Abstracts are listed in order of presentation.
Classifications: Clinical 1-6; Strategic 7-23; Statistical 24-28

1  The MATRICS Consensus Cognitive Battery (MCCB): Clinical and Cognitive Correlations

Sharon M. August, Jacqueline N. Kiwanuka, Samuel T. Kaiser, Leeka Hubzin, James M. Gold
Maryland Psychiatric Research Center, Dept of Psychiatry, University of Maryland School of Medicine.

Introduction: The MATRICS Consensus Cognitive Battery (MCCB) was developed to be an endpoint for clinical trials aiming to enhance cognition in schizophrenia. The battery is also used to characterize cognitive performance profiles. To date, there is little data in the literature concerning the clinical and cognitive correlates of the battery. The goal of this work is to begin to fill this gap in the literature.

Methods: A total of 119 patients with schizophrenia and 75 healthy volunteers were assessed. The patients and healthy volunteers all performed the MATRICS battery as part of a research protocol at the MPRC. Participants also received a WTAR, WRAT4, and WASI. Ratings on the BPRS, LOF, and SANS were performed with patients.

Results: Patients performed more poorly than healthy controls on each of the seven MATRICS cognitive domain scores and on the Overall Composite score (T = 30.06 patients and T = 50.15 controls, p=0.000). The MATRICS Overall Composite score was highly correlated with WASI IQ (r= 0.741, p=0.000 in patients, r=0.695, p=0.000 in controls). A discriminant function analysis revealed that the Processing Speed and Social Cognition domain scores were most sensitive to diagnostic status. After accounting for WASI IQ differences between groups, Processing Speed and Social Cognition domain scores still remained significant predictors of diagnosis. We also observed the Social Cognition domain to be negatively correlated (r=0.241, p=0.01) with BPRS disorganized syndrome when looking at correlations with BPRS symptom ratings. To examine functional outcome, we classified patients as having either good or poor vocational performance and having good or poor social function based on LOF rating. Good vocational performance and social function were defined by average scores 3 or 4 on the employment/work quality and social contacts/quality of relations items. Patients with good vocational performance scored significantly higher on the Processing Speed, Attention Vigilance, and Working Memory domains, as well as the Overall Composite scores. Patients with good social function scored significantly higher on the Social Cognition domain score.

Conclusions: The MATRICS battery is sensitive to the cognitive impairments observed in schizophrenia. The MATRICS domain scores are highly inter-correlated, and the overall score is highly correlated with IQ. There is evidence that the test is sensitive to important dimensions of outcome including vocational success and social function. Patients differing in social function differed on the Social Cognition Domain, providing evidence for the validity of the scale.

2  Paliperidone Palmitate vs Oral Antipsychotics for People With Schizophrenia Recently Released From Jail: Rationale and Methodology

Reuven Ferziger¹, Larry Alphs¹, Norris Turner¹, Lian Mao², Stephen Rodriguez¹, Lisa Dixon³, John Davis⁴, Joseph Hulihan¹

¹Ortho-McNeil Janssen Scientific Affairs, LLC, Titusville, NJ, USA, ²Johnson & Johnson Pharmaceutical Research and Development, LLC, Titusville, NJ, USA, ³University of Maryland School of Medicine, Baltimore, MD, USA, ⁴University of Illinois at Chicago, Chicago, IL, USA

Background: The overrepresentation of people with serious mental illness (SMI) in US jails and prisons is an important public mental health problem. After release from incarceration, successful community reentry is challenged by barriers to obtaining essential healthcare, social services, and financial resources, perpetuating a revolving door of incarceration followed by failed community reentry. Although untreated psychotic illness
may be an important variable, no studies have compared the effectiveness of psychopharmacologic treatments in individuals with schizophrenia following release from jail. A recently initiated clinical study compares a monthly, long-acting injectable antipsychotic with daily oral antipsychotics in delaying time to community reentry failure in patients with schizophrenia who have recently been released from jail. We report on the nature of the problem, the study rationale, and methodological challenges encountered in designing a psychopharmacology intervention trial that addresses public health and clinical dimensions of a problem frequently encountered by people with schizophrenia.

Methods: This 15-month randomized, open-label, multicenter, effectiveness study compares paliperidone palmitate with oral antipsychotics in subjects with schizophrenia recently released from jail and incarcerated at least twice within the preceding 24 months. Before randomization, the investigators/clinicians identify individually suitable oral antipsychotics from a list of 7 commonly prescribed oral antipsychotics. At randomization, subjects are assigned (1:1) to treatment with paliperidone palmitate or treatment with one of the prespecified oral antipsychotics. The primary endpoint is time to occurrence of a treatment failure event, defined as arrest, hospitalization, suicide, discontinuation of antipsychotic due to inadequate efficacy, safety, or tolerability; supplementation with another antipsychotic because of inadequate efficacy; or an increase in the level of services to prevent imminent psychiatric hospitalization.

Results: There is extensive site variation because of regional differences in mental health systems and available services for this population among the states and counties where the study sites are located. This both reflects the real world nature of the study and poses methodological challenges for the study design and analysis.

Conclusions: This is a comparative effectiveness study to determine whether time to treatment failure for people with schizophrenia recently released from jail differs between treatments with paliperidone palmitate, a long-acting atypical injectable antipsychotic, and commonly prescribed oral antipsychotics. Endpoints were chosen to address meaningful clinical and public health outcomes. Methodological approaches addressed here may be helpful to future studies at the intersection of clinical research and mental health policy.

Disclosers: Supported by Ortho-McNeil Janssen Scientific Affairs, LLC; L Alphs is an employee of Ortho-McNeil Janssen Scientific Affairs, LLC, and a Johnson & Johnson stockholder; R Ferziger is an employee of Ortho-McNeil Janssen Scientific Affairs, LLC, and a Johnson & Johnson stockholder; N Turner is an employee of Ortho-McNeil Janssen Scientific Affairs, LLC, and a Johnson & Johnson stockholder; L Mao is an employee of Johnson & Johnson Pharmaceutical Research and Development, LLC, and a Johnson & Johnson stockholder; S Rodriguez is an employee of Ortho-McNeil Janssen Scientific Affairs, LLC, and a Johnson & Johnson stockholder; Lisa Dixon is principal investigator and consultant for Ortho-McNeil Janssen Scientific Affairs; J Davis has nothing to disclose; J Hulihan is an employee of Ortho-McNeil Janssen Scientific Affairs, LLC, and a Johnson & Johnson stockholder.

Designing a Benefit-Risk Assessment of Maintenance Therapy with an Oral Versus a Long-Acting Injectable Agent in Schizophrenia

Michael Markowitz1, Bennett Levitan2, Dong Jing Fu1, Ibrahim Turkoz2, Larry Alphs1

1Ortho-McNeil Janssen Scientific Affairs, LLC, Titusville, NJ, USA 2Johnson & Johnson Pharmaceutical Research and Development, LLC, Titusville, NJ, USA

Introduction: Maintenance therapy in schizophrenia is an important clinical and public health concern, particularly because of the many clinical consequences and high healthcare costs when patients fall out of maintenance. Because of potential adverse events that occur with long-term schizophrenia treatment, maintenance requires careful balancing of efficacy/benefit and safety/risk. Schizophrenia benefit-risk particularly needs to address the many endpoints that characterize the patient’s experience and the impact of endpoint uncertainty. The Benefit:Risk Action Team (BRAT) framework, developed by the Pharmaceutical Research and Manufacturer’s Association (PhRMA), is a structured approach to benefit-risk assessment that helps address these issues. This work describes our solutions to key schizophrenia maintenance-specific challenges when applying the BRAT framework approach.

Methods: Initial steps in the BRAT framework focus on identification/definition of outcomes and identification/extraction of data. Clinical outcomes were identified from the literature and consultation with clinical experts. Efficacy outcomes included relapse, CGI, PSP, and PANSS. Safety outcomes included EPS, QT prolongation, syncope, weight gain, lipid abnormalities, and hyperprolactinemia. Outcome rates were developed from 2 double-blind placebo-controlled relapse studies with comparable (1) endpoint definitions, (2) study designs (run-in/transition, stabilization, double-blind maintenance, and open-label extension
phases), and (3) inclusion/exclusion criteria. Patient-level data were available from both studies. One study used an oral formulation and one used a long-acting injectable formulation.

**Challenges:** Three issues proved particularly challenging: (1) The stabilization and relapse criteria differed slightly between the studies. To allow for fair comparison, the study populations were adjusted to match stabilization criteria using patient-level data. (2) Many events occurred more frequently at the start of the maintenance period. To avoid bias due to an assumption of constant risk and to study benefit-risk evolution with time, assessments were defined at multiple time points. (3) Maintenance response in the placebo arm of the injectable study was markedly higher than in the oral study, possibly because the injectable formulation used in the previous stabilization period remained in the blood for several months, whereas the oral formulation was cleared in days. Also, a placebo effect may result from a formulation that requires an injection. For these reasons, placebo correction of the active arm response or number needed to treat/harm would be misleading. Instead, the benefit-risk assessment is being based on differences between the active arms for each formulation directly.

**Conclusion:** Benefit-risk analysis comparing oral antipsychotics with long-acting injectable maintenance therapy poses several methodologic challenges, which we address.

**Disclosers:**

**The Cognitive Effects of Long-Acting Injectable vs. Oral Second-Generation Antipsychotic Medication after an Initial Psychotic Episode**

Nuechterlein, KH 1,2, Subotnik, KL 1, Ventura, J 1, Gretchen-Doorly D 1, Casaus, LR 1, Luo, JS 1, Detore, NR 1, Victoria, A 1, Kurtz, AS 1

1UCLA Department of Psychiatry and Biobehavioral Sciences, Semel Institute for Neuroscience and Human Behavior, 2UCLA Department of Psychology

**Introduction:** Cognition has become a key intervention target in schizophrenia due to its notable links to everyday functioning. While numerous studies of the effects of second-generation antipsychotic medications on cognition have been completed, the consistency of medication delivery has been relatively ignored in this literature. Furthermore, few of the studies have focused on the initial period after onset of schizophrenia in which medication adherence is particularly problematic.

**Method:** In an ongoing UCLA study focused on the impact of medication adherence in the initial course of schizophrenia, we are comparing the efficacy of the long-acting injectable versus oral forms of risperidone in a 12-month, open-label, randomized controlled trial. The enrollment target is 110. Interim analyses for an NIMH grant renewal application were conducted with 61 recent-onset schizophrenia patients who were randomized to injectable vs. oral risperidone at least 6 months prior to these analyses, 57 of whom had completed assessments at the 6 month point. Repeated-measures and mixed model ANOVAs were conducted to examine the effects of medication administration mode and medication adherence.

**Results:** On the MATRICS Consensus Cognitive Battery (MCCB), long-acting injectable risperidone displayed significantly greater improvement than oral risperidone in verbal learning (5.0 vs. 0.5 T-score improvement, p = .023) and a similar nonsignificant tendency in working memory (p = .074) and the overall cognitive composite score (4 vs. 1 T-score improvement, p = .093). A dimension of adherence to risperidone across the long-acting injectable and oral groups was also considered to allow better sensitivity to the effects of medication adherence. Each patient’s adherence was rated on a 1-5 scale based on timeliness of injections for injectable medication, and pill counts, patient reports, plasma levels, and psychiatrist judgments for oral medication. Adherence was much better for injectable medication (p < .001). Better adherence was significantly predictive of improvement on the MCCB from baseline to 6 months in the overall composite score (p = .014) and the cognitive domains of working memory (p = .024) and visual learning (p = .049).

**Conclusions:** These findings support the view that consistency of use of a second-generation antipsychotic medication, risperidone, is a key factor in improving cognitive functioning among recently diagnosed
Effects of Desvenlafaxine on Cognitive Function in Outpatients with Major Depressive Disorder

Sujana Reddy, MD, MBA; Kristen Focht, MBA; Jeff Musgnung, MS; Chris Edgar, PhD, MS; Christine J. Guico-Pabia, MD, MPH, MBA; Keith Wesnes, BSc, PhD

Objective: This substudy of a larger clinical trial was conducted to assess the effects of desvenlafaxine 50 mg/day and placebo on cognitive functioning in a population of employed adults with major depressive disorder.

Methods: Eighty-one outpatients that were randomly assigned to double-blind treatment with either placebo (n=29) or desvenlafaxine 50 mg/d (n=52) for up to 84 days were enrolled in this study. The CDR System a set of computerised cognitive tests was used to assess patients in the domains of memory, attention, executive functioning and affective processing. The CDR System was developed for use in clinical trials to assess changes in cognitive functioning over time, and has been previously used to compare the cognitive effects of reboxetine and paroxetine in patients with depression.1 The CDR System was administered at baseline and days 14, 28, 56 and 84. Baseline-adjusted data (observed cases [OC] and last-observation carried forward [LOCF]) were subjected to analysis of covariance using baseline levels as covariates; data are summarized below as mean differences from placebo (95% confidence interval [CI]) and mean (standard error [SE]) changes from baseline.

Results: A significant overall main effect of treatment was observed for quality of working memory (OC: 0.07 SI units [0.016, 0.124], P=0.012; LOCF: 0.07 [0.029, 0.110], P=0.0008). Significant differences from placebo were observed as early as day 28 (OC: 0.11 SI units [0.033, 0.193], P=0.0058). Overall improvement from baseline (across all time points) was significant for desvenlafaxine (OC: 0.094 [0.02] SI units, P<0.0001; LOCF: 0.089 [0.01], P<0.0001), but not placebo (OC: 0.024 [0.02] SI units, P=0.274; LOCF: 0.019 [0.02], P=0.253), on the composite measure of quality of working memory. Overall improvement from baseline in speed of working memory was significant for desvenlafaxine (OC: -179.6 [45.4] msec, P=0.0002; LOCF: -175.16 [27.7], P<0.0001) and placebo (LOCF analysis only: -86.2 [37.3], P=0.022).


H3 Receptor Antagonist: A Potential Cognition Enhancer

Stacey Boyer, Donna Palumbo

Introduction: Impaired cognition and functional deterioration is observed in a number of neuropsychiatric diseases. H3 receptor antagonists and reverse agonists have the potential to function as cognitive enhancing agents given that preclinical models indicate that antagonists of presynaptic H3 receptors increase release of acetylcholine, norepinephrine, and glutamate, and increase attention, arousal and memory.

Methods: Effects of H3 antagonist [PF-03654746], on extracellular histamine levels in prefrontal cortex of freely moving rats were measured by microdialysis. An indirect response model with stimulation of histamine release was used to characterize the PK/PD relationship. In addition, a hypothetical stress compartment was included via a stimulatory function to account for the increased histamine response following vehicle injection. A microdialysis study was also carried out to examine dose–response of PF-03654746 on extracellular acetylcholine, dopamine and norepinephrine, release in prefrontal cortex of freely moving rats. Hippocampal theta oscillation was measured in electrophysiological studies utilizing field potential
(EEG) recordings from chloral hydrate anaesthetized rats. Novel object recognition with the compound dosed (SC 60 min prior to testing on each of 3 consecutive days) was also tested.

**Results:** SC administration of the H3 receptor elevated extracellular levels of histamine in prefrontal cortex of conscious rats. Additional microdialysis studies demonstrated the elevation of acetylcholine in prefrontal cortex in rats. The compound enhanced EEG theta oscillation consistent with increase in attention, arousal, and memory encoding. In addition, improved performance was noted in a novel recognition task, a model that assesses effects on non-spatial episodic working memory.

**Conclusions:** These data indicate that H3 antagonist (PF-03654746) increases extracellular histamine and acetylcholine in the prefrontal cortex of freely behaving rats. The compound also increases EEG theta oscillation and improves performance in novel object recognition. These data indicate that the H3 antagonist PF-03654746 may be useful in treatment of cognitive dysfunction.

---

**Randomization Strategies for Clustered Clinical Syndromes: Effects on Statistical Power**

Lebowitz BD¹, Wang Y², Duan N², Reynolds CF III³, Simon NM⁴, Zisook S¹, Shear MK²

¹University of California, San Diego, ²Columbia University, ³University of Pittsburgh, ⁴Massachusetts General Hospital

**Introduction:** Complicated Grief (CG) is representative of a number of clinical syndromes that occur in clusters. We needed to take clustering into account in the development of randomization strategies for a 4-site NIMH- and AFSP-supported RCT of CG monotherapies and combined treatments. Sampling of clusters is a commonly used strategy in survey sampling and randomization of clusters is often used in field trials of organization-level interventions in schools, hospitals, and communities. In these trials, all individuals in a cluster are offered the same intervention.

**Methods:** The trial is a 4-cell placebo controlled factorial design comparing the efficacy of pharmacotherapy (citalopram) and targeted psychotherapy (complicated grief therapy, CGT) alone and in combination. We chose a “split-plot” approach that will randomize therapy (CGT vs. no CGT) at the group level (couple, family, etc.), and randomize medication (medication vs. placebo) at the individual level. There is an established history of the split-plot design in agriculture and engineering.

**Results:** The split-plot design, compared to individual level randomization, maintains power for the pharmacotherapy aim of the trial but reduces power for the psychotherapy aim as a joint function of the degree of intercorrelation within the cluster and the proportion of clusters in the overall sample. With a sample size of 220 where within cluster correlation is .50, power is reduced from .84 to .81 as the proportion of clustered participants increases from 0 to .50. Increasing within cluster correlation to .90 reduces power from .84 to .74 as the proportion of clustered participants increases.

**Conclusion:** In the context of clustered eligibility for an RCT we considered, but rejected, block randomization of the pharmacotherapy condition as well. Block randomization would enhance statistical power and precision by blocking the randomization at the group level. This is equivalent to the common practice in animal experiments to block by litter. However, this option would introduce clinical and ethical complications. We chose to randomize individuals in the cluster to citalopram vs. placebo without blocking, i.e., the same way "singleton" participants are randomized. For the CGT condition, we may lose power and precision by randomization at the group level, compared to randomization at the individual level, due to intra-group correlation. The magnitude of the design effect or variance inflation is determined by the strength of the intra-group correlation. In an extreme scenario, if a husband and wife are perfectly correlated, there is no new information provided by the second participant.

**Funding:** NIMH Grants: 1R01 MH060783; 1R01 MH085288; 1R01 MH085297; 1R01 MH085308 and a grant from the American Foundation for Suicide Prevention
The Importance of Quality in Post-baseline Assessments in CNS Clinical Trials

Williams, JBW1,2 Kobak, KA3 Detke, M1,4,5

MedAvante, Inc.1, College of Physicians and Surgeons, Columbia University2, Center for Psychological Consultation3, Indiana University School of Medicine4, McLean Hospital, Harvard Medical School5

Introduction: Pressure to inflate baseline scores, “functional unblinding,” rater drift, and expectancy bias all contribute to trial failure. Inappropriate subjects are enrolled in a study when enrollment pressures cause inflated baseline severity scores. An increasing number of studies now include some method, such as independent blinded raters, for ensuring that the right subjects are entered into the trial. Post-baseline factors, however, can also affect outcomes. “Functional unblinding” and rater drift cause measurement noise that can reduce signal detection. Continuously-calibrated independent raters can avoid unblinding and rater drift. Expectancy bias can obscure a drug-placebo difference, but independent raters can limit subject-staff interactions and avoid expectancy bias.

Methods: Studies with ratings by both site raters and expert blinded remote raters can be evaluated to see how critical are continued blinding and continuous calibration post-baseline. In a trial of acute schizophrenia, independent remote blinded raters conducted the PANSS and site raters used the BPRS on the same 313 subjects. A study of 259 subjects with Parkinson’s psychosis included blinded independent central ratings in the US, and traditional site ratings in the OUS sites. Both sets of raters used subscales of the SAPS as the primary outcome measure. A negative trial of GAD had blinded independent raters evaluate 122 subjects admitted to the study by site raters’ SIGH-A baseline evaluations.

Results: In the schizophrenia trial, the independent blinded raters differentiated subjects on placebo, the active comparator, and one of two test arms throughout the study. Site raters did not separate either of the two test arms. In the Parkinson’s psychosis study, the test drug was separated from placebo by the central raters, but OUS site raters did not detect a signal. In the GAD trial remote blinded raters had a lower placebo response than site raters, independent of subject selection.

Conclusions: No one study answers the question of how critical is the quality of post-baseline assessments. However, data from several studies support the importance of the accuracy of assessments after subject selection. Continued vigilance and precision of ratings beyond baseline can increase the sensitivity of findings in a clinical trial and decrease placebo response rates. Blinding of raters to study protocol and visit number and independence from subjects’ sites minimizes expectancy bias by reducing the amount of time rating staff spends with subjects. Rater drift, even with experienced raters can be diminished only through continuous calibration of the cohort of raters.

Reconsidering the Criterion for Excluding Comorbid Substance Use Disorders from Clinical Trials for the Treatment of Depression

Lori L. Davis

Tuscaloosa VA Medical Center, Tuscaloosa, AL Alabama School of Medicine, Department of Psychiatry and Behavioral Neurobiology

Introduction: Many patients with major depressive disorder (MDD) present with recent or concurrent substance use disorders (SUD). These comorbid subjects are typically excluded from depression treatment trials, leaving a gap in understanding the treatment outcomes. Clinical trials are notorious for mandating a lengthy period of remission of SUDs as part of the eligibility criteria. This criterion often leads to more screen failures, which slows enrollment and limits generalizability. Are these criteria based on clear evidence that they serve to enhance the results or safety of a clinical trial? Or, are they based on clinical trial lore or bias?

Methods: Results from Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study and Depression Trials Network (DTN) studies are examined in an attempt to shed light on these questions. The outcomes between adult outpatients with MDD and those with MDD plus concurrent SUD who have participated in a clinical treatment trial are presented as a data-driven approach to reconsidering study entrance criteria in regards to concurrent SUD.

Results: In STAR*D, despite clear sociodemographic and clinical differences, there were no significant differences between groups in the time to achieve or rates of response to citalopram; however, those who endorsed both alcohol and drug use had significantly reduced rates of remission and increased times to reach
remission compared to the MDD group without SUD. Although numerically low, subjects with MDD and SUD had higher risk of psychiatric serious adverse events and hospitalization. Another STAR*D treatment outcomes analysis found that MDD subjects who had both concurrent anxiety disorder and SUD had the lowest rates of response and remission to citalopram treatment compared to the MDD-only group. In DTN COMED trial, despite baseline differences in lifetime suicidality and increased number of concurrent Axis I disorders, the presence of concurrent SUD in persons with chronic or recurrent MDD does not appear to negatively impact response or remission when treated with either single agent or combination antidepressants.

Conclusions: Despite baseline clinical and demographic differences, a concurrent drug or alcohol use disorder does not necessary impact the response to antidepressant treatment in clinical trials. However, the increased burden of having both an alcohol and drug use disorder or having a concurrent anxiety disorder and SUD may have impact on response and remission. Conceptualizing inclusion criteria that allow entrance of some, but not all, subjects with SUD into depression treatment trials is possible with these data.

Impact of Face to Face Assessment Time on Primary Outcome Measures in Randomized, Placebo-controlled Trials of Ziprasidone

Vanderburg, D1; Schwartz, J2; Wolf, P3; Fleischhacker, W4; Karayal, O1; Stewart, M2


Background: A trend for increased placebo response in CNS clinical trials has been noted over the past decade. One hypothesis generated by studying depression trials is that placebo response increased as a function of the number of assessments performed.1,2 One explanation is that increased face-to-face interaction time between the subject and site staff contributes to increased placebo response. Purpose of study: To investigate whether face-to-face assessment time had an impact on placebo response in clinical trials of ziprasidone.

Methods: Data from 11 placebo-controlled schizophrenia and bipolar trials were summarized at the protocol level. Mean changes from baseline (calculated as effect size [Cohen’s d]) on the primary efficacy endpoint were extracted. Face-to-face interaction time (FT) was estimated as follows: for each assessment in each study, the number of such assessments across all visits was determined; time per assessment was estimated using information from assessment manuals and published literature, or, if unavailable, by averaging two independent estimates from psychiatric trialists; the product of these factors was summed across all assessments. FT included only assessments requiring interpersonal interaction (e.g. BPRS, PANSS); other assessments (e.g. blood draws, cardiac testing) were excluded. Other trial design characteristics summarized included study start date (YEAR), study duration (SD), per-subject study duration (PSD), proportion of subjects in placebo group (PSP), and study indication (IND). Planned analyses included correlations and linear modeling using FT and other trial characteristics as predictor variables and the effect size of primary outcome measures as dependent variables.

Results: The correlation between FT and effect size was nonsignificant (.02). However, confounding of factors (e.g. the predictor variable IND accounted for up to 93% of the variability in both FT and effect size) and multicollinearity of predictor variables (e.g. the correlation coefficient for YEAR and IND was 0.96) precluded obtaining a valid result from regression modeling.

Discussion: The primary objective of this project was to examine the impact of FT on effect size of primary outcome measures in clinical trials of ziprasidone. Multicollinearity and confounding prevented valid testing of hypotheses using linear modeling. However, our strategy for creating the variable FT is of potential value in future research, and could be improved by using subject-level visit data. Strong correlations between trial-level predictor variables suggests that pooling data from multiple companies could help generate a data set of sufficient FT variability across indications to permit testing of the original hypothesis.

Can Objective Measures of Motion and Attention Distinguish Across Diagnoses?

Calvin Sumner, MD1, Birgit Amann, MD2, and Virginia Haynes, PhD3

1BioBehavioral Diagnostics Company, Westford, MA and Plymouth Meeting, PA, 2Behavioral Medical Center, Troy, MI, 3 i3 Global, Basking Ridge, NC

Introduction: The DSM-IV criteria for mood and anxiety disorders and attention-deficit/hyperactivity disorder (ADHD) share common symptoms that can complicate diagnosis, especially when these conditions are comorbid. Objective To evaluate the classification utility of the Quotient™ ADHD System in the assessment of children and adolescents referred to a psychiatric clinical practice for assessment and diagnosis.

Methods: This observational single site study approached patients aged 6 through 14 years old that were referred to the psychiatric clinical practice for evaluation to participate. Rating and assessment scales were administered consistent with the standard practice of the office. The Quotient™ ADHD system test was administered by designated clinicians who were trained to properly administer the test. It was recommended (but not required) that subjects withhold daily medication on the study visit day until after rating scales, assessments and Quotient™ ADHD System testing is completed. The investigator remained blinded to the Quotient™ ADHD System results until after a probable diagnosis was made, using the standard rating/assessment scales. Only after the probable diagnosis was documented did the investigator have access to the Quotient™ADHD System Patient Report (test results). ANOVA was used to compare mean scaled Attention and Motion scores and test for differences across diagnostic groups. Individual Quotient variables that contributed to a significant scaled score were tested for a linear trend across diagnostic groups using ANOVA. Tests of hypotheses were adjusted for multiplicity using Holm’s step-down procedure.

Results: One hundred subjects aged 6 to 14 consented to participate in this single visit study. Eighty-four subjects received a clinical diagnosis; there were 23 subjects with simple ADHD, 14 with a mood, anxiety or learning disorder, and 47 with ADHD and a comorbid mood or anxiety disorder. There was no significant difference in Quotient Attention scores across diagnostic groups; the mean Quotient Motion scores differed significantly across diagnostic groups (simple ADHD=64.0, comorbid ADHD=52.9, and mood/anx/learn=38.5, p=0.017). Diagnostic differences in the Quotient Motion scores were driven by the number of head movements and area of head motion. For both of these items, there was a linear trend with ADHD subjects demonstrating the greatest mean number of head movements (3739.65) or movement area (196.57 cm2) and subjects without ADHD demonstrating the lowest number of movements (1810.64) or movement area (77.21 cm2). The mean number of movements or movement area for subjects with comorbid ADHD was in between that of simple ADHD subjects and those with a mood, anxiety, or learning disorder but no ADHD diagnosis.

Conclusions: The motion scaled severity score from the Quotient system applied in a naturalistic clinic setting distinguished between subjects with ADHD and those without, even in the presence of comorbid mood and anxiety. The more sensitive of the motion measures were head movements and head movement area which were different for the simple ADHD, comorbid ADHD, and mood/anx/learn groups. This preliminary study suggests that objective measures such as Quotient, which more sensitively measure impairments or deficits in developmental motion control, may be useful in differentiating complex clinical cases. Larger studies with larger samples of patients who have psychiatric conditions other than ADHD are needed to further elucidate these findings.

WITHDRAWN

Reliability in Cognitive Testing: Does Rater Education Level Affect Data Quality?

Kolleen Hurley Fox1, Vicki G. Davis1, Richard S.E. Keefe2

1NeuroCog Trials, Inc., Durham, NC, USA 2Duke University Medical Center, Durham, NC, USA

Background: The Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) Project produced a battery of tests, the MATRICS Consensus Cognitive Battery (MCCB), designed to assess cognitive treatment effects in clinical trials of patients with schizophrenia. In validation studies, the MCCB demonstrated excellent reliability, minimal practice effects and large correlations with
measures of functional capacity. Study leads from sponsor companies utilizing the MCCB in their clinical trials often question the requisite education level of a rater to ensure quality data collection. As such, it seemed a worthy empirical question to examine existing MCCB battery databases to determine intra-rater reliability and to examine it by education level.

**Methods:** Data from two large studies of patients with schizophrenia were combined for the purpose of determining intra-rater reliability using MCCB data collected at screening and baseline. A total of 473 patients were assessed by 63 raters, 46 with a bachelor’s degree and 17 with either a Master’s or Doctorate degree. The time interval between these two visits was 4 weeks, with both visits occurring prior to the initiation of the treatment phase. Only patients who were assessed by the same rater at both visits are included in the analysis. The Intra-class correlation coefficient (ICC) between screening and baseline was calculated for the MCCB subtests, domains, and overall composite score by rater education level. In addition, intra-rater reliability was compared for the first and last patient assessed by each rater who completed at least 5 tests.

**Results:** ICCs for MCCB subtests, domains and overall composite did not appear to differ substantially by level of rater education. Raters with a Bachelor’s degree had a higher ICC for seven of the ten subtests, although absolute differences between education levels were quite small, ranging from 0.02 to 0.13. The ICC between screening and baseline for the overall composite score was 0.88 for raters with a Bachelor’s degree and 0.86 for those with a Master’s or Doctorate degree.

**Discussion:** Although study lead personnel from sponsor companies are often concerned with the necessary level of education by a rater to ensure quality data, our data indicate there is little difference between educational groups on intra-rater reliability. This should provide reassurance to sponsor companies that idiosyncratic rater variations are minimized regardless of educational background.

**Methodological Challenges of Demonstrating Cognitive Improvements with Broad Spectrum Agents: A Novel Approach**

Yona Geffen¹, Michael Davidson², Michael R. Hufford³, R. Michael Gendreau³, Srinivas G. Rao³, Rong Zablocki³, and Jay D. Kranzler³

¹BioLineRx, Ltd, Jerusalem, Israel. ²Tel Aviv University, Tel-Aviv, Israel. ³Cypress Bioscience, Inc., San Diego, USA

**Background:** There are numerous methodological challenges related to the development of treatments to address the cognitive impairments associated with schizophrenia. Current consensus guidance from FDA and the NIH recommends use of the MATRICS Consensus Cognitive Battery (MCCB) to assess potential cognitive improvement in stable schizophrenic patients. CYP-1020 is a potential broad-spectrum agent that combines dopamine antagonism with GABAergic activity. Because CYP-1020 appears to be acting in a broad spectrum manner, providing beneficial cognitive and antipsychotic effects soon after treatment has begun, we explored evaluating its therapeutic utility starting with acutely ill patients with schizophrenia.

**Methods:** Evidence for broad-spectrum activity is derived from the EAGLE study, which was conducted at approximately 40 sites in U.S., Romania and India. In this 6-week double blind study, 363 patients were randomized equally to treatment with 10 mg/day CYP-1020, 20-30mg/day of CYP-1020, risperidone (2-8mg/day) or placebo. The study was designed to demonstrate significant superiority of CYP-1020 to placebo on the total score of the PANSS. The effect on cognition as measured by the BACS was an exploratory end point.

**Results:** The total PANSS scores change (LOCF) indicated that treatment with CYP-1020 high dose (LS mean -23.6; 95% CI-28.4;18.8), was statistically superior to placebo (p=0.002); (LS mean -14.4; 95% CI: -19.1;9.7). Risperidone treatment also was associated with significant improvement (LS mean-26.2; 95% CI -31.0; -21.3) compared to placebo. Treatment with CYP-1020 yielded statistically superior cognition results as measured by the BACS composite score based on LOCF. The age- and gender-corrected change from baseline on the BACS composite score for the CYP-1020 high dose group was LSM=12.8, SE=1.58, as compared to a placebo group change of LSM=8.4, SE=1.63 and a change with risperidone treatment of LSM=8.2, SE=1.46 (p < 0.03 versus placebo and risperidone).

**Discussion:** In this acute study, CYP-1020 demonstrated comparable antipsychotic activity to risperidone versus placebo, while simultaneously demonstrating superior pro-cognitive results compared to both placebo and risperidone. Moving forward, a hybrid clinical trial design will assess antipsychotic efficacy and
cognitive functioning of CYP-1020, including both a 6-week acute treatment phase with CYP-1020, placebo or risperidone, followed by a 6-month chronic treatment phase of CYP-1020 or risperidone using both the PANSS and MCB. This trial should be able to more fully characterize CYP-1020’s long-term antipsychotic efficacy and impact on cognition.

Should We Ignore Family History as Valid MDD Indicators in Non-Western Countries?

Olga Rusak¹, Suzanne Edman¹, Bryce Kasuba², Daniel DeBonis¹, Gary S. Sachs¹ ³

¹Concordant Rater Systems, Boston, MA, ²Concordant Rater Systems, Havertown, PA, ³Massachusetts General Hospital, Boston, MA

**Background:** Utilization of study sites outside the US increases the need to evaluate entry criteria in the context of cultural variation. Family history is considered a validator of psychiatric diagnosis, but qualified study subjects in many non-Western cultures might have low rates of self-reported family history of mood disorder (sFHMD) due to a lack of social acceptance and low recognition of mood disorder diagnosis in their local health systems. We hypothesized that sFHMD would not impact study outcome in Russia, even when assessments were computer administered.

We examined the impact of sFHMD in a randomized clinical trial conducted exclusively in Russia. This study compared the efficacy of low (LD) and high doses (HD) of a putative antidepressant vs. Placebo (Pbo) in subjects with Major Depressive Disorder (MDD) across 15 sites in Russia. The study was stopped early due to futility.

**Methods:** The primary outcome measure in the double-blind parent study was the change from Baseline to Day 14 on the HAMD-17 scored by a trained site-based rater (ΔBL-D14 HAMD_SBR). A computer-administered diagnostic interview assessed sFHMD at screen, and a computer severity rating (HAMD_COMP) was obtained after each protocol-specified HAMD_SBR rating. Subjects were considered evaluable if they had knowledge of FHMD (for blood relatives) and completed the Baseline and Day 14 evaluations. The analysis plan called for ANOVA, comparing mean Placebo response on the primary outcome for sFHMD(+) vs. sFHMD(-), based on whether any blood relative was reported to have a mood disorder diagnosis at screen.

**Results:** Overall, no significant difference was found between active LD or HD compared to Pbo on the primary outcome.

**Table 1.** Comparison of HD, LD and Pbo responses on the primary outcome

<table>
<thead>
<tr>
<th></th>
<th>BL-D14 ± s. d.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HAMD_COMP</td>
</tr>
<tr>
<td>sFHMD(+)</td>
<td></td>
</tr>
<tr>
<td>n=10</td>
<td>-13.4±8</td>
</tr>
<tr>
<td>sFHMD(-)</td>
<td></td>
</tr>
<tr>
<td>n=20</td>
<td>-9.1±5.8</td>
</tr>
<tr>
<td>sFHMD(+)</td>
<td></td>
</tr>
<tr>
<td>n=10</td>
<td>-11.9±6.5</td>
</tr>
<tr>
<td>sFHMD(-)</td>
<td></td>
</tr>
<tr>
<td>n=20</td>
<td>-10.3±5.8</td>
</tr>
</tbody>
</table>

Of 138 subjects screened, 87 were evaluable including 32 (36.8%) sFHMD (+) and 55 (63.2%) sFHMD(-).

**Table 2.** Comparison of Pbo response in sFHMD(+) vs. sFHMD(-) subjects

<table>
<thead>
<tr>
<th></th>
<th>sFHMD (+) mean± s.d.</th>
<th>sFHMD(-) mean± s.d.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL-D14</td>
<td>-6.6±6.5</td>
<td>-11.9±6.5</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Conclusions: Contrary to our expectations, these results suggest that sFHMD is an important variable related to Placebo response and may be a useful part of diagnostic validation even in cultures where mood disorders may have lower rates of recognition and social acceptance. We speculate that the greater placebo response was for HAMD$_{SBR}$ compared to HAMD$_{COMP}$ is a reflection of interpersonal and cultural factors. Larger studies are needed determine if the rate of sFHMD(+) can be used as measure of sample quality for study sites in international multicenter trials.

References: [1]


16 Experienced and Inexperienced CGI Raters Underestimate the Impact of Negative Schizophrenia Symptoms on Functioning

Theresa A. Bromley PhD Marian A. Ormont MD

ePharmaSolutions

Introduction: The Clinical Global Impression of Severity and Change (CGI-S/I) are included as primary outcome measures in the majority of CNS trials with a variety of indications. CGI raters consider frequency, intensity and prevalence of symptoms and the impact of symptoms on functioning. This study targeted the negative symptoms of schizophrenia and subjects with acute positive symptoms were excluded. Thus, patients enrolled in this study demonstrated significant negative symptomatology but did not exhibit severe hallucinations, delusions, suspiciousness or hostility. Research indicates that negative symptoms are frequently not assessed as accurately as positive symptoms and may not be considered so predominantly in CGI ratings.

Methods: 103 US raters administered the CGI-S/I and the PANSS in a clinical trial. 100% of CGI raters were Principal or Sub-Investigators who held doctoral degrees level (MD., Ph.D., DO., PsyD). PANSS raters held degrees ranging from BA (19%) to a MA, MSW, RN (31%), to MD or PhD (50%). During the two years prior to the study, CGI raters reported <5 to >50 administrations and PANSS raters reported 5 to > 50 administrations. The CGI and PANSS were administered at 11 study visits. 1,981 rater submissions were reviewed by doctoral level clinicians. On 143 occasions CGI-S scores of 3 (Minimally ill) were given at Screening and Baseline ostensibly too low considering the functional level of the patients. All submissions were analyzed to compare the CGI-S with the PANSS Negative items score, Positive item score and PANSS Total Score; particularly for the cases with CGI-S scores of 3 at Screening and Baseline. The study targeted negative symptoms of Schizophrenia and therefore, CGI raters were predicted to be less sensitive to the functional detriments of negative symptoms of schizophrenia. None of the subjects in the study were employed full-time, married or living independently. CGI-S scores were predicted to correlate higher with PANSS Positive Item Subscale scores than with PANSS Negative Item Subscale Scores due to the lack of sensitivity to negative items.

Results: Preliminary data indicates that CGI-S scores at Screening and Baseline were more highly correlated with PANSS Positive items than PANSS Negative items; especially for CGI-S scores of 3 at Screening and Baseline.

Conclusions: Even highly experienced raters are not always sensitized to the functional detriments of negative symptoms of schizophrenia. CGI-S scores are more aligned with positive schizophrenia items than negative items. More training in relations to CGI-S scores and negative symptoms is warranted.

17 The Impact of a Data-Monitoring Program Implemented at Study Midpoint on PANSS Score Consistency
Background: It is well understood that there is a degradation of training impact over time. This drift can cause problems for both reliability and validity in clinical trials. In this study, we sought to determine what the impact of a coordinated program of data-monitoring would be in terms of score consistency on the PANSS (Positive and Negative Syndrome Scale) when implemented 12 months from the initial training period. Many studies (e.g., Müller & Szegedi, 2002) indicate that reliability can suffer if training and calibration is not conducted regularly. However this can be resource intensive. Data-monitoring provides targeted feedback for the raters that require it on an ongoing basis, yet the impact of such systems has not been studied in a systematic manner.

Methods: Raters were trained at the investigator meeting utilizing a standardized method consisting of didactic and applied techniques. There was no further training until a data-monitoring program was introduced at the mid-point in the study due to concerns about data integrity. A retrospective analysis of existing data was conducted at mid-point using computer-based algorithms designed to detect logical inconsistencies within the PANSS instrument. The results were used to provide targeted remediation. This system was then left in place for the remainder of the study.

Results: There was a significant reduction in inconsistent scores before and after data-monitoring was implemented. A chi-square (Pearson) test was conducted which indicated the difference in proportions is significant, $\chi^2(1, n = 2415) = 30.977, p < 0.001$.

Conclusions: The implementation of a data-monitoring system 12 months from the initial training uncovered a range of rater behaviors inconsistent with generation of reliable scores. In the analysis of the data obtained both before and after the data-monitoring was implemented we observed a relationship between those raters that generated very high numbers of inconsistencies and the patient’s subsequent failure to respond to treatment. Müller MJ, Szegedi A: Effects of Interrater Reliability of Psychopathologic Assessment on Power and Sample Size Calculations in Clinical Trials. J Clin Psychopharmacol; 2002; 22: 318-325

The Relationship Between Interview Quality and Scoring Accuracy: Evaluation of the Rater Applied Performance Scale

Danielle Popp¹, Michael Detke¹,2,3, Janet BW Williams¹,4

¹MedAvante, Inc., 2Indiana University School of Medicine, 3McLean Hospital & Harvard Medical School, 4College of Physicians and Surgeons, Columbia University

Introduction: Variable interview quality may be an important contributor to the high number of failed CNS clinical trials. Recent advances in the measurement of interviewer skill allow for examination of the effects of interview quality on signal detection1. Research suggests subjects with higher quality interviews are more likely to distinguish an effective drug from placebo2. The current study examines the relationships between interview quality and score accuracy to identify which areas of interview quality are most closely related to scoring accuracy.

Methods: Data were pooled from two ongoing clinical trials of Major Depressive Disorder (MDD). Site raters audiotaped their MADRS or HAM-D assessments. A group of calibrated, blinded independent clinician reviewers scored a subset of taped site rater assessments selected according to an a priori algorithm. Reviewers also rated interviews on 5 domains of the Rater Applied Performance Scale (RAPS): Adherence, Follow Up, Clarification, Neutrality, and Rapport. Independent reviewers’ scoring was completed and locked prior to revealing site rater scores.

Results: 1017 interviews across 2 MDD studies were analyzed. All total scores were standardized to z-scores. Scoring accuracy was computed as the difference between site rater and reviewer scores. 49.4% of interviews were categorized as Good/Excellent quality (average of 5 RAPS domains ≥3), 49.2% were Fair (2.0 to <3.0), and 1.5% were Unsatisfactory (<2). Intraclass correlations (ICCs) between site rater and independent reviewer scores were computed. Higher quality interviews had higher ICCs (Unsatisfactory=.39, Fair=.88, Good/Excellent=.96). For visits during which cut-off scores were applied (n=599), scoring accuracy was significantly correlated with all 5 RAPS domains and RAPS average score (all $r$’s>.097, all $p$’s<.018). A multiple regression was performed demonstrating that Adherence ($\beta$=.173, p<.001), Follow Up ($\beta$=.162, p<.001), and Clarification ($\beta$=.115, p<.05) significantly predicted scoring accuracy. Higher scores on each of
these domains predicted better scoring accuracy. At follow-up visits (n=417), scoring accuracy was significantly correlated with the overall RAPS average and 4 of the RAPS domains (all r’s>.098, all p’s<.046 excluding Adherence). Multiple regression revealed none of the RAPS domains significantly predicted scoring accuracy.

**Discussion:** Interview quality, as measured by the RAPS, is related to scoring accuracy across two MDD studies. Unsatisfactory quality interviews have lower scoring accuracy which may contribute to decreased signal detection and the high rate of CNS clinical trial failures. The RAPS is a useful tool for assessing interview quality and identifying which domains are more important for scoring accuracy.


**A Post-hoc Examination of Patient Characteristics: Implications for Enhanced Medical Surveillance**

Friedmann B, Riordan H, Kornsey EB, Williams K, Cox K, Murphy M

*Worldwide Clinical Trials, King of Prussia, PA*

**Introduction:** Recent studies in psychiatry have been marked by failed as well as negative trials. One contributory factor may be the enrollment of subjects that technically meet eligibility requirements including threshold scores for disease severity, who nevertheless are suboptimal candidates for study participation when viewed through a prism of historical and clinical information not captured by a study database.

**Methods:** The appropriateness of acutely exacerbated schizophrenic patients randomized to a completed study assessing the safety and efficacy of a novel antipsychotic was retrospectively, and independently, evaluated by three clinicians, blinded to country, site, patient identity, and treatment. During this process, medical history available through the study data base was amplified by examination of source documents addressing the subject’s background according to predefined criteria (e.g., living situation, employment history), clinical history (treatment, diagnosis, trends, medications), as well as information on trial procedures (e.g., PANSS interview process). Approximately 13% (26) of all randomized subjects were considered, randomly selected from participating centers. Two psychiatrists and one psychologist each with over ten years of psychiatric trial experience independently rendered an opinion regarding the subject’s appropriateness for inclusion, and all opinions were aggregated for final review.

**Results:** The three raters agreed on 88% (23/26) of subjects with an overall rater agreement of 92.3% for the 26 subjects for study appropriateness. A kappa value for multiple raters was considered to have “Very Good Agreement” according to Altman’s conventions (Fleiss kappa = .7912, SE = .11, 95% CI .56-1.0) reflecting high concordance. Three subjects were deemed to be inappropriately randomized by all three evaluators although all technically met inclusion/ exclusion criteria.

**Conclusions:** This exercise suggested that an investigator’s judgment regarding subject suitability was consistent with third party blinded reviewers presented with an enriched database. Nevertheless, inappropriate subjects can be identified in spite of nominal compliance with protocol eligibility requirements, and independently of conventional efforts to assure congruency in diagnostic and disease severity assessments. A more finely detailed suite of inclusion/ exclusion criteria coupled to “real-time” monitoring of subject characteristics by blinded clinicians would provide an index of site sophistication that could identify problematic sites where enhanced medical monitoring activity is warranted.

**Electronic Administration of the Columbia-Suicide Severity Rating Scale (eC-SSRS): Results from 14,937 Administrations**

Mundt, JC¹, Greist, JH², Federico, M³

¹Cashel Technology, Madison, WI; ²Healthcare Technology Systems, Madison, WI; ³ERT, Philadelphia, PA;
Introduction: Concern about possible treatment-emergent suicidality has prompted the Division of Psychiatry Products for the FDA to mandate prospective monitoring of suicidality in clinical trials. The Columbia-Suicide Severity Rating Scale (C-SSRS) is an accepted instrument for meeting this requirement. Procedural variance in the way clinical assessments are performed by human raters in randomized clinical trials (RCTs) has been a concern for many years. A fully-structured C-SSRS script, including standardized questions, follow-up prompts, error-handling routines, and scoring conventions was developed for eC-SSRS implementation using interactive voice response (IVR) technology. System feasibility and clinical validity has been demonstrated. The e-C-SSRS has since been incorporated into multiple RCTs by several study sponsors.

Methods: A total of 14,937 administrations of the eC-SSRS were completed between September 2009 and August 2010 in ten RCTs sponsored by four different pharmaceutical companies and the Epilepsy Study Consortium. Trials investigating treatments for major depression, PTSD, insomnia, and epilepsy were included in this analysis. eC-SSRS reports of active ideation with intention to act (with or without a fully-developed plan), reported behaviors preparatory to a suicide attempt, and/or any reported suicide attempts were classified as “positive” case findings.

Results: 3,263 patients completed baseline eC-SSRS assessments regarding lifetime suicidal ideation and/or behaviors; 11,674 follow-up assessments (“since last contact”) were also completed by 2,495 patients. The mean (+ SD) time between follow-up assessments was 12.9 ± 7.8 days. 871 (26.7%) baseline assessments and 170 (1.5%) follow-up assessments were classified as positive case findings. The mean time required to complete e-C-SSRS assessments was 3.9 ± 1.9 minutes. Positive case findings took significantly more time to complete than negative case findings during both baseline (7.8 ± 1.9 versus 3.7 ± 2.3 minutes p<0.001) and follow-up assessments (7.2 ± 2.0 versus 3.6 ± 1.3 minutes, p<0.001). 1,509 (46.2%) baseline assessments and 10,018 (85.8%) follow-up assessments reflected a complete absence of any suicidal ideation or behavior.

Conclusion: Preliminary data supported the feasibility of the eC-SSRS as a valid and effective means for prospectively monitoring suicidality in clinical trial research. Those data were limited with respect to the number of patients studied and the duration of the follow-up period. Accumulating data from ongoing RCTs, with more diverse patient populations monitored over longer intervals, continues to support the use of the eC-SSRS as a valid and reliable computer-automated assessment of suicidal ideation and behaviors.


Less is More: Analysis of Russian Recruitment

Suzanne Edman¹, Olga Rusał¹, Bryce Kasuba², Daniel DeBonis¹, Gary S. Sachs¹,³

¹Concordant Rater Systems, Boston, MA, ²Concordant Rater Systems, Havertown, PA, ³Massachusetts General Hospital, Boston, MA

Introduction: The goal of delivering faster, cheaper, and better clinical trials has increased interest in study sites outside the US, where there may be an abundant supply of qualified study subjects. We report post-hoc analyses that examine the impact of recruitment rate from a randomized clinical trial conducted exclusively in Russia. This study compared the efficacy of low (LD) and high doses (HD) of a putative antidepressant vs placebo (Pbo) in subjects with Major Depressive Disorder (MDD) across 15 sites in Russia. The study was stopped early due to futility.

Methods: The primary outcome measure in the double-blind parent study was change from baseline to day 14 HAMD-17, scored by a trained site-based rater (ΔBL-D14 HAMD_SBR). A computer-administered assessment (HAMD_COMP) was obtained from subjects after each HAMD_SBR rating. The analysis plan called for ANOVA, comparing mean drug-placebo separation on the primary outcome for high vs low-enrollment sites based on a median split of site enrollment and was repeated for ΔBL-D14HAMD_SBR and ΔBL-D14HAMD_COMP across treatment groups. If the ANOVA showed significant differences, pair-wise comparisons were made using Student’s t-test.

Results: Overall, no significant difference was found between active LD or HD compared to placebo on the primary outcome. The median enrollment rate was six subjects over eight months. Comparison groups were defined the sites as high-enrolling (>6 subjects) or low-enrolling (≤6 subjects).
### Table 1. Outcome for subjects in high- vs low-enrolling sites

<table>
<thead>
<tr>
<th></th>
<th>High-Enrolling Sites n= 70</th>
<th>Low-Enrolling Sites n= 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± s.d.</td>
<td>Mean ± s.d.</td>
</tr>
<tr>
<td>ΔBL-D14 HAMD&lt;sub&gt;SBR&lt;/sub&gt; p vs Pbo</td>
<td>-11.6 ±5.6 † ‡</td>
<td>-10.4 ±6.2 na</td>
</tr>
<tr>
<td>LD</td>
<td>-8.9 ±5.4 na</td>
<td>-13.0 ±4.6 .15</td>
</tr>
<tr>
<td>HD</td>
<td>-10.4 ±6.2 na</td>
<td>-13.6 ±6.9 .02</td>
</tr>
<tr>
<td>ΔBL-D14 HAMD&lt;sub&gt;COMP&lt;/sub&gt; p vs Pbo</td>
<td>-7.8 ±7.7 †</td>
<td>-12.4 ±5.0 .02</td>
</tr>
</tbody>
</table>

* ANOVA p < .05 † ANOVA p < .15 ‡ ANOVA p > .25

Separation of Active-Placebo ΔBL-D14 HAMD: None of the high-enrolling sites had a mean separation of Active-Placebo above the mean for low-enrolling sites, and four of the seven high-enrolling sites found greater improvement in P than LD or HD groups.

**Conclusions:** Results from this MDD study suggest that failure to detect significant separation between drug and placebo was caused by high placebo response at high-enrollment sites. Comparable results from ΔBL-D14 HAMD<sub>SBR</sub> and ΔBL-D14 HAMD<sub>COMP</sub> suggest a problem with sample characteristics at high-enrolling sites.

**References** [1] [2]

---

22 Ongoing Monitoring and Feedback Decreases Error Rates and Improves Internal Consistency of PANSS Ratings in an International Clinical Trial

David Daniel, M.D.<sup>1</sup>, Joan Busner, Ph.D.<sup>1</sup> and Cynthia McNamara, Ph.D.<sup>1</sup>

<sup>1</sup>United BioSource Corporation

**Background:** Diminution of placebo drug separation in multi-site international clinical trials in schizophrenia and other CNS disorders is a matter of urgent concern (Yang, Cusin and Fava, 2005; Kemp, Schooler and Kalali, 2008). We report the apparent effects of a program of ratings data monitoring, feedback and remediation on error rates and internal consistency of PANSS ratings in a double blind, placebo controlled international clinical trial.

**Method:** A cohort of US and international investigators were trained and certified in administration and scoring of the PANSS, CGI and other efficacy scales. At study onset, surveillance of clinician-rated outcomes data was enacted via a computer-driven system with daily clinical oversight. Predetermined PANSS, CGI and other data inconsistencies were flagged as proxies of potential rater error. Clinicians assessed each flag and, when indicated, contacted and remediated site raters. Internal consistency of the PANSS and the error rate for the PANSS and CGI ratings were tracked over one month measurement periods using an index score (number of flags corrected by number of possible flags given visits conducted and scales administered). Flag rates and internal consistency (Cronbach’s alpha) were compared at a one month measurement period near the onset of the trial versus a one month measurement period six months later as a means of assessing the effectiveness of the ratings data monitoring, feedback and remediation program.

**Results:** The flag rate index score for the initial month measurement period (.062) was statistically significantly higher than for the sixth month measurement period (.037) (Pearson chi-square =7.21, p < 0.008). In addition, Cronbach’s alpha reliability for the PANSS was statistically significantly higher for the
sixth month measurement period (r = .90) compared to the initial measurement period (r = .74) (z = 3.78, p =0.0001).

**Discussion:** In the context of an international clinical trial with close data monitoring, feedback and remediation, the error rate for the PANSS and CGI fell statistically significantly over a six month period, consistent with improvement in ratings technique. In addition, Cronbach’s alpha for the PANSS improved statistically significantly over time consistent with more appropriate use of the scale. The current findings with the PANSS and CGI are consistent with previous reports of reduction in error rates involving the MADRS, HAM-D and YMRS (Busner, Daniel and Bartko, Kalali and Spear, 2007)

### The Role of Effort in Neuropsychological Test Performance on the MATRICS Battery in Individuals with Schizophrenia

Gregory P. Strauss, Travis L. White, Jamie N. Seff, Tatyana M. Matveeva, Leeka Hubzin, Sharon M. August, Zuzana Kasanova, James M. Gold

*University of Maryland School of Medicine, Department of Psychiatry; Maryland Psychiatric Research Center*

**Introduction:** Clinicians and researchers have long wondered whether individuals with schizophrenia display poor neuropsychological test performance in part due to low effort. Poor effort might be expected to occur in only a minority of individuals with schizophrenia. In the current study we administered the Victoria Symptom Validity Test (VSVT) to determine whether a sub-set of patients exist who are characterized by low effort, and examined whether patients displaying low effort have greater severity of cognitive deficits and psychiatric symptoms.

**Methods:** Participants included 73 individuals with schizophrenia and 51 controls matched for age, gender, and ethnicity who completed the VSVT, MATRICS battery, Chapman Anhedonia Scales, and symptom interviews.

**Results:** Results indicated that 1/73 patients and 0 controls met the VSVT cut-off for malingering. However, when a statistically determined cut-off of “low effort” was applied 100% of controls and 81% of patients were characterized by normal effort, whereas 19% of patients displayed low effort. In comparison to normal effort patients and controls, low effort patients evidenced significantly greater impairment on the MATRICS battery, had greater physical anhedonia on the Chapman scale, and greater endorsement of Infrequency items on the Chapman scale.

**Conclusions:** These findings are consistent with the notion that the vast majority of individuals with schizophrenia put forth adequate effort on cognitive tests; however, a small percentage of patients display low effort, and these individuals tend to exhibit the poorest cognitive performance and higher levels of anhedonia. It is possible that these low effort patients are less intrinsically motivated to put forward higher levels of effort. Implications for neuropsychological testing in schizophrenia are discussed.

### Global Inter-rater Reliability, Scale Validity, and Local Perception: PANSS Ratings and Reactions from 4 Countries

Stacy Liechti¹, Ashleigh DeFries², Mark Opler³, Sean Lane⁴, Evgenia Ivanova⁵, Larry Yang⁶

¹ProPhase LLC, The PANSS Institute, ²ProPhase LLC, Johns Hopkins University, ³ProPhase LLC, The PANSS Institute, New York University, ⁴New York University, ⁵ProPhase LLC, The PANSS Institute, Moscow State University, ⁶The PANSS Institute, Columbia University

**Introduction:** There is a considerable body of literature suggesting that standardized rater training is critical to inter-rater reliability. With the expansion of clinical and research trials to locations worldwide, there have been challenges in adapting and translating instruments not only in terms of language but also culture. In this study we compare different cohorts of raters from different countries receiving training in the use of the PANSS. We attempted to determine if there was any consistent by-country impact on specific items, factors, or subscales. We queried raters about their perceptions of the instruments they are asked to use vis-à-vis their local patient populations.

**Methods:** The data set comes from standardized rater training events involving raters from four countries: India (n=83), Russia (n=59), the US (n=63), and Romania (n=76). Different groups of raters scored taped interviews of two schizophrenic patients using the Positive and Negative Syndrome Scale (PANSS). Scores were compared across countries and intra-class correlation coefficients (ICCs) and rater agreement with gold standard scores were evaluated. These results were viewed against global raters’ responses to questions about
how well PANSS items correlated to the presentation of symptoms among their patients.

**Results:** Raters from the US and Russia demonstrated a higher level of inter-rater consistency than India or Romania with ICCs of 0.883 and 0.835 respectively. For eight PANSS items, all four countries, raters demonstrated at least 80% agreement with the gold

**Conclusion:** The differences in rater performance across the four countries indicate that standardized rater training is broadly effective in cohorts selected to participate in clinical trials. However, further analysis indicated that there are some important differences in the way that different groups conceptualize items. This suggests a need to tailor training to those items that are less reliable in certain groups to ensure reliability and validity in the use of this instrument.

**Signal Detection in Adjunctive Therapy Trials for Partially Responsive Major Depressive Disorder**

Joan Busner, Ph.D. 1,2 Cynthia McNamara, Ph.D 1 Margot Oakley, B.S.N. 1 and Stuart Montgomery, M.D. 3

1 United BioSource Corporation, 2 Penn State College of Medicine, 3 Imperial College of Science, Technology and Medicine

**Introduction:** Achieving drug-placebo separation in MDD trials continues to be a daunting challenge despite the best efforts of the field. From the site perspective, adjunctive therapy trials pose additional complexities: there are typically two critical entry points and often a single-blind that must be maintained for months. In an attempt to improve data quality in an industry-sponsored randomized adjunctive antidepressant therapy Phase III trial that included an 8-week single-blind lead-in, we designed and enacted a multipronged training and surveillance program. We report preliminary data integrity results from the surveillance system of this still blinded ongoing trial.

**Method:** US investigators participated in diagnosis, efficacy scale, and interviewing skills training and certification. To minimize the cueing to subjects of an imminent change at mid-study randomization, investigators were also required to undergo training in techniques for maintaining the single-blind, including comportment of personnel at critical visits, and handling of difficult patient scenarios. At study onset, surveillance of clinician-rated outcomes data was enacted via a computer-driven system with daily clinical oversight. Predetermined data inconsistencies were flagged as proxies of potential rater error. Clinicians assessed each flag and when indicated contacted and remediated site raters. By design, the clinical contacts included scale conventions remediation as well as placebo response minimization coaching. We are tracking internal consistency of the primary efficacy measure (Montgomery Asberg Depression Rating Scale) using Cronbach’s alpha. We are tracking flag rates by month using an index score (number of flags corrected by number of possible flags given visits conducted and scales administered). We report an interim analysis comparing the initial month and nine month time points.

**Results:** 1) Cronbach’s alpha reliability was high for the first and ninth month ratings (r’s=.81 and .89, respectively). The improvement across time points was significant (Fisher r to z, p < .0002). 2) 1150 flags were generated during the initial month; 780 were assessed during month nine. The index rate decreased numerically (.048 to 0.039), but not significantly across the two time points.

**Conclusions:** An integrated training and surveillance program designed specifically for a complicated adjunctive therapy study design is feasible and is ongoing. Internal consistency of primary efficacy ratings can be measured during a trial and was shown to be high at outset and the month nine interim time point. At this interim analysis, internal consistency of the primary efficacy rating has improved beyond the high level seen at study onset, supporting the potential utility of the program. Data inconsistencies can be detected and addressed daily by clinicians. At this interim analysis, the flag index reduced numerically but not significantly. Indicators of study integrity can be assessed and ratings errors addressed during the course of an ongoing complicated trial. We hope to correlate surveillance findings with drug-placebo separation rate at study close.

**Signal Detection and Placebo Response in Schizophrenia: Parallels with Depression**

Mallinckrodt, CH; Zhang, L; Prucka, WR; and Millen, BA

Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285, USA

**Introduction:** Placebo response and the rate of failed clinical trials are increasing in schizophrenia,
resembling previous experience with antidepressant clinical trials. In depression, the percent of patients randomized to placebo was shown to be strongly associated with drug-placebo differences (signal detection).

Methods: We hypothesized that this factor would also be important in recent schizophrenia clinical trials. To test this hypothesis a database of acute schizophrenia placebo-controlled studies conducted between 1997 and 2008 was constructed. The database contained 27 studies, with 79 active treatment arms.

Results: As percentage of patients randomized to placebo increased, mean placebo improvement decreased (p=.047) and mean drug-placebo differences tended to increase (p=.166). The frequency of significant contrasts from studies with ≥ 25% randomized to placebo was 83.3%, compared with 58.3% in studies with < 25% randomized to placebo.

Conclusions: Caveats to these findings include limited data and confounding of potentially influential factors. These limitations prevent definitive conclusions. However, results are consistent with previous findings in depression where having a higher percent of patients randomized to placebo increased drug-placebo differences.

Signal Detection in Clinical Trials: A Post-Study Survey of Schizophrenia Trial Sites

A. Loebel\textsuperscript{1}, J. Cucchiaro\textsuperscript{1}, C. Siu\textsuperscript{2}, D. Daniel\textsuperscript{3}, A. Kalali\textsuperscript{4}

\textsuperscript{1}Sepracor, Inc., MA, USA; \textsuperscript{2}Data Power (DP), Inc., NJ, USA, \textsuperscript{3}United BioSource Corporation, MD, USA; \textsuperscript{4}Quintiles, Inc., NC, USA

Objective: Diminishing drug-placebo differences and increasing placebo responses have been observed in recent psychiatric clinical trials. The objective of this survey based study was to evaluate the relationship between investigative site characteristics and signal detection in schizophrenia clinical trials.

Methods: Principal investigators (PIs) at 42 US sites from two recently completed lurasidone, placebo-controlled schizophrenia trials received a questionnaire that covered questions regarding patient population and recruitment methods, research staff experience; rater consistency, and PI assessment of several factors potentially relevant to clinical trial outcome. Survey response reliability and validity were assessed by consistency checks. Site-specific drug-placebo effect sizes of PANSS total scores and their associations with site and PI practices were examined using multiple stepwise regression analysis weighted by the inverse variance of PANSS change score. Analyses were adjusted for significant confounding covariates, including CRO quality, duration of illness, or frequency of unstable residence (p<0.05).

Results: Of the 42 US study sites, 32 (76%) completed the survey. The PI responses showed a majority of the sites recruited patients outside their practices. A larger lurasidone-placebo effect size was observed at sites where subjects were more likely to be research-naïve (p<0.05) and had longer illness durations (p<0.05). Sites with less revenue derived from pharma-sponsored studies (p<0.01), fewer subjects recruited via advertisements (p<0.05), and which were academic/non-profit were associated with larger effect sizes. Other relevant factors at the site level included patient selection (diagnostic accuracy), ability to identify a reliable informant, the importance given by investigators to the need for control of placebo response and for investigator and rater experience, plus specific patient characteristics (history of response to non-standard (high dose) treatment, and history of alcohol or substance abuse). The PI rated “extremely important” contributors to trial success were sponsor and CRO quality, responsiveness, and involvement. Illness duration from survey responses was verified against the lurasidone trial database source data and showed a correlation trend (p<0.10), supporting the validity of survey responses.

Conclusion: Our findings suggest certain site characteristics can affect the likelihood of detecting treatment efficacy signals in schizophrenia trials. Further research is needed to confirm whether these findings are study-specific or generalizable to other clinical trials.

Classic Method, New Review: Standard Deviation of Difference and Standard Error of Measurement Difference in Reliable Change Index

Rong W. Zablocki\textsuperscript{1}, Vicki G. Davis\textsuperscript{2} and Richard S.E. Keefe\textsuperscript{2,3}

\textsuperscript{1}Cypress Bioscience Inc., San Diego, CA; \textsuperscript{2}NeuroCog Trials, Inc., Durham, NC; \textsuperscript{3}Duke University Medical Center, Durham, NC
**Purpose:** The Reliable Change Index (RCI) model has been widely utilized in neuropsychological research. Two formulas have been developed for RCI, one using the standard deviation of the difference (SD\textsubscript{diff}) and the other using the standard error of the measurement of the difference (SE\textsubscript{diff}). The purpose of the current review is to clarify the difference between the two formulas, conceptually illustrate the theory of RCI, and recommend its appropriate application.

**Interpretation:** RCI using the SD\textsubscript{diff} method is based on the distribution of score change from a control group and is comprised of both systematic and measurement error. RCI defined by the SE\textsubscript{diff} method is based on the distribution of true score change from a control group and involves only measurement error. Theoretically, if there is systematic error such as practice effect then SE\textsubscript{diff} will be less than or equal to SD\textsubscript{diff}. Therefore, using the larger SD\textsubscript{diff} as a criterion to define RCI will be more conservative than using SE\textsubscript{diff}, resulting in higher specificity with the SD\textsubscript{diff} method and higher sensitivity with the SE\textsubscript{diff} method.

**Assumptions:** The fundamental assumptions behind these methods are normally distributed data and sufficiently large and representative samples in both the hypothetical and control groups. In addition, the shape (dispersion) of the normal distribution for the comparison group is also important for the SD\textsubscript{diff} method; whereas the SE\textsubscript{diff} method requires the additional assumption that error components at the two time points are uncorrelated with equal variance.

**Recommendation:** For highly skewed data, transformation can be attempted to achieve a more normal distribution. If the control data are slightly skewed, remove the outliers before computing the RCI; if the hypothetical data are skewed, define the outliers as reliable change and assess the remaining data with RCI. When the distributions of score changes in the control group and the hypothetical group are inconsistent, the SE\textsubscript{diff} method is more appropriate. In practice, if SE\textsubscript{diff} in control group \geq SD\textsubscript{diff} in control group, the SD\textsubscript{diff} approach may be preferred.

**Key words:** Reliable Change Index, Standard Deviation of difference (SD\textsubscript{diff}), Standard Error of measurement difference (SE\textsubscript{diff})