedonia scale were subtracted from the Non-Cognitive scale, the significant effect disappeared. This shows that the finding that the Non-Cognitive scale is significantly different between R and NR is potentially a false positive. These methodological issues are important to consider, since the existence of these types of errors may (partly) explain why little (psychological) research is inconsistently replicated.

Importantly, one could wonder, if we would have handled the Bonferroni-corrected *p*-value in the main manuscript, whether we would have indicated the Non-Cognitive scale as a significant difference between R and NR. That is, given that both the Anhedonia scale and the Non-Cognitive scale were significant, yet the Anhedonia scale was part of the Non-Cognitive scale, it is a natural consequence to check whether the Non-Cognitive scale was driven by the Anhedonia scale (even if we would not have had the knowledge of the main manuscript), which in turn would have discarded the Non-Cognitive scale as a significant result. In short, in both the original manuscript and the current supplement it is likely that we would have

found the Anhedonia scale, but not the Non-Cognitive scale. Therefore, we cannot conclude whether a Discovery-Replication approach or a stricter Bonferroni-corrected approach is preferred, however, handling either of these two approaches does seem to decrease the chance of finding false positives.

Concluding, baseline anhedonia seems to be a replicable psychological scale that survives Bonferroni-corrected statistical testing and is significantly different between R and NR. Moreover, it confirms the finding that was found in the manuscript. However, baseline anhedonia scores only mildly influence treatment response, as was also indicated in the manuscript. The Non-Cognitive scale was also initially implicated as significantly different between R and NR at baseline. However, a correction for the Anhedonia items, which also comprised the Non-Cognitive scale, refuted this result. The current paper highlights that false positives are relatively easy obtained, emphasizing the importance of future replication or full-sample analyses with corrections for multiple testing.

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

6

A MULTICENTER EFFECTIVENESS TRIAL OF QEEG-*INFORMED* NEUROFEEDBACK IN ADHD: REPLICATION AND TREATMENT PREDICTION

Published as: Krepel, N., Egtberts, T., Sack, A. T., Heinrich, H., Ryan, M., Arns, M. (2020). A multicenter effectiveness trial of QEEG-*informed* neurofeedback in ADHD: Replication and treatment prediction. *NeuroImage: Clinical, 28*, 102399. doi: 10.1016/j.nicl.2020.102399

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

ABSTRACT

Introduction

Quantitative Electroencephalogram-(QEEG-) *informed* neurofeedback is a method in which standard neurofeedback protocols are assigned, based on individual EEG characteristics in order to enhance effectiveness. Thus far clinical effectiveness data have only been published in a small sample of 21 ADHD patients. Therefore, this manuscript aims to replicate this effectiveness in a new sample of 114 patients treated with QEEG-*informed* neurofeedback, from a large multicentric dataset and to investigate potential predictors of neurofeedback response.

Methods

A sample of 114 patients were included as a replication sample. Patients were treated with standard neurofeedback protocols (Sensori-Motor-Rhythm (SMR), Theta-Beta (TBR), or Slow Cortical Potential (SCP) neurofeedback), in combination with coaching and sleep hygiene advice. The ADHD Rating Scale (ADHD-RS) and Pittsburgh Sleep Quality Index (PSQI) were assessed at baseline, every 10th session, and at outtake. Holland Sleep Disorder Questionnaire (HSDQ) was assessed at baseline and outtake. Response was defined as ≥ 25% reduction (R25), ≥ 50% reduction (R50), and remission. Predictive analyses were focused on predicting remission status.

Results

In the current sample, response rates were 85% (R25), 70% (R50), and remission was 55% and clinical effectiveness was not significantly different from the original 2012 sample. Non-remitters exhibited significantly higher baseline hyperactivity ratings. Women who remitted had significantly shorter P300 latencies and boys who remitted had significantly lower IAFs.

Discussion

In the current sample, clinical effectiveness was replicated, suggesting it is possible to assign patients to a protocol based on their individual baseline QEEG to enhance signal-to-noise ratio. Furthermore, remitters had lower baseline hyperactivity scores. Likewise, female remitters had shorter P300 latencies, whereas boys who remitted have a lower IAF. Our data suggests initial specificity in treatment allocation, yet further studies are needed to replicate the predictors of neurofeedback remission.

INTRODUCTION

Neurofeedback is a promising non-pharmacological treatment that has been well investigated in the treatment of ADHD. Neurofeedback can be considered a multi-factorial treatment including components such as reinforcement, coaching and direct feedback on brain-activity, in particular electrical brain activity (electroencephalogram; EEG). Not all EEG frequencies being trained have been shown to be efficacious. For example, training of the posterior alpha rhythm (8 – 13 Hz) has failed to show clinical benefit in either hyperkinetic syndrome (Nall, 1973) and epilepsy (Rockstroh et al., 1993), suggesting some *specificity* in the EEG parameter trained for clinically effective neurofeedback. Therefore, three well-investigated protocols (Sensori-Motor-Rhythm; SMR, Theta-Beta; TBR and Slow Cortical Potential; SCP) have been proposed as 'standard neurofeedback protocols' (Arns, Heinrich, & Strehl, 2014). For these protocols meta-analyses have found support for clinical efficacy rated by parents (Cortese et al., 2016; Van Doren et al., 2019) as well as teachers (Cortese et al., 2016). Foremost, clinical benefit of neurofeedback is maintained – with a tendency for further improvement over time – over 6 – 12 months follow-up periods, approaching clinical benefit obtained with psychostimulant medication

(Van Doren et al., 2019). Holtmann and colleagues (2014) reported that SCP neurofeedback significantly decreased ADHD symptoms, however, when analyses were confined to probably blinded ratings, these effects were reduced to trend-level significance (Holtmann, Sonuga-Barke, Cortese, & Brandeis, 2014). A meta-analysis by Cortese et al. (2016) including also non-standard neurofeedback protocols reports that, when results are confined to probably blinded raters only, a previously significant result (as primarily reported by parents) becomes non-significant (although probably blinded ratings have some limitations, too (Van Doren et al., 2019)). Other studies also report contrasting support for the effectiveness of neurofeedback e.g. no difference between placebo and neurofeedback treatment, suggesting mechanisms of non-specificity (Logemann, Lansbergen, Van Os, Bocker, & Kenemans, 2010). Also, the benefits of neurofeedback for adults are still unclear, with mixed results (Mayer, Blume, Wyckoff, Brokmeier, & Strehl, 2016; Schönenberg et al., 2017). Therefore, the current study also aims to help build upon the body of knowledge currently investigating the effectiveness of neurofeedback. Next steps are: 1) to investigate the clinical effectiveness (also termed 'Clinical Utility'), or the applicability, feasibility, and usefulness of the intervention in practice (American Psychological Association, 2002) and 2) to enhance clinical efficacy of this neurofeedback technique and to identify moderators, mediators, and predictors of remission, which is the primary focus of this manuscript.

In a small proof-of-concept study in 2012 by Arns and colleagues the clinical effectiveness of Quantitative Electroencephalogram (QEEG)-*informed* neurofeedback was reported (Arns, Drinkenburg, & Kenemans, 2012). In its essence, QEEG-*informed* neurofeedback is based on patient assignment to one of the above three 'standard protocols', taking the signal-to-noise ratio from their individual EEG into account. For example, it has been reported that patients with high theta (low beta), and high theta/beta ratio (TBR) respond better to theta/beta neurofeedback (Gevensleben et al., 2009). However, high theta and a high TBR are not present in all children with ADHD but are consistently found in 1/3 of children with ADHD (Arns et al., 2013; Bussalb et al., 2019). Therefore, cases with high theta will be preferentially assigned to TBR-neurofeedback. In addition, the exact

theta frequency band that will be trained is individualized (i.e., 4 – 6 Hz, or 5 – 8 Hz) to increase signal-to-noise ratio and thus the specificity of the feedback. In cases with no clear excess of theta, patients will be treated with SMR or SCP neurofeedback, depending on trainability in that respective frequency band (i.e., in cases of excess 11 – 13 Hz Mu rhythm activity in sensori-motor regions over-lapping with the 12 – 15 Hz SMR band, SCP neurofeedback is preferred over SMR neurofeedback). In this way, virtually every patient will be treated with one of the 'standard neurofeedback' protocols. In addition, a second protocol can be added based on the presence of other EEG hypovigilance markers such as excess frontal alpha (Arns & Kenemans, 2014; Sander, Arns, Olbrich, & Hegerl, 2010) or spindling excessive beta, often associated with impulse control problems (Arns, Swatzyna, Gunkelman, & Olbrich, 2015). Arns, Drinkenburg, and Kenemans (2012) demonstrated that QEEG-*informed* neurofeedback was effective in decreasing ADHD symptoms and, importantly, response rates and effect sizes surpassed those of meta-analyses where one protocol was applied to the whole population (response rate of 76% (≥ 50% symptom reduction)) and large effect sizes for inattention and hyperactivity were observed (Arns, Drinkenburg, & Kenemans, 2012). However, these results still require replication.

In an effort to optimize treatment, predictors, moderators, and mediators of treatment success are often considered. Although not widely studied, some researchers have attempted to identify these. For example, EEG profiles have been proposed as a potential moderator of neurofeedback response in terms of clinical improvement. Specifically, EEG-subtypes may be independent of diagnostic status (Clarke et al., 2011; Clarke, Barry, McCarthy, & Selikowitz, 2001a; Clarke et al., 2003) and preselecting individuals for a particular type of neurofeedback based on their EEG profile may result in greater clinical improvements (Gevensleben et al., 2009). Other results from the NIMH-MTA trial, including three arms of treatment – behavioral, medication, and a combination thereof – found as moderators that youth with ADHD and comorbid anxiety disorder had a better response to behavioral or combined therapy (Hinshaw, Arnold, & Group, 2015). Similarly, the MTA trial also found that those with anxiety and comorbid conduct or oppositional defiant disorder respond-

ed better to combined therapy (Hinshaw et al., 2015). For adults, one medication study found that individuals that were younger, female, and had higher baseline scores had greater clinical improvements (Weiss et al., 2010). As mediators, Hinshaw and colleagues identified that, in combined therapy only, improved parenting skills over the course of treatment was linked to decreased aggressive and disruptive behavior in their children as well as increased social skills (Hinshaw et al., 2015). Interestingly, another study also highlighted the importance of parenting style for successful (combined EEG biofeedback and) methylphenidate treatment (Monastra et al., 2002).

Therefore, the purpose of this study is twofold. Firstly, the aim is to replicate the clinical effectiveness of QEEG-*informed* neurofeedback in clinical practice, as reported by Arns and colleagues in 2012. It was hypothesized that the effectiveness would not deviate significantly in the new sample relative to the 2012 results. Second, baseline clinical as well as neurophysiological variables (EEG and ERP) were examined as moderators, mediators (Baron & Kenny, 1986; Kraemer, Wilson, Fairburn, & Agras, 2002) and predictors of neurofeedback (non-)remission. A recent study by Arns and colleagues found that, in boys only, a lower IAF was indicative of MPH non-response (Arns et al., 2018), so in the current study IAF will also be examined. Analyses will be primarily focused on remission, rather than response. This was done because remission is considered a more clinically relevant endpoint as it implies a loss of diagnostic status (Steele et al., 2006), instead of merely a decrease of symptom presentation, and thus provides a clearer distinction between groups. However, to elucidate the effect of remission versus response, sensitivity analyses were also performed using response as a clinical endpoint, to further crystallize the (potential) differences between the response and remission and potential predictors thereof.

METHODS AND MATERIALS

Participants

The full sample consisted of 136 patients for the first analysis, 115 of which were acquired in the new sample and 21 that were already reported in Arns, Drinkenburg, and Kenemans (2012). This study was an open-label, naturalistic, multi-site study. Given the open-labelled nature of this study, treatment was performed as usual and the analyses were performed post-hoc. Therefore, this study was not reviewed by an independent ethics committee. Patient data were collected from five clinics, two in the Netherlands (neuroCare Group Nijmegen & neuroCare Group The Hague), one in Germany (neuroCare Group Munich) and two clinics in Australia (neuroCare Group Frenchs Forest and neuroCare Group Sydney). Data were collected between August 2008 and May 2018. Patients were screened for inclusion and included in case of an ADHD or ADD diagnosis (as confirmed by the MINI Diagnostic Interview or by a qualified clinician), or when ADHD-RS scores on either scale (ATT or HI) was equal to or higher than 6 (for adults a cut-off of 5 or higher was used, in line with current DSM-5 diagnostic requirements). The ADHD Rating Scale (ADHD-RS, (Kooij et al., 2008)) and the Pittsburgh Sleep Quality Index (PSQI, (Buysse, Reynolds III, Monk, Berman, & Kupfer, 1989)) were obtained at intake, every 10th session, and at outtake. If applicable, the Beck Depression Inventory (BDI-II-NL) and Depression, Anxiety, and Stress Scale (DASS) were assessed at intake, every 10th session, and at outtake as well. All patients signed an informed consent before treatment was initiated. In the case of children younger than 18, caregivers signed the informed consent form. Patients arrived at the clinic referral-based and received (partial) financial support from the government or health insurance, although the majority of treatments was self-paid.

QEEG

QEEG recordings were performed in accordance with the standardized methodology as developed by Brain Resource Ltd. (details of which can be found here (Arns et al., 2016)), of which reliability, validity, and across site-consistency has been published elsewhere (Clark et al., 2006; Paul et al., 2007; Williams et al., 2005). In short, a 26-channel recording based on the 10 – 20 electrode international system using the Quickcap was administered in a standardized room.

Data were referenced to averaged mastoids with a ground at AFz. Horizontal and vertical eye movements were controlled for. Skin resistance was < 10 kΩ for all electrodes. Data were offline corrected for EOG. The sampling rate was 500 Hz for all electrodes. A low pass filter above 100 Hz was applied prior to digitization. The EEG test battery consisted of nine tasks in total, three of which are considered in the current study: a 2-minute Eyes Open (EO) task, a 2-minute Eyes Closed (EC) task, and a 6-minute auditory oddball (ODDB) task.

ERP scoring is thoroughly described by van Dinteren, Arns, Jongsma, and Kessels (2014). ERPs were deduced from the ODDB task, in which a series of high- and low-pitched tones were quasi-randomly presented (the only constraint being that two high-pitched tones cannot occur right after each other), and the patient was asked to press a left- and right-handed button simultaneously at the high-pitched tones. ISI was 1 s. For ERP extraction, windows around the target stimuli of –300 ms to 700 ms were examined. Data were 25 Hz low-pass filtered and baselined to the relative 300 ms pre-stimulus window. Peak components were determined according to maximal response within specific latency intervals. This gave amplitudes and latencies for points N200 and P300 (Arns et al., 2008; Bahramali et al., 1999; Lim et al., 1999; Williams et al., 2005). In this study, the primary focus will be on P300.

IAF determination was based on prior studies (Arns, Drinkenburg, & Kenemans, 2012; Arns et al., 2018) and consisted of the following steps: 1) Fast Fourier Transform to both EO and EC conditions using 2000 ms segment epochs, 2) the difference between EO and EC power spectra was calculated (by subtracting EO from EC) in order to distinguish the alpha power (6 – 13 Hz) by its known suppression from EC to EO, and 3) the IAF was determined by identifying the maximum value between 6 and 13 Hz.

Neurofeedback treatment

Treatment of patients was identical to treatment as reported in 2012 and 2014 by Arns and colleagues (Arns, Drinkenburg, & Kenemans, 2012; Arns, Feddema, & Kenemans, 2014). In short, before treatment was started patients were assessed using the QEEG, through which

the choice for a QEEG-*informed* neurofeedback treatment protocol was derived. In some cases, neurofeedback protocol was adjusted according to the patient's needs. SMR neurofeedback was performed using a 12 – 15 Hz reward at central locations (C3, Cz, or C4). The TBR protocol consisted of a reward in the beta frequency range (e.g., 20 – 25 Hz) at midline sites Fz, FCz, or Cz, in addition to inhibition of theta power. The only difference with the procedure reported in 2012 and 2014, was that in the current sample neurofeedback treatment was complemented with sleep hygiene management and coaching.

The choice for a particular neurofeedback protocol was based on the QEEG assessed during EO and EC:

- Theta/(beta) protocol: when excess frontocentral slowing was observed. Only beta reward if beta was not elevated or beta spindles were not present. Only midline sites (Fz, FCz, Cz).
- SMR/SCP protocol: no clear QEEG deviations and/or sleep problems.
- Low-voltage EEG: SMR/SCP neurofeedback and/or alpha-uptraining during EC at Pz.
- Frontal Alpha protocol: when excess frontocentral alpha (mostly EO) was observed. Beta reward as per Theta/(beta) protocol. Only midline sites (Fz, FCz, Cz); mostly in adult ADHD.
- Beta-downtraining protocol: when beta spindles or excess beta was present, the specific frequency of this excess beta (spindles) was downtrained on the frontocentral site with maximal beta-spindle power.

All protocols employed EMG inhibits, where EMG (55 – 100 Hz) had to be kept below 5 – 10 μV.

Sessions were performed by a master's level psychologist specialized in neurofeedback, trained and accredited by the last author, and took place 2 – 3 times a week. 20 – 30-minute sessions were administered, offered in blocks of five minutes each, with a minimum one-minute break in between blocks. Threshold parameters were set to achieve 25 – 40% effective reinforcement. For SMR treatment, the time-above-threshold was set at 0.2 – 0.5 s. Equipment used to provide visual

and auditory feedback consisted of Brainquiry PET 4.0 (Brainquiry B.V., Nijmegen, the Netherlands) and BioExplorer software (CyberEvolution, Inc., Seattle, USA) for frequency neurofeedback. SCP Neurofeedback was provided using a Theraprax system (neuroConn, Ilmenau, Germany).

Data analysis

ADHD patients were categorized into four groups, according to outtake data, or the last available assessment (Last Observation Carried Forward, LOCF) (based on Arns, Drinkenburg, and Kenemans (2012)):

- *Response* (*R)*: either 25% (R25 (Steele et al., 2006)) or 50% (R50) or more reduction in ADHD-RS Inattention scale (ATT) or Hyperactivity/Impulsivity scale (HYP). Both criteria were used to ensure comparability with other studies (e.g. (Strehl et al., 2017)).
- *Remission*: remission (i.e., loss of diagnostic status) was defined as an ADHD-RS item mean of ≤ 1.00 (Steele et al., 2006; Swanson et al., 2001).
- *Drop-out* (*DO)*: when a patient did not take more than 20 sessions and could not be classified as a responder. In this case, the patient was not included in the analyses.
- *Non-responder* (*NR)*: a patient who had more than 20 sessions and did not meet the criteria for being a responder.

STATISTICS

To estimate the efficacy of QEEG-*informed* neurofeedback as a treatment for ADHD symptomatology, the response rates of the 2019 sample were compared to those of the 2012 sample, using Chi-square statistics. To study possible differences between the 2012 and 2019 sample as well as differences in response for children vs. adults, males vs. females, and protocol specific effects, a repeated measures ANOVA with Time (pre-, halfway-, and postintervention measurements) as a within-subject factor and Sample (2012 and 2019), Sex (female and male), Protocol (SMR, TBR, and other (specifically: SCP and protocols other than SMR/TBR)), and Age group (children and adults) as between-subject factors was performed. Only main effects of Time, Sample, Sex, Age Group and Protocol and interactions with

Time were considered. Lastly, baseline clinical and neurophysiological variables were examined for their value in predicting neurofeedback (non-)remission. In the current study, predictors are defined as variables that are associated with better or worse treatment outcome (followed from Hinshaw et al. (2015), in accordance with Kraemer et al. (2002)). For clinical variables, a GLM Univariate using a potential predictor as a dependent variable, age as a covariate, and Protocol (SMR, TBR, and other), Sex (female and male), and Remission (remission and no remission) were used as between-subject factors, was performed. For neurophysiological variables, the different components of ODDB ERPs were examined. This was done using a repeated measures ANOVA with Site (Fz, Cz, Pz) as a within-subject factor, and Protocol (SMR, TBR, and other), Sex (female and male), and Remission (remission and no remission) as between-subject factor, while covarying for age. A similar approach was taken for IAF, in which case IAF was examined using a repeated measures ANOVA using Site (Fz, FCz, Pz, Oz) as a within-subject factor and Protocol (SMR, TBR, and other), Sex (female and male), and Remission (remission and non-remission) as between-subject factors, while covarying for age. For TBR the same analyses as for IAF were performed, however, in the within-subject factor Site the sites Fz and Cz were used instead of Fz, FCz, Pz, and Oz. TBR analyses were also repeated for SMR and TBR protocols separately, given the probable selection bias because of QEEG-*informed* neurofeedback. Predictors were examined for their predictive utility by performing a discriminant analysis and investigating the receiver operator curve (ROC). A side-track of this study entails a possible association between hyperactivity and sleep breathing problems, based on Vollebregt et al. (2019, June 12). Vollebregt and colleagues found that children with sleep breathing problems exhibited increased levels of hyperactivity. In the current study, this association will be tested by performing a bivariate Spearman correlation between SBD and HYP. Potential mediator/moderator analyses were performed based on non-null findings. Mediator and moderator analyses were performed in accordance with Baron and Kenny (1986) and Kraemer et al. (2002). For mediation to occur, the following criteria should be met: 1) the independent variable should significantly affect the presumed mediator, 2) the presumed mediator should significantly affect the dependent variable, and 3) when

paths described in 1) and 2) are controlled for, the previously significant association between the independent and dependent variable should no longer exist (Baron & Kenny, 1986). Kraemer et al. (2002) added additional requirements for mediation in a clinical setting, being that 1) a mediator should measure a change or event during treatment, 2) the mediator must correlate with treatment choice, and 3) should have a main or interactive effect on the outcome. If mediation analyses were to be performed, partial correlations were run, while controlling for (one of) the potential mediator(s). On the other hand, Kraemer et al. (2002) describe a moderator of treatment efficacy such that a moderator 1) must be gathered at baseline or prior to randomization and 2) explains individual differences in treatment efficacy, meaning that the effect of treatment depends on the value of the moderator (Kraemer et al., 2002). In case of moderator analyses, the individual potential moderators and the interaction between the two ($moderator_A$ * $moderator_B$) were used in a linear regression as independent variables, while the variable of interest was used a dependent variable. In case of mediator analyses, partial correlations were run, correlating two out of three variables of interest, while controlling for the remaining variable. For all predictive analyses, only relevant Remission effects and interactions were considered. Effect sizes reported are Cohen's *d* and were calculated using the following formula: $d = \frac{m1 - m2}{\sqrt{\frac{(s1^2 + s2^2)}{2}}}$

Error bars represent ± 2*SE*. All analyses were performed in IBM SPSS Statistics 25 for Macintosh.

RESULTS

The full sample consisted of 136 patients for the first analysis, 114 (excluding 1 DO and the 21 already reported in Arns, Drinkenburg, and Kenemans (2012)) were included to replicate the initial response to treatment and outcomes were statistically compared to the results of Arns and colleagues in 2012). For further analyses the full sample was used. The demographics of the total sample, the 2019 and the 2012 sample can be found in Table 1 (opposite page). Note: given the clinical focus of the paper, medication usage was not controlled for.

Clinical outcome

Remission and response rates of the current sample (average age: 24.0; range 6 – 68 yrs; 76 males) were 54.8% remission, and 70.4% and 85.2% response for R50 and R25 criteria respectively. This was not significantly different (R50: $\chi^2(1) = 1.428$, $p = 0.232$) relative to the 2012 sample. Given clinical response was the same in both samples, the pooled remission and response rates in the full sample of 136 patients were 57.4% remission and 71.3% and 83.8% response for R50 and R25 criteria respectively.

	Total sample	2019 sample	2012 sample
Age	24.9 (14.9)	24.0 (14.6)	30.0 (16.2)
Number of sessions	32.3 (10.1)	32.0 (8.6)	33.6 (16.1)
Protocol (*n*, (%))			
SMR	84 (61.8)	69 (60.0)	15 (71.4)
TBR	27 (19.9)	25 (21.7)	2 (9.5)
Other	25 (18.4)	21 (18.3)	4 (19.0)
SCP	9	9	0
Males (*n*, (%))	89 (65.4)	76 (66.1)	13 (61.9)
Adults (*n*, (%))	80 (58.8)	66 (57.4)	14 (66.7)
ADHD total	12.4 (3.1)	12.5 (2.9)	11.5 (4.1)
ADHD total post	4.6 (4.6)	4.8 (4.7)	3.6 (3.6)
ADHD Hyperactivity (HYP)	5.5 (2.4)	5.6 (2.2)	4.7 (2.9)
ADHD Hyperactivity post	2.0 (2.3)	2.1 (2.3)	1.3 (2.3)
ADHD Inattention (ATT)	6.9 (1.8)	6.9 (1.7)	6.8 (2.0)
ADHD Inattention post	2.6 (2.7)	2.6 (2.8)	2.3 (2.2)
PSQI	7.7 (4.2)	7.4 (4.1)	9.6 (4.6)
PSQI post	4.6 (3.1)	4.5 (3.0)	5.6 (3.4)

Table 1: *Descriptive statistics for the total sample with means and* (SD)*, and separately for 2019 and 2012 sample. No significant differences were found* ($p \geq 0.055$)*.*

Moderating effects

A repeated measures ANOVA showed a significant effect of Time ($F(2,114) = 48.171$, $p < 0.001$; $d = 1.97$). No other significant interactions or main effects were observed, thus clinical response was not moderated by age-group, sex and neurofeedback protocol and no differences between the 2012 and current sample were found. These effects are visualized in Figure 1 on page 142.

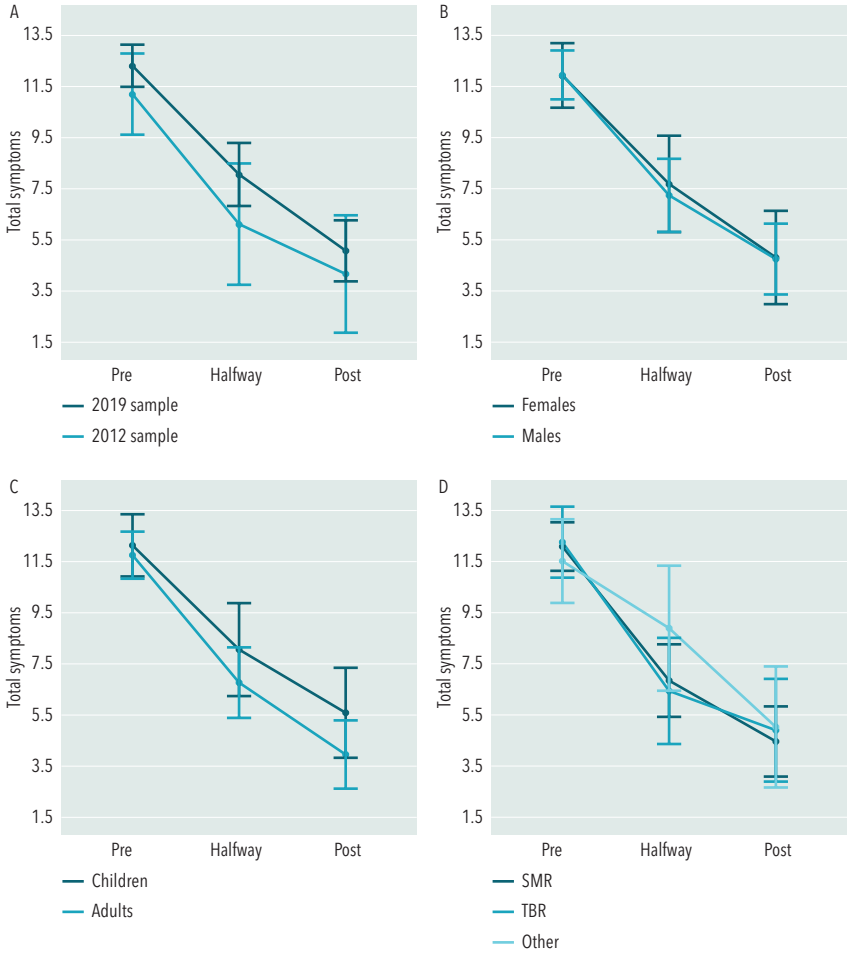


Figure 1: *A repeated measures ANOVA using Sample (2012 v. 2019), Sex (female v. male), Age group (children v. adults), and Protocol (SMR, TBR, other) as between-subject factors. Total ADHD-RS symptoms were used as a within-subject factor (pre-, halfway-, and post-measurements). The error bars represent ± 2*SE*. Analyses showed a significant effect of Time* ($F(2,114) = 48.171$, $p < 0.001$; $d = 1.97$)*, but no other significant interactions or main effects were observed.*

Predictors of neurofeedback (non-)remission

GLM Univariate analyses showed no significant main or interaction effects for ADHD total symptoms, nor for ATT, PSQI total score, HSDQ total score, insomnia, parasomnia, CRSD, hypersomnia, RLS-PLMD, or SBD ($p \geq 0.102$). However, for HYP there was a significant main effect of Remission ($F(1,114) = 5.095$, $p = 0.026$; $d = 0.56$). Thus,

remitters had lower HYP scores at baseline (Figure 2). Using HYP in a discriminant analysis yielded a significant model (p = 0.004; Wilks' Lambda = 0.934; Chi-square = 8.466; df = 1; AUC = 0.635).

For ERP variables, a repeated measures ANOVA showed no significant main or interaction effects for N200 amplitude and latency, and P300 amplitude. For P300 latency, a significant Site X Sex X Remission effect was found (F(1.713,154.200) = 3.235, p = 0.050), and a main effect of Remission (F(1,90) = 5.082, p = 0.027). There also was a significant main effect of Remission X Sex (F(1,90) = 3.958, p = 0.050). Splitting by Sex, in women there was a significant main effect of Remission (F(1,25) = 5.570, p = 0.026; d_{Fz} = 0.87, d_{Cz} = 0.85, d_{Pz} = 0.51; Figure 3), yet for men no such effect was observed. Using P300 latency at Fz in a discriminant analysis yielded a significant model (p = 0.025; Wilks' Lambda = 0.844; Chi-square = 5.007; df = 1; AUC = 0.743). Thus, female remitters had shorter P300 latencies.

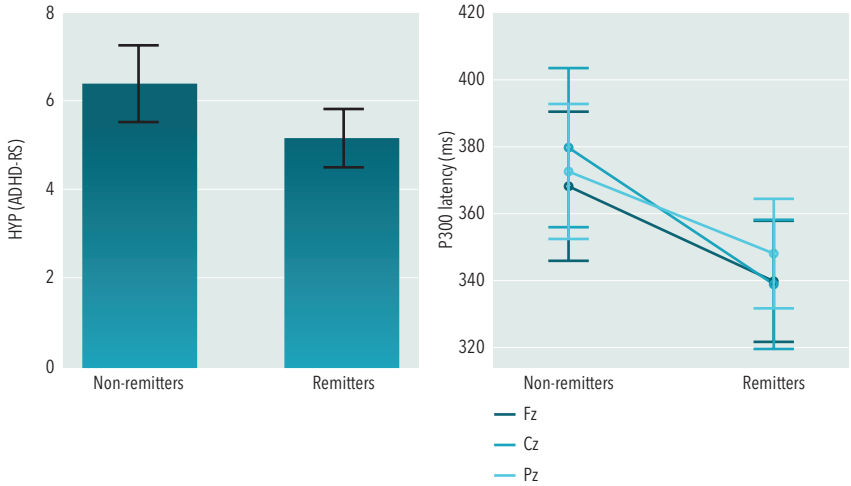


Figure 2 (left): *Bar graph of HYP scores, separated for remitters and non-remitters. A GLM Univariate analysis showed a significant main effect of Remission* (F(*1,114*) = *5.095,* p = *0.026;* d = *0.56).*

Figure 3 (right): *P300 latencies separated by remission. A repeated measures ANOVA showed that female remitters had a significantly shorter P300 latency* (F(*1,25*) = *5.570,* p = *0.026;* d_{Fz} = *0.87,* d_{Cz} = *0.85,* d_{Pz} = *0.51).*

For TBR, no significant results were obtained. Thus, TBR was not related to remission.

For IAF analyses, a repeated measures ANOVA yielded no significant results. Based on earlier work (Arns et al., 2018) and a directed hypothesis, the analysis was repeated in a selected sample of boys (average age: 11.1; age range: 6 – 18) only. The resulting sample consisted of 45 boys, three of which were excluded based on missing data (21 remitters, 21 non-remitters). A One-Way ANOVA showed no significant Age difference between remitters and non-remitters ($F(1,40) = 1.244$, $p = 0.271$). A repeated measures ANOVA using only Remission as a between-subject factor yielded a significant main effect of Remission ($F(1,37) = 4.534$, $p = 0.040$; $d_{Fz} = 0.78$, $d_{FCz} = 0.68$, $d_{Pz} = 0.42$, $d_{Oz} = 0.66$). Using IAF at Fz in a discriminant analysis yielded a significant model ($p = 0.019$; Wilks' Lambda = 0.863; Chi-square = 5.508; $df = 1$; AUC = 0.694). The IAF for remitters and non-remitters for Fz was 8.7 Hz vs. 9.7 Hz, respectively. This can be observed in Figure 4. This indicates that, in the group of boys only, remitters had a lower mean IAF.

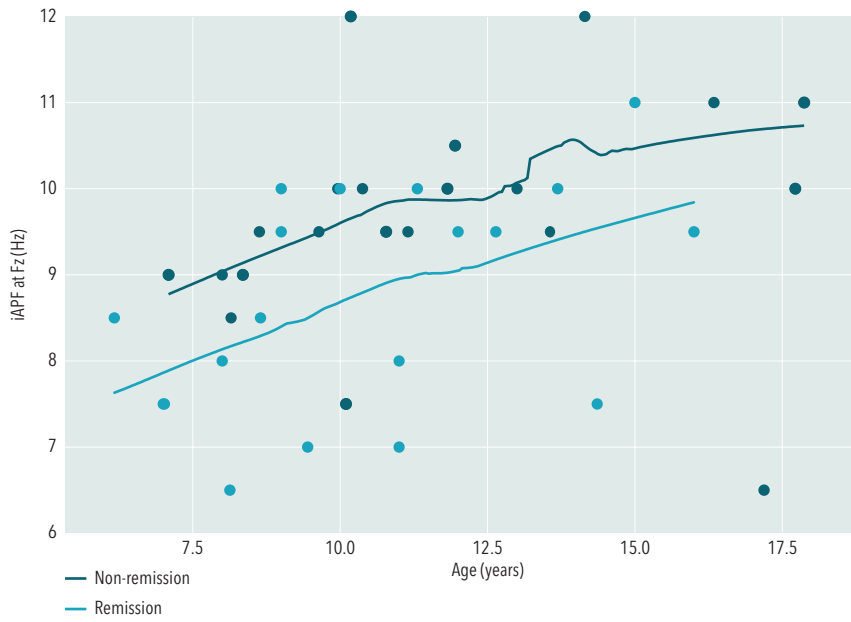


Figure 4: *A Loess-fit for IAF and Age, separated for Remission and Non-remission, for boys only. A repeated measures ANOVA showed a significant main effect of Remission* ($F(1,37) = 4.534$, $p = 0.040$; $d_{Fz} = 0.78$, $d_{FCz} = 0.68$, $d_{Pz} = 0.42$, $d_{Oz} = 0.66$).

Concluding, apart from HYP at baseline, no other clinical variables served as predictor for neurofeedback (non-)remission. From the ERP analyses P300 latency for women emerged as a predictor, however, no other components of P300 showed to be useful. TBR and IAF both showed to be not useful in predicting neurofeedback remission in the full sample. However, a subsample analysis showed a significant result for boys, where a slow IAF was associated with remission.

Post-hoc exploratory analysis

Based on the above results and earlier work indicating an association between HYP and SBD (Vollebregt et al., 2019, June 12), specific relations between variables were explored to further elucidate the direction of effects for HYP as a predictor.

In Vollebregt et al. (2019, June 12) a relation between hyperactivity and SBD in children was found, thus suggesting hyperactivity symptoms can be caused by SBD, and thus the association between SBD and reduced clinical response, could be mediated by the presence of HYP. To test this further in the current sample (children only), a bivariate Spearman correlation between HYP and HSDQ SBD was performed. A significant correlation was found ($r(35) = 0.353$, $p = 0.038$; $r^2 = 12.4\%$). Similarly, a bivariate Pearson correlation between HYP and clinical response showed to be significant ($r(50) = 0.314$, $p = 0.026$; $r^2 = 9.9\%$). However, a bivariate Pearson correlation between SBD (LOG-transformed) and clinical response was non-significant ($r(32) = 0.036$, $p = 0.844$; $r^2 = 0.1\%$). Because of the directionality and assumed working mechanism between SBD, HYP, and remission, a mediation analysis was performed. A partial correlation between HYP and clinical response, while controlling for SBD (LOG-transformed), yielded a significant correlation ($r(29) = 0.377$, $p = 0.037$; $r^2 = 14.2\%$). A partial correlation using SBD (LOG-transformed) and clinical response also showed a non-significant association ($r(29) = -0.076$, $p = 0.684$; $r^2 = 0.6\%$), leaving both associations unchanged. It was also tested whether children with or without SBD complaints had different outcomes on ATT or HYP %change. A Mann-Whitney *U* using SBD (with or without complaints) as a between-subject factor and ATT and HYP %change as dependent variables was performed. This yielded no significant results for ATT ($Mdn_{\text{remitters}} = 100.0$, $Mdn_{\text{non-remitters}} = 25.0$,

U = 127.5, z = −0.754, p = 0.451), nor for HYP ($Mdn_{\text{remitters}}$ = 100.0, $Mdn_{\text{non-remitters}}$ = 37.5, U = 120.5, z = −0.995, p = 0.320). Thus, even though that SBD seems to be related to hyperactivity and hyperactivity seems to be related to remission, there seems to be no interaction between HYP and SBD.

Sensitivity analyses
Given the overrepresentation of the SMR protocol in the current sample and the primary focus on remission, further analyses were performed to investigate the specificity of the obtained results. That is, focusing only on the significant results obtained in the main manuscript, analyses were repeated using Response (50%) as a between-subject factor instead of Remission. Analyses were also performed in the SMR group alone. The performed analyses are identical to the above. In the SMR-specific analyses, Remission was used as a between-subject factor. Details of the analyses can be found in the supplement.

Summarizing the results in the supplement, HYP and P300 did not emerge as predictors of non-response. For IAF, the difference was not significant, albeit the direction of the result was the same and the effect size was similar to the one observed in the main manuscript. The SMR analyses showed no significant effects, but the directions of the effects and effect sizes were similar to the ones observed in the main manuscript.

DISCUSSION

This paper aimed to replicate the clinical effectiveness of QEEG-*informed* neurofeedback, as reported in Arns, Drinkenburg, and Kenemans (2012). Also, potential moderators, mediators, and baseline behavioral and neurophysiological variables as predictors were examined of neurofeedback remission.

Clinical effectiveness of QEEG-*informed* neurofeedback was replicated, meaning that the current response and remission rates were not significantly different from those reported in 2012. Furthermore, hyperactivity emerged as a potential predictor of neurofeedback non-remission, specifically, non-remitters had higher baseline hyperactivity

scores. Additionally, females who had a faster P300 latency were more likely to be remitters, whereas boys who remitted had lower IAF as compared to those who did not. Lastly, SBD seemed to be significantly related to hyperactivity, however, hyperactivity does not seem to mediate the association between remission and SBD.

The effectiveness of this study yielded equal or larger effect sizes as reported by a meta-analysis that focused on neurofeedback randomized controlled trials (Cortese et al., 2016), and demonstrates similar remission rates and effect sizes compared to the NIMH-MTA Medication Management treatment (MTA Cooperative Group, 1999). While the design of the current study was an open-label trial, it provides important information regarding effectiveness or 'Clinical Utility' meaning the applicability, feasibility, and usefulness of the intervention in clinical practice. This construct is designed to assess the generalizability of the intervention into everyday clinical practice (American Psychological Association, 2002). For example, when considering clinical efficacy for methylphenidate in the treatment of ADHD as established in the MTA trial, remission rates of 56 – 68% were reported for the medication arms, while the results of the large international multicenter iSPOT-A effectiveness study yielded a 31% remission rate and a 33% smaller effect size for effectiveness obtained in clinical practice. Furthermore, in a study where the MTA medication algorithm was followed, a 44% smaller effect size was reported (Gelade et al., 2018), illustrating that clinical utility is equally important in the consideration of generalizability of clinical effects into clinical practice. Therefore, this study demonstrates that, across the five clinics involved, the effectiveness of neurofeedback translates well into practice. Potential reasons as to why the current study found greater effect sizes include the assumed specificity of QEEG-*informed* neurofeedback and the targeted frequency band and the increased emphasis on sleep hygiene management, however, the exact reasons should be investigated in further controlled studies.

This study suggests that non-remitters were characterized by higher hyperactivity scores at baseline, albeit this finding was not found in the sensitivity analysis for response (R50). This is most likely due to the definition of remission, requiring full symptom resolution in

absolute terms (item mean ≤ 1.0) opposed to response, which is a relative metric, and thus less sensitive to initial severity. The current result is in line with the notion symptoms of hyperactivity may be less sensitive to the effects of neurofeedback (Arns, de Ridder, Strehl, Breteler, & Coenen, 2009; Holtmann et al., 2014). Our results further indicated that SBD was significantly related to hyperactivity, and hyperactivity was associated with non-remission, yet SBD was not related to remission. Hyperactivity did not seem to act as a mediator in this working mechanism. This is in line with Chervin and Archbold (2001), who found that children with or without SBD scored equally high on hyperactivity. However, a recent meta-analysis found that people presenting symptoms of SBD are at an increased risk of developing complaints of inattention and hyperactivity, and therefore it is argued that people showing ADHD complaints should be screened for SBD (albeit the age groups only concerned children and adolescents and the overall effect showed a medium effect size (Hedges's g = 0.57) (Sedky, Bennett, & Carvalho, 2014).

For P300, prior studies have primarily focused on P300 amplitude, whereas P300 latency is less well studied. Although the majority of studies generally concern a small sample size, results seem to converge on a prolonged P300 latency in children with ADHD (Sanfins et al., 2017; Sunohara, Voros, Malone, & Taylor, 1997; Tsai, Hung, & Lu, 2012; Yamamuro, Ota, Iida, Nakanishi, Suehiro, et al., 2016), albeit support for this seems to be less clear in adults (Szuromi, Czobor, Komlósi, & Bitter, 2011). Interestingly, a recent study by Chi and colleagues found that parents with ADHD offspring had longer P300 latencies (Chi et al., 2019). P300 latency deviances may also not solely occur in ADHD (e.g., (Degabriele & Lagopoulos, 2009; Gao & Raine, 2009; Qiu, Tang, Chan, Sun, & He, 2014; Simons et al., 2011)). An important role in the presentation of P300 latency and amplitude is age, specifically, around the age of 16 P300 amplitude tends to decrease, whereas the latency tends to increase after the age of 22 (van Dinteren et al., 2014). Yet, since in P300 analyses age was used as a covariate, it is not expected that age might explain the current results. Regarding prognostics, normalization of ERP variables after pharmacological treatment has been reported (Ozdag, Yorbik, Ulas, Hamamcioglu, & Vural, 2004; Yamamuro, Ota, Iida, Nakanishi, Matsuura, et

al., 2016), yet P300 has not yet been evaluated as a predictor per se. As to why the current effect was specifically observed in women is not entirely clear. Some sex differences have been reported (Bakos et al., 2016; Nanova, Lyamova, Hadjigeorgieva, Kolev, & Yordanova, 2008), but a recent systematic review showed that the effect of sex on P300 latency is minimal to none (Melynyte, Wang, & Griskova-Bulanova, 2018). Also, sex specific concerns in the presentation of ADHD may be considered (Nussbaum, 2012). Importantly, given the above variance in available literature, this effect may be spurious and therefore requires thorough further investigation and replication.

Interestingly, in the current sample, boys who remitted to neurofeedback exhibited a lower frontal IAF, whereas in Arns et al. (2018) the opposite was found for treatment with methylphenidate. These results may indicate frontal IAF as a stratification biomarker to stratify, or differentially assign boys between two effective treatments (in this case low IAF implicates neurofeedback and high IAF implicates MPH), given the opposite association. However, further studies will need to prospectively test and replicate this as a possibility to further optimize and individualize ADHD treatments.

Concluding, the clinical effectiveness of QEEG-*informed* neurofeedback was replicated, and clinical benefit was the same for males vs. females, children vs. adults and irrespective of the protocol used. Hyperactivity, IAF, and P300 may serve as potential predictors of neurofeedback (non-)remission, although these findings still need to be replicated and tested for robustness.

Limitations

This study was based on a naturalistic, open-label design. While this can be viewed as a strength of the study (effectiveness, results translate into clinical practice), this is also a weakness of the study, since no control condition was used and effect sizes obtained are sometimes higher in such designs. This also means that potential non-specific mechanisms subjective to treatment as usual (e.g., structured environment, regular intervals of training) may have impacted clinical efficacy and thereby the current results. Future, randomized controlled studies should further investigate the added effect of assigning people

to an individualized neurofeedback protocol, such as the QEEG-*informed* neurofeedback presented here. Furthermore, patients in this study received treatment as usual, that included additional coaching and managing of sleep hygiene based on the patient's individual needs. Also, medication usage was not controlled for in the current analyses. Importantly, the majority of the current sample had already sought treatment for ADHD symptoms, yet had insufficient relief from their symptoms and therefore sought additional treatment options. Of the total (N = 136) sample, 43 patients did not use any medication at all. The remaining part used a combination of stimulant medication, sleep medication (among which melatonin), benzodiazepines, and antidepressant medication. To investigate potential medication effects, post-hoc analyses were repeated on the sample free of medication. The direction of the results remained unchanged, however, some of the results did not reach significance. Note that sample sizes were significantly reduced given the restriction of no medication, thereby complicating interpretation. Another limitation is that this study only considered baseline clinical and neurophysiological data. This means that changes in clinical assessment may have been the result of neurophysiological changes due to neurofeedback treatment (or vice versa). Indeed, Arns, Drinkenburg, and Kenemans (2012) found in their initial study that, after SMR treatment, P300 amplitude had increased and SMR power had decreased. Yet, this sample size was small and the current study does not have the necessary post EEG measurements to test this question. Future, well-powered studies entailing post EEG's should focus on this issue. Another issue (although perhaps not a limitation per se) is that the Contingent Negative Variation (CNV) was not considered in this study. The CNV was not considered because the SCP neurofeedback sample was small (n = 9) and, given that several studies have found the effect in SCP neurofeedback (Gevensleben et al., 2014; Heinrich, Gevensleben, Freisleder, Moll, & Rothenberger, 2004; Mayer et al., 2016), unsuitable for data analysis. Similarly, the CNV is typically extracted at more than 1000 ms after cue onset. Given that the oddball paradigm used in this study had an ISI of 1000 ms, this paradigm was unsuitable for CNV extraction. However, some studies have shown that the CNV shows potential to be used in clinical practice, and therefore future studies may investigate this issue further. Lastly, even though the total sample is 136 and thus sufficiently

statistically powered, zooming in on subgroups resulted in a substantial decrease in sample size, resulting in the smallest sample size of 11 (females, children).

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENTS

We would like to thank all neuroCare Group therapists, located in The Netherlands, Germany, and Australia, for support and collecting the data used in this study. In particular, we would like to thank, David Binnie (neuroCare Group Sydney), and Annika Simlacher (neuroCare Group München) for providing the data, and Rosalinde van Ruth, Iris van Oostrom, Vera Kruiver, Myrthe van Eerdt, Joris van Neijenhof, Renée Rouwhorst, and Wilco Jansen for collecting the data. We also acknowledge the support of Chris Spooner (Total Brain, Sydney, Australia) and Donna Palmer (Total Brain, Sydney, Australia) with the EEG and ERP data.

DISCLOSURES

MA is unpaid research director of the Brainclinics Foundation, a minority shareholder in neuroCare Group (Munich, Germany), reports options from Brain Resource (Sydney, Australia); and is a co-inventor on 4 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 Brainclinics received research funding from Brain Resource (Sydney, Australia), UrgoTech (Paris, France) and neuroCare Group (Munich, Germany), and equipment support from Brainsway, Deymed, neuroConn and Magventure.

MR is a shareholder in neuroCare Group.

SUPPLEMENTARY MATERIALS

Sensitivity analyses

Given the overrepresentation of the SMR protocol in the current sample and the primary focus on remission, further analyses were performed to investigate the specificity of the obtained results. That is, analyses were repeated using Response (50%) as a between-subject factor instead of Remission, and the analyses were also performed in the SMR group alone.

Effect sizes reported are Cohen's *d* and were calculated using the following formula: $d = \frac{m1 - m2}{\sqrt{\frac{(s1^2 + s2^2)}{2}}}$

Error bars represent ± 2*SE*. All analyses were performed in IBM SPSS Statistics 25 for Mac.

In a univariate analysis using Response (response and non-response) as a between-subject factor, age as a covariate, and HYP as a dependent variable, a non-significant main effect of Response was observed (Figure S1a (opposite page); $F(1,130) = .405$, $p = .526$; $d = -.10$). This is in contrast with the obtained finding in the main manuscript. For P300 latency, focusing on women only, a repeated measures ANOVA using Response (response and non-response) as a between-subject factor, age as a covariate, and Site (Fz, Cz, Pz) as a within-subject factor, showed no significant main effect of Response (Figure S1b; $F(1,30) = .039$, $p = .845$; $d_{Fz} = .16$, $d_{Cz} = -.21$, $d_{Pz} = -.14$), potentially indicating a severity-specific effect of P300 latency. For IAF analyses, confining to boys only, a repeated measures ANOVA using Response (response and non-response) as a between-subject factor and IAF (Fz, FCz, Pz, Oz) as a within-subject factor yielded a non-significant effect of Response (Figure S1c; $F(1,40) = 2.512$, $p = .121$; $d_{Fz} = .44$, $d_{FCz} = .29$, $d_{Pz} = .63$, $d_{Oz} = .57$), albeit the direction of the effect is similar.

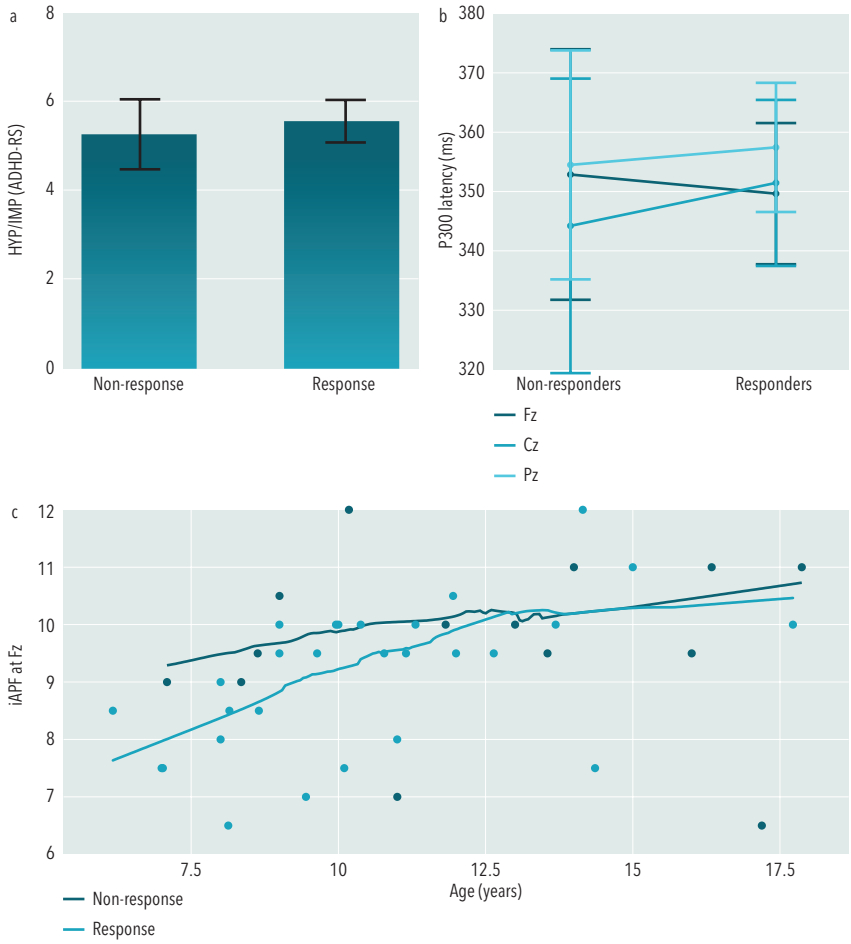


Figure S1a: *bar graph of HYP scores at baseline, separated by remission. A GLM Univariate showed that there was no significant difference between responders and non-responders* ($F(1,130) = 0.405$, $p = .526$; $d = -0.10$). ***S1b:*** *P300 latencies (Fz, Cz, Pz) for females only, separated by response. A repeated measures ANOVA showed a non-significant main effect of P300 latency* ($F(1,30) = 0.039$, $p = .845$; $d_{Fz} = 0.16$, $d_{Cz} = -0.21$, $d_{Pz} = -0.14$). ***S1c:*** *a Loess-fit for IAF and Age, separated for response and non-response, for boys only. A repeated measures ANOVA showed a non-significant main effect of Response* ($F(1,40) = 2.512$, $p = 0.121$; $d_{Fz} = 0.44$, $d_{FCz} = 0.29$, $d_{Pz} = 0.63$, $d_{Oz} = 0.57$).

Next, sensitivity analyses were performed in SMR group only. A GLM univariate using Remission (remission and no remission) as between-subject factors, age as a covariate, and HYP as a dependent variable yielded a significant effect of Remission (Figure S2a (page 154); $F(1,76) = 5.237$, $p = .025$; $d = .64$). In the case of P300 analyses, a re-

peated measures ANOVA using Remission (remission and no remission) as a between-subject factor and Site (Fz, Cz, and Pz) as a within-subject factor was performed in females only. A non-significant main effect of Remission was obtained (Figure S2b; $F(1,19) = 1.659$, $p = .213$; $d_{Fz} = .64$, $d_{Cz} = .61$, $d_{Pz} = .16$), although the direction of the effect remained the same. Corroborating the result of a lower IAF in boys who remit, a repeated measures ANOVA using Remission as a between-subject factor and Site (Fz, FCz, Pz, and Oz) as a within-subject factor showed a non-significant main effect of Remission (Figure S2c; $F(1,20) = 2.665$, $p = .118$; $d_{Fz} = .95$, $d_{FCz} = .82$, $d_{Pz} = .41$, $d_{Oz} = .66$), albeit the direction of the effect was identical to the one found in the main manuscript with a large effect size.

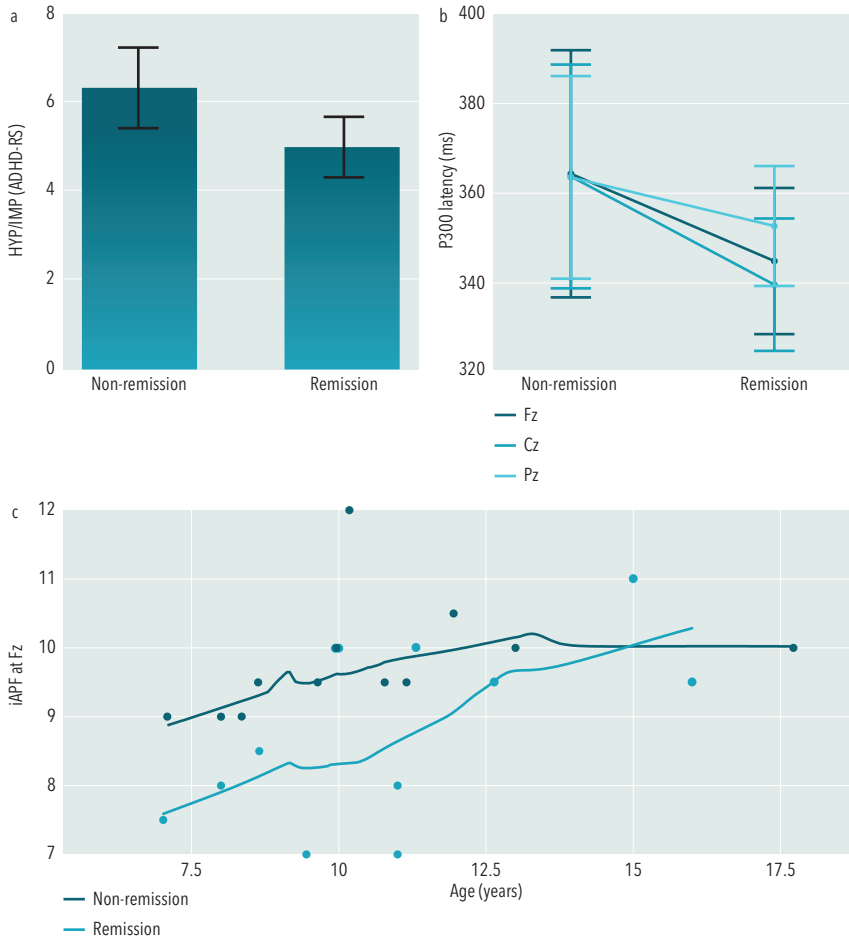


Figure S2a* (*opposite page*)**: *bar graph of HYP scores at baseline, separated by remission. A GLM Univariate showed that there was a significant difference between remitters and non-remitters* ($F(1,76) = 5.237$, $p = 0.025$; $d = 0.64$). ***S2b: *P300 latencies (Fz, Cz, Pz) for females only, separated by remission. A repeated measures ANOVA showed a non-significant main effect of P300 latency* ($F(1,19) = 1.659$, $p = 0.213$; $d_{Fz} = 0.64$, $d_{Cz} = 0.61$, $d_{Pz} = 0.16$). ***S2c:*** *a Loess-fit for IAF and Age, separated for remission and non-remission, for boys only. A repeated measures ANOVA showed a non-significant main effect of Remission* ($F(1,20) = 2.665$, $p = 0.118$; $d_{Fz} = 0.95$, $d_{FCz} = 0.82$, $d_{Pz} = 0.41$, $d_{Oz} = 0.66$).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2020.102399.

7

TO SPINDLE OR NOT TO SPINDLE: A REPLICATION STUDY INTO SPINDLING EXCESSIVE BETA AS A TRANSDIAGNOSTIC EEG FEATURE ASSOCIATED WITH IMPULSE CONTROL

Submitted as: Krepel, N., van Dijk, H., Sack, A. T., Swatzyna, R. J., and Arns, M. To Spindle or Not to Spindle: A Replication Study Into Spindling Excessive Beta as a Transdiagnostic EEG Feature Associated with Impulse Control (under review).

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

ABSTRACT

Background

Frontocentral spindling excessive beta (SEB), a spindle-like beta-activity observed in the electroencephalogram (EEG), has been transdiagnostically associated with more problems with impulse control and sleep maintenance. The current study aims to replicate and elaborate on these findings.

Methods

Individuals reporting sleep problems (n = 45) or attention-deficit/hyperactivity disorder (ADHD) symptoms (n = 105) were included. Baseline ADHD-Rating Scale (ADHD-RS), Pittsburgh Sleep Quality Index (PSQI), Holland Sleep Disorder Questionnaire (HSDQ), and EEG were assessed. Analyses were confined to adults with frontocentral SEB.

Results

Main effects of SEB showed more impulse control problems (d = .87) and false positive errors (d = .55) in individuals with SEB. No significant associations with sleep or interactions with Sample were observed.

Discussion

This study partially replicates an earlier study and demonstrates that individuals exhibiting SEB report more impulse control problems, independent of diagnosis. Future studies should focus on automating SEB classification and further investigate the transdiagnostic nature of SEB.

INTRODUCTION

Beta spindles or 'spindling excessive beta' (SEB), conceptualized as *"High frequency beta with a spindle morphology, often with an anterior emphasis"* (Johnstone, Gunkelman, & Lunt, 2005, p. 101), have not yet been thoroughly studied. An early study by Kubicki and Ascona (1983) described the presence of beta bursts over the frontal areas with a frequency ranging between 25 – 35 Hz and a maximum amplitude of 30 μV, and suggested these were reflective of sub-vigil beta or hypoarousal. A later study identified the presence of frontal excess beta in children diagnosed with ADHD and considered these to reflect an atypical ADHD group. The authors found that the group presenting SEB is characterized by higher levels of moodiness and proneness to temper tantrums (Clarke, Barry, McCarthy, & Selikowitz, 2001b). Studies suggest that SEB occurs in 13 – 20% of ADHD patients (Chabot & Serfontein, 1996; Clarke et al., 2001b), although similar percentage rates of individuals with frontocentral SEB have been reported in children with and without ADHD (Arns et al., 2008). Interestingly, high beta activity has usually been associated with hypervigilance. For example, it has been reported that individuals with complaints of insomnia show elevated levels of beta activity (based on absolute or relative power) around sleep onset (Perlis, Mer-

ica, Smith, & Giles, 2001; Perlis, Smith, Andrews, Orff, & Giles, 2001), possibly explained by central nervous system hyperarousal (Perlis, Merica, et al., 2001). Interestingly, individuals with insomnia as their primary complaint showed higher beta/gamma power at non-rapid eye movement (NREM) stages of sleep, whereas individuals who reported no sleep issues, or individuals whose complains of insomnia were secondary to depression, showed no such increases (Perlis, Smith, et al., 2001). This suggests that beta activity is positively associated with arousal, such that increased beta translates to increased arousal. Yet, some studies have reported findings that challenge this view. Strijkstra, Beersma, Drayer, Halbesma, and Daan (2003) found a positive association between frontocentral beta-2 (23 – 29 Hz) power and subjective sleepiness. Also, a study by Greneche et al. (2008), in which EEG was measured during a 24-hours sustained wakefulness period, found that individuals with obstructive sleep apnea (OSA) had increased waking delta, theta, and beta power compared to healthy controls. Interestingly, only in healthy individuals a negative association between alertness and beta power (among other bands) during this time period was found (Greneche et al., 2008). Another study (of which the sample only consisted of males) that focused on EEG changes in response to sleep deprivation reported increased beta power at central sites and, interestingly, beta power correlated positively with hours of wakefulness (Lorenzo, Ramos, Arce, Guevara, & Corsi-Cabrera, 1995). This leaves the question whether different types of beta serve different purposes; it has been suggested that desynchronized beta is related to hyperarousal and synchronized SEB is related to hypoarousal (Arns, Swatzyna, et al., 2015). This distinction can also be seen in a study on children diagnosed with ADHD and excess beta who present a degree of hypoarousal similar to excess theta (Clarke et al., 2013). The distinction in beta can also be seen in drug symptomatology. Benzodiazepines increase beta activity and are also known for their sedating effect (Blume, 2006). A recent animal study also highlighted that higher beta oscillations (15 – 35 Hz) behave differently depending on the animal's state (active wake or quiet wake) in which they are observed (Gronli, Rempe, Clegern, Schmidt, & Wisor, 2016). Challenging the view that beta has a unidimensional relationship with arousal, these findings open up doors to a more dynamic interpretation of beta activity.

Some studies suggest a genetic contribution of beta activity. A link between GABA-A receptor genes and beta power (subdivided in different frequency bins) was previously reported (Porjesz et al., 2002). Also, Zietsch et al. (2007) found support for heritability of power across different frequency bands, including beta, in a twin study. Given these findings, genetics may also influence the presence of SEB. A genetic component to the presence of SEB has been proposed by Kubicki and Ascona (1983), and Vogel (1970) observed potential support for an autosomal dominant mode of inheritance in family studies.

In 2015, using a Research Domain Criteria (RDoC) approach (Cuthbert & Insel, 2013), Arns and colleagues (2015) investigated SEB in relation to hyperactivity/impulsivity and sleep problems. It was found that problems with sleep maintenance and impulse control were higher in patients with frontocentral SEB. Importantly, the presence of SEB was not associated with having trouble falling asleep (Arns, Swatzyna, et al., 2015). The authors concluded that SEB may be regarded as a state marker, caused by sleep maintenance problems, and in turn be associated with more hyperactivity/impulsivity complaints (possibly as a vigilance-autostabilization behavior related to low vigilance (Arns, Gunkelman, Olbrich, Sander, & Hegerl, 2011; Arns & Kenemans, 2014)). However, these results still require replication and elaboration, which are the main aims of this manuscript. First, it will attempt to replicate the findings as reported by Arns and colleagues (2015). This will be done using a mixed dataset, consisting of clients reporting primary sleep problems or symptoms of ADHD. It was hypothesized that the presence of SEB is associated with complaints regarding impulse control and sleep maintenance. It was also expected that this association would be transdiagnostic and thus would be equally present in both the insomnia and ADHD groups (Arns, Swatzyna, et al., 2015).

METHODS AND MATERIALS

For both datasets the following assessments were conducted at baseline: ADHD-Rating Scale (ADHD-RS), Pittsburgh Sleep Quality Index (PSQI), Holland Sleep Disorder Questionnaire (HSDQ), and QEEG. Informed consent was obtained for all participants. Impulse control problems were operationalized identically to the original study using

items from the ADHD-RS. In addition, neuropsychologically-defined impulse control problems, measured by the amount of false positive errors on a Continuous Performance Task (CPT), were analyzed. The PSQI and HSDQ were systematically collected (different from the measurements in the original study). These scales were chosen to refine the associations with sleep maintenance problems since these questionnaires are well validated, in contrast to the three distinct items (part of the generic 300-item-screening questionnaire (CNC1020; EEG Professionals, The Netherlands)) used in the original study.

Dataset 1: Insomnia
Baseline EEG and behavioral data were gathered for an ongoing naturalistic, open-labelled study investigating the effects of SMR neurofeedback on sleep. The study included only patients that had a primary sleep problem and excluded any participants with primary psychiatric comorbidities that explained the sleep problem. The sample included patients between 18 – 65 years of age with a primary insomnia problem expressed as a sleep onset problem (latency (SOL) ≥ 30 minutes), sleep maintenance problem (wake after sleep onset (WASO) ≥ 30 minutes) or sleeping ≤ 6 hours per night. The sleep complaints should occur at least three times per week and be present for at least six months at time of intake. Medication usage was allowed if stable during the treatment. Exclusion criteria were comorbid medical or psychiatric complaints (as assessed using the MINI), recent parenthood, night shifts, students, pregnancy, excessive alcohol or caffeine usage, and diagnosis of a primary sleep disorder other than primary insomnia.

Dataset 2: ADHD
The ADHD sample was previously published in Krepel et al. (2020; Chapter 6) in an open-labelled, naturalistic multi-site study. Data were gathered at two different clinics specialized in neuromodulation treatment (neuroCare Group Nijmegen & neuroCare Group The Hague, The Netherlands).

QEEG
QEEG recording details were previously described elsewhere (e.g., Arns et al. (2016)) and were performed in accordance with the standardized methodology developed by Brain Resource Ltd., of which re-

liability and validity are published elsewhere (Clark et al., 2006; Paul et al., 2007; Williams et al., 2005). In short, using a 26-electrode EEG cap recording was performed based on the 10 – 20 international system. Data were referenced to averaged mastoids with a ground at AFz. Horizontal and vertical eye movements were controlled for, and skin resistance was < 10 kΩ for all electrodes. The sampling rate was 500 Hz. Prior to digitization, a low-pass filter of 100 Hz was applied. Data were corrected offline for EOG. Three tasks are recorded during the EEG: a 2-minute Eyes Open (EO), a 2-minute Eyes Closed (EC), and a 6-minute CPT. In the CPT, 125 letters were presented with an ISI of 2.5 seconds. Clients were asked to detect the occurrence of two consecutive identical letters. During the CPT, clients were asked to press the two buttons simultaneously (one under the left index finger and one under the right index finger).

STATISTICS

To determine the presence of SEB, the QEEGs of all clients were visually examined by the first and last author of this manuscript (NK and MA), blinded to diagnosis and behavioral scores. SEB presence was determined consistent with the definition proposed by Johnstone et al. (2005): *"High frequency beta with a spindle morphology, often with an anterior emphasis"* (Johnstone et al., 2005, p. 101) as well as the morphology as published by Clarke et al. (2001b). Both the raw EEG and quantitative EEG were inspected for SEB presence, and if applicable, peak frequency and maximum site of SEB were identified. The raw EEG was used for initial inspection, and the quantitative EEG was used to verify SEB presence using the following criteria: SEB should a) be in excess based on Z-scores, b) be present in the beta band (confined to 15 – 40 Hz), c) match the site of the observed SEB to the topography of the deviating Z-scores. Then, subjects were divided according to SEB presence: Category 0 (no SEB present), Category 1 (fast synchronous beta regularly present without a clear spindle morphology), or Category 2 (SEB present). These categories are in line with Clarke et al. (2001b), consisting of normal amplitude excess beta, high amplitude excess beta, and excess beta with frontal beta spindles. Examples of these three groups can be found in Figure 1 on page 164.

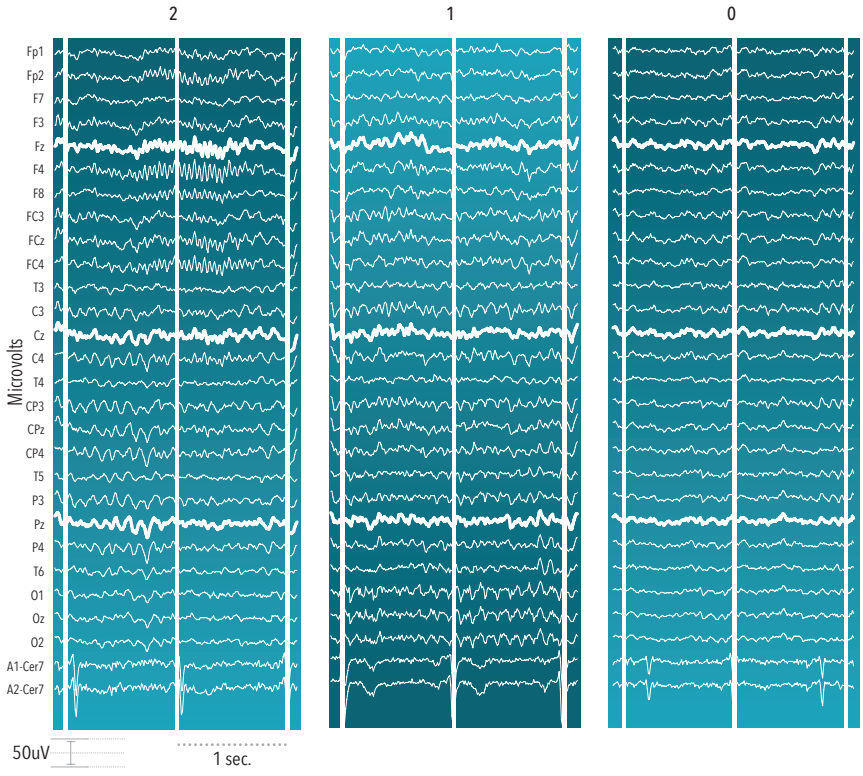


Figure 1: *Representative examples of what would be considered SEB (Category 2), synchronous beta (Category 1), and No-SEB (Category 0). The individual in category 2 shows SEB primarily in electrodes Fp2, Fz, F4, F8, and FC4 (to a lesser extent, SEB can also be observed in electrodes F3, FC3, FCz, Cz, and C4). This individual had a peak frequency at 22 Hz and the main site of SEB was identified at electrode Fz. Category 1 shows synchronous beta in electrodes F3, Fz, F4, FC3, FCz, C3, CP3, CPz and to a lesser extent FC4, C4, CP4, P3, Pz, and P4. Peak frequency as identified at 22 Hz at site Cz. Category 0 shows No-SEB.*

Using the ADHD-RS, an impulsivity (IMP) scale was created, consisting of item 19 (Blurt out answers), 21 (Difficulty waiting my turn), and 23 (Interrupt others), in line with Arns, Swatzyna, et al. (2015). Note that the IMP scale is a subscale of the ADHD-RS (which is composed of the hyperactivity/impulsivity (HYP) and Inattention (ATT) scale). However, it is calculated differently from the ATT and HYP scale, thus IMP cannot be compared to ATT or HYP. Also, given that the IMP scale is part of the HYP scale, HYP was not considered in this study. Behavioral differences were evaluated using a GLM Univariate, with a behavioral measure as a dependent variable, and SEB (No-SEB

and SEB) and Sample (Insomnia and ADHD) as between-subject factors. The objectively-measured CPT, False Positives (FP; a response was given when no response was required) and False Negatives (FN; no response was given when a response was required) were investigated in extension to the self-rated ADHD-RS, where specifically FP were considered to be indicative of impulse control problems. Other self-rated scales were used to investigate sleep problems. These included the PSQI including its components (Subjective Sleep Quality (SSQ); Sleep Latency (SL); Sleep Duration (SDu); Habitual Sleep Efficiency (HSE); Sleep Disturbances (SDi); Use of Sleep Medication (USM); Daytime Dysfunction (DD)) and the HSDQ and its components (insomnia, parasomnia, Circadian Rhythm Sleep Disorder (CRSD), hypersomnia, Restless Legs Syndrome/Periodic Limb Movement Disorder (RLS/PLMD), Sleep Breathing Disorder (SBD)). The *p*-value was set on 0.05. In case of non-normality, potential results were confirmed using a nonparametric Mann-Whitney U, using SEB as an independent variable, and if applicable, separated by Sample. Effect sizes are reported using Cohen's *d*.

Analyses were performed using Category 0 (n = 47) and 2 (n = 32) only, confined to frontocentral SEB and adults in line with Arns et al. (2015). Eight subjects with SEB at other sites were excluded. Two individuals were excluded because of EMG contamination in the EEG. Descriptive statistics can be found in Table 1. Sample differences in frontocentral or SEB presentation were tested using Chi-square. Frontocentral SEB representation did not differ between Samples ($\chi^2(1, N = 79) = 1.440$, $p = .230$).

Metric	Total (N = 79)	ADHD (n = 48)	Insomnia (n = 31)	p
Males (n (%))	37 (46.8)	28 (58.3)	9 (29.0)	.011
SEB (n (%))	32 (40.5)	22 (45.8)	10 (32.3)	.230

Table 1: *descriptive statistics of the sample considered in this study. The total* (N = *79) sample consisted of adult clients with* (n = *32) and without* (n = *47) frontocentral SEB. A significant difference between samples was found for Sex* ($\chi^2(1,N = 79) = 6.495$, p = *.011)). No significant difference between samples was found for frontocentral SEB representation* ($\chi^2(1,N = 79) = 1.440$, p = *.230)).*

RESULTS

A One-Way ANOVA showed no significant age differences between individuals with SEB (2) and No-SEB (0) ($F(1,77) = .099$, $p = .754$. No-SEB age range (yrs): 18 – 62, with average age: 38.7 (*SD* 13.2). SEB age range (yrs): 20 – 58, with average age: 37.8 (*SD* 12.0)). Therefore, age was not considered as a covariate in the analyses.

The obtained results are summarized in Table 2 on page 168. Analyses showed a significant main effect of SEB on IMP ($F(1,72) = 15.899$, $p < .001$, $d = .87$; Figure 2, opposite page), and FP on the CPT test ($F(1,66) = 4.051$, $p = .048$, $d = .55$; Figure 3). For IMP ($F(1,72) = 14.578$, $p < .001$, $d = 1.02$) a main effect of Sample was also found. No significant interactions between SEB and Sample were observed. FP were non-normally distributed, therefore non-parametric analyses were used to confirm the result. A Mann-Whitney U confirmed the result for the total sample ($U(n_{No\_SEB} = 42, n_{SEB} = 28) = 410.00$, $z = -2.293$, $p = .022$). Spearman's correlation showed no association between IMP and FP on the CPT test ($r(68)=.214$, $p =.079$, $r^2 = 4.6\%$).

Differences for SEB on sleep parameters were also examined (Table 2 on page 168). PSQI (SDi) most closely resembles the sleep maintenance problems reported before (albeit in the original study sleep maintenance problems were defined as awakenings accompanied by having trouble falling back asleep, whereas items on PSQI (SDi) solely reflects awakenings). No differences between SEB and No-SEB were observed on the total sample ($F(1,71) = 1.131$, $p = .291$, $d = -.20$) and results were in the opposite direction as in the original study. For ADHD only, a significant effect of PSQI (SL) was observed ($F(1,44) = 8.787$, $p = .005$, $d = -.87$).

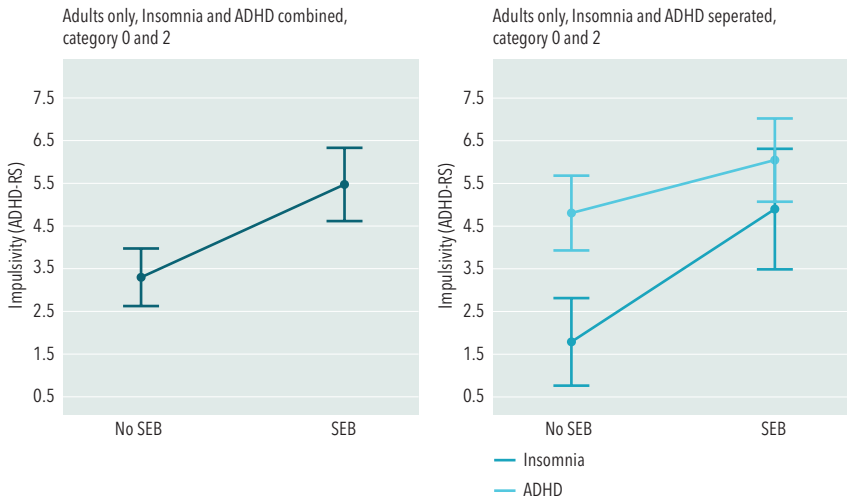


Figure 2: GLM Univariate using IMP as dependent variable and Sample and SEB as between-subject factors. A significant main effect of SEB was observed ($F(1,72) = 15.899$, $p < .001$, $d = .87$), *as well as a significant Sample effect* ($F(1,72) = 14.578$, $p < .001$, $d = 1.02$). *There were no significant Sample X SEB effects.*

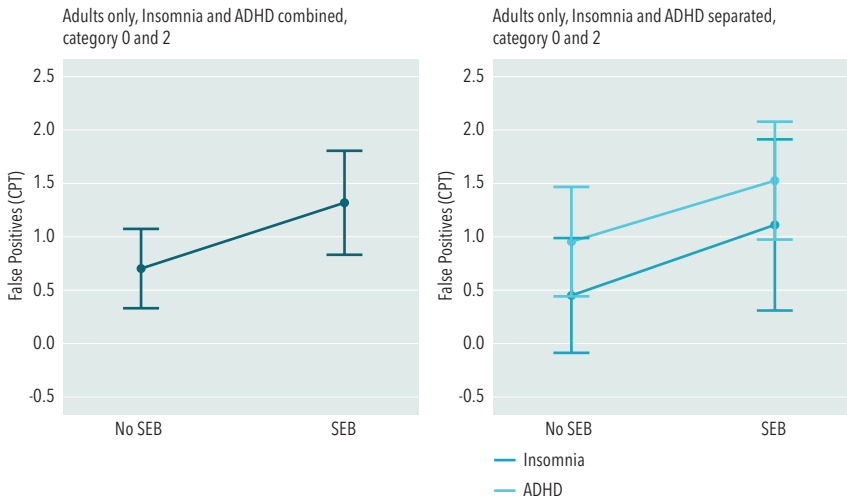


Figure 3: GLM Univariate using FP (CPT) as dependent variable and Sample (Insomnia and ADHD) and SEB (No-SEB and SEB) as between-subject factors. A significant main effect of SEB was observed ($F(1,66) = 4.051$, $p = .048$, $d = .55$). *No Sample effect was observed, nor was there a significant Sample X SEB effect. A Mann-Whitney U confirmed the result for the total sample* ($U(n_{No\_SEB} = 42, n_{SEB} = 28) = 376.500$, $z = -2.293$, $p = .022$).

	SEB present			No SEB present			*p*			*ES (d)*		
	T	A	I	T	A	I	T	A	I	T	A	I
ATT*#	5.5 (2.6)	6.4 (1.6)	3.6 (3.2)	5.1 (3.3)	7.2 (1.7)	2.2 (2.6)	.553	.107	.202	.15	-.48	.49
IMP*	5.7 (2.2)	6.0 (2.1)	4.9 (2.4)	3.5 (2.7)	4.8 (2.5)	1.8 (1.8)	<.001	.079	.001	.87	.53	1.46
FP_{WM}	1.4 (1.4)	1.5 (1.3)	1.1 (1.7)	0.7 (1.1)	1.0 (1.2)	0.5 (0.8)	.048	.148	.164	.55	.46	.50
FN_{WM}*	2.2 (2.0)	2.6 (2.2)	1.3 (1.4)	1.5 (1.7)	2.3 (1.9)	0.7 (0.9)	.260	.633	.133	.36	.15	.57
PSQI total*	9.8 (5.1)	7.9 (4.9)	14.1 (2.2)	11.4 (4.1)	9.2 (2.9)	14.3 (3.7)	.437	.285	.899	-.33	-.31	-.05
PSQI (SSQ)*	1.8 (0.9)	1.6 (0.9)	2.4 (0.7)	2.0 (0.7)	1.8 (0.7)	2.2 (0.6)	.826	.249	.465	-.20	-.34	.29
PSQI (SL)*#	1.6 (1.1)	1.2 (1.1)	2.4 (0.7)	2.1 (1.0)	2.1 (0.9)	2.1 (1.0)	.268	.005	.347	-.46	-.87	.40
PSQI (SDu)*	1.1 (1.1)	0.7 (0.9)	2.0 (0.9)	1.4 (1.2)	0.6 (0.8)	2.4 (0.7)	.396	.770	.181	-.25	.09	-.51
PSQI (HSE)*	1.1 (1.2)	0.5 (1.0)	2.3 (0.7)	1.5 (1.3)	0.7 (1.0)	2.4 (1.0)	.501	.512	.726	-.28	-.20	-.15
PSQI (SDi)	1.3 (0.5)	1.3 (0.6)	1.3 (0.5)	1.4 (0.5)	1.3 (0.5)	1.6 (0.6)	.291	.729	.142	-.20	.10	-.61
PSQI (USM)*	1.2 (1.4)	0.8 (1.3)	2.1 (1.0)	1.3 (1.4)	1.0 (1.3)	1.7 (1.4)	.677	.701	.406	-.03	-.11	.35
PSQI (DD)	1.5 (0.9)	1.5 (1.0)	1.6 (0.7)	1.7 (0.8)	1.6 (0.7)	1.8 (0.8)	.368	.622	.441	-.23	-.15	-.32
HSDQ total*	2.1 (0.5)	2.0 (0.6)	2.3 (0.4)	2.2 (0.5)	2.0 (0.4)	2.3 (0.5)	.817	.956	.794	-.17	-.02	-.11
Insomnia*	3.3 (1.2)	2.9 (1.2)	4.1 (0.8)	3.5 (0.9)	3.1 (0.8)	3.8 (0.8)	.876	.592	.382	-.16	-.19	.36
Parasomnia	1.5 (0.6)	1.5 (0.7)	1.4 (0.4)	1.5 (0.6)	1.5 (0.4)	1.5 (0.7)	.703	.834	.758	-.09	-.08	-.14
CRSD*	2.4 (1.0)	2.1 (0.9)	2.8 (1.0)	2.6 (0.8)	2.4 (0.6)	2.8 (1.0)	.673	.336	.863	-.24	-.35	.07
Hypersomnia	1.7 (0.7)	1.9 (0.7)	1.4 (0.7)	1.6 (0.7)	1.6 (0.8)	1.7 (0.6)	.947	.287	.358	.10	.38	-.37
RLS/PLMD	1.9 (0.8)	1.9 (0.9)	1.8 (0.7)	1.9 (0.8)	1.8 (0.7)	1.9 (0.9)	.992	.516	.587	.02	.23	-.23
SBD	1.6 (0.5)	1.6 (0.5)	1.7 (0.4)	1.9 (0.6)	1.9 (0.6)	1.8 (0.6)	.090	.076	.505	-.49	-.65	-.29

T Total sample,

A ADHD sample

I Insomnia sample.

Table 2: *overview of GLM Univariate analyses using a behavioral measure as dependent variable, and Sample (Insomnia and ADHD) and SEB (No-SEB and SEB) as between-subject factors. Significant* ($p \leq .05$) *Sample effects are indicated with *. Significant* ($p \leq .05$) *Sample X SEB interactions are indicated with #. Significant main effects of SEB can be found in IMP* ($F(1,72) = 15.899$, $p < .001$, $d = .87$) *and FP on the CPT test* ($F(1,66) = 4.051$, $p = .048$, $d = .55$). *For FP on the CPT test, a Mann-Whitney U confirmed the result for the total sample* ($U(n_{No\_SEB} = 42, n_{SEB} = 28) = 410.000$, $z = -2.293$, $p = .022$). *A significant main effect of SEB in the ADHD sample only was found on PSQI (SL)* ($F(1,44) = 8.787$, $p = .005$, $d = -.87$). *Note: the IMP scale is a subscale of the ADHD-RS (which compose HYP and ATT), but it is differently calculated than the ATT and HYP scale. Therefore, IMP cannot be compared to ATT and HYP. Also, since the IMP scale is part of the HYP scale, HYP is not considered.*

DISCUSSION

The current study reports a clear association between the presence of SEB and impaired levels of impulse control. This was found on both a self-rated as well as a neuropsychologically-defined scale. These effects were found in both an ADHD and an insomnia group, demonstrating that SEB represents a transdiagnostic feature related to impulse control problems. The current results replicate and extend on the earlier report (Arns, Swatzyna, et al., 2015). The effects observed

concerned large effect sizes, and while the association held for the two different operationalizations of impulse control (self-reported and FP errors), the correlation between these two operationalizations was not significant. However, the assocation of SEB with sleep maintenance problems could not be conceptually replicated, possibly due to the use of different sleep questionnaires.

An important additional finding in the current study was that the presence of SEB reflects a transdiagnostic EEG property (reflected by a lack of Sample and SEB interactions visualized in Figure 2 and Figure 3 on page 167). Remarkably, SEB presence was also related to more false positives errors on a CPT (d = .55). This means that SEB was associated with impulse control problems on a subjective as well as an objective scale. These results are found consistently across different disorders and pose the suggestion that SEB may be considered an RDoC (Insel et al., 2010), given the relation between SEB and impulse control problems seems to reflect a neurobehavioral correlate without being confined to a specific diagnosis. Yet, the association between impulse control problems and sleep maintenance problems was not apparent in the current study. Specifically, individuals showing SEB did not show to experience more sleep disturbances. An important note to this null-finding is that the questionnaire items in the current sample did not identically match the measures that showed to be significantly different in the original study, therefore, an accurate replication on this aspect could not be performed. No significant effects on sleep parameters were found, apart from SOL. In ADHD only, individuals showing SEB reported having less problems with falling asleep, yet for Insomnia as well as full sample there was no significant difference between individuals with and without SEB on SOL. These results are in line with the original study (Arns, Swatzyna, et al., 2015), specifically, the authors found that individuals with SEB did not differ from individuals showing no SEB on SOL. This is important because it is known that 70 – 80% of patients with ADHD have a delayed SOL, which may be related to their ADHD symptoms (Arns, Feddema, et al., 2014; Bijlenga et al., 2013; Bijlenga, Vollebregt, Kooij, & Arns, 2019; Konofal, Lecendreux, & Cortese, 2010). This suggests that a qualitatitvely different subgroup in ADHD can be identified in which impulse control problems are related to

SEB, but not to SOL. Interestingly, between subjective and objective measurements of sleep quality, there seem to be some differences, yet SOL is a metric that is different between ADHD and controls on subjective as well as objective measurements (Cortese, Faraone, Konofal, & Lecendreux, 2009; Diaz-Roman, Mitchell, & Cortese, 2018). In the currrent study, SOL problems were pronounced in ADHD (on average 35.6 (*SD* 24.2) minutes before falling asleep) and even more pronounced in Insomnia (on average 47.0 (*SD* 33.6) minutes before falling asleep).

The current study confined analyses to frontocentral SEB, as did the original study (Arns, Swatzyna, et al., 2015). When broadening analysis to include all SEB irrespective of site, results tended to be less pronounced or even disappear, which was also reported by Arns, Swatzyna, et al. (2015), thus suggesting site specificity for this association. Frontocentral SEB may be associated with impulse control problems, whereas SEB located elsewhere may have other behavioral correlates. A meta-analysis by Hart, Radua, Nakao, Mataix-Cols, and Rubia (2013), investigating fMRI studies in inhibition and attention in patients with ADHD, showed that for inhibition, lower activity in the right inferior frontal cortex, supplementary motor area (SMA), anterior cingulate cortex (ACC), and striato-thalamic areas was observed. Lower activity in these areas suggests a potential thalamo-cortical network that may be maintaining inhibition problems in patients with ADHD (Hart et al., 2013). The current results are in line with this notion and suggest a possible thalamo-cortical or thalamo-cingulate beta network that could be related to impulse control. Interestingly, another fMRI study found that, in boys, the right ventromedial prefrontal cortex (vmPFC) was a significant predictor of parent- and teacher-reported impulse control ratings. The authors also found a trend level effect for the right ACC and a negative correlation between impulse control ratings and right vmPFC volume (Boes et al., 2009). Future studies should investigate this further by combining CPT with neuroimaging methods such as fMRI or MEG, such that the objectively measured impulse control problems may be linked to a (dys)functional network involving the areas previously described.

Also, given the current transdiagnostic results, future studies should investigate the presence of SEB in disorders that are characterized by impulse control problems, such as pathological gambling, kleptomania, skin picking, and compulsive-impulsive shopping (Dell'Osso, Altamura, Allen, Marazziti, & Hollander, 2006; Grant & Potenza, 2004). The earlier study reported increased moodiness and temper tantrums in children with SEB (Clarke et al., 2001b), both of which seem to be in agreement with the underlying concept of impulse control problems. Hypothetically speaking, if SEB shows to be a transdiagnostic RDoC, as the current results seem to suggest, SEB and its relation to impulse control problems would be similar in various disorders. An association between impulse control disorders and obsessive-compulsive disorder (OCD) has also been studied (Dell'Osso et al., 2006). Interestingly, a study investigating the responsiveness of OCD patients to rTMS found that individuals with OCD showed increased levels of sleep disturbances. More so, individuals who did not respond to rTMS showed even higher levels of sleep disturbances compared to responders. Also, a model based on Circadian Rhythm Sleep Disorder (CRSD) could accurately predict rTMS non-response, whereas a model based on insomnia could not (Donse, Sack, Fitzgerald, & Arns, 2017). This further underlines the possible association between sleep and impulse control problems in a relevant subgroup.

Next steps

An important aspect of this paper is the detection of SEB, which currently can only be performed visually by expert ratings. This constriction poses some issues, and automatization of EEG feature detection may show to be a promissing venture for the future. Although focused on clinical diagnoses rather than EEG feature detection, Gemein et al. (2020) explain in their report that the evaluation of clinical EEGs is often time-consuming, requires years of training, and the diagnostic accuracy is limited by several aspects. These limitations include a dependency of training and experience of the evaluator, consistency of rating over time, different filter settings (e.g., the definition of targeted frequency bands), and unclear potential changes thresholding criteria (Gemein et al., 2020). Additionally, a study investigating interrater reliability on clinical EEG interpretation found that agreement among experts was moderate (Grant et al., 2014).

Automatization of feature detection in EEGs may help solve these limitations and contradictions. We propose that a similar case can be made for the detection of SEB, in that the current study can establish the foundation for future research and can suggest automatization of feature detection in EEGs. Given the initial results reported by Arns, Swatzyna, et al. (2015), the relation between impulse control problems, sleep, and possibly other territories, may shed light on symptom presentation in disorders in which the SEB-impulse control mechanism seems to be a contributing factor. Automated SEB detection will reduce SEB detection time in comparison to current detection methods (i.e., manual scoring) which allows for multiple advantages. These could include the use of larger samples and examining SEB in other labs' samples (possibly extending to multi-site findings), which are important factors in determining the replicability and robustness of a given finding (Maxwell et al., 2015; Simons, 2014).

Fernandez and Luthi (2020) highlight some ways automatization of spindle detection can be improved. Although this paper concerns sleep spindles (which are confined to a lower frequency range and usually are visible in NREM sleep), Fernandez and Luthi (2020) explain that spindle detection can be automated using a fixed thresholding approach (using a fixed frequency range, amplitude threshold, and duration threshold), an adaptive thresholding approach (a similar approach as in fixed thresholding but adjusted for possible external influences), a time-frequency analysis (using continuous wavelet analysis for simultaneous frequency and temporal occurrence of spindles), and intracranial recordings (Fernandez & Luthi, 2020). Machine learning-based detection may also show to be of use in the future. Accurate sleep spindle detection using machine learning-based detection methods is relatively well represented in the literature (e.g., (Chambon, Thorey, Arnal, Mignot, & Gramfort, 2018; Kulkarni et al., 2019; Sokolovsky, Guerrero, Paisarnsrisomsuk, Ruiz, & Alvarez, 2020)) and results are promising. Given the relative visual similarities between sleep spindles and SEB, future studies may consider machine learning as a way to automate SEB detection.

LIMITATIONS

While interpreting the results of this study one should keep in mind following limitations. Both the ADHD and Insomnia studies were open-labelled, therefore potential non-specific influences cannot be ruled out. ADHD data were gathered naturalistically. Medication usage was not controlled. Benzodiazepines and barbiturates are known to increase the presence of beta (Blume, 2006), which may have potentially influenced the current results. However, this does not seem likely since when analyses were repeated on individuals who were not using benzodiazepines or barbiturates, the results did not change. Furthermore, the scoring of SEB was limited insofar that some individuals were categorized as synchronous beta or indefinite SEB presence (synchronous beta without spindle morphology). These individuals were left out of current analyses. Future detection tools should aim to be developed so that doubtful cases can be accurately categorized into SEB or No-SEB.

CONCLUSION

This study has shown that individuals exhibiting frontocentral SEB show higher levels of impulse control problems. This finding showed to be true for a subscale of the self-rated ADHD-RS, as well as for performance on an objective CPT (measured by more false positive responses in individuals showing frontocentral SEB). The relation between sleep parameters and frontocentral SEB presentation could not be established. The results partially replicate earlier results communicated by Arns, Swatzyna, et al. (2015). Futures studies should aim to automate SEB detection and disentangle the association between frontocentral SEB, impulse control problems, sleep, and potentially other related factors.

ACKNOWLEDGEMENTS

We would like to acknowledge Jay Gunkelman for his initial teachings and inspiration on beta spindles, Helena Voetterl, Rosalinde van Ruth, Vera Kruiver, Tommy Egtberts, Eva de Valcke, Myrthe van Eerdt, Joris van Neijenhof, Monique Verhallen, Inge Janssen-Bouw-

meester, Renée Rouwhorst, Iris van Oostrom, Marleen Stam and Nicole van Merode for support and collecting the data used in this study. We also would like to acknowledge Rebecca White and Annie Bernica for reviewing and providing feedback on the manuscript.

DISCLOSURES

MA is unpaid chairman of the Brainclinics Foundation, a minority shareholder in neuroCare Group (Munich, Germany), and a co-inventor on 4 patent applications related to EEG, neuromodulation and psychophysiology, but receives no royalties related to these patents; Research Institute Brainclinics received research funding from Brain Resource (Sydney, Australia), Urgotech (France) and neuroCare Group (Munich, Germany), and equipment support from Deymed, neuroConn and Magventure.

NK, HvD, ATS, and RJS report no conflicts of interest.

8

SUMMARY AND GENERAL DISCUSSION

KEY FINDINGS

In this thesis, potential biomarkers in several psychiatric disorders were investigated. For these biomarkers, replicability and clinical relevance were also examined.

- Chapter 2 (Krepel et al., 2018) tested the replicability of results that were obtained in 2012 by Arns and colleagues. Depressed individuals that did not respond to rTMS showed to have a significantly lower IAF, a larger P300 amplitude, and more frontal theta. These results failed to replicate in a new sample of sufficient size. This study also proposed to make the dataset available to encourage other researchers to perform replication studies.

- Chapter 3 (Roelofs et al., 2021) reported a study that made use of the data-sharing proposal reported in Chapter 2 (Krepel et al., 2018). Corlier et al. (2019) showed in a group of depressed individuals treated with 10 Hz rTMS that the distance between IAF and 10 Hz was inversely and the IAF itself was linearly related to clinical improvement. Thus, the smaller the distance from IAF to 10, the greater the clinical improvement, and the higher the IAF, the greater the clinical improvement. In the current replication

attempt, it was found that the distance from IAF 10 Hz, but not the IAF itself (in line with Chapter 2 (Krepel et al., 2018)), was related to clinical improvement. This, again, only was visible in the group that was treated with 10 Hz rTMS and thus, only one of the two originally reported findings was replicated. The results suggest entrainment and synchronization as possible working mechanisms of rTMS in MDD. The association between clinical response and distance from IAF to 10 Hz showed to be quadratic (with a peak at approximately 10 Hz), rather than linear, which explains the findings reported in Chapter 2 (Krepel et al., 2018). Also, the montage (average reference versus linked ears) showed to have a significant impact on the results, as the results were only visible in the average reference montage. This also helps to explain the results of Chapter 2 (Krepel et al., 2018), as in that study a linked ears montage was used. Chapter 3.1 (Bailey et al., 2021) reported another collaborative study. The original study found that responders to rTMS treatment showed higher resting state EEG theta connectivity and lower alpha power (Bailey et al., 2019). Using the dataset proposed for data-sharing (described in Chapter 2 (Krepel et al., 2018)), these findings could not be replicated.

- In Chapter 4 (Krepel et al., in press) a biomarker of suicidal ideation was investigated. In a previous study, it was reported that female depressed patients reporting suicidal ideation exhibited frontal beta/gamma hypoactivity. This was attempted to be replicated in the large iSPOT-D sample. An explorative analysis, using a Region Of Interest based on the voxels that were significantly different between low-risk individuals and ideators in the original sample, was also performed. None of the analyses showed significant differences between low-risk individuals and ideators in the beta/gamma band.

- Chapter 5 (Krepel et al., 2019) investigated psychological features as potential predictors and the clinical relevance thereof. This was done using a Discovery-Replication approach. This means that 60% (i.e., the Discovery sample) of the total sample was used for exploratory analyses and the remaining 40% (i.e., the Replication sample) was used to confirm or refute the results obtained

in the Discovery sample. The results showed that anhedonia (measured as a subscale from the BDI), was significantly higher in depressed individuals that did not respond to rTMS. This was also confirmed in a total sample analysis using a Bonferroni-corrected *p*-value of 0.003. However, the NPV/PPV of anhedonia did not exceed the prespecified value of 75%, resulting in the conclusion that the results were not clinically relevant. This study also showed that out of 32 analyses, four initially positive results did not replicate.

- Chapter 6 (Krepel et al., 2020) investigated the effectiveness of QEEG-*informed* neurofeedback in treating symptoms of ADHD. QEEG-*informed* neurofeedback resulted in significantly fewer ADHD symptoms after treatment and response rates did not significantly differ from those reported by Arns, Drinkenburg, and Kenemans (2012). Potentials predictors of remission were identified; remitters showed lower hyperactivity/impulsivity scores, female remitters had shorter P300 latencies, and boy remitters had a lower IAF. These results still require further elaboration and investigation.

- In Chapter 7 (Krepel et al., under review) spindling excessive beta (SEB) and its association with impulse control and sleep maintenance were investigated. An initial study found that individuals presenting frontocentral SEB reported higher levels of impulse control problems as well as more sleep maintenance problems. In a new sample consisting of individuals reporting complaints of ADHD or sleep problems, it was confirmed that individuals showing frontocentral SEB had more self-rated impulse control problems and this was also reflected by more false positive responses on a Continuous Performance Task (CPT). This effect was independent of diagnosis, suggesting that SEB may be considered a transdiagnostic feature. No association with sleep was found.

THE USE OF BIOMARKERS IN PSYCHIATRY

BIOMARKERS IN DEPRESSION

A reoccurring topic has been the role of IAF in the treatment of depression. Chapter 2 (Krepel et al., 2018) showed that a low IAF as a linear predictor of rTMS non-response could not be replicated. Earlier studies showed that a low IAF was related to non-response in rTMS (Arns et al., 2010; Conca et al., 2000), medication (Ulrich, Renfordt, Zeller, & Frick, 1984), and stimulant medication in ADHD (Arns et al., 2018). As such, it was hypothesized that IAF could be a predictor of poor response to treatment in general. Yet, Chapter 3 (Roelofs et al., 2021) showed that not the IAF, but the distance from IAF to 10 Hz was related to treatment response. This was only visible in the group treated with 10 Hz rTMS and replicated an earlier report (Corlier et al., 2019). The results could be interpreted in line with the concepts of entrainment and synchronization. Fröhlich (2015) explains that entrainment is hypothesized to depend on the amplitude and frequency at which external stimulation is applied. When the stimulation frequency is close to the endogenously occurring frequency, the amplitude of stimulation can be kept low. When the stimulation frequency deviates from the naturally occurring frequency, stronger stimulation amplitudes are required. This is referred to as the Arnold Tongue (Fröhlich, 2015). The question remains whether this concept can be applied to rTMS treatment by adjusting rTMS stimulation frequency to IAF. Studies investigating the efficacy of αTMS (TMS applied at IAF) show mixed results. Some studies show no additional benefit of adjusting the frequency of rTMS to IAF (Arns et al., 2010; Jin & Phillips, 2014), yet Leuchter, Cook, et al. (2015) found that synchronized TMS (sTMS; a technique that delivers weak magnetic fields via electromagnetic coils to specific areas of the brain at the IAF) effectively reduces depressive symptoms and that individuals who were inadvertently treated at the incorrect IAF show less clinical response than patients who received sham treatment. Thus, although the results of Corlier et al. (2019) and Chapter 3 (Roelofs et al., 2021) imply that 10 Hz rTMS treatment is more effective in individuals whose IAF is closer to 10 Hz, it is still unclear whether rTMS delivered at other frequencies adjusted to IAF results in better clinical outcomes. Future studies may focus on this topic. Probably other parameters such as localization, intervals between rTMS ses-

sions, and stimulation protocols need to be considered to enhance the clinical effectiveness of rTMS.

Chapter 4 (Krepel et al., in press) showed that frontal beta/gamma hypoactivity could not be replicated as a biomarker of suicidal ideation in depressed females. Upon inspection of the literature, it is remarkable that there is not a lot of consensus of what might be considered a biomarker of suicidal ideation. EEG signatures have been found in theta (Lee, Jang, & Chae, 2017b), alpha asymmetry (although these findings also contradict each other: (Graae et al., 1996; Roh, Kim, Kim, Kim, & Lee, 2020), and gamma (Arikan et al., 2019), but consistent biomarkers have not yet emerged. Biomarker research may lead to an additional metric to assess suicide risk in an individual, especially since thoughts of suicide do not always proceed suicidal behavior, nor are they always shared with a health care professional (see Silverman and Berman (2014) for an elaborate report on suicide risk assessment). Therefore, biomarker research may help in identifying individuals who are at increased risk of suicidal ideation and should be considered in future research.

In Chapter 5 (Krepel et al., 2019) it was found that baseline anhedonia is significantly higher in individuals who do not respond to rTMS. However, this finding could not sufficiently predict rTMS response, therefore, the clinical relevance of this finding was considered to be low. There also was no sufficient effect for other baseline symptoms, which is in line with Fregni et al. (2006), who found that no baseline variables, other than age and refractoriness, significantly predicted rTMS response. Likewise, Chapter 5 (Krepel et al., 2019) showed that baseline depression severity did not differ between responders and non-responders. This is in line with Lisanby et al. (2009), who reported similarly in a randomized controlled 10 Hz rTMS treatment trial. Fitzgerald et al. (2016) also showed that although some baseline differences between responders and non-responders to rTMS exist, the relation to clinical response is not sufficient to base treatment decisions on. Thus, rTMS treatment seems to be an effective treatment for those reporting symptoms of depression, independent of the presented symptom profile. Considering that episode duration (Brakemeier et al., 2007; Garnaat, Fukuda, Yuan, & Carpenter, 2019) and refractori-

ness (Brakemeier et al., 2007; Fregni et al., 2006; Garnaat et al., 2019; Lisanby et al., 2009) have been related to less clinical improvement in rTMS treatment, it may be considered whether rTMS can be offered sooner in the treatment trajectory of depression than it is now.

BIOMARKERS IN ADHD

Chapter 6 (Krepel et al., 2020) identified several predictors of neurofeedback remission in a sample reporting symptoms of ADHD. Remitters showed lower levels of baseline hyperactivity (for the total sample), a shorter P300 latency (for females only), and a lower IAF (for boys only). Even though these results still need to be investigated further, they show promise for future treatment stratification. For example, Arns et al. (2018) found that male-adolescent non-responders to MPH had a lower IAF (8.1 Hz versus 9.2 Hz in responders), whereas Chapter 6 (Krepel et al., 2020) showed an opposite effect for neurofeedback (8.7 Hz in remitters versus 9.7 in non-remitters). On the full group level, no associations with IAF and neurofeedback remission were observed, which is in line with the original study (Arns, Drinkenburg, & Kenemans, 2012). This suggests that for individuals with a low IAF neurofeedback could be considered – as a low IAF has been associated with non-response to MPH (Arns et al., 2008; Arns et al., 2018). However, the results reported in Chapter 6 (Krepel et al., 2020) still require further investigation before treatment stratification based on these variables can occur.

TOWARDS TRANSDIAGNOSTIC RESEARCH: A TRANSDIAGNOSTIC BIOMARKER

Chapter 7 (Krepel et al., under review) resembled Chapter 4 (Krepel et al., in press) such that it concerned itself with baseline biomarkers and their behavioral correlates rather than focusing on predicting treatment outcome. In a sample consisting of individuals primarily reporting symptoms of ADHD or sleep problems, it was found that individuals exhibiting SEB showed higher levels of self-rated impulse control problems as well as more false positive responses on a CPT (and, interestingly, no correlation was observed between these two

concepts of impulsivity). These results were independent of diagnosis. The question remains whether SEB (and its association with impulse control) can be generalized to other disorders that are characterized by problems with impulse control (e.g., pathological gambling, kleptomania, and compulsive-impulsive shopping (Dell'Osso et al., 2006; Grant & Potenza, 2004)) and whether treatment options can be optimized with this knowledge. Importantly, this chapter has shown that it is possible to investigate biomarkers in a transdiagnostic manner, and this concept may apply to other areas of research. For example, suicidal ideation is considered a symptom of depression (American Psychiatric Association, 2013) but is also present in obsessive-compulsive disorder (Pellegrini et al., 2020), PTSD (Krysinska & Lester, 2010), schizophrenia (Chapman et al., 2015), and ADHD (Furczyk & Thome, 2014; Taylor et al., 2014). Suicidal ideation may thus be suitable for transdiagnostic research, which may help to identify suicidal ideation without being confined to one particular disorder.

SUMMARY

This section has assessed the main findings of this thesis. For depression, it was found that IAF was related to clinical improvement in depression, specifically, individuals with an IAF closer to 10 Hz responded better to 10 Hz rTMS treatment. Also, baseline psychological features do not seem to accurately predict antidepressant response to rTMS. Lastly, a biomarker of suicidal ideation could not be replicated, and thus consistent biomarkers of suicidal ideation have yet to be developed. For ADHD, it was found that QEEG-*informed* neurofeedback is an effective treatment for symptoms of ADHD, and potential predictors of neurofeedback remission that could potentially be used as stratification tools were identified. Lastly, a transdiagnostic EEG feature was found to be related to more self-rated impulse control problems as well as more false positive errors on a CPT.

THE ROLE OF REPLICATION IN SCIENTIFIC RESEARCH

Besides developing biomarkers in psychiatric treatment, this thesis also investigated the replicability of these biomarkers. The following sections will elaborate on multiple aspects of replication research, including the limitations and how these apply to the studies presented in this thesis, but also why and how replication research should still be up consideration as a mainstream concept in scientific practices.

LIMITATIONS OF REPLICATION RESEARCH

As this thesis has shown, replication research can help confirm, refute, or refine currently existing results. However, replication also has its flaws and thus these should be considered as such. For example, although more replication research seems to be the adequate answer to the Replication Crisis, it has been proposed that this may not be the case as direct replication is limited in its contribution to confirming the existence of a finding and direct and conceptual replications are insufficiently able to tackle systematic errors (Feest, 2019). Likewise, it has been argued that direct replications identical to the original study can never occur (given the time variable) but aiming for similarity between studies leaves room for subjective considerations (Feest, 2019). A similar argument has been made by Brandt et al. (2014), who argue in the 'Replication Recipe' that a replication study can never follow the methods of the original study – given geographical and temporal issues – but the original methods should be taken as a starting point. It is also unclear 'how many' replications are sufficient to convince the (scientific) community that a certain effect exists (Lamal, 1990), although meta-analyses may provide an estimate. Moher, Tetzlaff, Tricco, Sampson, and Altman (2007) reported that, on average, therapeutic systematic reviews report 16 studies involving 1,112 participants. However, as pointed out in Nature Research Highlights: Social Selection, it is difficult to find a balance between replication and original studies (Chawla, 2016). Tying into this discussion is a recent commentary by Coles, Tiokhin, Scheel, Isager, and Lakens (2018, January 17) who suggest that the costs of a replication study may vary from study to study and that decision theory may help in deciding whether the costs of a replication study are in balance with its benefits.

Zwaan et al. (2017) also describe six 'Concerns' about replication. The first concern ties into the context in which a replication study takes place (Concern I). It is explained that context is everchanging and thus replication can never be subject to direct replication. Context, in this description, may vary from historical or geographical factors to lab conditions (among others). The theoretical value of direct replication can also be questioned (Concern II). The quality of a replication study depends on the quality of the original study – meaning that, if an original study is flawed, it follows that a replication study is flawed. Even disregarding the possibility that an original study is flawed; a replication study can be flawed by itself as well. Concern III discusses the practical limitations of replications. In some cases (e.g., in large studies or studies that examine rare events such as natural disasters), replication attempts are impractical or impossible to achieve. Relating to this is that replication rates may vary as a result of levels of difficulty in the original study, leading to biased replication rates. Likewise, studies that are easier to replicate may undergo the consequences of unsuccessful replication sooner or more rigorously than studies that are difficult to replicate. Concern IV discusses how replications may distract the field from tackling bigger problems (Zwaan et al., 2017), such as studies on seemingly unimportant topics, media releases of research that is primarily consumer-oriented and is communicated as such, questionable research practices (QRPs), and questionable publication practices (Coyne, 2016). Coyne (2016) also describes that financial incentives (by undeclared conflicts of interests) may influence the quality of the research, yet this research may still receive a lot of attention in the media. Concern V describes the influence of replications on the reputations of the original or replication author. In case of a non-replication, concerns regarding the validity of the original study are directed towards the original author, potentially harming the scientist's reputation. Conversely, in the case of successful replication, replication authors may not be given enough credit – given it is just a replication. Concern VI states that the interpretation of what is considered to be a 'successful' replication can differ from researcher to researcher, meaning that there currently is no consensus on what the standard in replication attempts represents (Zwaan et al., 2017).

Some of the discussed limitations of replication apply to this thesis as well. For example, although some of the results presented in this thesis show to be replicable and possibly can be applied in clinical practice, analyses were performed post-hoc using naturalistic data. This means that, to claim clinical superiority of a biomarker-based approach over evidence-based treatments, randomized controlled trials or head-to-head comparisons to TAU are needed. As such, although the results presented in this thesis are robust, they are limited in establishing the existence of a finding, in line with a concern about replication research raised by Feest (2019). Future studies should address this issue.

Another example in which a non-replication is not a flawless method to test the existence of a particular finding can be observed in Chapter 2 (Krepel et al., 2018) and Chapter 3 (Roelofs et al., 2021). First, a low IAF was identified as a potential predictor of rTMS non-response in a depressed sample (Arns, Drinkenburg, Fitzgerald, et al., 2012), but this effect could not be replicated (Krepel et al., 2018; Chapter 2). Yet, the results of Chapter 3 (Roelofs et al., 2021) reported a significant relation between the distance from IAF to 10 Hz (not the frequency of IAF itself) and BDI percentage change. Thus, individuals whose IAF were closer to 10 Hz had a greater clinical response, an effect only visible in the group treated with 10 Hz rTMS. Chapter 3 (Roelofs et al., 2021) helped explain the null-findings observed in Chapter 2 (Krepel et al., 2018). For example, in Chapter 2 (Krepel et al., 2018) analyses were performed using a linked-ears reference, whereas the results in Chapter 3 (Roelofs et al., 2021) were based on an average reference montage. Performing the same analyses using a linked-ears montage did not yield significant results. Chapter 2 (Krepel et al., 2018) also focused on the total sample (combining all stimulation protocols), whereas the results in Chapter 3 (Roelofs et al., 2021) were only observed in the group treated with 10 Hz rTMS. Likewise, in Chapter 2 (Krepel et al., 2018) the association between IAF and clinical improvement was assumed to be linear, whereas Chapter 3 (Roelofs et al., 2021) showed that this relation was quadratic. This can be observed in Figure 1 (opposite page). Figure 1a shows a linear association between IAF and BDI percentage change for the original sample (Arns, Drinkenburg, Fitzgerald, et al., 2012). Figure 1b shows the replication sample using the same correlational analyses and it shows that the association is different from the

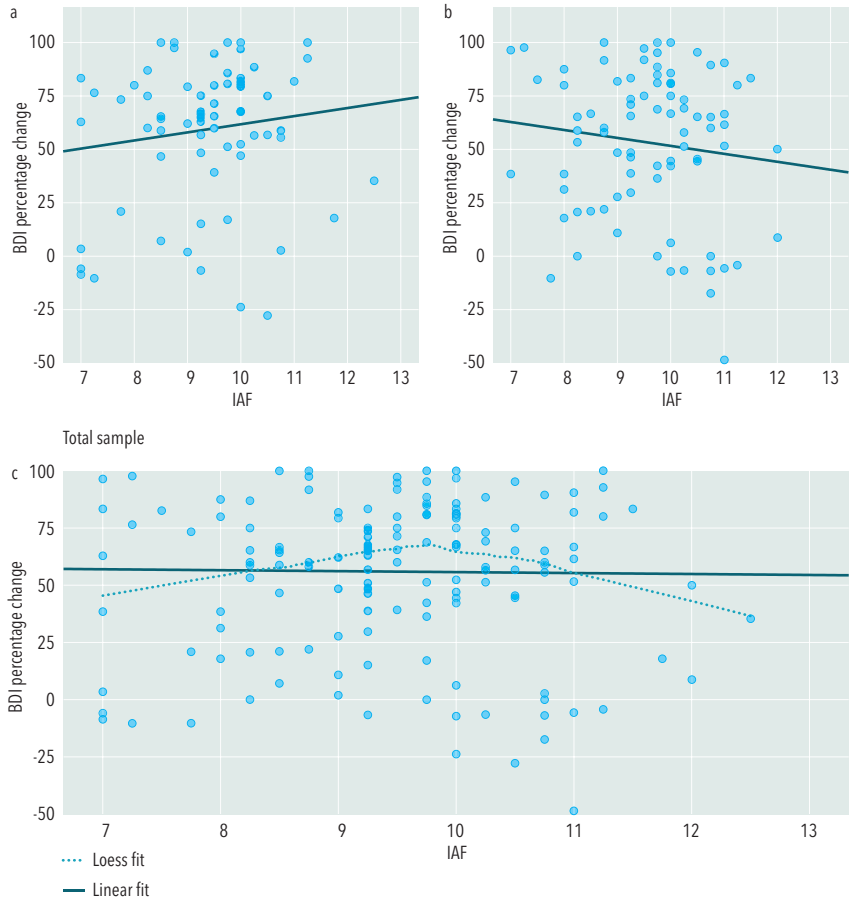


Figure 1a-c: *the original study (a) showed a positive association between IAF and BDI percentage change, whereas the replication sample showed a negative association (b). Thus, an overall linear association was non-apparent (c). However, further inspection of the data showed a quadratic rather than a linear association between IAF and BDI percentage change (c). Note: the sample used in this figure is 10 Hz and 1 Hz sample combined. The finding reported in Chapter 3 (Roelofs et al., 2021) was based on the 10 Hz unilateral sample only.*

original sample. Then, Figure 1c shows that the relation between IAF and BDI percentage change is not linear but quadratic (shown using linear fits (Figures 1a and 1b) and a Loess fit (Figure 1c), respectively). Lastly, the distribution of IAF may have influenced the results as well. Specifically, there is a relative difference in the representation of individuals with an IAF < 10 Hz (82% in the original sample versus 68.3% in

the replication sample). This skews the correlational results and may obscure that there is a quadratic rather than a linear association. This example shows that it is important to investigate why a non-replication occurred, as it may aid the development and refinement of scientific findings. Note: the sample used in this figure is 10 Hz and 1 Hz sample combined. The finding reported in Chapter 3 (Roelofs et al., 2021) was based on the 10 Hz unilateral sample only. Results based on the total sample were non-significant.

A PROBLEM OF REPLICATION OR A REPLICATION OF PROBLEMS?

As has been discussed, there are limitations to the applicability of replication research. Relating to this is a study by Kunert (2016), who used the same dataset as in Open Science Collaboration (2015). Kunert (2016) reported that internal successful replication is not predictive of successful independent replication, possibly be due to QRPs – including selective sampling, publication bias, and post-hoc hypothesis generation (also called Hypothesizing After the Results are Known, or HARKing (Kerr, 1998)). Tying into this is the notion of 'researcher degrees of freedom', meaning that if a researcher experiences pressure to publish significant results, increased combinations of analytic options may be attempted until a significant result is obtained (Zwaan et al., 2017). Also emphasized by Protzko et al. (2020, September 10), transparency and adhering to high methodological standards in conducting research are critical in achieving high replication rates.

Although QRPs are problematic for research, non-replications may occur for reasons other than QRPs – such as false positives. The level of alpha of 0.05 translates to incorrectly rejecting the null-hypothesis in 1 out of 20 cases (not taking into consideration the problem of multiple comparisons) (Schmidt, 2009) and thus a non-replication may simply occur because the original result rested on a false positive. In fact, Francis (2012) explains that in low or moderately powered studies non-replications should occur, and a lack of non-replications reported in the literature may be indicative of a publication bias. Another reason for non-replications is that the previously observed effect may not be apparent anymore. For example, Diener and Biswas-Diener (2015) describe an example in which a survey adminis-

tered to people in the 1950s may produce different outcomes than if the survey were administered in the 2010s, simply because the initial effect was not enduring. An example of such an effect may be the decreasing ES of TBR between controls and individuals diagnosed with ADHD. Arns et al. (2013) found a significant negative relation between the year of publication and the reported ES of TBR between ADHD and controls, meaning that the ES tended to get smaller over time (measured between 1999 and 2012). This effect could be explained by an increase in TBR in controls, rather than a decrease in TBR in individuals diagnosed with ADHD (Arns et al., 2013).

REPLICATION AS AN INTEGRAL PART OF SCIENCE

As was shown in the above sections, replication studies have limitations and thus replication studies should be performed and interpreted cautiously. However, ever since our first non-replication was published (Krepel et al., 2018; Chapter 2) we have been attempting to increase the rate by which replications are being performed. Through collaborative efforts, replicating our own results, and submitting the results for publication, we hope to have shown that replication studies (or other forms of methodological control) have the potential to become more mainstream. Indeed, as shown in the section 'Limitations of replication research' Zwaan et al. (2017) accurately describe six issues with replication research, yet the authors also excellently rebut these Concerns. For example, the Concern that direct replications can never occur because of everchanging contexts forfeits the purpose of falsifiability of a study. Studies that are context-dependent such that they cannot be subjected to replication research are, by definition, unfalsifiable and the falsifiability of a result is essential for it to be considered 'scientific' (Zwaan et al., 2017). Likewise, generating post-hoc, context-dependent explanations in cases of non-replications may always be possible and as such renders a finding unfalsifiable. Also, relying on post-hoc explanations rejects the possibility of the occurrence of a false positive (Zwaan et al., 2017). The authors conclude that replication research should still be up for consideration as being a routine aspect of psychological research and that, even though the Concerns are valid and should be taken seriously, the benefits of replication research still outweigh the costs.

Several options exist to make replication research an integral part of science. For example, Open Science Collaboration (2012) is a platform in which researchers offer their interests, skills, and available resources to increase the replicability of psychology research. #EEGManyLabs is another example in which large collaborative efforts attempt to increase the replication rate, specifically focused on EEG research (Pavlov et al., 2020, November 27). Another example is the International Consortium On Neuromodulation – Biomarker Discovery (ICON-DB), which is elaborated on in Chapter 3 (Roelofs et al., 2021) and Chapter 3.1 (Bailey et al., 2021). This consortium aims to test the robustness of biomarkers by sharing knowledge, skills, and datasets. Even in cases where collaboration is not possible, replication research can still be achieved by, for example, splitting the total sample (if large enough) in a Discovery (for exploratory analyses) and a Replication (for replication analyses) sample (Krepel et al., 2019; Chapter 5).

SUMMARY

This section has shown that replication research has its limitations, and these should be considered as such. The consideration of replication studies should be done with care, as the different approaches to study set-up, research practices, and interpretation of studies may influence the results. Nevertheless, replication studies usually show to be a valid method to test the robustness of scientific findings. Replication research – internal or in collaboration with other labs – may therefore be considered to be part of mainstream scientific practices.

SCIENCE FROM THE CLINIC, TO THE CLINIC: ASSESSING CLINICAL RELEVANCE

Besides the development and replicability of biomarkers, the clinical relevance of these findings was also assessed, and the current section will shortly elaborate on this aspect. Examples of assessing clinical relevance can be found throughout this thesis. Chapter 5 (Krepel et al., 2019) concluded that anhedonia as a predictor of rTMS non-response was not clinically relevant in terms of its predictive power, even though anhedonia showed to be significantly different between responders and

non-responders (Krepel et al., 2019; Chapter 5). Likewise, in Chapter 6 (Krepel et al., 2020) the effectiveness of QEEG-*informed* neurofeedback was assessed using remission and not response. Remission is indicative of minimal symptom presence or a loss of diagnostic status and it has been suggested that remission in ADHD may be related to improved functional outcomes in emotional and academic areas (Steele et al., 2006). Likewise, a secondary analysis of the MTA trial by Swanson and colleagues (2001) showed that using an outcome measure similar to remission increased clinical decision-making precision (Swanson et al., 2001). Considering remission in depression, McIntyre and O'Donovan (2004) argue that there is a cost to not achieving full remission in depression. Patients who do not achieve full remission have a higher chance of relapsing and experience more severe and chronic courses (McIntyre & O'Donovan, 2004). Thus, future studies could consider using remission instead of response as a clinical endpoint.

Although clinical relevance seems a straightforward concept, it may mean different things in different settings. Kazdin (1999) refers to clinical significance as "*...the practical or applied value or importance of the effect of an intervention–that is, whether the intervention makes a real (e.g., genuine, palpable, practical, noticeable) difference in everyday life to the clients or to others with whom the clients interact."* (Kazdin, 1999, p. 332) and that it may vary depending on the type of problems and goals of treatment. For example, when a researcher defines response as a minimum of 50% symptom decrease over the course of treatment (for example, Chapter 5 (Krepel et al., 2019)), a patient is not considered a responder when BDI scores decrease from 40 (meaning a 'severe' depression) to 21 ('moderate' depression), yet these results may be highly valuable for the patient. Other factors that may be considered in determining clinical relevance is the level of impairment (which is related to day-to-day functioning and seeking treatment but not necessarily to symptom severity), and the perspective from which clinical relevance is determined (Kazdin, 1999). Likewise, an effect that is statistically significant does not necessarily have to be clinically meaningful (Ranganathan, Pramesh, & Buyse, 2015). For example, Chapter 3 (Roelofs et al., 2021) showed a statistically significant association between the distance from IAF to 10 Hz and clinical improvement, yet the correlation coefficient and explained variance were relatively small. As such, al-

though IAF may be considered as an optimization parameter in rTMS treatment, other factors should also be considered to make a clinically meaningful impact.

It is important to consider and define the clinical relevance of findings, as potential findings may influence clinical decision making. For example, a study by Hinds and colleagues (2002) showed that different assessments of the Symptom Distress Scale (SDS) may lead to different clinical decisions. They found that making clinical decisions based on the total SDS score alone may differ when only specific items (for example, appetite, sleep, or pain) are considered. At first sight, this conclusion may seem trivial, yet the clinical implication is noteworthy. It indicates that patients who report symptoms in a specific area, but who do not surpass the cut-off on the total score, may not receive treatment or support for these symptoms. The authors conclude that the discrepancy between a summation of scores and the individual contributions of symptom-specific items is important to be aware of and argue that in some cases a symptom-specific assessment and treatment may be more appropriate (Hinds, Schum, & Srivastava, 2002). An example of a symptom-specific domain can be found in Chapter 5 (Krepel et al., 2019) where a subset of items from the BDI representing anhedonia, but not the total score of the BDI, showed to be different between responders and non-responders to rTMS treatment. Similarly, although this chapter did not concern itself with predicting treatment outcome, Chapter 7 (Krepel et al., under review) showed that impulsivity (reflected by a subset of individual items retrieved from the ADHD-RS), but not hyperactivity or inattention assessed on the total scale, was higher in individuals exhibiting SEB.

As such, clinical relevance may mean different things in different situations and studies, yet it should be considered in studies as the results may impact contexts other than the study and research community.

CONCLUSION

In this thesis, multiple topics have been investigated. For depression, it was shown that a low IAF, more frontal theta, and a larger P300 amplitude could not be replicated as predictors of rTMS non-response

(Krepel et al., 2018; Chapter 2). It was also demonstrated that the distance from the IAF to 10 Hz, but not the IAF itself, was related to clinical response. This effect was only visible in the group treated with 10 Hz rTMS (Roelofs et al., 2021; Chapter 3). It was also shown that higher theta connectivity and lower alpha power could not be replicated as predictors of rTMS response (Bailey et al., 2021; Chapter 3.1), nor could frontal beta/gamma hypoactivity be replicated as a biomarker of suicidal ideation (Krepel et al., in press; Chapter 4). It was also shown that non-responders to rTMS treatment reported higher levels of anhedonia at baseline, yet the clinical relevance of this finding was not evident (Krepel et al., 2019; Chapter 5). For ADHD, the effectiveness of QEEG-*informed* neurofeedback was replicated, showing that about half of all patients reporting symptoms of ADHD were in remission after treatment. Additionally, remitters showed lower levels of baseline hyperactivity (for the total sample), a shorter P300 latency (for females only), and a lower IAF (for boys only) (Krepel et al., 2020; Chapter 6). Lastly, in a replication study, it was found that individuals with frontocentral SEB reported higher levels of self-rated impulsivity, also reflected by more false positive errors on a CPT. These results showed to independent of diagnosis (Krepel et al., under review; Chapter 7).

Nearing the end of this thesis, a solid conclusion wrapping up the main message may be expected. Yet, given the emphasis on replication in this thesis, this may be out of place. As this thesis has hopefully shown, plenty of progress is being made in psychiatric research, yet a lot of work needs to be done before any statements can be made about the usefulness of biomarkers in an attempt to move towards stratified psychiatry and eventually personalized medicine. As of yet, no biomarkers have emerged that seem to be standing on solid, unshaken grounds. One important reason is that relatively little replication research is being performed. Even when replications studies are being performed, the results thereof are not always considered for publication. The taboo that rests on replication research is an undeserved one and hopefully, this thesis has shown that replication research is useful and interesting. Perhaps, in the future, more replication research may contribute to the development of a sound framework in which patients suffering from psychiatric issues are treated in an optimized, effective, and fitted-to-the-individual way.

REFERENCES

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nat Med, 22*(11), 1248-1255. doi:10.1038/nm.4190

Adrian, E. D., & Matthews, B. H. C. (1934). The Berger Rhythm: Potential Changes from the Occipital Lobes in Man. *Brain, 57,* 355-385.

Albouy, P., Weiss, A., Baillet, S., & Zatorre, R. J. (2017). Selective Entrainment of Theta Oscillations in the Dorsal Stream Causally Enhances Auditory Working Memory Performance. *Neuron, 94,* 193-206. doi:10.1016/j.neuron.2017.03.015

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (Fifth ed.). Arlington, VA.

American Psychological Association. (2002). Criteria for Evaluating Treatment Guidelines. *American Psychologist, 57*(12), 1052-1059. doi:10.1037//0003-066X.57.12.1052

Anastasiadou, M. N., Christodoulakis, M., Papathanasiou, E. S., Papacostas, S. S., Hadjipapas, A., & Mitsis, G. D. (2019). Graph Theoretical Characteristics of EEG-Based Functional Brain Networks in Patients With Epilepsy: The Effect of Reference Choice and Volume Conduction. *Front Neurosci, 13,* 221. doi:10.3389/fnins.2019.00221

Arikan, M. K., Gunver, M. G., Tarhan, N., & Metin, B. (2019). High-Gamma: A biological marker for suicide attempt in patients with depression. *J Affect Disord, 254,* 1-6. doi:10.1016/j.jad.2019.05.007

Arikan, M. K., Metin, B., & Tarhan, N. (2018). EEG gamma synchronization is associated with response to paroxetine treatment. *J Affect Disord, 235,* 114-116. doi:10.1016/j.jad.2018.04.041

Arns, M., Bruder, G., Hegerl, U., Spooner, C., Palmer, D. M., Etkin, A., . . . Gordon, E. (2016). EEG alpha asymmetry as a gender-specific predictor of outcome to acute treatment with different antidepressant medications in the randomized iSPOT-D study. *Clin Neurophysiol, 127*(1), 509-519. doi:10.1016/j.clinph.2015.05.032

Arns, M., Conners, C. K., & Kraemer, H. C. (2013). A Decade of EEG Theta/Beta Ratio Research in ADHD: A Meta-Analysis. *Journal of Attention Disorders, 17*(5), 374-383. doi:10.1177/1087054712460087

Arns, M., de Ridder, S., Strehl, U., Breteler, M., & Coenen, A. (2009). Efficacy of Neurofeedback Treatment in ADHD: the Effects on Inattention, Impulsivity and Hyperactivity: a Meta-Analysis. *Clinical EEG and Neuroscience, 40*(3), 180-189.

Arns, M., Drinkenburg, W., & Kenemans, J. L. (2012). The effects of QEEG-informed neurofeedback in ADHD: an open-label pilot study. *Appl Psychophysiol Biofeedback, 37*(3), 171-180. doi:10.1007/s10484-012-9191-4

Arns, M., Drinkenburg, W. H., Fitzgerald, P. B., & Kenemans, J. L. (2012). Neurophysiological predictors of non-response to rTMS in depression. *Brain Stimul, 5*(4), 569-576. doi:10.1016/j.brs.2011.12.003

Arns, M., Etkin, A., Hegerl, U., Williams, L. M., DeBattista, C., Palmer, D. M., . . . Gordon, E. (2015). Frontal and rostral anterior cingulate (rACC) theta EEG in depression: implications for treatment outcome? *Eur Neuropsychopharmacol, 25*(8), 1190-1200. doi:10.1016/j.euroneuro.2015.03.007

Arns, M., Feddema, I., & Kenemans, J. L. (2014). Differential effects of theta/beta and SMR neurofeedback in ADHD on sleep onset latency. *Front Hum Neurosci, 8*, 1019. doi:10.3389/fnhum.2014.01019

Arns, M., & Gordon, E. (2014). Quantitative EEG (QEEG) in psychiatry: diagnostic or prognostic use? *Clin Neurophysiol, 125*(8), 1504-1506. doi:10.1016/j.clinph.2014.01.014

Arns, M., Gordon, E., & Boutros, N. N. (2015). EEG Abnormalities Are Associated With Poorer Depressive Symptom Outcomes With Escitalopram and Venlafaxine-XR, but Not Sertraline: Results From the Multicenter Randomized iSPOT-D Study. *Clinical EEG and Neuroscience, 48*(1), 33-40. doi:10.1177/1550059415621435

Arns, M., Gunkelman, J., Breteler, M., & Spronk, D. (2008). EEG Phenotypes Predict Treatment Outcome to Stimulants in Children with ADHD. *Journal of Integrative Neuroscience, 7*(3), 421-438. doi:10.1142/S0219635208001897

Arns, M., Gunkelman, J., Olbrich, S., Sander, C., & Hegerl, U. (2011). EEG Vigilance and Phenotypes in Neuropsychiatry: Implications for Intervention. In *Neurofeedback and Neuromodulation Techniques and Applications* (pp. 79-123).

Arns, M., Heinrich, H., & Strehl, U. (2014). Evaluation of neurofeedback in ADHD: the long and winding road. *Biol Psychol, 95,* 108-115. doi:10.1016/j.biopsycho.2013.11.013

Arns, M., & Kenemans, J. L. (2014). Neurofeedback in ADHD and insomnia: Vigilance stabilization through sleep spindles and circadian networks. *Neuroscience and Biobehavioral Reviews, 44,* 183-194. doi:http://dx.doi.org/10.1016/j.neubiorev.2012.10.006

Arns, M., & Olbrich, S. (2014). Two EEG channels do not make a 'quantitative EEG (QEEG)': a response to Widge, Avery and Zarkowski (2013). *Brain Stimul, 7*(1), 146-148. doi:10.1016/j.brs.2013.09.009

Arns, M., Spronk, D., & Fitzgerald, P. B. (2010). Potential differential effects of 9 Hz rTMS and 10 Hz rTMS in the treatment of depression. *Brain Stimulation, 3*(2), 124-126. doi:10.1016/j.brs.2009.07.005

Arns, M., Swatzyna, R. J., Gunkelman, J., & Olbrich, S. (2015). Sleep maintenance, spindling excessive beta and impulse control: an RDoC arousal and regulatory systems approach? *Neuropsychiatric Electrophysiology, 1*(5), 1-11. doi:10.1186/s40810-015-0005-9

Arns, M., Vollebregt, M. A., Palmer, D., Spooner, C., Gordon, E., Kohn, M., . . . Buitelaar, J. K. (2018). Electroencephalographic biomarkers as predictors of methylphenidate

response in attention-deficit/hyperactivity disorder. *European Neuropsychopharmacology, 28*(8), 881-891. doi:10.1016/j.euroneuro.2018.06.002

Baehr, E., Rosenfeld, J. P., Baehr, R., & Earnest, C. (1998). Comparison of two EEG asymmetry indices in depressed patients vs. normal controls. *International Journal of Psychophysiology, 31,* 89-92.

Baeken, C., Marinazzo, D., Wu, G. R., Van Schuerbeek, P., De Mey, J., Marchetti, I., . . . De Raedt, R. (2014). Accelerated HF-rTMS in treatment-resistant unipolar depression: Insights from subgenual anterior cingulate functional connectivity. *World J Biol Psychiatry, 15*(4), 286-297. doi:10.3109/15622975.2013.872295

Bagby, R. M., Quilty, L. C., Segal, Z. V., McBride, C. C., Kennedy, S. H., & Costa, P. T. (2008). Personality and Differential Treatment Response in Major Depression: A Randomized Controlled Trial Comparing Cognitive-Behavioural Therapy and Pharmacotherapy. *The Canadian Journal of Psychiatry, 53*(6), 361-370.

Bagby, R. M., Ryder, A. G., & Cristi, C. (2002). Psychosocial and clinical predictors of response to pharmacotherapy for depression. *Journal of Psychiatry & Neuroscience: JPN, 27*(4), 250-257.

Bahramali, H., Gordon, E., Lagopoulos, J., Lim, C. L., Li, W., Leslie, J., & Wright, J. (1999). The Effects of Age on Late Components of the ERP and Reaction Time. *Experimental Aging Research, 25*, 69-80. doi:10.1080/036107399244147

Bailey, N. W., Hoy, K. E., Rogasch, N. C., Thomson, R. H., McQueen, S., Elliot, D., . . . Fitzgerald, P. B. (2018). Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders. *Brain Stimul, 11*(1), 190-203. doi:10.1016/j.brs.2017.10.015

Bailey, N. W., Hoy, K. E., Rogasch, N. C., Thomson, R. H., McQueen, S., Elliot, D., . . . Fitzgerald, P. B. (2019). Differentiating responders and non-responders to rTMS treatment for depression after one week using resting EEG connectivity measures. *J Affect Disord, 242,* 68-79. doi:10.1016/j.jad.2018.08.058

Bailey, N. W., Krepel, N., van Dijk, H., Leuchter, A. F., Vila-Rodriguez, F., Blumberger, D. M., . . . Fitzgerald, P. B. (2021). Resting EEG theta connectivity and alpha power to predict repetitive transcranial magnetic stimulation response in depression: A non-replication from the ICON-DB consortium. *Clin Neurophysiol, 132*(2), 650-659. doi:10.1016/j.clinph.2020.10.018

Bakhshayesh, H., Fitzgibbon, S. P., Janani, A. S., Grummett, T. S., & Pope, K. J. (2019). Detecting synchrony in EEG: A comparative study of functional connectivity measures. *Comput Biol Med, 105,* 1-15. doi:10.1016/j.compbiomed.2018.12.005

Bakos, S., Töllner, T., Trinkl, M., Landes, I., Bartling, J., Grossheinrich, N., . . . Greimel, E. (2016). Neurophysiological Mechanisms of Auditory Information Processing in Adolescence: A Study on Sex Differences. *Developmental Neuropsychology, 41*(3), 201-214. doi:10.1080/87565641.2016.1194840

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. doi:10.1038/s41562-017-0189-z

Benschop, L., Baeken, C., Vanderhasselt, M.-A., Van de Steen, F., Van Heeringen, K., & Arns, M. (2019). Electroencephalogram Resting State Frequency Power Characteristics of Suicidal Behavior in Female Patients With Major Depressive Disorder. *J Clin*

Psychiatry, 80(6). doi:10.4088/JCP.18m12661
Berlim, M. T., McGirr, A., Beaulieu, M. M., Van den Eynde, F., & Turecki, G. (2013). Are neuroticism and extraversion associated with the antidepressant effects of repetitive transcranial magnetic stimulation (rTMS)? An exploratory 4-week trial. *Neurosci Lett, 534*, 306-310. doi:10.1016/j.neulet.2012.12.029
Berlim, M. T., van den Eynde, F., Tovar-Perdomo, S., & Daskalakis, Z. J. (2014). Response, remission and drop-out rates following high-frequency repetitive transcranial magnetic stimulation (rTMS) for treating major depression: a systematic review and meta-analysis of randomized, double-blind and sham-controlled trials. *Psychol Med, 44*(2), 225-239. doi:10.1017/S0033291713000512
Bijlenga, D., van der Heijden, K. B., Breuk, M., van Someren, E. J., Lie, M. E., Boonstra, A. M., . . . Kooij, J. J. (2013). Associations between sleep characteristics, seasonal depressive symptoms, lifestyle, and ADHD symptoms in adults. *J Atten Disord, 17*(3), 261-275. doi:10.1177/1087054711428965
Bijlenga, D., Vollebregt, M. A., Kooij, J. J. S., & Arns, M. (2019). The role of the circadian system in the etiology and pathophysiology of ADHD: time to redefine ADHD? *ADHD Attention Deficit and Hyperactivity Disorders, 11,* 5-19. doi:10.1007/s12402-018-0271-z
Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther, 69*(3), 89-95. doi:10.1067/mcp.2001.113989
Blumberger, D. M., Vila-Rodriguez, F., Thorpe, K. E., Feffer, K., Noda, Y., Giacobbe, P., . . . Downar, J. (2018). Effectiveness of theta burst versus high-frequency repetitive transcranial magnetic stimulation in patients with depression (THREE-D): a randomised non-inferiority trial. *The Lancet, 391*(10131), 1683-1692. doi:10.1016/s0140-6736(18)30295-2
Blume, W. T. (2006). Drug Effects on EEG. *J Clin Neurophysiol, 23*(4), 306-311.
Boes, A. D., Bechara, A., Tranel, D., Anderson, S. W., Richman, L., & Nopoulos, P. (2009). Right ventromedial prefrontal cortex: a neuroanatomical correlate of impulse control in boys. *Soc Cogn Affect Neurosci, 4*(1), 1-9. doi:10.1093/scan/nsn035
Bollimunta, A., Mo, J., Schroeder, C. E., & Ding, M. (2011). Neuronal Mechanisms and Attentional Modulation of Corticothalamic Alpha Oscillations. *The Journal of Neuroscience, 31*(13), 4935-4943. doi:10.1523/JNEUROSCI.5580-10.2011
Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency. *British Journal of Psychiatry, 186,* 11-17.
Brakemeier, E. L., Luborzewski, A., Danker-Hopfe, H., Kathmann, N., & Bajbouj, M. (2007). Positive predictors for antidepressive response to prefrontal repetitive transcranial magnetic stimulation (rTMS). *J Psychiatr Res, 41*(5), 395-403. doi:10.1016/j.jpsychires.2006.01.013
Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.
Bruder, G. E., Sedoruk, J. P., Stewart, J. W., McGrath, P. J., Quitkin, F. M., & Tenke, C. E. (2008). Electroencephalographic alpha measures predict therapeutic response to a selective serotonin reuptake inhibitor antidepressant: pre- and post-treatment findings. *Biol Psychiatry, 63*(12), 1171-1177. doi:10.1016/j.biopsych.2007.10.009
Bruder, G. E., Stewart, J. W., Tenke, C. E., McGrath, P. J., Leite, P., Bhattacharya, N., & Quitkin, F. M. (2001). Electroencephalographic and Perceptual Asymmetry Differ-

ences between Responders and Nonresponders to an SSRI Antidepressant. *Biological Psychiatry, 49*(5), 416-425.
Brunelin, J., Jalenques, I., Trojak, B., Attal, J., Szekely, D., Gay, A., . . . Poulet, E. (2014). The efficacy and safety of low frequency repetitive transcranial magnetic stimulation for treatment-resistant depression: the results from a large multicenter French RCT. *Brain Stimul, 7*(6), 855-863. doi:10.1016/j.brs.2014.07.040
Bussalb, A., Collin, S., Barthélemy, Q., Ojeda, D., Bioulac, S., Blasco-Fontecilla, H., . . . Mayaud, L. (2019). Is there a cluster of high theta-beta ratio patients in Attention Deficit Hyperactivity Disorder? *Clinical Neurophysiology.* doi:https://doi.org/10.1016/j.clinph.2019.02.021
Buysse, D. l. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). T he Pittsburgh Sleep Quality Index: A New Instrument for Psychiatric Practice and Research. *Psychiatry Research, 28*, 193-213.
Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the Replicability of Social Science Experiments in Nature and Science. *Nat Hum Behav, 2*(9), 637-644. doi: 10.1038/s41562-018-0399-z
Cantisani, A., Koenig, T., Stegmayer, K., Federspiel, A., Horn, H., Muller, T. J., . . . Walther, S. (2016). EEG marker of inhibitory brain activity correlates with resting-state cerebral blood flow in the reward system in major depression. *Eur Arch Psychiatry Clin Neurosci, 266*(8), 755-764. doi:10.1007/s00406-015-0652-7
Carpenter, L. L., Janicak, P. G., Aaronson, S. T., Boyadjis, T., Brock, D. G., Cook, I. A., . . . Demitrack, M. A. (2012). Transcranial magnetic stimulation (TMS) for major depression: a multisite, naturalistic, observational study of acute treatment outcomes in clinical practice. *Depress Anxiety, 29*(7), 587-596. doi:10.1002/da.21969
Chabot, R. J., & Serfontein, G. (1996). Quantitative Electroencephalographic Profiles of Children with Attention Deficit Disorder. *Biol Psychiatry, 40,* 951-963.
Chambon, S., Thorey, V., Arnal, P., Mignot, E., & Gramfort, A. (2018). A Deep Learning Architecture To Detect Events In EEG Signals During Sleep. *arXiv preprint arXiv:1807.05981v1.*
Chapman, C. L., Mullin, K., Ryan, C. J., Kuffel, A., Nielssen, O., & Large, M. M. (2015). Meta-analysis of the association between suicidal ideation and later suicide among patients with either a schizophrenia spectrum psychosis or a mood disorder. *Acta Psychiatr Scand, 131*(3), 162-173. doi:10.1111/acps.12359
Chawla, D. S. (2016). How many replication studies are enough? *Research Highlights: Social Selection.* Retrieved from *https://www.nature.com/news/how-many-replication-studies-are-enough-1.19461*
Chervin, R. D., & Archbold, K. H. (2001). Hyperactivity and Polysomnographic Findings in Children Evaluated for Sleep-Disordered Breathing. *Sleep, 24*(3), 313-320.
Chi, M. H., Chu, C. L., Lee, I. H., Hsieh, Y. T., Chen, K. C., Chen, P. S., & Yang, Y. K. (2019). Altered Auditory P300 Performance in Parents with Attention Deficit Hyperactivity Disorder Offspring. *Clin Psychopharmacol Neurosci, 17*(4), 509-516. doi:10.9758/cpn.2019.17.4.509
Clark, C. R., Paul, R. H., Williams, L. M., Arns, M., Fallahpour, K., Handmer, C., & Gordon, E. (2006). Standardized assessment of cognitive functioning during development and aging using an automated touchscreen battery. *Arch Clin Neuropsychol, 21*(5), 449-467. doi:10.1016/j.acn.2006.06.005
Clarke, A. R., Barry, R. J., Dupuy, F. E., Heckel, L. D., McCarthy, R., Selikowitz, M., & Johnstone, S. J. (2011). Behavioural differences between EEG-defined subgroups

of children with Attention-Deficit/Hyperactivity Disorder. *Clin Neurophysiol, 122*(7), 1333-1341. doi:10.1016/j.clinph.2010.12.038
Clarke, A. R., Barry, R. J., Dupuy, F. E., McCarthy, R., Selikowitz, M., & Johnstone, S. J. (2013). Excess beta activity in the EEG of children with attention-deficit/hyperactivity disorder: a disorder of arousal? *Int J Psychophysiol, 89*(3), 314-319. doi:10.1016/j.ijpsycho.2013.04.009
Clarke, A. R., Barry, R. J., McCarthy, R., & Selikowitz, M. (2001a). EEG-defined subtypes of children with attention-deficit/hyperactivity disorder. *Clinical Neurophysiology, 112,* 2098-2105.
Clarke, A. R., Barry, R. J., McCarthy, R., & Selikowitz, M. (2001b). Excess beta activity in children with attention-deficit?hyperactivity disorder: an atypical electrophysiological group. *Psychiatry Research, 103,* 205-218.
Clarke, A. R., Barry, R. J., McCarthy, R., & Selikowitz, M. (2002). EEG differences between good and poor responders to methylphenidate and dexamphetamine in children with attention-deficit/hyperactivity disorder. *Clinical Neurophysiology, 113,* 194-205.
Clarke, A. R., Barry, R. J., McCarthy, R., Selikowitz, M., Clarke, D. C., & Croft, R. J. (2003). EEG activity in girls with attention-deficit/hyperactivity disorder. *Clinical Neurophysiology, 114*(2), 319-328. doi:10.1016/s1388-2457(02)00364-4
Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018, January 17). The costs and benefits of replication studies. *PsyArXiv.* doi:10.31234/osf.io/c8akj
Conca, A., Swoboda, E., König, P., Koppi, S., Beraus, W., Künz, A., . . . Weiß, P. (2000). Clinical Impacts of Single Transcranial Magnetic Stimulation (sTMS) as an Add-on Therapy in Severely Depressed Patients under SSRI Treatment. *Hum. Psychopharmacol. Clin. Exp., 15*, 429-438.
Cook, I. A. (2008). Biomarkers in Psychiatry: Potentials, Pitfalls, and Pragmatics. *Primary Psychiatry, 15*(3), 54-59.
Corlier, J., Carpenter, L. L., Wilson, A. C., Tirrell, E., Gobin, A. P., Kavanaugh, B., & Leuchter, A. F. (2019). The relationship between individual alpha peak frequency and clinical outcome with repetitive Transcranial Magnetic Stimulation (rTMS) treatment of Major Depressive Disorder (MDD). *Brain Stimulation, 12,* 1572-1578. doi:10.1016/j.brs.2019.07.018
Cortese, S., Adamo, N., Del Giovane, C., Mohr-Jensen, C., Hayes, A. J., Carucci, S., . . . Cipriani, A. (2018). Comparative efficacy and tolerability of medications for attention-deficit hyperactivity disorder in children, adolescents, and adults: a systematic review and network meta-analysis. *The Lancet Psychiatry, 5*(9), 727-738. doi:10.1016/s2215-0366(18)30269-4
Cortese, S., Faraone, S. V., Konofal, E., & Lecendreux, M. (2009). Sleep in children with attention-deficit/hyperactivity disorder: meta-analysis of subjective and objective studies. *J Am Acad Child Adolesc Psychiatry, 48*(9), 894-908. doi:10.1097/CHI.0b013e3181ac09c9
Cortese, S., Ferrin, M., Brandeis, D., Holtmann, M., Aggensteiner, P., Daley, D., . . . on behalf of the European ADHD Guidelines Group (EAGG). (2016). Neurofeedback for Attention-Deficit/Hyperactivity Disorder: Meta-Analysis of Clinical and Neuropsychological Outcomes From Randomized Controlled Trials. *Journal of the American Academy of Child & Adolescent Psychiatry, 55*(6), 444-455.
Costa, P. T., & McCrae, R. R. (1992). Revised NEO Personality Inventory and NEO Five-Factor Inventory: Professional Manual. *Psychological Assessment Resources, Inc.,*

Lutz, FL.
Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychol, 4*(1), 28. doi:10.1186/s40359-016-0134-3
Croughan, J. L., Secunda, S. K., Katz, M. M., Robins, E., Mendels, J., Swann, A., & Harris-Larkin, B. (1988). Sociodemographic and Prior Clinical Course Characteristics Associated with Treatment Response in Depressed Patients. *Journal of Psychiatric Research, 22*(3), 227-237.
Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *J Affect Disord, 159,* 118-126. doi:10.1016/j.jad.2014.02.026
Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine, 11*(126). doi:10.1186/1741-7015-11-126
d'Elia, G., & Perris, C. (1973). Cerebral functional dominance and depression: An analysis of EEG amplitude in depressed patients. *Acta Psychiatrica Scandinavica, 49*(3), 191-197.
Degabriele, R., & Lagopoulos, J. (2009). A review of EEG and ERP studies in bipolar disorder. *Acta Neuropsychiatrica, 21*, 58-66. doi:10.1111/j.1601-5215.2009.00359.x
Dell'Osso, B., Altamura, A. C., Allen, A., Marazziti, D., & Hollander, E. (2006). Epidemiologic and clinical updates on impulse control disorders: a critical review. *Eur Arch Psychiatry Clin Neurosci, 256*(8), 464-475. doi:10.1007/s00406-006-0668-0
Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods, 134*(1), 9-21. doi:10.1016/j.jneumeth.2003.10.009
Deng, Y., McQuoid, D. R., Potter, G. G., Steffens, D. C., Albert, K., Riddle, M., . . . Taylor, W. D. (2018). Predictors of recurrence in remitted late-life depression. *Depress Anxiety, 35*(7), 658-667. doi:10.1002/da.22772
DeRubeis, R. J., Siegle, G. J., & Hollon, S. D. (2008). Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. *Nat Rev Neurosci, 9*(10), 788-796. doi:10.1038/nrn2345
Diaz-Roman, A., Mitchell, R., & Cortese, S. (2018). Sleep in adults with ADHD: Systematic review and meta-analysis of subjective and objective studies. *Neurosci Biobehav Rev, 89*, 61-71. doi:10.1016/j.neubiorev.2018.02.014
Diener, E., & Biswas-Diener, R. (2015). The replication crisis in psychology. In R. Biswas-Diener & E. Diener (Eds.), *Noba textbook series: Psychology.* Champaign, IL: DEF publishers.
Dinga, R., Schmaal, L., Penninx, B., van Tol, M. J., Veltman, D. J., van Velzen, L., . . . Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *Neuroimage Clin, 22,* 101796. doi:10.1016/j.nicl.2019.101796
Dolsen, M. R., Cheng, P., Arnedt, J. T., Swanson, L., Casement, M. D., Kim, H. S., . . . Deldin, P. J. (2017). Neurophysiological correlates of suicidal ideation in major depressive disorder: Hyperarousal during sleep. *J Affect Disord, 212,* 160-166. doi:10.1016/j.jad.2017.01.025
Donse, L., Padberg, F., Sack, A. T., Rush, A. J., & Arns, M. (2018). Simultaneous rTMS and psychotherapy in major depressive disorder: Clinical outcomes and predictors from a large naturalistic study. *Brain Stimul, 11*(2), 337-345. doi:10.1016/j.brs.2017.11.004

Donse, L., Sack, A. T., Fitzgerald, P. B., & Arns, M. (2017). Sleep disturbances in obsessive-compulsive disorder: Association with non-response to repetitive transcranial magnetic stimulation (rTMS). *J Anxiety Disord, 49*, 31-39. doi:10.1016/j.janxdis.2017.03.006

Downar, J., Geraci, J., Salomons, T. V., Dunlop, K., Wheeler, S., McAndrews, M. P., . . . Giacobbe, P. (2014). Anhedonia and reward-circuit connectivity distinguish nonresponders from responders to dorsomedial prefrontal repetitive transcranial magnetic stimulation in major depression. *Biol Psychiatry, 76*(3), 176-185. doi:10.1016/j.biopsych.2013.10.026

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., . . . Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med, 23*(1), 28-38. doi:10.1038/nm.4246

Dunlap, K. (1926). The experimental methods of psychology. In *C. Murchison* (*Ed.)* (pp. 331-353). Worchester: Clark University Press.

Essau, C. A., Lewinsohn, P. M., Seeley, J. R., & Sasagawa, S. (2010). Gender differences in the developmental course of depression. *J Affect Disord, 127*(1-3), 185-190. doi:10.1016/j.jad.2010.05.016

Faraone, S. V., Spencer, T., Aleardi, M., Pagano, C., & Biederman, J. (2004). Meta-analysis of the efficacy of methylphenidate for treating adult attention-deficit/hyperactivity disorder. *J Clin Psychopharmacol, 24*(1), 24-29. doi:10.1097/01.jcp.0000108984.11879.95

Farnsworth, B. (2019, July 12th). EEG vs. MRI vs. fMRI - What are the Differences? Retrieved from *https://imotions.com/blog/eeg-vs-mri-vs-fmri-differences/*

Fava, M., Alpert, J. E., Carmin, C. N., Wisniewski, S. R., Trivedi, M. H., Biggs, M. M., . . . John Rush, A. (2004). Clinical correlates and symptom patterns of anxious depression among patients with major depressive disorder in STAR*D. *Psychological Medicine, 34*(7), 1299-1308. doi:10.1017/s0033291704002612

Fava, M., Rush, A. J., Alpert, J. E., Balasubramani, G. K., Wisniewski, S. R., Carmin, C. l. N., . . . Trivedi, M. H. (2008). Difference in Treatment Outcome in Outpatients With Anxious Versus Nonanxious Depression: A STAR*D Report. *American Journal of Psychiatry, 165,* 342-351.

Feest, U. (2019). Why Replication Is Overrated. *Philosophy of Science, 86,* 895-905.

Fernandez, L. M. J., & Luthi, A. (2020). Sleep Spindles: Mechanisms and Functions. *Physiol Rev, 100*(2), 805-868. doi:10.1152/physrev.00042.2018

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33,* 503-513.

Fitzgerald, P. B., Brown, T. L., Marston, N. A. U., Daskalakis, Z. J., de Castella, A., & Kulkarni, J. (2003). Transcranial Magnetic Stimulation in the Treatment of Depression. *Archives of General Psychiatry, 60*, 1002-1008.

Fitzgerald, P. B., Hoy, K., Daskalakis, Z. J., & Kulkarni, J. (2009). A randomized trial of the anti-depressant effects of low- and high-frequency transcranial magnetic stimulation in treatment-resistant depression. *Depress Anxiety, 26*(3), 229-234. doi:10.1002/da.20454

Fitzgerald, P. B., Hoy, K. E., Anderson, R. J., & Daskalakis, Z. J. (2016). A STUDY OF THE PATTERN OF RESPONSE TO rTMS TREATMENT IN DEPRESSION. *Depress Anxiety, 33*(8), 746-753. doi:10.1002/da.22503

Fitzgerald, P. J., & Watson, B. O. (2018). Gamma oscillations as a biomarker for major depression: an emerging topic. *Transl Psychiatry, 8*(1), 177. doi:10.1038/s41398-018-

0239-y
Fitzgibbon, M. L., Cella, D. F., & Sweeney, J. A. (1988). Redundancy in measures of depression. *J Clin Psychol, 44*(3), 372-374. doi: 10.1002/1097-4679(198805)44:3<372::aid-jclp2270440310>3.0.co;2-e
Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant Drug Effects and Depression Severity: A Patient-Level Meta-analysis. *JAMA, 303*(1), 47-53.
Fox, M. D., Buckner, R. L., White, M. P., Greicius, M. D., & Pascual-Leone, A. (2012). Efficacy of transcranial magnetic stimulation targets for depression is related to intrinsic functional connectivity with the subgenual cingulate. *Biol Psychiatry, 72*(7), 595-603. doi:10.1016/j.biopsych.2012.04.028
Frances, A. (2014). RDoC is necesssary, but very oversold. *World Psychiatry, 13*(1), 47-49. doi:10.1002/wps.20102
Francis, G. (2012). The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science, 7*(6), 585-594. doi:10.1177/1745691612459520
Fregni, F., Marcolin, M. A., Myczkowski, M., Amiaz, R., Hasey, G., Rumi, D. O., . . . Pascual-Leone, A. (2006). Predictors of antidepressant response in clinical trials of transcranial magnetic stimulation. *Int J Neuropsychopharmacol, 9*(6), 641-654. doi:10.1017/S1461145705006280
Fröhlich, F. (2015). Experiments and models of cortical oscillations as a target for noninvasive brain stimulation. *Progress in Brain Research, 222,* 41-73. doi:10.1016/bs.pbr.2015.07.025
Furczyk, K., & Thome, J. (2014). Adult ADHD and suicide. *Atten Defic Hyperact Disord, 6*(3), 153-158. doi:10.1007/s12402-014-0150-1
Gao, Y., & Raine, A. (2009). P3 event-related potential impairments in antisocial and psychopathic individuals: A meta-analysis. *Biological Psychiatry, 82,* 199-210. doi:10.1016/j.biopsycho.2009.06.006
Garnaat, S. L., Fukuda, A. M., Yuan, S., & Carpenter, L. L. (2019). Identification of clinical features and biomarkers that may inform a personalized approach to rTMS for depression. *Personalized Medicine in Psychiatry, 17-18*, 4-16. doi:10.1016/j.pmip.2019.09.001
Gaynes, B. N., Lloyd, S. W., Lux, L., Gartlehner, G., Hansen, R. A., Brode, S., . . . Lohr, K. N. (2014). Repetitive transcranial magnetic stimulation for treatment-resistant depression: a systematic review and meta-analysis. *J Clin Psychiatry, 75*(5), 477-489; quiz 489. doi:10.4088/JCP.13r08815
Gelade, K., Janssen, T. W. P., Bink, M., Twisk, J. W. R., van Mourik, R., Maras, A., & Oosterlaan, J. (2018). A 6-month follow-up of an RCT on behavioral and neurocognitive effects of neurofeedback in children with ADHD. *Eur Child Adolesc Psychiatry, 27*(5), 581-593. doi:10.1007/s00787-017-1072-1
Gemein, L. A. W., Schirrmeister, R. T., Chrabaszcz, P., Wilson, D., Boedecker, J., Schulze-Bonhage, A., . . . Ball, T. (2020). Machine-learning-based diagnostics of EEG pathology. *NeuroImage, 220,* 117021. doi:10.1016/j.neuroimage.2020.117021
George, M. S., Lisanby, S. H., Avery, D., McDonald, W. M., Durkalski, V., Pavlicova, M., . . . Sackeim, H. A. (2010). Daily Left Prefrontal Transcranial Magnetic Stimulation Therapy for Major Depressive Disorder. *Archives of General Psychiatry, 67*(5), 507-516.
George, M. S., & Post, R. M. (2011). Daily Left Prefrontal Repetitive Transcranial Magnetic Stimulation for Acute Treatment of Medication-Resistant Depression. *Am J Psychiatry, 168*(4), 356-364.

Gevensleben, H., Albrecht, B., Lutcke, H., Auer, T., Dewiputri, W. I., Schweizer, R., . . . Rothenberger, A. (2014). Neurofeedback of slow cortical potentials: neural mechanisms and feasibility of a placebo-controlled design in healthy adults. *Front Hum Neurosci, 8*, 990. doi:10.3389/fnhum.2014.00990

Gevensleben, H., Holl, B., Albrecht, B., Schlamp, D., Kratz, O., Studer, P., . . . Heinrich, H. (2009). Distinct EEG effects related to neurofeedback training in children with ADHD: A randomized controlled trial. *International Journal of Psychophysiology, 74,* 149-157. doi:10.1016/j.ijpsycho.2009.08.005

Goldberg, L. R. (1990). An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology, 59*(6), 1216-1229.

Gollan, J. K., Hoxha, D., Chihade, D., Pflieger, M. E., Rosebrock, L., & Cacioppo, J. (2014). Frontal alpha EEG asymmetry before and after behavioral activation treatment for depression. *Biol Psychol, 99,* 198-208. doi:10.1016/j.biopsycho.2014.03.003

Graae, F., Tenke, C., Bruder, G., Rotheram, M.-J., Piacentini, J., Castro-Blanco, D., . . . Towey, J. (1996). Abnormality of EEG Alpha Asymmetry in Female Adolescent Suicide Attempters. *Biol Psychiatry, 40,* 706-713.

Grant, A. C., Abdel-Baki, S. G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., . . . Omurtag, A. (2014). EEG interpretation reliability and interpreter confidence: a large single-center study. *Epilepsy Behav, 32,* 102-107. doi:10.1016/j.yebeh.2014.01.011

Grant, J. E., & Potenza, M. N. (2004). Impulse control disorders: clinical characteristics and pharmacological management. *Ann Clin Psychiatry, 16*(1), 27-34. doi:10.1080/10401230490281366

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A New Method For Off-line Removal Of Ocular Artifact. *Electroencephalography and clinical neurophysiology, 55,* 468-484.

Greenhill, L., Kollins, S., Abikoff, H., McCracken, J., Riddle, M., Swanson, J., . . . Cooper, T. (2006). Efficacy and safety of immediate-release methylphenidate treatment for preschoolers with ADHD. *J Am Acad Child Adolesc Psychiatry, 45*(11), 1284-1293. doi:10.1097/01.chi.0000235077.32661.61

Greneche, J., Krieger, J., Erhardt, C., Bonnefond, A., Eschenlauer, A., Muzet, A., & Tassi, P. (2008). EEG spectral power and sleepiness during 24 h of sustained wakefulness in patients with obstructive sleep apnea syndrome. *Clin Neurophysiol, 119*(2), 418-428. doi:10.1016/j.clinph.2007.11.002

Griffith, J. W., Zinbarg, R. E., Craske, M. G., Mineka, S., Rose, R. D., Waters, A. M., & Sutton, J. M. (2010). Neuroticism as a common dimension in the internalizing disorders. *Psychol Med, 40*(7), 1125-1136. doi:10.1017/S0033291709991449

Gronli, J., Rempe, M. J., Clegern, W. C., Schmidt, M., & Wisor, J. P. (2016). Beta EEG reflects sensory processing in active wakefulness and homeostatic sleep drive in quiet wakefulness. *J Sleep Res, 25*(3), 257-268. doi:10.1111/jsr.12380

Haegens, S., Cousijn, H., Wallis, G., Harrison, P. J., & Nobre, A. C. (2014). Inter- and intra-individual variability in alpha peak frequency. *NeuroImage, 92,* 46-55. doi:10.1016/j.neuroimage.2014.01.049

Hagemann, D., Naumann, E., & Thayer, J. F. (2001). The quest for the EEG reference revisited: A glance from brain asymmetry research. *Psychophysiology, 38,* 847-857.

Haller, M., Donoghue, T., Peterson, E., Varma, P., Sebastian, P., Gao, R., . . . Voytek, B. (2018). Parameterizing neural power spectra. *bioRxiv.* doi:10.1101/299859

Hammen, C. (2005). Stress and depression. *Annu Rev Clin Psychol, 1,* 293-319. doi:10.1146/annurev.clinpsy.1.102803.143938

Harmer, C. J. (2014). Neural Predictors of Treatment Response in Depression. *Current*

Behavioral Neuroscience Reports, 1(3), 125-133. doi:10.1007/s40473-014-0021-2
Hart, H., Radua, J., Nakao, T., Mataix-Cols, D., & Rubia, K. (2013). Meta-analysis of functional magnetic resonance imaging studies of inhibition and attention in attention-deficit/hyperactivity disorder: exploring task-specific, stimulant medication, and age effects. *JAMA Psychiatry, 70*(2), 185-198. doi:10.1001/jamapsychiatry.2013.277
Hayward, R. D., Taylor, W. D., Smoski, M. J., Steffens, D. C., & Payne, M. E. (2013). Association of five-factor model personality domains and facets with presence, onset, and treatment outcomes of major depression in older adults. *Am J Geriatr Psychiatry, 21*(1), 88-96. doi:10.1016/j.jagp.2012.11.012
Heinrich, H., Gevensleben, H., Freisleder, F. J., Moll, G. H., & Rothenberger, A. (2004). Training of slow cortical potentials in attention-deficit/hyperactivity disorder: evidence for positive behavioral and neurophysiological effects. *Biol Psychiatry, 55*(7), 772-775. doi:10.1016/j.biopsych.2003.11.013
Helfrich, R. F., Knepper, H., Nolte, G., Strüber, D., Rach, S., Herrmann, C. S., . . . Engel, A. K. (2014). Selective Modulation of Interhemispheric Functional Connectivity by HD-tACS Shapes Perception. *PLoS Biol, 12*(12), e1002031. doi:10.1371/journal.pbio.1002031
Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. *Br J Clin Psychol, 44*(Pt 2), 227-239. doi:10.1348/014466505X29657
Hinds, P. S., Schum, L., & Srivastava, D. K. (2002). Is Clinical Relevance Sometimes Lost in Summative Scores? *Western Journal of Nursing Research, 24*(4), 345-353.
Hinshaw, S. P., Arnold, L. E., & Group, F. t. M. C. (2015). ADHD, Multimodal Treatment, and Longitudinal Outcome: Evidence, Paradox, and Challenge. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(1), 39-52. doi:10.1002/wcs.1324.
Hiyoshi, T., Kambe, D., Karasawa, J., & Chaki, S. (2014). Differential effects of NMDA receptor antagonists at lower and higher doses on basal gamma band oscillation power in rat cortical electroencephalograms. *Neuropharmacology, 85,* 384-396. doi:10.1016/j.neuropharm.2014.05.037
Holtmann, M., Sonuga-Barke, E., Cortese, S., & Brandeis, D. (2014). Neurofeedback for ADHD: a review of current evidence. *Child Adolesc Psychiatr Clin N Am, 23*(4), 789-806. doi:10.1016/j.chc.2014.05.006
Hoy, K. E., Segrave, R. A., Daskalakis, Z. J., & Fitzgerald, P. B. (2012). Investigating the relationship between cognitive change and antidepressant response following rTMS: a large scale retrospective study. *Brain Stimul, 5*(4), 539-546. doi:10.1016/j.brs.2011.08.010
Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annu Rev Clin Psychol, 6,* 155-179. doi:10.1146/annurev.clinpsy.3.022806.091532
Hyvärinen, A. (1999). Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks, 10*(3), 626-634.
Hyvärinen, A., & Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks, 13*(4-5), 411-430.
Insel, T. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. *Am J Psychiatry, 171*(4), 395-397.
Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research Domain Criteria (RDoc): Toward a New Classification Framework for Research on Mental Disorders. *Am J Psychiatry, 167*(7), 748-751. doi:10.1176/appi.ajp.2010.09091379

Insel, T., & Wang, P. S. (2010). Rethinking Mental Illness. *JAMA, 303*(19), 1970-1971.
Ionescu, D. F., Luckenbaugh, D. A., Niciu, M. J., Richards, E. M., Slonena, E. E., Vande Voort, J. L., . . . Zarate, C. A., Jr. (2014). Effect of baseline anxious depression on initial and sustained antidepressant response to ketamine. *J Clin Psychiatry, 75*(9), e932-938. doi:10.4088/JCP.14m09049
Iseger, T. A., Korgaonkar, M. S., Kenemans, J. L., Grieve, S. M., Baeken, C., Fitzgerald, P. B., & Arns, M. (2017). EEG connectivity between the subgenual anterior cingulate and prefrontal cortices in response to antidepressant medication. *Eur Neuropsychopharmacol, 27*(4), 301-312. doi:10.1016/j.euroneuro.2017.02.002
Janssen, T. W., Bink, M., Gelade, K., van Mourik, R., Maras, A., & Oosterlaan, J. (2016). A Randomized Controlled Trial Investigating the Effects of Neurofeedback, Methylphenidate, and Physical Activity on Event-Related Potentials in Children with Attention-Deficit/Hyperactivity Disorder. *J Child Adolesc Psychopharmacol, 26*(4), 344-353. doi:10.1089/cap.2015.0144
Jin, Y., Kemp, A. S., Huang, Y., Thai, T. M., Liu, Z., Xu, W., . . . Potkin, S. G. (2012). Alpha EEG guided TMS in schizophrenia. *Brain Stimul, 5*(4), 560-568. doi:10.1016/j.brs.2011.09.005
Jin, Y., & Phillips, B. (2014). A pilot study of the use of EEG-based synchronized Transcranial Magnetic Stimulation (sTMS) for treatment of Major Depression. *BMC Psychiatry, 14*(13).
Jin, Y., Potkin, S. G., Kemp, A. S., Huerta, S. T., Alva, G., Thai, T. M., . . . Bunney, W. E., Jr. (2005). Therapeutic effects of individualized alpha frequency transcranial magnetic stimulation (alphaTMS) on the negative symptoms of schizophrenia. *Schizophrenia Bulletin, 32*(3), 556-561. doi:10.1093/schbul/sbj020
Joffe, R. T., Bagby, R. M., & Levitt, A. (1993). Anxious and Nonanxious Depression. *American Journal of Psychiatry, 150,* 1257-1258.
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci, 23*(5), 524-532. doi:10.1177/0956797611430953
Johnstone, J., Gunkelman, J., & Lunt, J. (2005). Clinical Database Development: Characterization of EEG Phenotypes. *Clinical EEG and Neuroscience, 36*(2), 99-107.
Jones, N. C., Reddy, M., Anderson, P., Salzberg, M. R., O'Brien, T. J., & Pinault, D. (2012). Acute administration of typical and atypical antipsychotics reduces EEG gamma power, but only the preclinical compound LY379268 reduces the ketamine-induced rise in gamma power. *International Journal of Neuropsychopharmacology, 15,* 657-668. doi:10.1017/S1461145711000848
Jylhä, P., & Isometsä, E. (2006). The relationship of neuroticism and extraversion to symptoms of anxiety and depression in the general population. *Depress Anxiety, 23*(5), 281-289. doi:10.1002/da.20167
Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry, 17*(12), 1174-1179. doi:10.1038/mp.2012.105
Kazdin, A. E. (1999). The Meanings and Measurement of Clinical Significance. *Journal of Consulting and Clinical Psychology, 67*(3), 332-339.
Kemp, A. H., Gordon, E., Rush, A. J., & Williams, L. M. (2008). Improving the Prediction of Treatment Response in Depression: Integration of Clinical, Cognitive, Psycho physiological, Neuroimaging, and Genetic Measures. *CNS Spectrums, 13*(12), 1066-1086.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217.
Khan, A., Leventhal, R. M., Khan, S. R., & Brown, W. A. (2002). Severity of Depression and Response to Antidepressants and Placebo: An Analysis of the Food and Drug Administration Database. *Journal of Clinical Psychopharmacology, 22*(1), 40-45.
Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B. J., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490. doi:100i.101187/170/.21157175/2541559214858918082120522
Klimesch, W., Sauseng, P., & Gerloff, C. (2003). Enhancing cognitive performance with repetitive transcranial magnetic stimulation at human individual alpha frequency. *European Journal of Neuroscience, 17*(5), 1129-1133. doi:10.1046/j.1460-9568.2003.02517.x
Konofal, E., Lecendreux, M., & Cortese, S. (2010). Sleep and ADHD. *Sleep Med, 11*(7), 652-658. doi:10.1016/j.sleep.2010.02.012
Kooij, J. J. S., Boonstra, A. M., Swinkels, S. H. N., Bekker, E. M., de Noord, I., & Buitelaar, J. K. (2008). Reliability, Validity, and Utility of Instruments for Self-Report and Informant Report Concerning Symptoms of ADHD in Adult Patients. *Journal of Attention Disorders, 11*(4), 445-458. doi:10.1177/1087054707299367
Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychol Bull, 136*(5), 768-821. doi:10.1037/a0020327
Kozel, F. A., Trivedi, M. H., Wisniewski, S. R., Miyahara, S., Husain, M. M., Fava, M., . . . Rush, A. J. (2008). Treatment Outcomes for Older Depressed Patients With Earlier Versus Late Onset of First Depressive Episode: A Sequenced Treatment Alternatives to Relieve Depression (STAR*D) report. *American Journal of Geriatric Psychiatry, 16*(1), 58-64.
Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and Moderators of Treatment Effects in Randomized Clinical Trials. *Archives of General Psychiatry, 59,* 877-883. doi:10.1001/archpsyc.59.10.877
Krepel, N., Egtberts, T., Sack, A. T., Heinrich, H., Ryan, M., & Arns, M. (2020). A multicenter effectiveness trial of QEEG-informed neurofeedback in ADHD: Replication and treatment prediction. *NeuroImage: Clinical, 28,* 102399. doi:10.1016/j.nicl.2020.102399
Krepel, N., Rush, A. J., Iseger, T. A., Sack, A. T., & Arns, M. (2019). Can psychological f eatures predict antidepressant response to rTMS? A Discovery-Replication approach. *Psychological Medicine, 50*(2), 264-272. doi:10.1017/S0033291718004191
Krepel, N., Sack, A. T., Kenemans, J. L., Fitzgerald, P. B., Drinkenburg, W. H., & Arns, M. (2018). Non-replication of neurophysiological predictors of non-response to rTMS in depression and neurophysiological data-sharing proposal. *Brain Stimul, 11*(3), 639-641. doi:10.1016/j.brs.2018.01.032
Krysinska, K., & Lester, D. (2010). Post-traumatic stress disorder and suicide risk: a systematic review. *Arch Suicide Res, 14*(1), 1-23. doi:10.1080/13811110903478997
Kubicki, S., & Ascona, A. (1983). Bases Genetiques des Modeles Beta de l'EEG du Sujet Adulte. *Rev. E.E.G. Neurophysiol., 13*, 245-252.
Kuk, A. Y., Li, J., & Rush, A. J. (2010). Recursive subsetting to identify patients in the STAR*D: a method to enhance the accuracy of early prediction of treatment outcome and to inform personalized care. *J Clin Psychiatry, 71*(11), 1502-1508.

doi:10.4088/JCP.10m06168blu
Kulkarni, P. M., Xiao, Z., Robinson, E. J., Jami, A. S., Zhang, J., Zhou, H., . . . Chen, Z. (2019). A deep learning approach for real-time detection of sleep spindles. *J Neural Eng, 16*(3), 036004. doi:10.1088/1741-2552/ab0933
Kumar, G., Steer, R. A., Teitelman, K. B., & Villacis, L. (2002). Effectiveness of Beck Depression Inventory–II Subscales in Screening for Major Depressive Disorders in Adolescent Psychiatric Inpatients. *Assessment, 9*(2), 164-170.
Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychon Bull Rev, 23*(5), 1631-1638. doi:10.3758/s13423-016-1030-9
Lamal, P. A. (1990). On the Importance of Replication. *Journal of Social Behavior and Personality, 5*(4), 31-35.
Langguth, B., Wiegand, R., Kharraz, A., Landgrebe, M., Marienhagen, J., Frick, U., . . . Eichhammer, P. (2007). Pre-treatment anterior cingulate activity as a predictor of antidepressant response to repetitive transcranial magnetic stimulation (rTMS). *Neuroendocrinology Letters, 28*(5), 633-638.
Lansbergen, M. M., Arns, M., van Dongen-Boomsma, M., Spronk, D., & Buitelaar, J. K. (2011). The increase in theta/beta ratio on resting-state EEG in boys with attention-deficit/hyperactivity disorder is mediated by slow alpha peak frequency. *Prog Neuropsychopharmacol Biol Psychiatry, 35*(1), 47-52. doi:10.1016/j.pnpbp.2010.08.004
Lee, S. M., Jang, K.-I., & Chae, J.-H. (2017a). Electroencephalographic Correlates of Suicidal Ideation in the Theta Band. *EEG and Clinical Neuroscience Society (ECNS), 48*(5), 316-321. doi:10.117/1550059417692083
Lee, S. M., Jang, K.-I., & Chae, J.-H. (2017b). Electroencephalographic Correlates of Suicidal Ideation in the Theta Band. *Clinical EEG and Neuroscience, 48*(5), 316-321. doi:10.1177/1550059417692083
Lei, X., & Liao, K. (2017). Understanding the Influences of EEG Reference: A Large-Scale Brain Network Perspective. *Frontiers in Human Neuroscience, 11*. doi:10.3389/fnins.2017.00205
Leon, A. C., & Heo, M. (2009). Sample Sizes Required to Detect Interactions between Two Binary Fixed-Effects in a Mixed-Effects Linear Regression Model. *Comput Stat Data Anal, 53*(3), 603-608. doi:10.1016/j.csda.2008.06.010
Leuchter, A. F., Cook, I. A., Feifel, D., Goethe, J. W., Husain, M., Carpenter, L. L., . . . George, M. S. (2015). Efficacy and Safety of Low-field Synchronized Transcranial Magnetic Stimulation (sTMS) for Treatment of Major Depression. *Brain Stimul, 8*(4), 787-794. doi:10.1016/j.brs.2015.05.005
Leuchter, A. F., Cook, I. A., Jin, Y., & Phillips, B. (2013). The relationship between brain oscillatory activity and therapeutic effectiveness of transcranial magnetic stimulation in the treatment of major depressive disorder. *Front Hum Neurosci, 7,* 37. doi:10.3389/fnhum.2013.00037
Leuchter, A. F., Hunter, A. M., Krantz, D. E., & Cook, I. A. (2015). Rhythms and blues: modulation of oscillatory synchrony and the mechanism of action of antidepressant treatments. *Ann N Y Acad Sci, 1344*, 78-91. doi:10.1111/nyas.12742
Leventhal, A. M., Chasson, G. S., Tapia, E., Miller, E. K., & Pettit, J. W. (2006). Measuring hedonic capacity in depression: a psychometric analysis of three anhedonia scales. *J Clin Psychol, 62*(12), 1545-1558. doi:10.1002/jclp.20327
Li, J., Kuk, A. Y. C., & Rush, A. J. (2012). A practical approach to the early identification of antidepressant medication non-responders. *Psychological Medicine, 42,* 309-316.

doi:10.1017/S0033291711001280
Lim, C. L., Gordon, E., Rennie, C., Wright, J. J., Bahramali, H., Li, W. M., . . . Morris, J. G. L. (1999). Dynamics of SCR, EEG, and ERP activity in an oddball paradigm with short interstimulus intervals. *Psychophysiology, 36*, 543-551.
Lisanby, S. H., Husain, M. M., Rosenquist, P. B., Maixner, D., Gutierrez, R., Krystal, A., . . . George, M. S. (2009). Daily left prefrontal repetitive transcranial magnetic stimulation in the acute treatment of major depression: clinical predictors of outcome in a multisite, randomized controlled clinical trial. *Neuropsychopharmacology, 34*(2), 522-534. doi:10.1038/npp.2008.118
Logemann, H. N., Lansbergen, M. M., Van Os, T. W., Bocker, K. B., & Kenemans, J. L. (2010). The effectiveness of EEG-feedback on attention, impulsivity and EEG: a sham feedback controlled study. *Neurosci Lett, 479*(1), 49-53. doi:10.1016/j.neulet.2010.05.026
Lombard, J., & Doraiswamy, P. M. (2013). What is the role of pharmacogenetics in clinical psychiatry? *Expert Opin Drug Metab Toxicol, 9*(1), 1-4. doi:10.1517/17425255.2013.733696
Lorenzo, I., Ramos, J., Arce, C., Guevara, M. A., & Corsi-Cabrera, M. (1995). Effect of Total Sleep Deprivation on Reaction Time and Waking EEG Activity in Man. *Sleep, 18*(5), 346-354.
Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., . . . Wagenmakers, E.-J. (2019). JASP: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software, 88*(2). doi:10.18637/jss.v088.i02
Lubar, J. F. (1991). Discourse on the Development of EEG Diagnostics and Biofeedback for Attention-Deficit/Hyperactivity Disorders. *Biofeedback and Self-Regulation, 16*(3), 201-225.
Maia, C. R., Cortese, S., Caye, A., Deakin, T. K., Polanczyk, G. V., Polanczyk, C. A., & Rohde, L. A. (2017). Long-Term Efficacy of Methylphenidate Immediate-Release for the Treatment of Childhood ADHD. *J Atten Disord, 21*(1), 3-13. doi:10.1177/1087054714559643
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science, 7*(6), 537-542. doi:10.1177/1745691612460688
Marcus, S. M., Young, E. A., Kerber, K. B., Kornstein, S., Farabaugh, A. H., Mitchell, J., . . . Rush, A. J. (2005). Gender differences in depression: findings from the STAR*D study. *J Affect Disord, 87*(2-3), 141-150. doi:10.1016/j.jad.2004.09.008
Marshall, L., Helgadóttir, H., Mölle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature, 444*. doi:10.1038/nature05278
Martin, G. N., & Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Front Psychol, 8*, 523. doi:10.3389/fpsyg.2017.00523
Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am Psychol, 70*(6), 487-498. doi:10.1037/a0039400
Mayer, K., Blume, F., Wyckoff, S. N., Brokmeier, L. L., & Strehl, U. (2016). Neurofeedback of slow cortical potentials as a treatment for adults with Attention Deficit-/Hyperactivity Disorder. *Clinical Neurophysiology, 127*(2), 1374-1386. doi:10.1016/j.clinph.2015.11.013
McGirr, A., Van den Eynde, F., Chachamovich, E., Fleck, M. P., & Berlim, M. T. (2014). Personality dimensions and deep repetitive transcranial magnetic stimulation

(DTMS) for treatment-resistant depression: a pilot trial on five-factor prediction of antidepressant response. *Neurosci Lett, 563,* 144-148. doi:10.1016/j.neulet.2014.01.037
McGonagle, K. A., & Kessler, R. C. (1990). Chronic Stress, Acute Stress, and Depressive Symptoms. *American Journal of Community Psychology, 18*(5), 681-706.
McIntyre, R. S., & O'Donovan, C. (2004). The Human Cost of Not Achieving Full Remission in Depression. *Can J Psychiatry, 49*, 10S-16S.
McMakin, D. L., Olino, T. M., Porta, G., Dietz, L. J., Emslie, G., Clarke, G., . . . Brent, D. A. (2012). Anhedonia Predicts Poorer Recovery Among Youth With Selective Serotonin Reuptake Inhibitor Treatment–Resistant Depression. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(4), 404-411.
Melynyte, S., Wang, G. Y., & Griskova-Bulanova, I. (2018). Gender effects on auditory P300: A systematic review. *International Journal of Psychophysiology, 133*, 55-65. doi:10.1016/j.ijpsycho.2018.08.009
Miller, R. (2007). Theory of the normal waking EEG: from single neurones to waveforms in the alpha, beta and gamma frequency ranges. *Int J Psychophysiol, 64*(1), 18-23. doi:10.1016/j.ijpsycho.2006.07.009
Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Med, 4*(3), e78. doi:10.1371/journal.pmed.0040078
Monastra, V., Linden, M., Green, G., Phillips, A., Lubar, J. F., VanDeusen, P., . . . Fenger, T. N. (1999). Assessing Attention Deficit Hyperactivity Disorder via Quantitative Electroencephalograhy: An Initial Validation Study. *Neuropsychology, 13*(3), 424-433.
Monastra, V., Monastra, D. M., & George, S. (2002). The Effects of Stimulant Therapy, EEG Biofeedback, and Parenting Style on the Primary Symptoms of Attention-Deficit/Hyperactivity Disorder. *Applied Psychophysiology and Biofeedback, 27*(4), 231-249.
MTA Cooperative Group, (2004). National Institute of Mental Health Multimodal Treatment Study of ADHD Follow-up: 24-Month Outcomes of Treatment Strategies for Attention-Deficit/Hyperactivity Disorder. *Pediatrics, 113*(4), 754-761. doi:10.1542/peds.113.4.754
Mulder, R. T. (2002). Personality Pathology and Treatment Outcome in Major Depression: A Review. *American Journal of Psychiatry, 159*, 359-371.
Nall, A. (1973). Alpha Training and the Hyperkinetic Child -is it Effective? *Academic Therapy, 9*(1), 5-19. doi:10.1177/105345127300900101
Nanova, P., Lyamova, L., Hadjigeorgieva, M., Kolev, V., & Yordanova, J. (2008). Gender-specific development of auditory information processing in children: An ERP study. *Clinical Neurophysiology, 119*, 1992-2003. doi:10.1016/j.clinph.2008.05.002
Neuliep, J. W., & Crandall, R. (1990). Editorial Bias Against Replication Research. *Journal of Social Behavior and Personality, 5*(4), 85-90.
Newcorn, J. H., Kratochvil, C. J., Allen, A. J., Casat, C. D., Ruff, D. D., Moore, R. J., & Michelson, D. (2008). Atomoxetine and Osmotically Released Methylphenidate for the Treatment of Attention Deficit Hyperactivity Disorder: Acute Comparison and Differential Response. *Am J Psychiatry, 165*(6), 721-730.
Nigbur, R., Ivanova, G., & Sturmer, B. (2011). Theta power as a marker for cognitive interference. *Clin Neurophysiol, 122*(11), 2185-2194. doi:10.1016/j.clinph.2011.03.030
Notbohm, A., Kurths, J., & Herrmann, C. S. (2016). Modification of Brain Oscillations via Rhythmic Light Stimulation Provides Evidence for Entrainment but Not for Superposition of Event-Related Responses. *Frontiers in Human Neuroscience, 10*. doi:10.3389/fnhum.2016.00010

Nussbaum, N. L. (2012). ADHD and Female Specific Concerns: A Review of the Literature and Clinical Implications. *Journal of Attention Disorders, 16*(2), 87-100. doi:10.1177/1087054711416909

O'Reardon, J. P., Solvason, H. B., Janicak, P. G., Sampson, S., Isenberg, K. E., Nahas, Z., . . . Sackeim, H. A. (2007). Efficacy and safety of transcranial magnetic stimulation in the acute treatment of major depression: a multisite randomized controlled trial. *Biol Psychiatry, 62*(11), 1208-1216. doi:10.1016/j.biopsych.2007.01.018

Olbrich, S., van Dinteren, R., & Arns, M. (2015). Personalized Medicine: Review and Perspectives of Promising Baseline EEG Biomarkers in Major Depressive Disorder and Attention Deficit Hyperactivity Disorder. *Neuropsychobiology, 72*, 229-240. doi:10.1159/000437435

Onton, J., Delorme, A., & Makeig, S. (2005). Frontal midline EEG dynamics during working memory. *NeuroImage, 27*(2), 341-356. doi:10.1016/j.neuroimage.2005.04.014

Open Science Collaboration. (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspect Psychol Sci, 7*(6), 657-660. doi:10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Ozdag, M. F., Yorbik, O., Ulas, U. H., Hamamcioglu, K., & Vural, O. (2004). Effect of methylphenidate on auditory event related potential in boys with attention deficit hyperactivity disorder. *International Journal of Pediatric Otorhinolarygology, 68*, 1267-1272. doi:10.1016/j.ijporl.2004.04.023

Park, S. C., Kim, J. M., Jun, T. Y., Lee, M. S., Kim, J. B., Yim, H. W., & Park, Y. C. (2017). How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. *Nord J Psychiatry, 71*(3), 217-222. doi:10.1080/08039488.2016.1265584

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science, 7*(6), 528-530. doi:10.1177/1745691612465253

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect Psychol Sci, 11*(4), 539-544. doi:10.1177/1745691616646366

Paul, R. H., Gunstad, J., Cooper, N., Williams, L. M., Clark, C. R., Cohen, R. A., . . . Gordon, E. (2007). Cross-cultural assessment of neuropsychological performance and electrical brain function measures: additional validation of an international brain database. *Int J Neurosci, 117*(4), 549-568. doi:10.1080/00207450600773665

Pavlov, Y. i. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S. Y., Beste, C., . . . Mushtaq, F. (2020, November 27). #EEGManyLabs: Investigating the Replicability of Influential EEG Experiments. *PsyArXiv.* doi:10.31234/osf.io/528nr

Pellegrini, L., Maietti, E., Rucci, P., Casadei, G., Maina, G., Fineberg, N. A., & Albert, U. (2020). Suicide attempts and suicidal ideation in patients with obsessive-compulsive disorder: A systematic review and meta-analysis. *J Affect Disord, 276,* 1001-1021. doi:10.1016/j.jad.2020.07.115

Perlis, M. L., Merica, H., Smith, M. T., & Giles, D. E. (2001). Beta EEG activity and insomnia. *Sleep Med Rev, 5*(5), 363-374. doi:10.1053/smrv.2001.0151

Perlis, M. L., Smith, M. T., Andrews, P. J., Orff, H., & Giles, D. E. (2001). Beta/Gamma EEG Activity in Patients with Primary and Secondary Insomnia and Good Sleeper Controls. *Sleep, 24*(1), 110-117.

Petrosino, N. J., Zandvakili, A., Carpenter, L. L., & Philip, N. S. (2018). Pilot Testing of Peak Alpha Frequency Stability During Repetitive Transcranial Magnetic Stimulation. *Front Psychiatry, 9*, 605. doi:10.3389/fpsyt.2018.00605

Phillips, K. G., Cotel, M. C., McCarthy, A. P., Edgar, D. M., Tricklebank, M., O"Neill, M. J., . . . Wafford, K. A. (2012). Differential effects of NMDA antagonists on high frequency and gamma EEG oscillations in a neurodevelopmental model of schizophrenia. *Neuropharmacology, 62*, 1359-1370. doi:10.1016/j.neuropharm.2011.04.006

Pizzagalli, D., Pascual-Marqui, R. D., Nitschke, J. B., Oakes, T. R., Larson, C. L., Abercrombie, H. C., . . . Davidson, R. J. (2001). Anterior Cingulate Activity as a Predictor of Degree of Treatment Response in Major Depression: Evidence From Brain Electrical Tomography Analysis. *Am J Psychiatry, 158*(3), 405-415.

Pizzagalli, D. A. (2011). Frontocingulate Dysfunction in Depression: Toward Biomarkers of Treatment Response. *Neuropsychopharmacology REVIEWS, 36*, 183-206. doi:10.1038/npp.2010.166

Pizzagalli, D. A. (2014). Depression, Stress, and Anhedonia: Toward a Synthesis and Integrated Model. *Annual Review of Clinical Psychology, 10,* 393-423. doi:10.1146/annurev-clinpsy-050212-185606

Pogosyan, A., Gaynor, L. D., Eusebio, A., & Brown, P. (2009). Boosting Cortical Activity at Beta-Band Frequencies Slows Movement in Humans. *Current Biology, 19*, 1637-1641. doi:10.1016/j.cub.2009.07.074

Polanía, R., Nitsche, M. A., Korman, C., Batsikadze, G., & Paulus, W. (2012). The Importance of Timing in Segregated Theta Phase-Coupling for Cognitive Performance. *Current Biology, 22*(14), 1314-1318. doi:10.1016/j.cub.2012.05.021

Popper, K. R. (1959/2002). *The Logic of scientific discovery*. London: Routledge.

Porjesz, B., Almasy, L., Edenberg, H. J., Wang, K., Chorlian, D. B., Foroud, T., . . . Begleiter, H. (2002). Linkage disequilibrium between the beta frequency of the human EEG and a GABAA receptor gene locus. *Proc Natl Acad Sci U S A, 99*(6), 3729-3733. doi:10.1073/pnas.052716399

Prasad, S., Harpin, V., Poole, L., Zeitlin, H., Jamdar, S., Puvanendran, K., & Group, S. S. (2007). A multi-centre, randomised, open-label study of atomoxetine compared with standard current therapy in UK children and adolescents with attention-deficit/hyperactivity disorder (ADHD). *Curr Med Res Opin, 23*(2), 379-394. doi:10.1185/030079906X167309

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., . . . Schooler, J. W. (2020, September 10). High Replicability of Newly-Discovered Social-behavioral Findings is Achievable. *PsyArXiv*. doi:10.31234/osf.io/n2a9x

Qin, Y., Xu, P., & Yao, D. (2010). A comparative study of different references for EEG default mode network: the use of the infinity reference. *Clinical Neurophysiology, 121*(12), 1981-1991. doi:10.1016/j.clinph.2010.03.056

Qiu, Y., Tang, Y., Chan, R. C. K., Sun, X., & He, J. (2014). P300 Aberration in First-Episode Schizophrenia Patients: A Meta-Analysis. *PLoS One, 9*(6).

Quilty, L. C., De Fruyt, F., Rolland, J. P., Kennedy, S. H., Rouillon, P. F., & Bagby, R. M. (2008). Dimensional personality traits and treatment outcome in patients with major depressive disorder. *J Affect Disord, 108*(3), 241-250. doi:10.1016/j.jad.2007.10.022

Ranganathan, P., Pramesh, C. S., & Buyse, M. (2015). Common pitfalls in statistical analysis: Clinical versus statistical significance. *Perspect Clin Res, 6*(3), 169-170. doi:10.4103/2229-3485.159943

Rockstroh, B., Elbert, T., Birbaumer, N., Wolf, P., Düchting-Röth, A., Reker, M., . . . Dich-

gans, J. (1993). Cortical self-regulation in patients with epilepsies. *Epilepsy Research, 14*, 63-72.
Roediger III, H. L. (2012). Psychology's Woes and a Partial Cure: The Value of Replication. Retrieved from *https://www.psychologicalscience.org/observer/psychologys-woes-and-a-partial-cure-the-value-of-replication*
Roelofs, C. L., Krepel, N., Corlier, J., Carpenter, L. L., Fitzgerald, P. B., Daskalakis, Z. J., . . . Arns, M. (2021). Individual alpha frequency proximity associated with repetitive transcranial magnetic stimulation outcome: An independent replication study from the ICON-DB consortium. *Clin Neurophysiol, 132*(2), 643-649. doi:10.1016/j.clinph.2020.10.017
Roh, S. C., Kim, J. S., Kim, S., Kim, Y., & Lee, S. H. (2020). Frontal Alpha Asymmetry Moderated by Suicidal Ideation in Patients with Major Depressive Disorder: A Comparison with Healthy Individuals. *Clin Psychopharmacol Neurosci, 18*(1), 58-66. doi:10.9758/cpn.2020.18.1.58
Romei, V., Bauer, M., Brooks, J. L., Economides, M., Penny, W., Thut, G., . . . Bestmann, S. (2016). Causal evidence that intrinsic beta-frequency is relevant for enhanced signal propagation in the motor system as shown through rhythmic TMS. *NeuroImage, 126*, 120-130. doi:10.1016/j.neuroimage.2015.11.020
Rosellini, A. J., & Brown, T. A. (2011). The NEO Five-Factor Inventory: latent structure and relationships with dimensions of anxiety and depressive disorders in a large clinical sample. *Assessment, 18*(1), 27-38. doi:10.1177/1073191110382848
Rosenthal, R. (1979). The "File Drawer Problem" and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638-641.
Rostami, R., Kazemi, R., Nitsche, M. A., Gholipour, F., & Salehinejad, M. A. (2017). Clinical and demographic predictors of response to rTMS treatment in unipolar and bipolar depressive disorders. *Clinical Neurophysiology, 128*, 1961-1970. doi:10.1016/j.clinph.2017.07.395
Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., . . . Fava, M. (2006). Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several A STAR*D Report. *American Journal of Psychiatry, 163*, 1905-1917.
Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution, 1*(2), 114-117. doi:10.1111/j.2041-210X.2010.00014.x
Sander, C., Arns, M., Olbrich, S., & Hegerl, U. (2010). EEG-vigilance and response to stimulants in paediatric patients with attention deficit/hyperactivity disorder. *Clinical Neurophysiology, 121*, 1511-1518. doi:10.1016/j.clinph.2010.03.021
Sanfins, M. D., Hatzopoulos, S., Torre, O. H. D., Donadon, C., Skarzynski, P. H., & Colella-Santos, M. F. (2017). Methylphenidate effects on P300 responses from children and adolescents. *International Journal of Pediatric Otorhinolarygology, 96*, 152-155. doi:10.1016/j.ijporl.2017.01.034
Sauseng, P., Hoppe, J., Klimesch, W., Gerloff, C., & Hummel, F. C. (2007). Dissociation of sustained attention from central executive functions: local activity and interregional connectivity in the theta range. *Eur J Neurosci, 25*(2), 587-593. doi:10.1111/j.1460-9568.2006.05286.x
Saveanu, R., Etkin, A., Duchemin, A. M., Goldstein-Piekarski, A., Gyurak, A., Debattista, C., . . . Williams, L. M. (2015). The international Study to Predict Optimized Treatment in Depression (iSPOT-D): outcomes from the acute phase of antidepressant

treatment. *J Psychiatr Res, 61*, 1-12. doi:10.1016/j.jpsychires.2014.12.018
Schaffer, C. E., Davidson, R. J., & Saron, C. (1983). Frontal and parietal electroencephalogram asymmetry in depressed and nondepressed subjects. *Biological Psychiatry, 18*(7), 753-762.
Scheeringa, R., Bastiaansen, M. C., Petersson, K. M., Oostenveld, R., Norris, D. G., & Hagoort, P. (2008). Frontal theta EEG activity correlates negatively with the default mode network in resting state. *Int J Psychophysiol, 67*(3), 242-251. doi:10.1016/j.ijpsycho.2007.05.017
Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology, 13*(2), 90-100. doi:10.1037/a0015108
Schmidt, S. L., Iyengar, A. K., Foulser, A. A., Boyle, M. R., & Fröhlich, F. (2014). Endogenous Cortical Oscillations Constrain Neuromodulation by Weak Electric Fields. *Brain Stimulation, 7*, 878-889. doi:10.1016/j.brs.2014.07.033
Schönenberg, M., Wiedemann, E., Schneidt, A., Scheeff, J., Logemann, A., Keune, P. M., & Hautzinger, M. (2017). Neurofeedback, sham neurofeedback, and cognitive-behavioural group therapy in adults with attention-deficit hyperactivity disorder: a triple-blind, randomised, controlled trial. *The Lancet Psychiatry, 4*(9), 673-684. doi:10.1016/s2215-0366(17)30291-2
Schuch, J. J., Roest, A. M., Nolen, W. A., Penninx, B. W., & de Jonge, P. (2014). Gender differences in major depressive disorder: results from the Netherlands study of depression and anxiety. *J Affect Disord, 156*, 156-163. doi:10.1016/j.jad.2013.12.011
Schutter, D. J. (2009). Antidepressant efficacy of high-frequency transcranial magnetic stimulation over the left dorsolateral prefrontal cortex in double-blind sham-controlled designs: a meta-analysis. *Psychol Med, 39*(1), 65-75. doi:10.1017/S0033291708003462
Schutter, D. J. (2010). Quantitative review of the efficacy of slow-frequency magnetic brain stimulation in major depressive disorder. *Psychol Med, 40*(11), 1789-1795. doi:10.1017/S003329171000005X
Sedky, K., Bennett, D. S., & Carvalho, K. S. (2014). Attention deficit hyperactivity disorder and sleep disordered breathing in pediatric populations: A meta-analysis. *Sleep Medicine Reviews, 18*, 349-356. doi:10.1016/j.smrv.2013.12.003
Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology, 69*, 487-510. doi:10.1146/annurev- psych- 122216- 011845
Silverman, M. M., & Berman, A. L. (2014). Suicide risk assessment and risk formulation part I: a focus on suicide ideation in assessing suicide risk. *Suicide Life Threat Behav, 44*(4), 420-431. doi:10.1111/sltb.12065
Sim, M. K., & Tsoi, W. F. (1992). The Effects of Centrally Acting Drugs on the EEG Correlates of Meditation. *Biofeedback and Self-Regulation, 17*(3), 215-220.
Simon, G. E., & Perlis, R. H. (2010). Personalized Medicine for Depression: Can We Match Patients With Treatments? *American Journal of Psychiatry, 167*, 1445-1455.
Simons, C. J. P., Sambeth, A., Krabbendam, L., Pfeifer, S., van Os, J., & Riedel, W. J. (2011). Auditory P300 and N100 components as intermediate phenotypes for psychotic disorder: Familial liability and reliability. *Clinical Neurophysiology, 122*, 1984-1990. doi:10.1016/j.clinph.2011.02.033
Simons, D. J. (2014). The Value of Direct Replication. *Perspect Psychol Sci, 9*(1), 76-80. doi:10.1177/1745691613514755

Smit, C. M., Wright, M. J., Hansell, N. K., Geffen, G. M., & Martin, N. G. (2006). Genetic variation of individual alpha frequency (IAF) and alpha power in a large adolescent twin sample. *International Journal of Psychophysiology, 61*(2), 235-243. doi:10.1016/j.ijpsycho.2005.10.004

Sokolovsky, M., Guerrero, F., Paisarnsrisomsuk, S., Ruiz, C., & Alvarez, S. A. (2020). Deep Learning for Automated Feature Discovery and Classification of Sleep Stages. *IEEE/ACM Trans Comput Biol Bioinform, 17*(6), 1835-1845. doi:10.1109/TCBB.2019.2912955

Speckmann, E.-J., Elger, C. E., & Gorji, A. (2011). Neurophysiologic Basis of EEG and DC Potentials. In *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (Sixth ed.): Lippincott Williams & Wilkins, a Wolters Kluwer business.

Spencer, T., Biederman, J., Wilens, T., Doyle, R., Surman, C., Prince, J., . . . Faraone, S. (2005). A large, double-blind, randomized clinical trial of methylphenidate in the treatment of adults with attention-deficit/hyperactivity disorder. *Biol Psychiatry, 57*(5), 456-463. doi:10.1016/j.biopsych.2004.11.043

Spijker, J., Bijl, R. V., de Graaf, R., & Nolen, W. A. (2001). Determinants of poor 1-year outcome of DSM-III-R major depression in the general population: results of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Acta Psychiatrica Scandinavica, 103*, 122-130.

Steele, M., Jensen, P. S., & Quinn, D. M. P. (2006). Remission Versus Response as the Goal of Therapy in ADHD: A New Standard for the Field? *Clinical Therapeutics, 28*(11), 1892-1908. doi:10.1016/j.clinthera.2006.11.006

Strehl, U., Aggensteiner, P., Wachtlin, D., Brandeis, D., Albrecht, B., Arana, M., . . . Holtmann, M. (2017). Neurofeedback of Slow Cortical Potentials in Children with Attention-Deficit/Hyperactivity Disorder: A Multicenter Randomized Trial Controlling for Unspecific Effects. *Frontiers in Human Neuroscience, 11*. doi:10.3389/fnhum.2017.00135

Strijkstra, A. M., Beersma, D. G. M., Drayer, B., Halbesma, N., & Daan, S. (2003). Subjective sleepiness correlates negatively with global alpha (8–12 Hz) and positively with central frontal theta (4–8 Hz) frequencies in the human resting awake electroencephalogram. *Neuroscience Letters, 340*(1), 17-20. doi:10.1016/s0304-3940(03)00033-8

Suffin, S. C., & Emory, W. H. (1995). Neurometric subgroups in attentional and affective disorders and their association with pharmacotherapeutic outcome. *Clinical Electroencephalography, 26*(2), 76-83.

Sunohara, G. A., Voros, J. G., Malone, M. A., & Taylor, M. J. (1997). Effects of methylphenidate in children with attention deficit hyperactivity disorder: a comparison of event-related potentials between medication responders and non-responders. *International Journal of Psychophysiology, 27*, 9-14.

Swanson, J. M., Kraemer, H. C., Hinshaw, S. P., Arnold, L. E., Conners, C. K., Abikoff, H. B., . . . Wu, M. (2001). Clinical Relevance of the Primary Findings of the MTA: Success Rates Based on Severity of ADHD and ODD Symptoms at the End of Treatment. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(2), 168-179.

Szuromi, B., Czobor, P., Komlósi, S., & Bitter, I. (2011). P300 deficits in adults with attention deficit hyperactivity disorder: a meta-analysis. *Psychological Medicine, 41*, 1529-1538. doi:10.1017/S0033291710001996

Tabb, K. (2015). Psychiatric Progress and the Assumption of Diagnostic Discrimination.

Philosophy of Science, 82, 1047-1058.
Taylor, M. R., Boden, J. M., & Rucklidge, J. J. (2014). The relationship between ADHD symptomatology and self-harm, suicidal ideation, and suicidal behaviours in adults: a pilot study. *Atten Defic Hyperact Disord, 6*(4), 303-312. doi:10.1007/s12402-014-0139-9
Tenke, C. E., Kayser, J., Pechtel, P., Webb, C. A., Dillon, D. G., Goer, F., . . . Bruder, G. E. (2017). Demonstrating test-retest reliability of electrophysiological measures for healthy adults in a multisite study of biomarkers of antidepressant treatment response. *Psychophysiology, 54*(1), 34-50. doi:10.1111/psyp.12758
Thase, M. E. (2014). Using biomarkers to predict treatment response in major depressive disorder: evidence from past and present studies. *Dialogues Clin Neurosci, 16*, 539-544.
The MTA Cooperative Group. (1999). A 14-Month Randomized Clinical Trial of Treatment Strategies for Attention-Deficit/ Hyperactivity Disorder. *Arch Gen Psychiatry, 56*, 1073-1086.
Thut, G., Veniero, D., Romei, V., Miniussi, C., Schyns, P., & Gross, J. (2011). Rhythmic TMS Causes Local Entrainment of Natural Oscillatory Signatures. *Curr Biol, 21*(14), 1176-1185. doi:10.1016/j.cub.2011.05.049
Treadway, M. T., & Zald, D. H. (2011). Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neurosci Biobehav Rev, 35*(3), 537-555. doi:10.1016/j.neubiorev.2010.06.006
Trentini, C. M., de Freitas Xavier, F. M., Chachamovich, E., da Rocha, N. S., Hirakata, V. N., & de Almeida Fleck, M. P. (2005). The influence of somatic symptoms on the performance of elders in the Beck Depression Inventory (BDI). *Revista Brasileira de Psiquiatria, 27*(2), 119-123.
Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., . . . STAR*D Study Team. (2006). Evaluation of Outcomes With Citalopram for Depression Using Measurement-Based Care in STAR*D: Implications for Clinical Practice. *American Journal of Psychiatry, 163*, 28-40.
Tsai, M.-L., Hung, K.-L., & Lu, H.-H. (2012). Auditory Event-related Potentials in Children With Attention Deficit Hyperactivity Disorder. *Pediatrics and Neonatology, 53*, 118-124. doi:10.1016/j.pedneo.2012.01.009
Uher, R., Dernovsek, M. Z., Mors, O., Hauser, J., Souery, D., Zobel, A., . . . Farmer, A. (2011). Melancholic, atypical and anxious depression subtypes and outcome of treatment with escitalopram and nortriptyline. *J Affect Disord, 132*(1-2), 112-120. doi:10.1016/j.jad.2011.02.014
Ulrich, G., Renfordt, E., Zeller, G., & Frick, K. (1984). Interrelation between changes in the EEG and psychopathology under pharmacotherapy for endogenous depression. A contribution to the predictor question. *Pharmacopsychiatry, 17*(6), 178-183.
van der Vinne, N. (2020). *EEG Biomarkers in Depression: Paving the Way for Stratified Psychiatry*. Synaeda Psycho Medisch Centrum, Research Institute Brainclinics, University of Twente, Nijmegen, The Netherlands.
van der Vinne, N., Vollebregt, M. A., Rush, A. J., Eebes, M., van Putten, M. J. A. M., & Arns, M. (2021). EEG biomarker informed prescription of antidepressants in MDD: a feasibility trial. *European Neuropsychopharmacology*. doi:10.1016/j.euroneuro.2020.12.005
van der Vinne, N., Vollebregt, M. A., van Putten, M., & Arns, M. (2017). Frontal alpha asymmetry as a diagnostic marker in depression: Fact or fiction? A meta-analysis. *Neuroimage Clin, 16*, 79-87. doi:10.1016/j.nicl.2017.07.006

van der Vinne, N., Vollebregt, M. A., van Putten, M., & Arns, M. (2019). Stability of frontal alpha asymmetry in depressed patients during antidepressant treatment. *Neuroimage Clin, 24*, 102056. doi:10.1016/j.nicl.2019.102056

van Dijk, H., deBeus, R., Kerson, C., Roley-Roberts, M. E., Monastra, V. J., Arnold, L. E., . . . Arns, M. (2020). Different Spectral Analysis Methods for the Theta/Beta Ratio Calculate Different Ratios But Do Not Distinguish ADHD from Controls. *Applied Psychophysiology and Biofeedback*. doi:10.1007/s10484-020-09471-2

van Dinteren, R., Arns, M., Jongsma, M. L., & Kessels, R. P. (2014). P300 development across the lifespan: a systematic review and meta-analysis. *PLoS One, 9*(2), e87347. doi:10.1371/journal.pone.0087347

van Dinteren, R., Arns, M., Kenemans, L., Jongsma, M. L., Kessels, R. P., Fitzgerald, P., . . . Williams, L. M. (2015). Utility of event-related potentials in predicting antidepressant treatment response: An iSPOT-D report. *Eur Neuropsychopharmacol, 25*(11), 1981-1990. doi:10.1016/j.euroneuro.2015.07.022

Van Doren, J., Arns, M., Heinrich, H., Vollebregt, M. A., Strehl, U., & Loo, S. K. (2019). Sustained effects of neurofeedback in ADHD: a systematic review and meta-analysis. *European Child & Adolescent Psychiatry, 28*, 293-305. doi:https://doi.org/10.1007/s00787-018-1121-4

van Straten, A., Hill, J., Richards, D. A., & Cuijpers, P. (2015). Stepped care treatment delivery for depression: a systematic review and meta-analysis. *Psychol Med, 45*(2), 231-246. doi:10.1017/S0033291714000701

Venkatasubramanian, G., & Keshavan, M. S. (2016). Biomarkers in Psychiatry - A Critique. *Ann Neurosci, 23*(1), 3-5. doi:10.1159/000443549

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., & Pennartz, C. M. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage, 55*(4), 1548-1565. doi:10.1016/j.neuroimage.2011.01.055

Vinckier, F., Gourion, D., & Mouchabac, S. (2017). Anhedonia predicts poor psychosocial functioning: Results from a large cohort of patients treated for major depressive disorder by general practitioners. *Eur Psychiatry, 44*, 1-8. doi:10.1016/j.eurpsy.2017.02.485

Vitiello, B., Severe, J. B., Greenhill, L. L., Arnold, L. E., Abikoff, H. B., Bukstein, O. G., . . . Cantwell, D. P. (2001). Methylphenidate Dosage for Children With ADHD Over Time Under Controlled Conditions: Lessons From the MTA. *J. Am. Acad. Child Adolesc. Psychiatry, 40*(2), 188-196.

Vogel, F. (1970). The Genetic Basis of the Normal Human Electroencephalogram (EEG). *Humangenetik, 10*, 91-114.

Vollebregt, M. A., Arns, M., Monastra, V., Roley-Roberts, M. E., Arnold, L. E., & Hollway, J. l. A. (2019, June 12). Shedding light on the etiology of ADHD: The influence of disturbed sleep on symptomatology. doi:https://doi.org/10.31234/osf.io/bue3v

Vollebregt, M. A., van Dongen-Boomsma, M., Slaats-Willemse, D., Buitelaar, J. K., & Oostenveld, R. (2015). How the Individual Alpha Peak Frequency Helps Unravel the Neurophysiologic Underpinnings of Behavioral Functioning in Children With Attention-Deficit/Hyperactivity Disorder. *Clin EEG Neurosci, 46*(4), 285-291. doi:10.1177/1550059414537257

Vosskuhl, J., Strüber, D., & Herrmann, C. S. (2018). Non-invasive Brain Stimulation: A Paradigm Shift in Understanding Brain Oscillations. *Frontiers in Human Neuroscience, 12*. doi:10.3389/fnhum.2018.00211

Weiss, M. D., Gibbins, C., Goodman, D. W., Hodgkins, P. S., Landgraf, J. M., & Faraone, S. V. (2010). Moderators and Mediators of Symptoms and Quality of Life Outcomes in an Open-Label Study of Adults Treated for Attention-Deficit/Hyperactivity Disorder. *Journal of Clinical Psychiatry, 71*(4), 381-390. doi:10.4088/JCP.08m04709pur

Whitham, E. M., Pope, K. J., Fitzgibbon, S. P., Lewis, T., Clark, C. R., Loveless, S., . . . Willoughby, J. O. (2007). Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clinical Neurophysiology, 118*, 1877-1888. doi:10.1016/j.clinph.2007.04.027

Whitton, A. E., Deccy, S., Ironside, M. L., Kumar, P., Beltzer, M., & Pizzagalli, D. A. (2018). EEG source functional connectivity reveals abnormal high- frequency communication among large-scale functional networks in depression. *Biol Psychiatry Cogn Neurosci Neuroimaging, 3*(1), 50-58. doi:10.1016/j.bpsc. 2017.07.001

Widge, A. S., Avery, D. H., & Zarkowski, P. (2013). Baseline and treatment-emergent EEG biomarkers of antidepressant medication response do not predict response to repetitive transcranial magnetic stimulation. *Brain Stimul, 6*(6), 929-931. doi:10.1016/j.brs.2013.05.001

Widge, A. S., Bilge, M. T., Montana, R., Chang, W., Rodriguez, C. I., Deckersbach, T., . . . Nemeroff, C. B. (2019). Electroencephalographic Biomarkers for Treatment Response Prediction in Major Depressive Illness: A Meta-Analysis. *Am J Psychiatry, 176*(1), 44-56. doi:10.1176/appi.ajp.2018.17121358

Widge, A. S., Ellard, K. K., Paulk, A. C., Basu, I., Yousefi, A., Zorowitz, S., . . . Eskandar, E. N. (2017). Treating refractory mental illness with closed-loop brain stimulation: Progress towards a patient-specific transdiagnostic approach. *Exp Neurol, 287*(Pt 4), 461-472. doi:10.1016/j.expneurol.2016.07.021

Wilens, T., McBurnett, K., Stein, M., Lerner, M., Spencer, T., & Wolraich, M. (2005). ADHD treatment with once-daily OROS methylphenidate: final results from a long-term open-label study. *J Am Acad Child Adolesc Psychiatry, 44*(10), 1015-1023. doi:10.1097/01.chi.0000173291.28688.e7

Williams, L. M., Rush, A. J., Koslow, S. H., Wisniewski, S. R., Cooper, N. J., Nemeroff, C. B., . . . Gordon, E. (2011). International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials, 12*, 4. doi:10.1186/1745-6215-12-4

Williams, L. M., Simms, E., Clark, C. R., Paul, R. H., Rowe, D., & Gordon, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker". *Int J Neurosci, 115*(12), 1605-1630. doi:10.1080/00207450590958475

Wium-Andersen, I. K., Vinberg, M., Kessing, L. V., & McIntyre, R. S. (2017). Personalized medicine in psychiatry. *Nord J Psychiatry, 71*(1), 12-19. doi:10.1080/08039488.2016.1216163

Wong, E. H., Yocca, F., Smith, M. A., & Lee, C. M. (2010). Challenges and opportunities for drug discovery in psychiatric disorders: the drug hunters' perspective. *Int J Neuropsychopharmacol, 13*(9), 1269-1284. doi:10.1017/S1461145710000866

Yamamuro, K., Ota, T., Iida, J., Nakanishi, Y., Matsuura, H., Uratani, M., . . . Kishimoto, T. (2016). Event-related potentials reflect the efficacy of pharmaceutical treatments in children and adolescents with attention deficit/hyperactivity disorder. *Psychiatry Research, 242*, 288-294.

Yamamuro, K., Ota, T., Iida, J., Nakanishi, Y., Suehiro, Y., Matsuura, H., . . . Kishimoto, T. (2016). Event-Related Potentials Correlate with the Severity of Child and Adoles-

cent Patients with Attention Deficit/Hyperactivity Disorder. *Neuropsychobiology, 73*, 131-138. doi:10.1159/000444490

Yong, E. (2018, November 19). Psychology's Replication Crisis Is Running Out of Excuses. Retrieved from *https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/*

Zaehle, T., Lenz, D., Ohl, F. W., & Herrmann, C. S. (2010). Resonance phenomena in the human auditory cortex: individual resonance frequencies of the cerebral cortex determine electrophysiological responses. *Experimental Brain Research, 203*, 629-635. doi:10.1007/s00221-010-2265-8

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *NeuroImage, 53*(4), 1197-1207. doi:10.1016/j.neuroimage.2010.06.041

Zietsch, B. P., Hansen, J. L., Hansell, N. K., Geffen, G. M., Martin, N. G., & Wright, M. J. (2007). Common and specific genetic influences on EEG power bands delta, theta, alpha, and beta. *Biol Psychol, 75*(2), 154-164. doi:10.1016/j.biopsycho.2007.01.004

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making Replication Mainstream. *Behav Brain Sci*, 1-50. doi:10.1017/S0140525X17001972

ENGLISH SUMMARY

Personalized medicine aims to use biomarkers to disentangle psychiatric disorders. This could be either in a diagnostic (i.e., using biomarkers to diagnose disorders) or in a prognostic (i.e., using biomarkers to predict treatment outcome) way. This latter approach was one of the focal points of this thesis and it attempted to identify replicable biomarkers that are associated with clinical improvement.

In Chapter 2 (Krepel et al., 2018), Chapter 3 (Roelofs et al., 2021), Chapter 3.1 (Bailey et al., 2021), Chapter 4 (Krepel et al., in press), and Chapter 5 (Krepel et al., 2019), depression was addressed. The symptoms of depression are a depressed mood, diminished interest or pleasure, weight changes, slowing down of thoughts and a reduction of physical movement, fatigue, feelings of guilt or worthlessness, diminished ability to think or concentrate (or indecisiveness), and suicidal ideation (based on the DSM-V). Currently, treatments for depression mainly consist of psychotherapy, pharmacotherapy, or a combination of both. A lot of research is performed to provide new treatment options, one of which is repetitive Transcranial Magnetic Stimulation (rTMS). rTMS has shown to be effective in the treatment

of depression and aims to alleviate depressive symptoms by stimulating the brain (usually the dorsolateral prefrontal cortex (DLPFC)) with magnetic pulses.

Chapter 2 showed a non-replication of predictors of non-response to rTMS. The original study showed that more frontal theta, a larger P300 amplitude, and a lower individual alpha frequency (IAF) were associated with non-response to rTMS. In a newly acquired sample, these findings could not be replicated. In this paper, we also offered to share our database consisting of 196 depressed individuals treated with rTMS in combination with psychotherapy for replication studies. This data-sharing proposal led to multiple replication papers in collaboration with other labs and research groups.

Chapter 3 reported a collaborative effort, in which a finding that was reported by another lab was attempted to be replicated. The researchers from the original study found that the distance from the IAF to 10 Hz was related to clinical response, expressed as a negative association between these variables. Thus, a smaller IAF-10 Hz distance was related to a greater clinical response (measured by symptom decrease in percentage). This result was successfully replicated. Chapter 3.1 showed another collaboration, this time reporting a non-replication. The original study reported higher theta connectivity and low alpha power in individuals who responded to rTMS treatment, but these effects could not be replicated.

In Chapter 4 another collaborative effort was described. The original study showed that depressed females who show a higher suicide risk (including suicide ideators) exhibited less frontal beta/gamma activity. In the replication study, using a large, multisite sample, these effects could not be replicated.

Chapter 5 investigated whether baseline psychological features were associated with response to rTMS treatment. This was done using a Discovery-Replication approach, in which a large sample was split into a Discovery sample (in which exploratory analyses were performed) and Replication sample (in which results that were obtained in the exploratory analyses were tested for replicability). It was found

that anhedonia was significantly higher in individuals who did not respond to rTMS treatment, yet anhedonia could not sufficiently predict rTMS response. Therefore, the clinical relevance of this finding was considered to be low.

Another topic discussed in this thesis is attention-deficit/hyperactivity disorder (ADHD). ADHD can be diagnosed when several different criteria are met. An individual should report on several symptoms of inattention and/or hyperactivity/impulsivity, and some of these should be present before the age of 12. These symptoms should occur in multiple settings (e.g., school, work, at home, etcetera) and interfere with daily functioning.

In Chapter 6 (Krepel et al., 2020), QEEG-*informed* neurofeedback was investigated as a treatment for ADHD. QEEG-*informed* neurofeedback is an approach in which the Quantitative Electroencephalogram (QEEG) report is used to determine the neurofeedback protocol. The original study showed that QEEG-*informed* neurofeedback effectively can reduce symptoms of ADHD. In a newly acquired sample, this effect could be replicated. Potential predictors of neurofeedback remission were identified, among which were lower levels of hyperactivity (for the total sample), shorter P300 latencies (for women only), and a lower IAF (for boys only).

In Chapter 7 (Krepel et al., under review), spindling excessive beta (SEB) and its association with impulse control problems was investigated. The original study reported that individuals with frontocentral SEB reported experiencing more self-rated impulse control problems as well as more sleep maintenance problems. Chapter 7 (Krepel et al., under review) showed, using a heterogeneous sample, that individuals with frontocentral SEB reported experiencing more self-rated impulse control problems. This was also reflected by more false positive responses on a Continuous Performance Task (CPT). No associations with sleep were found. These results were independent of diagnoses and suggest that the development of transdiagnostic biomarkers is possible.

NEDERLANDSE SAMENVATTING

Gepersonaliseerde zorg richt zich erop om het gebruik van biomarkers in te zetten ten behoeve van het begrijpen van psychiatrische aandoeningen. Dit kan op een diagnostische (oftewel diagnoses maken op basis van biomarkers) en prognostische (oftewel biomarkers inzetten om behandeluitkomst te voorspellen) wijze. Deze laatste aanpak was één van de onderzoeksthema's van deze thesis, waarbij geprobeerd werd om repliceerbare biomarkers te identificeren die geassocieerd zijn met klinische verbeteringen.

In Hoofdstuk 2 (Krepel et al., 2018), Hoofdstuk 3 (Roelofs et al., 2021), Hoofdstuk 3.1 (Bailey et al., 2021), Hoofdstuk 4 (Krepel et al., in press), en Hoofdstuk 5 (Krepel et al., 2019) werd depressie geadresseerd. De symptomen van een depressie omvatten een sombere stemming, verminderde interesse of plezier, veranderingen in het gewicht, vertraagd denken en een vermindering van fysieke activiteit, vermoeidheid, gevoelens van schuld of waardeloosheid, een verminderd denk- of concentratievermogen (of besluiteloosheid), en suïcidale gedachten (gebaseerd op de DSM-V). Huidige behandelingen van depressie bestaan met name uit psychotherapie, medicatie, of een combinatie hiervan. Veel onderzoek wordt besteed aan het vin-

den van nieuwe behandelopties, en één daarvan is repetitieve Transcraniële Magnetische Stimulatie (rTMS). rTMS is effectief gebleken bij de behandeling van depressie en richt zich erop om symptomen van depressie te verminderen middels het stimuleren van het brein (gewoonlijk de dorsolaterale prefrontale cortex (DLPFC)) met magnetische pulsen.

Hoofdstuk 2 liet een non-replicatie van voorspellers van non-respons op een rTMS behandeling zien. De originele studie vond dat meer frontale theta, een grotere P300 amplitude, en een lagere individuele alfa frequentie (IAF) geassocieerd waren met niet reageren op een rTMS behandeling. Deze bevindingen konden niet worden gerepliceerd in een nieuwe onderzoekspopulatie. In deze publicatie is ook een voorstel om onze database, bestaande uit 196 depressieve personen die elk behandeld zijn met rTMS in combinatie met psychotherapie, te delen met andere wetenschappers die hun eigen bevindingen willen repliceren. Dit voorstel heeft tot meerdere samenwerkingen geleid.

Hoofdstuk 3 toonde een samenwerking, waarbij een bevinding van een ander lab werd getracht te repliceren. De onderzoekers van de originele studie vonden dat de afstand van 10 Hz naar de IAF was geassocieerd met klinische verbetering, wat te zien was als een negatieve associatie tussen de variabelen. Dat wil zeggen, een kleinere afstand tot 10 Hz was gerelateerd aan een betere respons op rTMS. Dit resultaat was succesvol gerepliceerd. Hoofdstuk 3.1 liet ook een samenwerking zien, die ditmaal een non-replicatie rapporteerde. De originele studie vond dat meer theta connectiviteit en lagere alfa power waren geassocieerd met reageren op een rTMS behandeling, maar deze effecten konden niet worden gerepliceerd.

In Hoofdstuk 4 is een andere samenwerking beschreven. De originele studie vond dat vrouwen met een verhoogd suïcide risico een frontale hypoactiviteit in de beta/gamma band toonden. In een grote onderzoekspopulatie, verzameld op meerdere plekken, kon dit effect niet worden gerepliceerd.

Hoofdstuk 5 onderzocht of psychologische variabelen konden wor-

den geassocieerd met de klinische respons op een rTMS behandeling. Dit werd gedaan middels een Exploratie-Replicatie aanpak, waarbij een grote onderzoekspopulatie was gesplitst in een Exploratie (waarin exploratieve analyses werden verricht) en een Replicatie (waarin resultaten die waren bevonden in de exploratieve analyses werden getest voor repliceerbaarheid) onderzoekspopulatie. Anhedonie was geïdentificeerd als zijnde significant hoger in individuen die niet reageerden op een rTMS behandeling. Echter, anhedonie kon niet voldoende het behandelsucces voorspellen, en dus werd de klinische relevantie van deze bevinding als laag bevonden.

Een ander onderwerp behandeld in dit proefschrift is attention-deficit/hyperactivity disorder (ADHD). ADHD kan worden gediagnosticeerd aan de hand van meerdere criteria. Een persoon moet meerdere klachten rapporteren op het gebied van aandachtstekort en/of hyperactiviteit/impulsiviteit, en sommige van deze klachten moeten aanwezig zijn voor de leeftijd van 12 jaar. De klachten moeten in meerdere omgevingen voorkomen (b.v., op school, werk, thuis, enzovoorts) en interfereren met dagelijks functioneren.

In Hoofdstuk 6 (Krepel et al., 2020), QEEG-*geïnformeerde* neurofeedback was onderzocht als een behandeling voor ADHD. QEEG-*geïnformeerde* neurofeedback is een werkwijze waarbij het Quantitative Electroencephalogram (QEEG) rapport wordt gebruikt om een bijpassend neurofeedback protocol te kiezen. De originele studie vond dat dit een effectieve manier is om symptomen van ADHD te verminderen. In een nieuwe onderzoekspopulatie kon dit effect worden gerepliceerd. Ook werden potentiële voorspellers van neurofeedback remissie gevonden, waaronder minder klachten van hyperactiviteit/impulsiviteit (voor de totale onderzoekspopulatie), een kortere P300 latentie (alleen voor vrouwen), en een verlaagde IAF (voor jongens).

In Hoofdstuk 7 (Krepel et al., under review) is de relatie tussen spindling excessive beta (SEB) en problemen met impulscontrole onderzocht. De originele studie rapporteerde dat personen met frontocentrale SEB meer zelf-gerapporteerde problemen met impulscontrole alsook meer doorslaapproblemen lieten zien. Hoofdstuk 7 vond, gebruikmakend van een heterogene onderzoekspopulatie, dat personen

met frontocentrale SEB meer zelf-gerapporteerde problemen met impulscontrole hadden. Dit werd ook gereflecteerd door meer fout positieve reacties bij een Continuous Performance Task (CPT). Er werden geen associaties met slaap gevonden. De huidige resultaten waren onafhankelijk van diagnose en suggereren dat het mogelijk is om transdiagnostische biomarkers te ontwikkelen.

IMPACT PARAGRAPH

This thesis studied the potential of biomarkers to be used in psychiatry, with a focus on the robustness and clinical relevance of these biomarkers. These topics are important for the development of stratified psychiatry and eventually personalized medicine. Stratified psychiatry, in this thesis considered as an interim step between the current treatment system and personalized medicine, aims to provide information on treatment success for subgroups of people who are identically diagnosed. For example, treatments for depression involve (but are not limited to) psychotherapy, pharmacological treatment, and repetitive Transcranial Magnetic Stimulation (rTMS). In stratified psychiatry, biomarkers provide information on what subgroup of depressed individuals may respond best to either of the aforementioned (or a combination of these) treatments, thereby circumventing the current trial-and-error approach. Personalized medicine goes one step further in optimizing treatment to the individual. It is a framework through which each individual reporting psychiatric complaints is assessed and treatment is allocated according to the assessed biomarkers. Biomarkers show the potential to make more informed treatment decisions and thereby increase the clinical effectiveness of psychiatric treatment. This is

important because, currently, the treatment of psychiatric disorders rests on a one-size-fits-all approach and clinical efficacy is limited.

The research presented in this thesis found that some biomarkers show the potential to be further developed as stratification tools in stratified psychiatry or personalized medicine. For example, it was found that depressed individuals with an individual alpha frequency (IAF) closer to 10 Hz respond better to 10 Hz rTMS treatment than individuals whose IAF is further away from 10 Hz. It was also found that boys reporting symptoms of attention-deficit/hyperactivity disorder (ADHD) with a low IAF respond well to QEEG-*informed* neurofeedback, whereas previous reports investigating IAF in methylphenidate treatment showed the opposite effect. In addition, spindling excessive beta (SEB), a feature observed in the electroencephalogram (EEG), is indicative of impulse control problems and this feature was similarly present in two groups who were diagnosed differently – yielding this a transdiagnostic EEG feature.

Importantly, besides developing biomarkers, this thesis focused on assessing the replicability of these biomarkers – for a biomarker that does not replicate, and thus does not show to be robust, cannot be used in clinical practice. As such, the biomarkers mentioned in the previous paragraph successfully survived replication. Yet, this thesis also identified biomarkers that could not be replicated. For example, a low IAF, more frontal theta, and a larger P300 amplitude could not be replicated as predictors of rTMS non-response. Likewise, less frontal beta/gamma activity could not be replicated as a biomarker of suicidal ideation in females reporting symptoms of depression. Also, a specific relation between SEB and sleep maintenance problems was not observed, as was reported in the original study. The importance of assessing the replicability of scientific findings is multifold. Not only does it test the existence of a finding, but it also helps establish the foundation on which future research can be based. Related to this is the importance of reporting null-findings and non-replications, as not reporting these findings can result in a skewed, unreliable representation of (the robustness of) a scientific finding. As such, performing and reporting on replication studies is vital for the progress and reliability of scientific findings.

As described before, replication research can be used to confirm or refute certain results, but its applicability to scientific practice extends beyond this dichotomy. Replication research can help explain previously contradictory findings, sculpt and refine existing work, increase methodological soundness and transparency, and establish collaborations between researchers and research facilities. These aspects are not only important for internal research groups but extend to the scientific community as a whole. For example, working together with other researchers is accompanied by different perspectives on the same topic, different ideas, and different knowledge frameworks through which a given scientific finding can be explained. This all helps in theory building and accelerating scientific progress. On a more practical level, working together with other research facilities may encourage data sharing, thereby creating larger and multi-site samples, resulting in representative datasets with high power. More so, sharing the responsibility of reporting the results of replication attempts increases the transparency and reliability of the representation of scientific findings in the literature. All these aspects are essential to consider while evaluating the results reported by scientific studies.

Another focus of this thesis has been the assessment of clinical relevance besides statistical significance. Clinical relevance is important to consider in mental health research, as the majority of the research in this research area is aimed at improving mental healthcare. Yet, many studies are based on statistical significance alone without assessing clinical relevance. This is problematic, for a statistically significant finding may not be clinically relevant. For example, clinically irrelevant findings cannot be used to base clinical decisions on or have a limited impact on the treatment or the individual seeking treatment. This thesis focused on clinical relevance by assessing the predictive value of findings and by focusing on remission rather than response. For example, it was shown that anhedonia was higher in depressed individuals who did not respond to rTMS treatment and this effect was replicated. Yet, the predictive value of this finding was relatively low – yielding a clinically irrelevant finding. Contrary, a study on the effectiveness of QEEG-*informed* neurofeedback as a treatment for ADHD reported that approximately half of the indi-

viduals achieved remission (indicative of experiencing minimal to no symptoms) after treatment. Assessing clinical relevance helps to establish the usefulness and impact of scientific findings in clinical practice and should be considered in studies focusing on improving psychiatric treatment.

This thesis has also shown that it is possible to base scientific research on heterogeneous samples. One issue of the one-size-fits-all approach is that diagnoses are primarily based on subjective reports from a patient, yet a lot of overlap in symptoms may exist between groups. Also, a lot of variation in symptom profiles within disorders exists. Individuals that are currently diagnosed with the same disorder may therefore not share similar symptoms, nor does it mean that individuals that have different diagnoses share no symptoms at all. Focusing on diagnoses and confining research to these diagnoses may result in groups that do not share similar pathological mechanisms, which may complicate research. Migrating the focus from diagnoses to shared constructs or concepts may be important for the development of robust biomarkers that correlate to behavioral profiles and transcend diagnostics. The concept of transdiagnostic research also applies to research incorporating healthy controls. Biomarkers that distinguish healthy controls from patients may not always be clinically useful, as clinical decision making rarely comes down to distinguishing a healthy control from a patient. Rather, biomarkers may help distinguish patients whose symptoms are similar, but whose diagnoses (and potentially treatment) are different (for example, in the case of unipolar and bipolar depression). Research in heterogeneous samples provides the first step in this direction, as transdiagnostic research focuses on shared constructs without being confined to one particular disorder.

CURRICULUM VITAE

Noralie Krepel was born on the 4th of November in 1993 in Dordrecht, The Netherlands. She started her academic education at University College Maastricht, where she received her Bachelor of Arts with honors in 2015. During her bachelor, Noralie attended an honors program in sciences at the University of Washington in Seattle, United States of America. Noralie continued her education at Maastricht University, at which she obtained her Master of Science in Psychology, specialization Neuropsychology in 2016.

After she finished her education, Noralie began an extra internship at Kempenhaeghe (Academic Center for Epileptology, Center for Sleep Medicine, and Center for Neurological Learning and Development Disabilities), gaining experience in clinical care as well as in research. Then, Noralie applied for a PhD position at Brainclinics Foundation, of which this thesis is the result.

During her time as a PhD student, Noralie attended multiple international conferences at which she gave various presentations. In 2020, Noralie also was part of the local organizing committee of the 4th European Conference of Brain Stimulation in Psychiatry.

Noralie Krepel werd op 4 november 1993 geboren in Dordrecht. Ze begon haar opleiding aan University College Maastricht, waar zij haar Bachelor of Arts met honors verkreeg in 2015. Tijdens haar bachelor is Noralie ook op uitwisseling geweest naar Seattle, Verenigde Staten van Amerika, waar zij een honors programma in de wetenschappen volgde. Zij vervolgde haar opleiding bij Universiteit Maastricht, waaraan ze in 2016 haar Master of Science in Psychology, specialization Neuropsychology behaalde.

Na haar opleiding is Noralie een extra stage begonnen bij Kempenhaeghe (Academisch Centrum voor Epileptologie, Centrum voor Slaapgeneeskunde, en Centrum voor Neurologische Leer- en ontwikkelingsstoornissen), waar zij ervaring in de kliniek en het onderzoek opdeed. Hierna heeft Noralie gesolliciteerd naar de functie als promovenda bij Stichting Brainclinics Foundation, waarvan het huidige proefschrift het resultaat is.

Tijdens haar promotietraject heeft Noralie meerdere internationale congressen bijgewoond waar zij verschillende presentaties gaf. Daarnaast was Noralie in 2020 ook onderdeel van het lokaal organiserende comité van de 4e European Conference of Brain Stimulation in Psychiatry.

LIST OF PUBLICATIONS

PUBLISHED

Krepel, N., Benschop, L., Baeken, C., Sack, A. T., Arns, M. (in press). An EEG Signature of Suicidal Behavior in Female Patients with Major Depressive Disorder? A Non-Replication. *Biological Psychology.*

Bailey, N., Krepel, N., van Dijk, H., Leuchter, A., Vila-Rodriquez, F., Blumberger, D., Wilson, A., Daskalakis, Z., Carpenter, L., Corlier, J., Arns, M., Fitzgerald, P. (2021). Resting EEG Theta Connectivity and Alpha Power to Predict rTMS Response in Depression: A Non-Replication from the ICON-DB Consortium. *Clinical Neurophysiology, 132*(2), 650-659. doi: 10.1016/j.clinph.2020.10.018

Roelofs, C., Krepel, N., Corlier, J., Carpenter, L., Fitzgerald, P., Daskalakis, Z., Tendolkar, I., Wilson, A., Downar, J., Bailey, N., Blumberger, D., Vila-Rodriquez, F., Leuchter, A., Arns, M. (2021). Individual Alpha Frequency Proximity Associated with rTMS Outcome: An Independent Replication Study from the ICON-DB Consortium. *Clinical Neurophysiology, 132*(2), 643-649. doi: 10.1016/j.clinph.2020.10.017

Krepel, N., Egtberts, T., Sack, A. T., Heinrich, H., Ryan, M., Arns, M. (2020). Multicenter Effectiveness Trial of QEEG-Informed Neurofeedback in ADHD: Replication and Treatment Prediction. *NeuroImage: Clinical, 28*, 102399. doi: 10.1016/j.nicl.2020.102399

Krepel, N., Rush, A. J., Iseger, T. A., Sack, A. T., Arns, M. (2019). Can psychological features predict antidepressant response to rTMS? A Discovery-Replication approach. *Psychological Medicine, 50*(2), 264-272. doi: 10.1017/S0033291718004191

Krepel, N., Sack, A. T., Kenemans, J. L., Fitzgerald, P. B., Drinkenburg, W. H., Arns, M. (2018). Non-replication of neurophysiological predictors of non-response to rTMS in depression and neurophysiological data-sharing proposal. *Brain Stimulation, 11*(3), 639-641. doi: 0.1016/j.brs.2018.01.032

IN PREPARATION/SUBMITTED

Krepel, N., van Dijk, H., Sack, A. T., Swatzyna, R. J., Arns, M. To Spindle or Not to Spindle: A Replication Study Into Spindling Excessive Beta as a Transdiagnostic EEG Feature Associated with Impulse Control. *Under review.*

OTHER PUBLICATIONS

Wu, W., Zhang, Y., Jiang, J., Lucas, M. V., Fonzo, G. A., Rolle, C. E., Cooper, C., Chin-Fatt, C., Krepel, N., Cornelssen, C. A., Wright, R., Toll, R. T., Trivedi, H. M., Monuszko, K., Caudle, T. L., Sarhadi, K., Jha, M. K., Trombello, J. M., Deckersbach, T., Adams, P., McGrath, P. J., Weissman, M. M., Fava, M., Pizzagalli, D. A., Arns, M., Trivedi, M. H., Etkin, A. (2020). An electroencephalographic signature predicts antidepressant response in major depression. *Nature Biotechnology, 38*(4), 439-447. doi: 10.1038/s41587-019-0397-3

Iseger, T. A., Vollebregt, M. A., Krepel, N., Arns, M. (2019). Heart rate variability related to season of birth: A replication study. *Psychophysiology, 56*(10), e13419. doi: 10.1111/psyp.13419

CONFERENCE CONTRIBUTIONS AND OTHER PRESENTATIONS

PRESENTATIONS

Kunnen psychologische kenmerken in depressie rTMS respons voorspellen? Een exploratie-replicatie methode. *NedKAD-conferentie*, October 2018, Amersfoort, The Netherlands.

Clinical and neurophysiological predictors of rTMS response in major depressive disorder: Robustness and clinical relevance. *3rd European Conference on Brain Stimulation in Psychiatry* – From Mechanisms to Medicine, October 2018, Lyon, France.

Can psychological features predict antidepressant response to rTMS? A Discovery-Replication approach. *20th Biennial International Pharmaco-EEG Society IPEG Meeting*, November 2018, Zürich, Switzerland.

Clinical and neurophysiological predictors of rTMS response in major depressive disorder: Robustness and clinical relevance. *3rd International Brain Stimulation Conference,* February 2019, Vancouver, Canada.

Predicting rTMS treatment response in MDD using EEG and clinical markers: Robustness, replication and refinement. *7th International Conference on Non-invasive Brain Stimulation,* November 2020, virtual.

POSTER PRESENTATIONS

Non-replication of neurophysiological predictors of non-response to rTMS in depression and neurophysiological data-sharing proposal. *Society of Biological Psychiatry's 2018 Annual Meeting*, May 2018, New York, United States of America. For abstract see: *Biological Psychiatry* 2018;82:S129-S455.