Personality and Social Psychology

Validity and reliability of electroencephalographic frontal alpha asymmetry and frontal midline theta as biomarkers for depression

CHRISTIAN GOLD, 1 JÖRG FACHNER 2 and JAAKKO ERKKILÄ 1

1Grieg Academy Music Therapy Research Center (GAMUT), Uni Health, Uni Research, Bergen, Norway
2Finnish Center of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Jyväskylä, Finland

INTRODUCTION

Clinical endpoints, surrogate endpoints, and biomarkers

An important distinction between types of outcomes in clinical trials is between clinical endpoints and surrogate endpoints or biomarkers. A clinical endpoint is a characteristic that is of direct importance for the patient, reflecting “how a patient feels or functions, or how long a patient survives” (De Gruttola, Clax, DeMets et al., 2001). In the case of depression, the clinical endpoint is usually the score on a standardized depression rating scale. In contrast, a biomarker is an objectively measured indicator of normal or pathological processes (De Gruttola et al., 2001). How well this indicator reflects a clinical endpoint may vary. A biomarker is called a surrogate endpoint if it is intended to substitute for a clinical endpoint (De Gruttola et al., 2001). The potential benefit of using surrogate endpoints is that they may indicate clinical change early, enabling trials to be shorter and smaller than if using clinical endpoints. Thus, the quality of a biomarker is related not only to mechanisms of disease but also to the mechanisms of treatment. However, this advantage is bought at the expense of additional uncertainty if the relation between surrogate and clinical endpoint is imperfect – which it usually is. There are examples of treatments that showed a beneficial effect on surrogate endpoints but no effect or even harm on clinical endpoints (Psaty, Weiss, Furberg et al., 1999). There are also other uses of biomarkers other than to substitute for a clinical endpoint. If used together with other indicators, they may help to better understand the mechanisms in which a treatment works or the pathogenic processes of the condition.

EEG measures as biomarkers for depression and anxiety

Frontal alpha asymmetry. In mental health, it has been suggested that certain measures derived from quantitative electroencephalography (QEEG) may serve as biomarkers for depression (Leiser, Dunlop, Bowly & Devilbiss, 2011). Frontal alpha asymmetry (FAA) has been studied most extensively, both in clinical and non-clinical studies (Jakobi, 2009; Thibodeau, Jorgensen & Kim, 2006). When correlating depression test scores with EEG highest values were found at mid-frontal electrodes (Thibodeau et al., 2006). Power in the alpha-band was originally thought to be inversely related to activity in terms of energy consumption and brain activity (Lindsley & Wicke, 1974; Oakes, Pizzagalli, Hendrick et al., 2004) but is now interpreted to reflect active inhibition rather than idling (Cooper, Croft, Dominey, Burgess & Gruzelie, 2003). In any case, findings seem to suggest that depressed individuals have decreased left anterior cortical activity (or more accurately relative right-sided frontal activity, because the asymmetry could be due to increased right-sided or decreased left-sided activity, or a combination of both; Thibodeau et al., 2006).

In spite of the extensive clinical and non-clinical research on FAA, there is no agreement as to its potential as a biomarker for depression. Some studies have found no association between FAA and depressive symptom severity, for example, Vuga, Fox, Cohn, George, Levenstein & Kovacs (2006, p. 114), concluding that “EEG asymmetry reflects a stable individual difference that is robust to variation in clinical status”. Two recent meta-analyses have attempted to summarize the evidence for FAA as a marker for depression. Thibodeau et al. (2006) found on average small to
moderate correlation of FAA with depression ($r = 0.26$ in 26 studies of adults with depression) and anxiety ($r = 0.17$ in eight studies on adults), but with high degrees of heterogeneity across studies that could not be fully explained. One source of variation was the scalp site. F3–F4 was most commonly measured (22 studies on adult depression) and showed the most substantial effect ($r = 0.26$). Another meta-analysis of 19 clinical and non-clinical studies also found a moderate correlation between FAA, measured at F3–F4, and depression score ($r = 0.19$; Jakobi, 2009, p. 52). However, publication bias might be an issue (Thibodeau et al., 2006). A recent study using a large clinically depressed sample still found no conclusive evidence, but provided further hypotheses concerning different reference electrodes and possible sex-specific differences (Stewart, Bismark, Towers, Coan & Allen, 2010).

Frontal midline theta. Frontal midline (FM) theta has been suggested as a potential marker for anxiety (Sueutsugi, Mizuki, Ushijima et al., 2000), but the literature on FM theta is much more limited than for FAA. FM theta is discussed as being a correlate of heightened mental effort and sustained attention observable in states of low-level awareness (Mitchell, McNaughton, Flanagan & Kirk, 2008). It is often examined in the context of task-related EEG activity but is also evident in rest and even sleep (Inanaga, 1998). It may help to distinguish between high and low anxious personality in pleasant/unpleasant tasks and can be enhanced in clients when receiving anxiolytic medication (for a review see Mitchell et al., 2008). Mitchell et al. (2008, p. 170) reviewed correlational evidence that “FM theta was more likely in extrovert, less neurotic, and less anxious subjects”. However, the non-clinical nature of the samples and the small sample sizes mean that those observations should at present be viewed as hypotheses rather than as confirmed scientific fact. Whether FM theta in rest EEG may prove useful as a biomarker for anxiety or depression is unknown, but it may be hypothesized that low levels of FM theta are associated with higher levels of anxiety.

Objectives and hypotheses of this study

The present paper aimed to determine to what extent FAA and FM theta are valid and reliable biomarkers for the severity of depression and anxiety in adults with a diagnosis of depression. Specifically, we aimed to examine: (1) mean values of FAA and FM theta values, compared to the general population; (2) internal consistency and test-retest reliability of FAA and FM theta measures over 3 months; and (3) correlation of FAA and FM theta values with standardized psychiatric measures of depression and anxiety.

Our main hypothesis was that if FAA and FM theta are valid biomarkers for the severity of depression/anxiety, they should be significantly correlated with the severity of depression/anxiety as measured using standardized psychiatric tests. Because of the high comorbidity between depression and anxiety and the strong overlap between depressive and anxiety symptoms, we chose to examine correlations of both FAA and FM theta with both depression and anxiety. We hypothesized a positive correlation between FAA and depression/anxiety (see Method section on the direction of the FAA metric) and a negative correlation between FM theta and depression/anxiety. In order for these measures to be regarded as valid biomarkers, this should apply to people with any severity level of depression (mild, moderate, or severe depression) and regardless of medication and other treatments received.

METHOD

Participants

Participants were referred to the study site through specialized outpatient centers in central Finland, according to inclusion criteria more fully described by Ekkilä, Gold, Fachner, Ala-Ruona, Punkanen & Vanhala (2008) and Ekkilä, Punkanen, Fachner et al. (2011). They were also recruited through newspaper advertisements. The sample was thus not randomly selected but clinically referred, and all participants were in outpatient treatment at the time of their inclusion. The study was approved by the ethical board of the Central Finland Health Care District, and patients signed informed consent before entering the study. A total of 79 adults (62 female; age range 18 to 50, M = 35.6, SD = 9.8) were recruited and agreed to participate in the study. All had a clinical diagnosis of depression, which was confirmed independently by a psychiatrist and a nurse with special training in depression assessment. The Structured Clinical Interview for DSM III R (Mini-SCID) (Spitzer, Williams, Gibbon & First, 1992) was used in health centers and polyclinics for diagnosing depression. In addition, a clinical expert with a specific training in diagnosing depression assessed all participants before inclusion, using the Montgomery and Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979). The specific ICD-10 diagnostic codes, based on MADRS levels, were F32.0 (mild depressive episode, n = 23, 29%), F32.1 (moderate depressive episode, n = 36, 46%), and F32.2 (severe depressive episode, n = 20, 25%). The mean time since they were first diagnosed with depression was 5.5 years. Sixty three (80%) also had comorbid anxiety (as indicated by a HADS-A score of 8 or more). Of the 79 participants, 57 (72%) currently used any antidepressant medication and 20 (25%) used any other medication. More specifically, 46% used selective serotonin reuptake inhibitors (SSRIs), 18% serotonin and noradrenaline reuptake inhibitors (SNRIs), 3% monoamine oxidase inhibitors (MAOIs), 5% tricyclic antidepressants, 5% noradrenergic and specific serotonergic antidepressants (NaSSAs), 3% medication for bipolar disorder, 22% benzodiazepines, 13% anxiolytic medication and 5% antipsychotics. All participants were offered short-term psychotherapy and 33 (42%) were also offered music therapy (Ekkilä et al., 2011). Five participants (6%) were left-handed.

Four participants (5%) did not have usable EEG data due to technical failures and were excluded from the analysis, so that 75 were available with complete data at intake. At 3 months, 12 of the original 79 participants could not be followed up and EEG recording failed in five cases (including the four where it also failed at intake). Therefore, 75 participants were available with complete data for the correlation analysis, and 62 were available for the analysis of test-retest reliability. A correlation of $r = 0.30$ is generally regarded as a medium effect size (Cohen, 1988), and the available sample size provided sufficient power ($1 - \beta = 75\%$) to identify a correlation of this magnitude in a two-sided test (calculated using function pwr.t.test of R package pwr).

Psychiatric assessments

Psychiatric tests were taken at 0 and 3 months and were administered by a psychiatric nurse. Symptoms of depression were measured with the Montgomery and Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979). The MADRS consists of 10 items, and the total score varies between 0 and 60. High joint reliability (0.76 to 0.95) and sensitivity to change have been shown in several studies. Predictive validity for major depressive disorder has been demonstrated, and cut-off scores have been defined for severe, moderate, and mild forms of depression (very severe 44, severe 31, moderate 25, mild 15, recovered 7; Rush, First & Blacker, 2008).

Anxiety was evaluated by the anxiety subscale (HADS-A) of the Hospital Anxiety and Depression Scale (HADS-A; Zigmond & Snaith, 2012).
The HADS-A is a widely used, valid and reliable self-report scale (Bjelland, Dahl, Haug & Neckelmann, 2002; Hermann, 1997) consisting of 7 items. Scores can range from 0 to 21 (0–7 normal, 8–10 mild, 11–14 moderate and 15–21 severe), and higher scores indicate more anxiety. Internal consistency (Cronbach’s alpha 0.83) has been demonstrated for the Finnish version (Aro, Ronkainen, Storskrubb et al., 2004).

**EEG assessment**

**Recording technology.** The EEG was recorded from 32 scalp Ag/AgCl electrodes (Active pin-type electrodes; Biosemi Headcap, SignaGel; electrodes were placed at Fp1, Fp2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz, O2) using a BioSemi Active II amplifier system utilizing the Active View 6.05 recording software. Active electrodes voltage as displayed in the recording offset were (as recommended from BioSemi) below 25 mV at each scalp electrode. EEG and artifact signals (EOG, EMG, ECG) were amplified with a band pass of 0–417 Hz by BioSemi Active-Two amplifiers and sampled at 2048 Hz. To re-reference the data to a linked ear electrode pair two active electrodes were placed at the mastoids of the participants. Mastoid voltage differences were not exceeding 10 mV.

**Procedure and recording setting.** EEG was recorded in a dry and sound-proof but not magnetically or radio frequency shielded room. The EEG investigator and his assistant as well as the client were in the room. After filling out a short form where subjects were asked for date of birth, self-report of handedness and current medication (including frequency and amount of medication), they were sitting in recumbent or upright position in a comfortable armchair. A video protocol of the client was recorded for EEG artifact monitoring. Subjects were informed about this use and no one disagreed to be recorded for this purposes. Subjects were instructed to keep their eyes closed and relax for 5 minutes while recording the ongoing rest EEG. Using only an eyes closed condition was justified because previous research had shown no difference between eyes-open and eyes-closed conditions (Henriques & Davidson, 1991; Stewart et al., Bismark, 2010). Daytime of the EEG recording was not standardized.

**EEG data processing.** The 32 channels rest EEG, mastoid reference and artifact recordings (EOG, EMG and ECG) were imported into the Neuroguide (NG) EEG analysis software (version 2.5.6). The advantage and a reason to choose this FDA-approved clinical EEG analysis software is an integrated EEG normative database (Thatcher, Walker, Biver, North & Curtin, 2003) that allows to compute z-scores of the clients’ EEG and to compare their data in terms of normality. NG includes calibrated amplifier characteristics of BIOSEMI Hardware to enable comparability to the normative database, and we matched our EEG recording technology, recording procedures, and data processing to the procedures that were used in creating the normative database to enable a meaningful comparison of values.

Reflecting the discussion on the reference problem in FAA the scalp electrode Cz was chosen as collection reference (Blackhart, Minnix & Kline, 2006; Hagemann, Naumann & Thayer, 2001) for the import of the raw EEG data. Then the EEG was re-referenced to averaged linked mastoids using a montage set including both mastoid channels recorded with the raw EEG. NG does examples the EEG data to 128 Hz and filters (high pass cut-off = 1 Hz and low pass cut-off = 55 Hz) the raw data with a 5th Order Butterworth filter. The continuous data were screened for eye movement artifacts and drowsiness by using NG’s artifact processing toolbox that also computes split-half and test-retest reliability scores within one recording for the selected EEG data.

Power Spectral Analysis was performed for the EEG selections with a Fast Fourier Transform (FFT). For the FFT the EEG-selections were epoched into 2 seconds that would result for the re-sampled EEG in 256 digital timepoints on a frequency range from 0.5 to 55 Hz with a resolution of 0.5 Hz using a cosine taper window. Based on this the raw EEG data from each channel were transformed into predefined frequency ranges.

Regarding FAA and FM theta the resulting absolute power values from electrodes Fp1/Fp2, F3/F4, F7/F8 and Fz were selected from the standard theta (4–8Hz) and alpha ranges (8–12Hz). No pre-treatment analysis of the subject’s individual alpha frequency as proposed by Doppelmayr, Klimesch, Pachinger & Ripper (1998) was conducted. For the computation of population-normalized z-scores, each of the amplitudes within the frequency range for each channel of the 2 second epochs was log-transformed to better approximate a normal distribution and the low pass filter was set to 40 Hz. The average of edited epochs used of the 5 minutes rest recordings had a time length of 114 seconds (SD = 48, range 39 to 239) at intake and 133 seconds (SD = 47, range 23 to 241) at 3 months.

**Statistical analyses**

**FAA metric.** FAA was calculated in NG, which uses the lateralization coefficient formula \( 200 \times (L - R) / (L + R) \). This measure is a simple linear transformation of the metric \((R - L) / (R + L)\) (Davidson, 1988), but with a different scaling and, important for interpretation, reversed sign. With this formula, a positive value indicates higher amplitude in the left hemisphere compared to the right hemisphere, which is interpreted as less activity (or more inhibition of activity) in the left hemisphere. In other words, left frontal hypoeactivation (or inhibition of activation) is indicated by positive values, and according to theory a positive correlation with depression scores would be expected.

**Population database.** Single electrode values as well as FAA scores were transformed to z-scores, again using NG, based on a large \( n = 678 \), age-matched as well as sex- and condition-matched (eyes open/closed) normative database (Thatcher, Biver & North, 2009; Thatcher et al., 2003). Because the range of values in that database was larger than in the present study (i.e., with estimates available for people aged 2 months to 82 years, of both genders, with eyes open and closed), valid z-scores could be obtained for all participants in our sample. The z-distribution is a normal distribution with \( M = 0 \) and \( SD = 1 \) (also known as the standard normal distribution). Using the normative database, each participant’s values were thus transformed to a corresponding value from the z-distribution. From statistical theory, it is known that about two thirds of the population will have values between –1 and 1, and about 95% will fall in the range between –2 and 2 (Upton & Cook, 2002). This property greatly facilitates interpretation of values, with values in the range of +/-2 SDs commonly being regarded as normal (as a commonly known example, an intelligence quotient between 70 and 130, corresponding to a z-value between –2 and 2 through linear transformation, is commonly regarded as within the normal range into which 95% of all people will fall).

**Descriptive statistics.** We first examined the descriptive statistics of the sample for any preliminary confirmation or disconfirmation of our hypotheses. We used geometric means for raw values of absolute power because these values were (expectedly) skewed. (This is equivalent to taking the arithmetic mean of log values and exponentiating this back to the original scale.) For normally distributed data, arithmetic means and standard deviations were calculated and presented. Normality of distributions was achieved for values from single electrodes through natural log-arithmetic transformation. FAA scores calculated as described above were found to be normally distributed without any further transformation. Psychiatric test scores were also normally distributed without any transformation.

**Reliability statistics.** Internal consistency of FAA and FM theta measures was calculated using Cronbach’s alpha. Test-retest reliability over 3 months was calculated using intraclass correlation coefficients (ICC) in R package irr (Gamer, Lemon & Fellows, 2007). Of the various types of ICC, we used two-way random effects models as the most conservative choice, consistency type as the most adequate for test-retest reliability.

© 2012 The Authors.
single measures. The single measures ICC is more adequate in this context than average measures and also represents a more conservative choice than the average measures ICC. It will generally be lower and corresponds to the ICC1 reported in Tomarken, Davidson, Wheeler & Kinney (1992). The results concerning test-retest reliability were interpreted as lower bounds of the true reliability because some participants received treatment between the measurements. This additional source of variation could therefore reduce the estimates. This was considered tolerable because reliability was not the main focus of this paper, but rather a necessary preliminary investigation leading up to the analysis of validity which was the main focus; furthermore, if a high reliability estimate was obtained in spite of possible error variation due to treatment responses, it could safely be concluded that reliability was in fact high.

Correlations between EEG measures and psychiatric tests. The main hypothesis of this study concerned the presence of a significant and relevant correlation between EEG measures and psychiatric tests. Values of these measures at intake were examined using Pearson correlation coefficients with 95% confidence intervals. Graphical methods (scatterplots) were also used and presented to further examine the possibility of any non-linear association. We interpreted the magnitude of correlations according to established guidelines (Cohen, 1988) as "small" \( r = 0.10 \), "medium" \( r = 0.30 \), and "large" \( r = 0.50 \). All statistical and graphical analyses were carried out in R (R Development Core Team, 2008).

**RESULTS**

**Means of raw and population-normalized z-values**

Descriptive statistics of the EEG data at intake are shown in Table 1. Scores of psychiatric tests were \( M = 23.67 \) (SD = 7.10) for MADRS and \( M = 10.67 \) (SD = 3.77) for HADS-A. Geometric means of absolute power values (left column) indicated little difference between corresponding electrodes on the right and left hemisphere. The same can be seen for arithmetic means of log-transformed values. Clinical interpretation of these values is facilitated in the z-transformed values of all measures (right column). As noted above, we know from statistical theory that two thirds of the general population would fall between \(-1\) and 1, and 95% between \(-2\) and 2 on these z-values. The z-values seen in the right column of Table 1 all fall clearly within the range of \((-1, 1\) and are therefore all fairly representative of the general population. Two aspects should be noted: (1) The three columns of the table all show the same data in different units – raw scores and their log- and z-transformations, respectively; and (2) This analysis was descriptive in nature, not inferential. Even if a slight deviation of z-scores from zero might turn out statistically significant, this would not change the interpretation that all z-values were within the healthy or normal range. Thus, no relevant deviation of this clinical sample from the general population was found on FAA or FM theta measures.

**Table 1. Alpha and theta amplitudes in adults diagnosed with depression**

<table>
<thead>
<tr>
<th>Raw data, not normally distributed</th>
<th>Transformed data, normally distributed</th>
<th>Population-normalized z-scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alpha</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fp1</td>
<td>9.58</td>
<td>2.26 (0.89)</td>
</tr>
<tr>
<td>Fp2</td>
<td>9.87</td>
<td>2.29 (0.88)</td>
</tr>
<tr>
<td>Fp1–Fp2</td>
<td>–2.50 (5.96)</td>
<td>–0.02 (0.41)</td>
</tr>
<tr>
<td>F3</td>
<td>10.59</td>
<td>2.36 (0.91)</td>
</tr>
<tr>
<td>F4</td>
<td>11.02</td>
<td>2.40 (0.92)</td>
</tr>
<tr>
<td>F3–F4</td>
<td>–3.37 (7.34)</td>
<td>0.17 (0.54)</td>
</tr>
<tr>
<td>F7</td>
<td>7.54</td>
<td>2.02 (0.84)</td>
</tr>
<tr>
<td>F8</td>
<td>7.54</td>
<td>2.02 (0.84)</td>
</tr>
<tr>
<td>F7–F8</td>
<td>–0.07 (12.43)</td>
<td>0.17 (0.51)</td>
</tr>
<tr>
<td><strong>Theta</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>9.78</td>
<td>2.28 (0.51)</td>
</tr>
<tr>
<td>Fz</td>
<td>12.18</td>
<td>2.50 (0.52)</td>
</tr>
<tr>
<td>F4</td>
<td>10.07</td>
<td>2.31 (0.52)</td>
</tr>
</tbody>
</table>

*Notes: Valid n = 75. For raw data (\( \mu V^2 \)), the geometric mean was calculated because it is more applicable as a measure of central tendency than the arithmetic mean when data are skewed. Skewed distributions were expected and observed for these data. Data in this column were interpreted as lower bounds of the true reliability because some participants received treatment between the measurements. This additional source of variation could therefore reduce the estimates. This was considered tolerable because reliability was not the main focus of this paper, but rather a necessary preliminary investigation leading up to the analysis of validity which was the main focus; furthermore, if a high reliability estimate was obtained in spite of possible error variation due to treatment responses, it could safely be concluded that reliability was in fact high.*

**Table 2. Test-retest reliability of FAA and FM theta over 3 months**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data (95% CI)</td>
<td>0.44 (0.22 to 0.62) **</td>
<td>0.35 (0.11 to 0.55) **</td>
<td>0.46 (0.23 to 0.63) **</td>
<td>0.61 (0.42 to 0.74) **</td>
<td>0.89 (0.81 to 0.93) **</td>
<td>0.89 (0.82 to 0.93) **</td>
<td>0.85 (0.77 to 0.91) **</td>
<td>0.90 (0.83 to 0.94) **</td>
</tr>
<tr>
<td>Population-normalized z-score ICC (95% CI)</td>
<td>0.71 (0.56 to 0.81) **</td>
<td>0.33 (0.09 to 0.53) **</td>
<td>0.60 (0.41 to 0.74) **</td>
<td>0.68 (0.51 to 0.79) **</td>
<td>0.88 (0.81 to 0.92) **</td>
<td>0.88 (0.81 to 0.93) **</td>
<td>0.85 (0.77 to 0.91) **</td>
<td>0.88 (0.81 to 0.92) **</td>
</tr>
</tbody>
</table>

*Notes: Valid n = 62. Raw FM theta values were log–transformed. ICC – intraclass correlation (two–way random effects, consistency, single measures). CI – confidence interval. ** p < 0.01.*
FAA measure helped to improve test-retest reliability of raw values (ICC 0.61) and provided a good compromise for z-scores (ICC 0.68). For comparison and completeness, test-retest reliability of the psychiatric tests was found to be acceptable (MADRS: ICC 0.47; HADS-A: ICC 0.59). As noted above, test-retest reliability coefficients have to be regarded as lower bounds of the true reliability because some participants received treatment between the measurements.

**Correlation of EEG measures and psychiatric tests**

Relations between EEG measures and psychiatric tests were examined statistically (Table 3) and graphically (Figure 1). Of the 12 correlations analyzed, only one reached a moderate strength of association (HADS-A and FAA at F7–F8, \( r = 0.33, p < 0.01 \); Table 3). The other associations ranged between no effect and a small effect (\( r \) ranging from –0.13 to 0.17; Table 3) and were all non-significant, even without correcting for multiple testing. Scatterplots excluded the possibility of a strong but non-linear association (shown in Figure 1 for FAA \( z \)-values; similar results were obtained for FM theta and for raw values of all EEG measures). For comparison, the correlation between MADRS and HADS-A in this sample was 0.54, indicating that clinical anxiety and clinical depression are much stronger indicators of each other than the EEG biomarkers are of any of those.

Correlations of FAA measures with psychiatric tests tended to be positive (ranging from 0.03 to 0.33), as was hypothesized; however, this correlation was more clearly seen with anxiety than with depression. Correlation coefficients of FM theta with psychiatric tests tended to be negative (\( r \) ranging from –0.05 to –0.13), again in the hypothesized direction. As a sensitivity analysis, we also examined through multiple regression models whether the relationship between EEG and psychiatric tests was influenced by medication status, and found no such indication. Finally, in an exploratory analysis we also repeated the correlation analyses separately by gender (as suggested by Stewart et al., 2010) and found no differences.

**DISCUSSION**

**Findings**

FAA and FM theta have been suggested as potential biomarkers for depression and anxiety, but the evidence these theories are

---

**Table 3. Correlation between FAA, FM theta, and psychiatric tests**

<table>
<thead>
<tr>
<th>Depression (MADRS)</th>
<th>Anxiety (HADS–A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAA: Fp1–Fp2</td>
<td>0.06 (0.17 to 0.28)</td>
</tr>
<tr>
<td>FAA: F3–F4</td>
<td>0.03 (0.20 to 0.26)</td>
</tr>
<tr>
<td>FAA: F7–F8</td>
<td>0.17 (0.06 to 0.38)</td>
</tr>
<tr>
<td>FM theta: F3</td>
<td>–0.05 (–0.28 to 0.18)</td>
</tr>
<tr>
<td>FM theta: F4</td>
<td>–0.07 (–0.29 to 0.16)</td>
</tr>
</tbody>
</table>

**Notes:** Valid n = 75. All EEG measures are population-normalized \( z \)-values. CI – confidence interval. ** \( p < 0.01 \).

---

*Fig. 1. Correlations between FAA \( z \)-scores and psychiatric tests. Note. Alpha asymmetry \( z \)-scores are based on a normative database. Positive values indicate greater amplitude (less activation) in the left hemisphere. * \( p < 0.05 \); ** \( p < 0.01 \).*
based on stem primarily from basic research, ranging from animal studies to emotion theory research (Harmon-Jones, Gable & Peterson, 2009; Mitchell et al., 2008). Clinical research in the area is sparse and has often been limited by small sample sizes. The present study examined the reliability and validity of these EEG measures in a reasonably large, clinically referred, generalizable sample including mild, moderate and severe cases and people with and without medication and other treatments. We were primarily interested in examining validity but also examined reliability as a necessary but not sufficient precondition of validity. We found sufficient reliability but limited validity.

Internal consistency of FAA and FM theta, as measured on different locations, was high. This may indicate that the brain processes occurring at different locations are closely linked and/ or that the same brain processes occurring at one site are strong enough to be measurable at different electrodes. Especially the three electrodes of FM theta were so closely related that it would be redundant to analyze each of them separately. In order to avoid unnecessary multiple comparisons, they might best be replaced by a combined measure of FM theta. Test-retest reliability was also high (and similar as in previous studies, e.g., Tomarken, Davidson, Wheeler & Kinney, 1992). FM theta tended to be more reliable than FAA; population-normalized z-scores more than raw values; and combined measures more than individual measures.

In terms of the validity of FAA and FM theta to indicate severity of depression and anxiety, results were mixed and inconclusive at best. First, population-normalized z-values were on average very close to zero, indicating that it may not be possible to distinguish people with depression from the general population using these measures alone. Second, most of the predicted correlations with psychiatric tests were too weak to reach statistical significance, even with a reasonably large clinical sample that had sufficient power to detect moderately strong correlations. Most observed correlations were in the “small” range (Cohen, 1988), indicating that any association that might be present may not be strong enough to be clinically relevant. This did not change when we attempted to statistically control for medication in a sensitivity analysis. The only correlation where the point estimate was in the “medium” range (r = 0.33 for FAA at F7–F8 and anxiety) also had a confidence interval wide enough to include the possibility of only a small effect.

Figure 2 presents our findings in the context of current evidence, using the results of Thibodeau et al.’s (2006) meta-analysis. For FAA at F3–F4, the most commonly studied parameter so far, our results did not confirm the previous positive findings. The most likely explanation is that the meta-analysis may have overestimated the true correlation because of publication bias. For FAA measured at other electrodes, our results confirmed the previous null findings (with some tendencies for F7–F8 but near-zero correlation for Fp1–Fp2). For the less well-studied correlation of FAA with anxiety measures our results overlap with Thibodeau et al.’s, and add incremental evidence that this correlation might be stronger at F7–F8 than at other scalp sites. In summary, we found some limited evidence of an association between EEG measures and psychopathology. The association is far from deterministic, and there seem to be many other factors that are influencing the values of FAA and FM theta as well.

Limitations

In this study, we limited ourselves to the most commonly used measures of FAA and FM theta in resting EEG. We chose the linked mastoids reference because it has been described as the measure of choice, especially in FAA (Hagemann et al., 2001), and used the FAA metric based on the difference divided by the sum (as opposed to the difference of log scores) because is the most commonly used and accepted measure (Blackhart et al., 2006; Davidson, 1988). We cannot exclude the possibility that another type of measure might have produced a clearer association with psychiatric tests. In particular the choice of a reference remains a thorny issue in FAA research (Stewart et al., 2010). Circadian and seasonal influences on depression (Allen, Iacono, Depue & Arbisi, 1993) were not a focus of our analysis, and we did not control for the time of day when measurements were taken. Briere, Forest, Chouinard & Godbout (2003) reported lateralization effects of daytime and gender, and Peterson and Harmon-Jones (2009) showed FAA to be influenced by daytime and season, that is, they reported a relatively greater right frontal asymmetry during autumn mornings. We cannot exclude the possibility that clearer associations with psychiatric measures may have emerged had all these potential confounders been taken into account. However, as noted by Thatcher et al. (2003), randomly varying time of day may only increase variance but does not cause bias.

One might also argue that newer technologies with higher spatial resolution might be more sensitive, but we chose EEG because it is a well-established and relatively inexpensive technology with an extensive research base, and to our knowledge, also

\[ r (95\% CI) \]

Fig. 2. Comparison of FAA results with previous meta-analysis. Note. Estimates and 95% CIs of correlations between FAA and psychiatric tests from this study and Thibodeau et al.’s (2006) meta-analysis, respectively.
the only brain imaging technology that offers clinically relevant, large normative databases for standardization purposes. New indices and biomarkers based on both resting and task-related EEG are still being developed (Leuchter, Cook, Gilmer, et al., 2009; Segrave, Thomson, Cooper, Crote, Sheppard & Fitzgerald, 2010), presenting somewhat of a moving target for researchers seeking to confirm the validity of EEG-based biomarkers. In summary, our goal was to examine the validity of the most commonly used, not the most recently proposed EEG biomarkers of depression and anxiety. Even with a relatively large sample, it was necessary to limit the number of comparisons due to the problem of type I error accumulation in multiple testing.

Sample size is always an issue, even though our sample was much larger than what is commonly encountered in clinical studies in this field. An even larger sample could have detected smaller correlations more clearly and with greater power, but the question would then arise how clinically relevant such small correlations would be. The characteristics of our sample emphasized generalizability and therefore people with severe depression or people already receiving medication were not excluded. This was in order to be able to make valid inferences to a clinically referred population. In this context, the relatively large proportion of women in our sample might reflect not only a higher prevalence of depression in women, but also a greater propensity of women to seek treatment for depression.

Finally, the present study addressed primarily correlation of depression severity with EEG measures in people with depression. A comparison with a matched control group of people without depression was not intended in this study. Although comparison with the population database implied such a comparison, potential differences in technical equipments might have masked an existing difference.

Implications for practice

Clinicians – therapists or referrers – may wish to have an objective measure of change or an early indicator of later change to assess or predict clinical change in their clients. In music therapy, for example, some studies have used EEG-based measures (Fachner, Gold & Erkkiälä, 2012; O’Kelly & Magee, 2012) or other physiological parameters (Hasbeck, 2012; Teckenberg-Jansson, Huotilainen, Pulkki, Lipasnen & Järvenpää, 2011) either alone or alongside psychological measures to indicate clinical change, whereas others have relied on psychological assessments alone (Albornoz, 2011; Gattino, Riesgo, Longo, Leite & Faccini, 2011). The findings of this study suggest that the EEG measures examined should not be used as exclusive or primary measures for such purposes. Traditional clinical assessments that rely mainly on human judgment still serve that purpose much better. EEG measures such as FAA and FM theta may be used as an additional source of information. The present study has shown them to be reliable measures, which is the first precondition for using them; but their clinical validity is still unclear. Yet, given that mental functioning is a complex phenomenon, EEG measures might add interesting information if used critically and in addition to established psychiatric assessments. Researchers deciding to use them as outcome measures should probably continue to carefully monitor the correlation with more established outcomes.

Implications for research

Further research is needed to investigate potential biomarkers for depression. Both FAA and FM theta have an empirical basis from previous research to support their use. It might be that certain aspects of depression and/or anxiety are more closely related to FAA and FM theta than others. Given that an important basis for the theory on FAA comes from emotion research (Harmon-Jones et al., 2009; Tomarken, Davidson, Wheeler & Doss, 1992), it might be, for example, that FAA is more related to emotional than to cognitive aspects of depression. We did not aim to examine any such subclasses or subgroups in this study, but where we did we were unable to replicate previous suggestions (that FAA may be more relevant in women; Stewart et al., 2010). Overall, this shows how little is still understood about this phenomenon.

Similarly, an established distinction in the measurement of anxiety is between state and trait anxiety. Given that EEG recordings reflect current mental processes at the time of the recording, this distinction might be useful to investigate. Our results were based on rest EEG, and applying the same measures during specific tasks (e.g., measuring FM theta during a task involving effort or attention) might be worthwhile. EEG responses while listening to different pieces of music were also recorded as another part of this study and will be analyzed separately (Erkkiälä et al., 2008).

Also the relations among different electrode pairs (we examined the three most commonly used pairs) provide a useful area of future research. In our study, it appeared that F7–F8 was more clearly linked to pathological processes than Fp1–Fp2 or F3–F4, but this finding requires replication and further elaboration. Can the neuropsychological meaning of these different sites be clarified further? Is FAA a one-dimensional or a multidimensional phenomenon? The internal consistency of FAA measures found in our study – high but not very high – still leaves both interpretations open. Other technologies that enable locating brain processes more precisely, such as fMRI or PET, may also help to investigate these questions. Other potentially useful biomarkers based on EEG recordings (e.g., Leuchter, Cook, Gilmer et al., 2009) may also be investigated further.

Finally, it has been stated above that a biomarker should reflect not only relevant pathological mechanisms but also the mechanisms of change of a specific treatment. If this is already non-trivial for pharmacological treatments whose active ingredients are relatively easily defined (Leuchter, Cook, Gilmer et al., 2009; Leuchter, Cook, Marangell, et al., 2009), it becomes even more complex for behavioural therapies (such as music therapy; Erkkiälä et al., 2008). This opens up a very wide field of research. Such investigation into mechanisms of change would benefit if a clear association with pathological processes was established first.

CONCLUSION

Frontal alpha asymmetry and FM theta are reliable measures, but their potential as biomarkers for depression and anxiety remains unclear. Clinical interpretation is difficult and clinical trials using such biomarkers should be interpreted with care because associations with psychopathology are at best of moderate strength. As
of today, these measures do not seem to be suitable as surrogate endpoints intended to replace more traditional clinical assessments. If used, they should be used in addition rather than instead of standardized psychiatric or psychological assessments and the correlation between different measures should be observed critically. Further research into the effectiveness of EEG-based biomarkers is needed.

This study was supported by the European Union, Contract No 028570 (NEST), and the Center of Excellence in interdisciplinary music research, SA 20/510/2007. The sponsors had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication. The authors would like to thank Robert Thatcher for technical solutions, Milka Lenninen and Tuomas Eerola for help with methodological questions.

NOTES

1It should be noted that the process of selecting a measure remains subjective. A biomarker is “objective” only in the sense that it is not easily influenced by social expectancy bias and similar biases that may be encountered in clinical assessments. That is, a measure can only be objective once it is decided which measure to use.

2These two metrics can easily be transformed into each other by multiplying or dividing by (~200). Because this is a linear transformation, the magnitude of correlation estimates will not be affected by the choice of one of these two metrics.

REFERENCES


Received 27 July 2012, accepted 9 October 2012