Critical Evaluation of Auditory Event-Related Potential Deficits in Schizophrenia: Evidence From Large-Scale Single-Subject Pattern Classification

Andres H. Neuhaus*14, Florin C. Popescu2, Johannes Rentzsch1, and Jürgen Gallinat1

1Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Berlin, Germany; 2Competence Center IT4Energy, Fraunhofer Institute for Open Communication Systems FOKUS, Berlin, Germany

*To whom correspondence should be addressed; Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Campus Benjamin Franklin, Eschenallee 3, 14050 Berlin, Germany; tel: 49-30-8445-8412, fax: 49-30-8445-8393, e-mail: andres.neuhaus@charite.de

Event-related potential (ERP) deficits associated with auditory oddball and click-conditioning paradigms are among the most consistent findings in schizophrenia and are discussed as potential biomarkers. However, it is unclear to what extent these ERP deficits distinguish between schizophrenia patients and healthy controls on a single-subject level, which is of high importance for potential translation to clinical routine. Here, we investigated 144 schizophrenia patients and 144 matched controls with an auditory click-conditioning/oddball paradigm. P50 and N1 gating ratios as well as target-locked N1 and P3 components were submitted to conventional general linear models and to explorative machine learning algorithms. Repeated-measures ANOVAs revealed significant between-group differences for the oddball-locked N1 and P3 components but not for any gating measure. Machine learning-assisted analysis achieved 77.7% balanced classification accuracy using a combination of target-locked N1 and P3 amplitudes as classifiers. The superiority of machine learning over repeated-measures analysis for classifying schizophrenia patients was in the range of about 10% as quantified by receiver operating characteristics. For the first time, our study provides large-scale single-subject classification data on auditory click-conditioning and oddball paradigms in schizophrenia. Although our study exemplifies how automated inference may substantially improve classification accuracy, our data also show that the investigated ERP measures show comparably poor discriminatory properties in single subjects, thus illustrating the need to establish either new analytical approaches for these paradigms or other paradigms to investigate the disorder.

Key words: N1/N100/P3/P300/event-related potentials/gating/click conditioning/auditory oddball/schizophrenia/machine learning

Introduction

The implementation of clinical tests to aid diagnostic processes is a central desiderate of biological psychiatry. Much effort has been devoted to the optimization of the diagnostic process of schizophrenia as reflected in the evolution of criteria from Bleuler’s descriptive psychopathology to diagnostic taxonomies of Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) and International Classification of Diseases, 10th revision (ICD-10).1–3 However, despite the hope to establish a biologically supported diagnostic system,4 reliable biomarkers are still lacking.

Deficits of certain auditory event-related potential (ERP) components are among the most consistent findings in schizophrenia. These include the P3 component that is usually elicited by applying the oddball paradigm as well as the gating of the P50 component in a click-conditioning paradigm. Decrement of P3 amplitude5,6 and P50 gating are long-standing findings in schizophrenia research,7,8 and there is good evidence that these deficits are independent of duration and severity of illness.9 Based on findings from genetic association and family studies, it has been suggested that the P50 gating and the P3 reflect promising endophenotype candidates for schizophrenia.10–13 However, despite considerable efforts to establish these ERP components as markers of schizophrenia, conclusive evidence of marker properties is missing.

Reasons for the large gap between a vast number of significant studies in the field of biological psychiatry and the absence of established biomarkers were recently delineated.14 The authors proposed several reasons such as the missing gold standard, significance chasing, approximate replications, and extreme comparisons. While it is beyond the scope of any single study to define a new diagnostic
gold standard, it seems worthwhile to address the remaining critiques with an appropriate study design. In this context, the criticism of “significance chasing” is of special interest. The mutual consent of accepting a maximum of 5% \(\alpha\) error probability allowed for establishing a scientific standard in medical research. On the other hand, this advancement seems to be accompanied by a relative neglect of the importance of effect sizes: once study results survive the statistical threshold, acceptance rates for concomitant publications rise dramatically, thus leading to publication biases in favor of significant findings that may be hard to replicate or of negligible clinical significance.\(^{15,16}\)

The present study addresses this particular criticism in the context of replicated electrophysiological signatures of schizophrenia, ie, the P50 gating and the oddball-associated P3 response. The goal of this study was to quantify the diagnostic accuracy of these electrophysiological measures as clinical differentiators. We employed a standard protocol to evoke gating and oddball responses in 288 subjects, thereby also addressing the criticisms of approximate replications and extreme comparisons. We applied repeated-measures ANOVAs and state-of-the-art machine learning algorithms to allow for comparing classical statistical methods and automated extraction of experimental data. Ultimately, this approach offers a classification accuracy of each approach that might answer the question why these electrophysiological measures, while often replicated, have not yet been translated to clinical practice.

Methods

Subjects

One hundred forty-four schizophrenic patients (36 females and 108 males) participated in this study. They met DSM-IV criteria for schizophrenia, were clinically stable as assessed with the Positive and Negative Syndrome Scale,\(^{17}\) and had no psychiatric disorder other than schizophrenia and nicotine abuse/dependence. Histories of severe medical or neurological disorders or of electroconvulsive therapy were exclusion criteria. All patients were recruited at the Department of Psychiatry, Charité University Medicine Berlin, Germany. All patients received antipsychotic medication (\(N = 88\) second generation; \(N = 24\) first generation; \(N = 32\) combination). One hundred forty-four healthy subjects (36 females and 108 males) were recruited as controls via newspaper advertisements and matched for age and gender. None of these participants had a history of substance abuse other than tobacco smoking, a history of psychiatric axis I disorder according to DSM-IV, including that they had never received any psychopharmacological treatment, a history of severe medical or neurological disorder, or a family history of psychiatric illness. All controls were examined by a psychiatrist prior to the study. All participants were right-handed and reported normal audition. Normal intelligence was confirmed with a German multiple choice vocabulary test.\(^{18}\) Clinical and demographic data of patients and controls are summarized in table 1. All subjects gave written informed consent before participating in this study. The study protocol was approved by the ethics committee of the Charité University Medicine, Berlin, Germany, and the study was conducted in accordance with the Declaration of Helsinki and its amendments.

Procedure and Task Design

Subjects were seated in a slightly reclined chair with a head rest in a sound-attenuated and electrically shielded room. Auditory stimuli were presented binaurally by calibrated headphones. We employed a combined auditory click-conditioning/oddball paradigm to assess gating and target detection. Stimuli were presented in pseudo-randomized order with frequent nontarget (175 double clicks, ie, 1 ms square waves with 500 ms stimulus onset asynchrony, presented at 109 dB) and rare target stimuli (55 sine waves at 1000 Hz with 40 ms duration including 10 ms rise and 10 ms fall at 83 dB) with a mean interstimulus interval of 2.8 seconds. Subjects were instructed to close their eyes and to respond to target stimuli by

Table 1. Summary of Demographic and Clinical Data

<table>
<thead>
<tr>
<th></th>
<th>Schizophrenia</th>
<th>Controls</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N) (female/ male)</td>
<td>144 (36/108)</td>
<td>144 (36/108)</td>
<td>—</td>
</tr>
<tr>
<td>Age (y) (range)</td>
<td>31.51 ± 11.0 (18–65)</td>
<td>32.42 ± 8.9 (19–65)</td>
<td>ns</td>
</tr>
<tr>
<td>Education (y) (range)</td>
<td>12.67 ± 2.5 (8–18)</td>
<td>14.16 ± 1.9 (10–18)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>IQ</td>
<td>99.84 ± 13.0</td>
<td>105.41 ± 12.1</td>
<td>.003</td>
</tr>
<tr>
<td>DOI (y)</td>
<td>6.87 ± 8.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PANSS positive scale</td>
<td>18.50 ± 6.4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PANSS negative scale</td>
<td>20.72 ± 6.8</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PANSS general scale</td>
<td>33.61 ± 11.2</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*Note: IQ, intelligence quotient; DOI, duration of illness; PANSS, Positive and Negative Syndrome Scale; ns, nonsignificant. Between-group differences were assessed with \(t\)-test for independent samples.*
pressing a response button as fast and as accurately as possible.

Electroencephalography Acquisition and ERP Construction

Electroencephalography (EEG) was recorded with 32 Ag/AgCl electrodes internally referenced to Cz using an electrode cap. The electrodes were positioned according to the extended 10/20 system with additional mastoid and electro-oculographic electrodes. Impedances were kept below 5 kΩ. EEG was continuously digitized at 500 Hz with a Neuroscan SynAmps. Analog band-pass filters were set at 0.16 and 100 Hz.

EEG analysis was conducted with Brain Vision Analyzer 1.05 (Brain Products). Ocular artifact correction was performed using independent component analysis. Oddball data were re-referenced against linked mastoids and digitally filtered at 20 Hz (low pass). Gating data were re-referenced to common average and filtered at 10 Hz (high pass), 50 Hz (notch), and 70 Hz (low pass) for P50 construction; and at 0.5 Hz (high pass) and 20 Hz (low pass) for N1 construction. We then applied an artifact criterion of ≥80 µV at any electrode and tagged artifacts for later removal. Next, data were segmented according to stimulus class and relative to stimulus onset (−350 to 800 ms); for the P50 analysis, segmentation was done for the first click to include both stimuli. Response accuracy and reaction time constrained usage of target epochs that analyzed if correct responses were recorded within 100–1500 ms following a target. After rejecting artifact-contaminated segments, ERPs were baseline corrected and averaged.

All ERP components were determined as baseline-to-peak amplitudes. P50 and N1 responses to click-conditioning stimuli were identified at Fz and Cz. P50 was determined between 40 and 70 ms (S1) and between 540 and 570 ms (S2) with a visual control post hoc (blind to diagnosis) to exclude misclassification of late P30 components. N1 was picked between 70 and 130 ms (S1) and between 570 and 630 ms (S2). Gating ratios were calculated as amplitude second stimulus/amplitude first stimulus; to avoid extraordinary high or even negative amplitudes, we chose to truncate the maximum ratio value to 2 (no gating) and to set the minimum at 0 (complete gating). Following target stimuli, target-locked N1 was identified at Fz and Cz between 70 and 130 ms poststimulus, and P3 was identified at Fz, Cz, and Pz as a prominent positive deflection at 250–450 ms. Whole-scalp ERP data of all components were extracted for machine learning analysis. Latencies and amplitudes were extracted for all electrodes, thus resulting in 2 (latency and amplitude) * 32 (channels) = 64 “features” for every ERP component.

The mean number of ERP segments in response to standard stimuli that were accepted for final analyses of gating measures were 154.85 ± 20.5 for schizophrenia patients and 166.49 ± 14.2 for control subjects (P < .001). Corresponding target ERP segments amounted to 46.20 ± 7.8 for schizophrenia patients and 49.91 ± 5.5 for controls (P < .001).

Repeated-Measures Analysis

Statistical calculations were conducted with SPSS Statistics 19.0 (IBM). Demographic and behavioral data were analyzed with t-tests for independent samples. Hypothesis-driven ERP analysis was performed with repeated-measures ANOVAs. For the P50 and N1 gating ratios as well as for the target-locked N1 amplitude, “Electrodes” (Fz and Cz) served as a 2-level within-subjects factor, while “Group” (schizophrenia, control) and “Sex” (female and male) served as 2-level between-subjects factors. The P3 ANOVA model was comparable but contained a 3-level-subjects factor “Electrode” (Fz, Cz, and Pz). Mauchly’s test assured that the sphericity assumption was not violated. Partial eta squared (η²) served as an estimator of effect size. Post hoc tests were performed as t-tests for paired or independent samples, as appropriate. All dependent variables were submitted to receiver operating characteristic (ROC) analyses to estimate the diagnostic fit when applying variables as clinical classifiers. All tests were performed as 2-tailed tests with an α level set at P < .01 due to the high statistical power of our sample.

Machine-Learning Pattern Classification

The machine learning discovery methods involve 3 separate stages of analysis: feature selection, hypothesis ranking, and validation. In brief, the whole dataset is split into a 75% training subset and a 25% validation subset. Using the training subset only, ERP features are ranked according to the strength of difference between groups and the most promising features are selected for further analysis. These features are grouped into different combinations, each analyzed with several algorithms, again using only the 75% training subset. The most promising feature-classifier combinations (“hypotheses”) are then reanalyzed using the 25% validation subset to confirm the hypothesis derived from the 75% training subset. The bulk of this analysis was done in Matlab (MathWorks) with help from the open-source packages Weka (data mining software in Java, available online at http://www.cs.waikato.ac.nz/ml/weka/), LibSVM (library for support vector machines [SVMs], available online at http://www.csie.ntu.edu.tw/~cjlin/libsvm/), and bolasso.m (bootstrap-enhanced least absolute shrinkage operator, online at http://code.google.com).

The first 2 stages involve a randomly chosen subset of the entire dataset, which comprises 75% of the data, ie, the training subset. In the first stage (feature selection), features are ranked according to each feature’s strength of
difference between control and patient groups as measured by the Kolmogorov-Smirnov (KS) test. In order to provide an intuitive measure of strength of difference (other than the \( P \) value of the KS test significance), we also calculated the classification value of the linear discriminant analysis (LDA) classifier, which is the simplest means of classifying a one-dimensional value. The justification for providing such an error value is to provide an approximate lower bound for possible final classification accuracy values of more complex, multidimensional classifiers.

The next step of analysis involves hypothesis ranking, where hypotheses refer to a candidate feature set analyzed with a specific algorithm, eg, 4 ERP features analyzed with naive Bayes. The hypotheses are formed as such: a candidate feature combination set is constructed and consists of exhaustive combinatorial subsets with up to 5 maximal features, to which is added an element consisting of all strong features, thus resulting in \( 2^5 = 32 \) candidate feature combinations. A product of the candidate feature combination set and the classifier set is formed to provide the hypothesis set. The classifier set is a list of common classification algorithms used in machine learning practice and is composed of linear and quadratic discriminants (LDA/quadratic discriminant analysis and their diagonal variants); SVMs with radial basis, linear, polynomial, and multilayer perceptron kernels; naive Bayes; \( k \)-nearest neighbors with Euclidean and cosine distance measures; and Mahalanobis with different types of regularization of the covariance estimator (no regularization, model consistent Lasso estimation through the bootstrap [BLOASSO], minimum description length [MDL], and robust regression). The hypothesis set is thus formed by these 15 classifiers plus a trivial classifier that assumes a candidate feature combination set is constructed and consists of exhaustive combinatorial subsets with up to 5 maximal features, to which is added an element consisting of all strong features, thus resulting in \( 2^5 = 32 \) candidate feature combinations. A product of the candidate feature combination set and the classifier set is formed to provide the hypothesis set. The classifier set is a list of common classification algorithms used in machine learning practice and is composed of linear and quadratic discriminants (LDA/quadratic discriminant analysis and their diagonal variants); SVMs with radial basis, linear, polynomial, and multilayer perceptron kernels; naive Bayes; \( k \)-nearest neighbors with Euclidean and cosine distance measures; and Mahalanobis with different types of regularization of the covariance estimator (no regularization, model consistent Lasso estimation through the bootstrap [BLOASSO], minimum description length [MDL], and robust regression). The hypothesis set is thus formed by these 15 classifiers plus a trivial classifier that assumes a candidate feature combination set is constructed and consists of exhaustive combinatorial subsets with up to 5 maximal features, to which is added an element consisting of all strong features, thus resulting in \( 2^5 = 32 \) candidate feature combinations. A product of the candidate feature combination set and the classifier set is formed to provide the hypothesis set. The classifier set is a list of common classification algorithms used in machine learning practice and is composed of linear and quadratic discriminants (LDA/quadratic discriminant analysis and their diagonal variants); SVMs with radial basis, linear, polynomial, and multilayer perceptron kernels; naive Bayes; \( k \)-nearest neighbors with Euclidean and cosine distance measures; and Mahalanobis with different types of regularization of the covariance estimator (no regularization, model consistent Lasso estimation through the bootstrap [BLOASSO], minimum description length [MDL], and robust regression). The hypothesis set is thus formed by these 15 classifiers plus a trivial classifier that assumes

evaluating on the remaining 1/10th, repeated 10 times and averaged) classification error called the predicted error is used to a final ranking of hypotheses in the second stage of analysis.

In the third and final step (validation), the hypotheses are evaluated (best-to-worst) on the 25% validation subset, which until now has not played part in any fitting or optimization.

**Results**

**Behavioral Performance**

Schizophrenia patients showed a mean accuracy of 96.09 ± 8.9%, while controls gave 97.70 ± 7.0% correct responses (not significant).

**Repeated-Measures Analysis**

Mean ERP amplitudes and SDs stratified across groups are summarized in [Table 2].

For the click-conditioning part of our paradigm, repeated-measures ANOVA did not indicate a significant main effect of “Group” for P50 (\( P = .239 \)) or N1 gating ratios (\( P = .806 \)). Likewise, neither “Sex” nor “Electrodes” yielded any significant main effects or interactions. Central midline ERP waveforms in response to click-conditioning stimuli are depicted in [Figure 1] for illustrative purposes.

For the oddball part of our paradigm, several significant main effects and interactions emerged for both target-locked N1 and the P3 response. Importantly, analysis of the target-locked N1 showed a significant main effect of “Group” (\( F(1,284) = 26.29; \ P < .001; \eta^2 = .085 \)), indicating a larger mean target-locked N1 response in controls (−13.23 ± 5.7 \( \mu \)V) than in schizophrenia patients (−9.98 ± 6.6 \( \mu \)V; \( T(286) = 4.49; \ P < .001 \)). Next, a significant main effect of “Electrode” was found (\( F(1,284) = 83.86; \ P < .001; \eta^2 = .239 \)) that was driven by a larger negativity at Cz (−13.0 ± 7.3 \( \mu \)V) compared to Fz (−10.21 ± 6.2 \( \mu \)V; \( T(287) = 9.87; \ P < .001 \)). Last, a significant “Electrode”×“Group” interaction (\( F(1,284) = 14.10; \ P < .001; \eta^2 = .047 \))

**Table 2.** Post Hoc Statistics of Event-Related Potential Component Amplitudes

<table>
<thead>
<tr>
<th>Schizophrenia</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>P50 gating ratio, Fz</td>
<td>0.77</td>
</tr>
<tr>
<td>P50 gating ratio, Cz</td>
<td>0.61</td>
</tr>
<tr>
<td>N1 gating ratio, Fz</td>
<td>0.59</td>
</tr>
<tr>
<td>N1 gating ratio, Cz</td>
<td>0.56</td>
</tr>
<tr>
<td>Target N1, Fz (( \mu )V)</td>
<td>−9.08</td>
</tr>
<tr>
<td>Target N1, Cz (( \mu )V)</td>
<td>−10.89</td>
</tr>
<tr>
<td>P3, Fz (( \mu )V)</td>
<td>6.41</td>
</tr>
<tr>
<td>P3, Cz (( \mu )V)</td>
<td>9.54</td>
</tr>
<tr>
<td>P3, Pz (( \mu )V)</td>
<td>12.26</td>
</tr>
</tbody>
</table>

*Note: ns, nonsignificant. All differences assessed with \( t \)-test for independent samples with Bonferroni correction.*
demonstrated greater between-electrode differences in healthy controls (Fz: $-11.35 \pm 5.5 \, \mu V$; Cz: $-15.12 \pm 6.8 \, \mu V$) than in schizophrenia patients (Fz: $-9.08 \pm 6.6 \, \mu V$; Cz: $-10.78 \pm 7.1 \, \mu V$), although both post hoc t-tests indicated significant between-electrode differences in both groups (both $P < .001$), owing to the sample size.

The P3 ANOVA model revealed a significant main effect of ‘Group’ ($F(1,284) = 9.08; \, P = .003; \, \eta^2 = 0.031$), driven by a larger mean P3 response in controls (11.39 $\pm 5.8 \, \mu V$) than in schizophrenia patients (9.40 $\pm 5.5 \, \mu V$; $T(286) = -3.01; \, P = .003$). A significant main effect of ‘Electrode’ ($F(2,572) = 252.19; \, P < .001; \, \eta^2 = 0.469$) was driven by a posterior-anterior amplitude gradient with highest amplitudes at Pz (13.74 $\pm 6.0 \, \mu V$), followed by Cz (10.31 $\pm 6.7 \, \mu V$), and Fz (7.14 $\pm 6.4 \, \mu V$; for all t-tests, $P < .001$). Target ERP waveforms are illustrated in figure 2.

Machine Learning Analysis

Feature selection was done to identify each feature’s strength to differentiate between control and patient groups. The identified strong features are illustrated in figure 3. Following feature selection and hypothesis ranking, the top 5 hypotheses identified are shown in table 3, in terms of prediction and validation errors, along with the expected SD of the binomial distribution based on the validation error and the number of samples in the validation set. Note that the predicted and validated errors are within the margin of error. The results revealed cross-validation errors of 22.2%–23.6% among the top 5 hypotheses. Here, naive Bayes achieved a classification accuracy of 22.2% using only target-locked ERP responses, with discriminant features being P3 voltage at O2 and N1 voltage at P4. Of note, chosen features were identified by virtually all machines and are not simply the first N-ranked features, but subsets of features.

ROC Analysis

ROC analyses indicated a disadvantageous fit of both P50 and N1 gating ratios with a mean area under the curve (AUC) of approximately 0.5 (see table 4), ie, in close proximity to the no-discrimination line. Target-locked

![Fig. 1. Grand average event-related potentials at Cz in response to click-conditioning stimuli. Stimulus onset is at 0 (solid line) and 500 ms (dotted line) for conditioning and test stimuli, respectively. Upper panel: P50 component; lower panel: N1 component.](image-url)
Large-Scale Classification of Schizophrenia

N1 amplitudes as well as parietal P3 amplitude achieved predictive models that deviated significantly from the no-discrimination line. The 2 ERP features identified by the naive Bayes algorithm were Z transformed and merged to a composite classifier that achieved a diagnostic classification accuracy clearly superior to the parietal P3 component (see figure 4).

Discussion

We investigated ERPs elicited by auditory click-conditioning and oddball stimuli and estimated the diagnostic accuracy of these measures in a large sample of schizophrenia patients and healthy controls. ERPs associated with oddball stimuli were significantly different between diagnostic groups and were also selected by automated single-subject classification. ROC analysis indicated classification accuracies of around 65% for the “classic” parietal P3 component (the best discriminator according to repeated-measures analysis) and around 74% for a composite score of oddball-associated ERPs. The incremental value of the present study is thus constituted by 2 results: for the first time, we provide diagnostic classification accuracies for the classic P3 component and for a composite measure derived from oddball data in a large and representative sample; these representative ERP data demonstrate that frequently reported ERP measures associated with auditory click-conditioning and oddball paradigms show comparably poor discriminatory properties in single subjects.

The obtained classification accuracy of around 74% is in the range of previous electrophysiological studies that reported 69%, 20 77%, 21 and 79% 22 classification accuracy. Of note, another study using auditory and visual oddball paradigms in a small sample of schizophrenia patients and controls arrived at a classification accuracy of around 72%. 23 Regarding classification accuracy in the present study, many other types of classifiers may have been tested, which might yield error rates with favorable percentages. However, some classifiers may work better by pure chance, even if cross-validated scores were used. 24 The goal of the present study was to investigate which

Fig. 2. Grand average event-related potentials to target stimuli (0 ms) at midline electrodes.
features are critical for establishing the difference among groups. For this purpose, the naive Bayes classifier, using no a priori knowledge and working on the data itself, found a very compact dimensionality of features ($N = 2$).

This classifier not only provided the strongest differences among classes but also led to the exclusion of all other features that differentiated less among groups and thus can be regarded as redundant with respect to classification.

Fig. 3. Strong features found in training set. The x-axis in each column is a nonlinear transformation of the basic value shown [2/$\pi$ atan(x/2)], chosen as to provide a common visual axis for features occurring on different scales and ranges. The curves and the intensity levels show a (smoothed) cumulative probability density function (CDF) for each class, along with the distribution of the absolute difference between these 2 (the maximum value of which is the Kolmogorov-Smirnov test statistic). The right column of the figure depicts the CDF difference for strong features. This CDF difference thus provides visual information about which feature types may provide complementary information about the diagnostic or subject state.

Table 3. Classification Accuracy, Measured as Prediction and Validation Error Rates

<table>
<thead>
<tr>
<th>Machine</th>
<th>PE</th>
<th>VE ± SD</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>22.65%</td>
<td>22.22 ± 4.9%</td>
<td>N1 at P4</td>
<td>P3 at O2</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>LDA</td>
<td>25.03%</td>
<td>23.61 ± 5.0%</td>
<td>N1 at FC2</td>
<td>P3 at O2</td>
<td>N1 at P4</td>
<td>—</td>
</tr>
<tr>
<td>QDA</td>
<td>24.60%</td>
<td>23.61 ± 5.0%</td>
<td>N1 at FC2</td>
<td>P3 at O2</td>
<td>N1 at P4</td>
<td>P3 at P4</td>
</tr>
<tr>
<td>QDA</td>
<td>24.61%</td>
<td>23.61 ± 5.0%</td>
<td>N1 at FC2</td>
<td>P3 at O2</td>
<td>N1 at P4</td>
<td>—</td>
</tr>
<tr>
<td>D-LDA</td>
<td>24.34%</td>
<td>23.61 ± 5.0%</td>
<td>N1 at FC2</td>
<td>P3 at O2</td>
<td>N1 at P4</td>
<td>—</td>
</tr>
</tbody>
</table>

Note: PE, prediction error; VE, validation error; NB, naive Bayes; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; D-LDA, diagonal linear discriminant analysis. Tested hypotheses are ranked according to validation error including SD. Within our machine learning approach, a hypothesis consists of a classifier (“Machine”) that is tested on a specific feature set (“Features 1–4”). For each machine learning classifier, prediction and validation error rates are given. ERP features (only amplitudes) are coded as: ERP component (N1 to target stimuli or P3) and electrode.
Large-Scale Classification of Schizophrenia

With respect to the implications of our main result, 2 critical issues have to be discussed that cast doubt on the implementation of the paradigms investigated here and currently widespread analytical approaches. First, we were not able to replicate the gating deficit usually reported for schizophrenia subjects, which may be due to several reasons. Most of our patients were medicated with atypical antipsychotics that have been reported to be associated with less attenuated or even normalized P50 gating. Next, most of the patients in the study were of the paranoid subtype, which may be related with better P50 gating compared to other subtypes; a subgroup analysis would have led to considerably smaller group sizes that would have severely limited our current machine learning approach. Last, and maybe most important, it has been shown that P50 gating has a limited retest-reliability. Further supporting this idea, there is a wide range of P50 gating values reported in the literature for both schizophrenia patients and healthy subjects, and many studies failed to find a significantly reduced P50 ratio in schizophrenia patients compared with healthy controls. Problems with test-retest reliability have been noted frequently, especially with ratio values that introduce a disadvantageous signal-to-noise ratio to both the numerator and the denominator.

Second, the single-subject classification accuracy of the conventional parietal P3 amplitude that is used in a wide range of studies was comparably low although ANOVA revealed a highly reliable significant difference between schizophrenia patients and control subjects. This contradiction between an apparently substantial and replicable result, on the one hand, and a comparably low clinical usefulness, on the other hand, illustrates an important reason why psychiatric research has not yet been able to identify biomarker of schizophrenia. As has been discussed before, study designs are usually geared towards existing, significant data that are either replicated or are thought to be replicable. The significance of a biological finding, however, does not necessarily mean that the respective effect is clinically important.

Our data nicely illustrate this line of thoughts by simultaneously calculating and contrasting (“high”) statistical significance and (relatively low) clinical discriminatory capacity. Although our study demonstrates how automated inference may improve classification accuracy, the final error rate clearly illustrates the need to establish either optimized analytical approaches for oddball and click-conditioning paradigms, novel paradigms, or other data assessment methods that are more suitable to investigate the disorder. The average ERP is thought to capture only that part of neuronal activity that is both time locked and phase locked to the stimulus, which may lead to a loss of a large amount of stimulus-related brain activity. Hence, new signal processing methods like single-trial classification and modeling of event-related brain dynamics using time-frequency or blind source separation analysis may be favorable for modeling complex pathophysiological network processes. Next, new paradigms or the investigation of other cognitive domains, eg, theory of mind or emotion recognition, could prove highly beneficial for discriminating schizophrenia and control groups. Last, magnetic resonance imaging methods have shown to hold some promise for differentiating between schizophrenia and mentally healthy subjects.

Some limitations have to be acknowledged. We used a different protocol to measure P50 gating with a shorter intertrial interval between click pairs that infrequently

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>SE</th>
<th>P</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>P50 gating ratio, Fz</td>
<td>0.503</td>
<td>0.034</td>
<td>.932</td>
<td>(0.436–0.570)</td>
</tr>
<tr>
<td>P50 gating ratio, Cz</td>
<td>0.523</td>
<td>0.034</td>
<td>.491</td>
<td>(0.456–0.591)</td>
</tr>
<tr>
<td>N1 gating ratio, Fz</td>
<td>0.436</td>
<td>0.034</td>
<td>.058</td>
<td>(0.369–0.502)</td>
</tr>
<tr>
<td>N1 gating ratio, Fz</td>
<td>0.470</td>
<td>0.034</td>
<td>.381</td>
<td>(0.403–0.537)</td>
</tr>
<tr>
<td>Target N1, Fz</td>
<td>0.620</td>
<td>0.033</td>
<td>&lt;.001</td>
<td>(0.555–0.686)</td>
</tr>
<tr>
<td>Target N1, Cz</td>
<td>0.681</td>
<td>0.032</td>
<td>&lt;.001</td>
<td>(0.619–0.743)</td>
</tr>
<tr>
<td>P3, Fz</td>
<td>0.574</td>
<td>0.034</td>
<td>.29</td>
<td>(0.508–0.641)</td>
</tr>
<tr>
<td>P3, Cz</td>
<td>0.569</td>
<td>0.034</td>
<td>.044</td>
<td>(0.502–0.635)</td>
</tr>
<tr>
<td>P3, Pz</td>
<td>0.654</td>
<td>0.032</td>
<td>&lt;.001</td>
<td>(0.591–0.718)</td>
</tr>
<tr>
<td>Composite classifier</td>
<td>0.737</td>
<td>0.029</td>
<td>&lt;.001</td>
<td>(0.680–0.794)</td>
</tr>
</tbody>
</table>

Note: AUC, area under the curve.

Fig. 4. Receiver operating characteristic curves illustrating P3 amplitude and the composite measure derived from naive Bayes machine learning model. Area under the curve (AUC) was 0.654 for the midline parietal P3 and 0.737 for the composite feature.

With respect to the implications of our main result, 2 critical issues have to be discussed that cast doubt on the implementation of the paradigms investigated here and currently widespread analytical approaches. First, we were not able to replicate the gating deficit usually reported for schizophrenia subjects, which may be due to several reasons. Most of our patients were medicated with atypical antipsychotics that have been reported to be associated with less attenuated or even normalized P50 gating. Next, most of the patients in the study were of the paranoid subtype, which may be related with better P50 gating compared to other subtypes; a subgroup analysis would have led to considerably smaller group sizes that would have severely limited our current machine learning approach. Last, and maybe most important, it has been shown that P50 gating has a limited retest-reliability. Further supporting this idea, there is a wide range of P50 gating values reported in the literature for both schizophrenia patients and healthy subjects, and many studies failed to find a significantly reduced P50 ratio in schizophrenia patients compared with healthy controls. Problems with test-retest reliability have been noted frequently, especially with ratio values that introduce a disadvantageous signal-to-noise ratio to both the numerator and the denominator.

Second, the single-subject classification accuracy of the conventional parietal P3 amplitude that is used in a wide range of studies was comparably low although ANOVA revealed a highly reliable significant difference between schizophrenia patients and control subjects. This contradiction between an apparently substantial and replicable result, on the one hand, and a comparably low clinical usefulness, on the other hand, illustrates an important reason why psychiatric research has not yet been able to identify biomarker of schizophrenia. As has been discussed before, study designs are usually geared towards existing, significant data that are either replicated or are thought to be replicable. The significance of a biological finding, however, does not necessarily mean that the respective effect is clinically important.

Our data nicely illustrate this line of thoughts by simultaneously calculating and contrasting (“high”) statistical significance and (relatively low) clinical discriminatory capacity. Although our study demonstrates how automated inference may improve classification accuracy, the final error rate clearly illustrates the need to establish either optimized analytical approaches for oddball and click-conditioning paradigms, novel paradigms, or other data assessment methods that are more suitable to investigate the disorder. The average ERP is thought to capture only that part of neuronal activity that is both time locked and phase locked to the stimulus, which may lead to a loss of a large amount of stimulus-related brain activity. Hence, new signal processing methods like single-trial classification and modeling of event-related brain dynamics using time-frequency or blind source separation analysis may be favorable for modeling complex pathophysiological network processes. Next, new paradigms or the investigation of other cognitive domains, eg, theory of mind or emotion recognition, could prove highly beneficial for discriminating schizophrenia and control groups. Last, magnetic resonance imaging methods have shown to hold some promise for differentiating between schizophrenia and mentally healthy subjects.

Some limitations have to be acknowledged. We used a different protocol to measure P50 gating with a shorter intertrial interval between click pairs that infrequently
also included an additional infrequent sine wave stimulus as opposed to the classic P50 gating protocol that usually comprises longer intertrial intervals of about 8 seconds or more. However, we explicitly addressed this paradigmatic variation in a previous study and did not find any P50 gating differences between our and the classic protocol, thus contradicting the assumption that long interstimulus intervals are necessary to elicit a marked sensory gating phenomenon for P50 and N100 auditory responses. Next, we did not aim to provide sensitivity and specificity measures in a broad sense. This study aimed at differentiating only 2 clinical states, thus we cannot draw any conclusion about discriminatory capacity of our measures in a clinical sample that would include several different psychiatric diagnoses. Last, and inherent to virtually any large-scale studies of schizophrenia, all of our patients were medicated with a wide range of antipsychotic agents. Given that gating and oddball measures seem to be impaired already in very early stages including the prodromal phase, antipsychotic medication may distort the native brain responses to auditory stimuli, which in turn might have impacted on ERP features in our study, especially because some antipsychotic agents like clozapine have been reported to normalize oddball-related ERPs like the P3 to a certain extent. This hypothetical confound can only be controlled for by a replication study in unmedicated schizophrenia subjects.

In conclusion, the present study aimed at characterizing the classification accuracy of both gating and oddball-related ERP measures in schizophrenia at the single-subject level. Contrary to expectation based on the literature, we only found moderate discriminatory capacities of those ERP measures under investigation. Although our approach also illustrates the potential of machine learning algorithms for identifying biomarkers that are independent of, and thus might be employed in addition to, clinical assessments, new analytical approaches, and/or new paradigms are clearly needed in the search for biomarkers of schizophrenia.

Acknowledgments

The authors thank all participants of this study for contributing their time and effort. The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

3. World Health Organization. Tenth Revision of the International Classification of Diseases, Chapter V (F):


46. Kiehl KA, Liddle PF. An event-related functional magnetic resonance imaging study of an auditory oddball task in schizophrenia. *Schizophr Res*. 2001;48:159–171.


