ISCTM 2009 Autumn Conference

5-6 October 2009

Loews Coronado Bay – San Diego, CA

Poster Abstracts

Abstracts are listed in order of presentation.
Study Design: 1-12; Rater Scale/Performance: 13-23; Cognition: 24-29

1 **Differential Response to Treatment Across Countries in a Randomized Clinical Trial of Ziprasidone and Haloperidol in Patients With Bipolar Mania**

Eduard Vieta, MD, PhD;[1] Elizabeth Pappadopulos, PhD;[2] Francine Mandel PhD;[2] Ilise Lombardo PhD;[2] Antony Loebel MD[3]

[1]Clinical Institute of Neuroscience, Hospital Clinic, University of Barcelona, IDIBAPS, Barcelona, Spain; [2]Pfizer Inc. New York, NY, USA; [3]Dainippon Sumitomo Pharma America, Inc, Fort Lee, NJ USA.

Introduction: International trials are designed to reduce inter-country differences in operation, but regional prescribing practices and cultural differences may affect outcomes. Results from a study on the treatment of acute mania in the United States (US), Russia, and India revealed country variations in outcomes and adverse events. The placebo response was highest in the US and post hoc analyses examined demographic disparities, and differences in outcome and discontinuations in patients treated with ziprasidone or haloperidol.

Methods: Data from a 12-week, double-blind, 2-part study in 438 adults with acute bipolar mania were analyzed. Patients received flexibly dosed ziprasidone (80–160 mg/d), haloperidol (8–30 mg/d), or placebo for the first 3 weeks, followed by maintenance treatment with ziprasidone (40–160 mg/d) or haloperidol (8–30 mg/d) for 9 weeks. Baseline values, discontinuations, adverse events, and Mania Rating Scale (MRS) scores were assessed by country.

Results: Mean weight at baseline was significantly higher in the US (82.4 kg) and Russia (73.9 kg) than in India (57.1 kg). The mean dose of ziprasidone at week 3 was 128.4 mg in India, 121.8 mg in Russia, and 126.5 mg in the US and the mean dose of haloperidol was 20.7 mg/d in India, 15.2 mg/d in Russia and 15.3 mg/d in the US. Baseline MRS scores were higher in India (34.4) than in Russia (28.0) or the US (23.8). MRS change at week 3 was higher in India (ziprasidone –11.8; haloperidol –19.1) than in the US (ziprasidone –11.7; haloperidol –13.2) and Russia (ziprasidone –8.1; haloperidol –13.6). Fewer patients discontinued in Russia (ziprasidone 44.6%; haloperidol 21.8%) than in India (ziprasidone 67.8%; haloperidol 52.2%) or the US (ziprasidone 62.5%; haloperidol 85.1%). Mean time to discontinuation with ziprasidone was 63.8 ± 4.0 (Russia), 43.8 ± 3.5 (India), and 46.2 ± 4.6 days (US) and with haloperidol was 74.9 ± 4.0 (Russia), 60.5 ± 4.1 (India), and 32.9 ± 5.1 (US).

Conclusions: Differences among these countries in discontinuations, time to discontinuation, and drug treatments highlights the need for research into the cultural differences and health care systems available in these countries and their impact on clinical trials results.

Source of Funding: Pfizer Inc.

2 **SAFTEE: Specific Inquiry Yields More Relevant Side Effects**

Nina R. Schooler, Ph.D.,[1.] Jerome Levine, M.D.,[2.] Joanne B. Severe, M.S., [3.] Adam Haim, Ph.D., [3.] Leslie Citrome, M.D.[2.]

[1.]SUNY Downstate Medical Center, [2.] Nathan Kline Inistitite & NYU School of Medicine, [3.] National Institute of Mental Health

Background: Detection of adverse events is a high priority in drug development programs and clinical trials but a point of continuing debate is whether it is "better" to elicit events with general inquiry (GI) or specific inquiry (SI). The argument against specific questions is that they elicit events that are less relevant and less severe than events elicited by general questions. The SAFTEE (Systematic Assessment for Treatment Emergent Events) includes both and allows test of this proposition. Using data from a multi-center schizophrenia RCT, we compare "signal" (events associated with antipsychotic medications) to "noise" (events not associated) generated by SAFTEE GI and SI.

Methods: Two psychiatrists (JL and LC) independently judged whether each term in SAFTEE represents an AE "signal" or "noise." At the baseline RCT assessment, each participant was asked the GI questions, followed by the SI (in a review of systems format). Frequencies and severity of signal and noise reported

events were compared between GI and SI. [This is 15 words less than above paragraph]

Results: Of 298 event terms, there was agreement on 291. Seven were categorized based on consensus; 167 were judged signal and 131 noise. 337 RCT participants reported 2169 events; 1930 were signal and 239 were noise. A significantly higher percentage of signal events were reported by SI elicitation (95 vs. 70%). GI reported signal events were significantly more severe.

Discussion: Most events reported were signal. Specific questioning did not yield AEs unrelated to treatment refuting the proposition that detailed inquiry elicits irrelevant events that clouds understanding of medication AE profiles. Although AEs detected by GI were more severe, limiting the questioning would have missed many events. We recommend that future studies use the SI review of systems approach and dispense with the introductory GI.

3 ## Factorial Clinical Trials for Hybrid (Explanatory and Pragmatic) Studies: Design of "Optimizing Treatment for Complicated Grief"

Naihua Duan, Ph.D.,[1] Barry D. Lebowitz, Ph.D.,[2] Charles F. Reynolds III, M.D.,[3] M. Katherine Shear, M.D.,[1] Naomi M. Simon, M.D.,[4] Sidney Zisook, M.D.,[5]

[1]Columbia University, [2]UC San Diego, [3]University of Pittsburgh, [4]Massachusetts General Hospital, [5]UC San Diego

Introduction: Schwartz and Lellouch (1967) described the distinction between explanatory and pragmatic clinical trials. The combination of the two paradigms can be valuable in studies of combination therapies using factorial designs, as illustrated in the design of our on-going study, Optimizing Treatment for Complicated Grief.

Methods: We employed a 2x2 factorial design, with treatment arms (1) Medication with clinical management, (2) Placebo with clinical management, (3) Medication with clinical management + CGT, and (4) Placebo with clinical management + CGT; where Medication denotes the use of citalopram, and CGT denotes the use of Complicated Grief Treatment, a therapy designed specifically for the treatment of complicated grief (CG). Our first study aim compares Arms (1) and (2) to evaluate the efficacy of Medication for the treatment of CG from the explanatory perspective. Our second study aim compares Arms (3) and (4) to evaluate the impact of Medication in the presence of CGT, also from the explanatory perspective. Our third study aim compares Arms (3) and (1) to evaluate the impact of CGT in the presence of Medication, from the pragmatic perspective.

Result: We chose to interpret the medication aims (aims 1 and 2) as explanatory, due to the lack of established evidence for the efficacy of medication for the treatment of CG; and the CGT aim (aim 3) as pragmatic, considering existing evidence for the efficacy of CGT for the treatment of CG, e.g., in Shear et al. (2005). These interpretations led to important design decisions. First, double-blinded, placebo controlled comparisons are used for the medication aims, while open-label comparisons are used for the CGT aim with attention "bias" considered part of the treatment bundle for the naturalistic delivery of CGT under the pragmatic paradigm. Second, we employ parallel but distinct assessment schedules appropriate for each aim, with a fixed-time schedule (irrespective of treatment completion) for the explanatory aims, and a variable time schedule for the pragmatic aim.

Conclusion: Factorial clinical trials can be used to study multiple research questions simultaneously in the same study. It is important to accommodate unique design needs for various components of such hybrid studies.

Funding: NIMH Grants:1R01 MH060783; 1R01 MH085288; 1R01 MH085297; 1R01 MH085308

4 ## Factors Associated with Functional Recovery Among 224 Bipolar-I Disorder Patients Followed From Illness-Onset

Cruz N,[a,b] Khalsa H-MK,[a] Baldessarini RJ,[a] Vieta E,b Tohen M [a,c]

a. Department of Psychiatry, Harvard Medical School; International Consortium for Bipolar Disorder Research, McLean Division of Massachusetts General Hospital, Boston, MA; b. Bipolar Disorder Research Unit, University of Barcelona, Spain; c. Department of Psychiatry, University of Texas Health Sciences Center, San, Antonio, TX

Introduction: Functional impairment in bipolar disorder (BPD) is high prevalent. We hypothesized that functional recovery would be less likely with more:[a] total percent-time ill; and [b] time in depressive and mixed-state illness.

Methods: Type I BPD patients (N=224) were followed systematically from first-lifetime hospitalization prospectively every 6–12 months for 2 years and rated as functionally recovered or not, defined as achieving or exceeding both highest premorbid vocational and independent-living status. Times in specific

morbid states and other clinical and demographic factors were assessed using regression modeling.

Results: Functional recovery at 6, 12, and 24 months, respectively, was achieved by 32%, 35%, and 40% of patients. Among those not recovered by two years (60%), %-of-weeks in major illness-episodes ranked:mixed (5.6%) ≥ depression (5.4%) > mania-hypomania (3.9%) >> psychosis (0.1%); minor morbidity ranked:dysthymia (11.6%) > subsyndromal mixed (4.5%) > mild hypomania (3.9%) >> mild or questionable psychosis (0.8% of weeks). In preliminary bivariate regression, functional recovery was associated with fewer total episodes/year ($p=0.001$) and less time in depressive-dysthymic-mixed (D-type) episodes ($p=0.016$), as predicted. In addition, more time in mania ($p<0.0001$), younger onset-age ($p=0.002$), lower baseline symptom ratings ($p=0.016$), and being married ($p=0.022$), but not education ($p=0.651$), were associated with recovery. Multivariate logistic-regression found 3 factors at least weakly associated with recovery:lower baseline symptom score ($p=0.03$), fewer episodes/year ($p=0.03$), and being married ($p=0.08$).

Conclusions: More morbidity, especially in depressive or dysphoric-mixed states, was associated with less functional recovery among BP-I patients followed prospectively from onset.

5 ## Using Computational Neuropharmacology to Help Develop Switching Guidelines With Long Acting Antipsychotics in Schizophrenia.

Hugo Geerts[1,2] Athan Spiros[1] Robert Carr[1]

[1]In Silico Biosciences, 686 Westwind Dr, Berwyn, PA19312; [2]University of Pennsylvania, School of Medicine, Philadelphia, PA19104

Introduction: Release of the active agent in long-acting formulations of antipsychotics takes time and special attention is needed during the transition period when tapering off the basal oral medication. Because many antipsychotics affect many receptors at clinically relevant doses, the use of traditional PK/PD modeling, based upon relations between plasma level and pharmacodynamic effect, can sometimes yield erroneous results, especially when there are non-linear interactions of the two compounds at the same receptors.

Methods: We have developed a mechanistic Computational Neuropharmacology approach that is well calibrated and validated for schizophrenia using actual published clinical trials with 24 neuroleptic drugs for both Positive And Negative Symptoms in Schizophrenia (PANSS) total scales and Extra-pyramidal symptoms (EPS) liability. Actual functional brain concentrations at specific clinical doses are derived using a computer simulation of the competition between the neurotransmitter, the drug and its metabolite and a radio-active tracer to fit reported clinical PET imaging studies of tracer displacement at the relevant receptor. The computer model is based upon preclinical physiology from rodents and primates, but parametrized using actual human patient imaging, EEG, postmortem and clinical data and includes the physiology of 29 different targets.

Results: Based upon the PK profile and their unique pharmacological profile of the different drugs, we simulate their interaction in the Computational Neuropharmacology platform for a variety of doses and treatment schedules. We present a number of unexpected examples for both positive and negative synergy between two compounds during the transition period.

Conclusion: This approach could potentially be used to identify the best individualized 'titration' algorithm when switching from oral medication to a long-lasting formulation in a clinical trial. It could potentially identify possible issues around motor side-effects or lack of protection against exarcerbations during the transition phase, therefore possibly increasing the signal-to-noise ratio in clinical trials.

6 ## The Impact of Geography on Precision in Clinical Trials: Experience with Placebo and an Active Control in a Phase 2a Study in Patients with Acute Symptoms of Schizophrenia

Eric Watsky, Jeffrey H. Schwartz, Charlotte Kremer

Pfizer Inc

Introduction: To evaluate the effect of clinical trial site location in the US and outside of the US on the primary endpoint placebo change from baseline, active-placebo separation, and placebo responders in a Phase 2a clinical trial in schizophrenia. This exploratory analysis was based on the observation in the study of greater apparent treatment separation at US sites and recent reports of antipsychotic clinical trials (Patil 2007, Kemp 2008) that have highlighted active-placebo separation at sites located outside of the US.

Methods The ITT population consisted of 163 patients with acute symptoms of schizophrenia randomized to 1 of 3 doses of an experimental agent, to aripiprazole, or to placebo and treated for 21 days. A mixed

effects model was used to test the PANSS change from baseline to Day 21 and the comparison of active control to placebo PANSS change from baseline. A Fisher's exact test was applied to responders to treatment based on ≥30% PANSS change from baseline. Post hoc analyses were conducted with the original study design specification of alpha=0.1.

Results: The placebo group PANSS change from baseline to Day 21 for the pooled ex-US sites was statistically significant ($p<0.001$). The comparison of the active control to placebo change from baseline to Day 21 was significant for the pooled US sites ($p<0.1$) but not for the pooled ex-US sites. The percentage of subjects on placebo at US sites who demonstrated a ≥30% change on the PANSS was numerically lower than for subjects at ex-US sites.

Conclusion: This post-hoc exploratory analysis contrasts with other recent datasets regarding the assay sensitivity provided by US compared with ex-US sites and may have implications for site selection for the conduct of clinical trials in schizophrenia.

7 Trends in Placebo Response and Effect Size in Schizophrenia: A Meta-Analysis

Dan Davis PhD[1], Chip Hunter PhD[1], Barb Echevarria PhD[1], John Peloian MA[1], Kia Critteneden PhD[1], Bruce Wampold, Ph.D[1], Kenneth Kobak Ph.D[1]

[1]*MedAvante Research Institute*

Introduction: Anecdotal reports and a growing body of empirical evidence suggest apparent diminishing drug-placebo differences and a growing placebo response rate in Schizophrenia randomized controlled trials over time even to previously established comparator drugs.

Methods: A search was conducted using PubMed, PhRMA, FDA, Cochrane, and pharmaceutical company databases for published and unpublished studies covering the dates of 1980 through the present. This yielded 286 studies that were reviewed in their entirety to determine if they met inclusion criteria to the meta-analysis. Studies that included diagnoses other than Schizophrenia, (e.g., Schizoaffective, Schizotypal), were not included, nor were studies that did not present sufficient data to calculate an effect size. Twenty-seven published and unpublished studies met these criteria and were included in the analysis.

Results: A metaregression found that the main effect of decreasing separation of experimental drug from placebo due to increasing placebo effect size over time was not significant, and in fact, separation was increasing slightly (Slope -.012 Q=3.18 p=.074). A slight increase in placebo effect size over time was not significant (Slope = -.00053, Q= .0068, p = .93396). Sample sizes showed a significant increase over time for both placebo (F=10.40, p=.003) and treatment (F=6.48, p=.017) groups. Data were reanalyzed pooling active comparators in studies where comparators were utilized; however, no significant or meaningful differences emerged.

Conclusion: The study found a small increase in placebo effect and a small increase in effect size over time, neither of which was statistically significant. An important limitation was that over 90% of studies included in the analysis were positive trials, and this may not be reflective of all schizophrenia trials. The literature search was exhaustive and thus may reflect the lack of negative research outcomes reported. A large number of studies were also excluded because of insufficient data; had sufficient data been reported it is possible their inclusion may have altered the findings. Thus, a major secondary finding was that negative studies are rarely reported, and many studies are not adequately reported, thus limiting the validity and generalizability of the results.

8 Clinical Trial Design for the Assessment of Efficacy in the Treatment of Persistent Negative Symptoms of Schizophrenia

Pilar Cazorla,[1] Larry Alphs,[2] Robert W. Buchanan,[3] Nina Schooler,[4] Wilden Hollander,[1] Jun Zhao,[1] Phillip Phiri,[1] Armin Szegedi,[1] John Panagides[1]

[1]*Schering-Plough, Summit, NJ;* [2]*Ortho-McNeil Janssen, Titusville, NJ;* [3]*Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, MD;* [4]*State University of New York–Downstate Medical Center, Brooklyn, NY*

Introduction: The design of studies to establish drug efficacy for persistent negative symptoms in people with schizophrenia requires rigorous inclusion criteria to minimize the potential effect of confounding factors such as changes in positive, depressive, and extrapyramidal symptoms. To address regulatory, scientific, and ethical concerns, a trial should include participants with stable, persistent, negative symptoms; establish optimal treatment regimens; and identify appropriate endpoints and statistical methods. Different designs are necessary to assess antipsychotic monotherapy, combined or adjunctive antipsychotic pharmacotherapy, or to assess noninferiority versus superiority. We describe the design and

outcomes of an antipsychotic monotherapy study in people displaying persistent negative symptoms of schizophrenia.

Methods: A randomized, double-blind, flexible-dose study comparing 2 active drugs (Drug A vs Drug B) was designed. To avoid ethical issues involving long-term use of placebo in people with persistent negative symptoms, a placebo control was not included. Participants who completed that 26-week core study could enter a 26-week extension. The primary instrument for rating efficacy was the 16-item Negative Symptom Assessment scale (NSA-16). Changes on NSA-16 total score (Drug A vs Drug B) from baseline of the core study to the end of the extension were analyzed using a mixed model for repeated measures.

Results: Of 481 randomized participants, 349 completed the core study, 306 entered the extension, and 266 completed the extension (ie, a total of 1 year of treatment). Least squares mean ± SE changes in NSA-16 total score were –16.9±0.98 (from 61.7±0.85 at baseline of the core study) for Drug A vs –15.4±0.85 (from 60.4±0.74) for Drug B (P=0.23). Changes on the Quality of Life Scale, a secondary outcome measure, were 18.7±1.64 (from 45.1±1.63 at baseline of the core study) for Drug A vs 16.4±1.4 (from 47.7±1.42) for Drug B (P=0.28). Positive and depressive symptoms showed minimal change.

Conclusions: This trial was rationally designed to assess the effectiveness of treatment of the persistent negative symptoms of schizophrenia. Within this framework, 2 active agents produced statistically similar reductions in negative symptoms with long-term use. The minimal changes in positive and depressive symptoms suggest that this outcome may represent a primary treatment effect.

9 A Pattern Recognition Matrix for Placebo-Response in Schizophrenia

Mark Opler, PhD, MPH;[1,2] Guillermo Di Clemente, PhD, MSW[3]

[1]ProPhase LLC; [2]New York University; [3]CRONOS CCS

Introduction: CNS clinical trials present significant methodological and logistic challenges. High failure rates of Phase 2 and Phase 3 studies, weak signal detection, high placebo response rates, and treatment by country interactions are common concerns. There is a need for real-time capability to detect inconsistencies in efficacy outcome measures and to predict individual-level or group-level placebo response. Training and calibration to improve inter-rater reliability is one method; it is widely reported in the literature that the more training raters receive during the course of a trial, the less rater drift is observed and the more likely a trial will not fail (e.g., Mulsant et al., 2002). Placebo response has been identified as a problem across indications (e.g., Kemp et al, 2008) and it is reasonable to assume that even very high inter-rater reliability will be undermined if large numbers of placebo responders are enrolled in a trial.

Methods: A pattern recognition matrix for placebo-response in schizophrenia was developed based on a Phase II study of schizophrenia conducted in the US. A data monitoring algorithm based on the Positive and Negative Syndrome Scale (PANSS) was retrospectively applied to the unblinded data and score patterns for the placebo responders versus placebo non-responders were analyzed.

Results: A total of 35 placebo responders (those who completed all 8 study visits) and 35 randomly selected placebo non-responders were compared. For certain score patterns during the first 2 visits, patients were significantly likely to be placebo responders. Within this sample of patients who were randomized to placebo, those who demonstrated the pattern within the initial study visits were approximately three times more likely (OR=2.9, p=0.027) to demonstrate a placebo response.

Conclusion: This initial finding suggests that this method could aid in the detection of placebo response early in a trial. In concert with a data-monitoring process we would expect more robust signal detection. The use of the same system, coupled with ongoing training has demonstrated significant improvements in reliability, increasing the ICC by 19% in a 3-month period. If deployed at startup, this method provides a cost-effective way of managing the data quality in RCTs.

10 Methodology of Time to Onset of Response Characterization in the Proof-of-Concept Trial for a Combination Drug in MDD

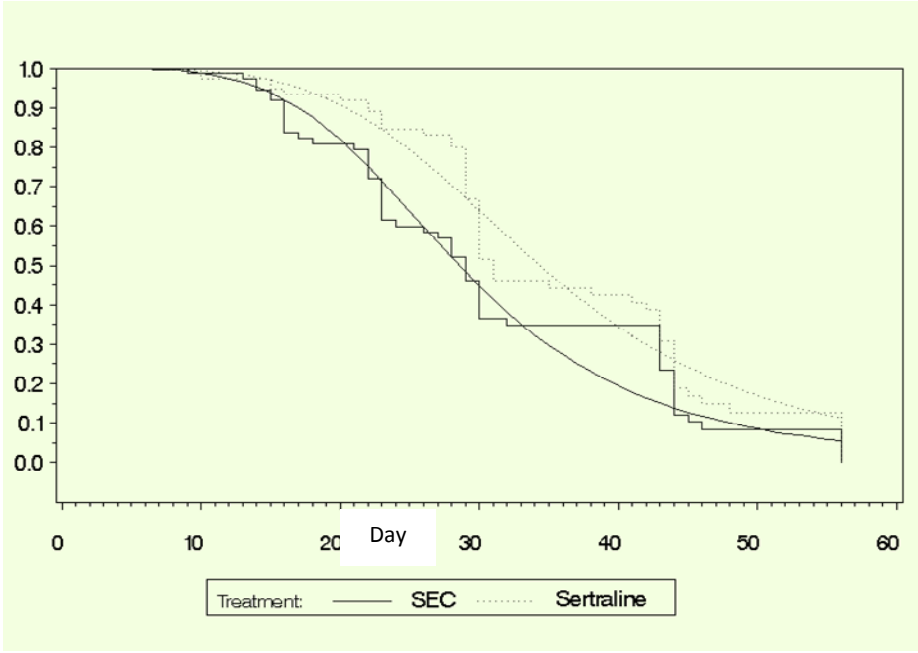
T. Ramey,[2] T. Stiger,[1]A.Banerjee[3] Y.A. Aleksandrovsky[4], A.S.Avedisova[4], U.Kalassalu[6],V.N. Krasnov [4] ,G.G.Neznamov[4], N.G.Neznanov[4], A.Okhapkin[4], A.B. Smulevich[4] , S.Sobolov[5] ; V.A.Totchilov[4], A. Dugar.[1]

[1]Pfizer Specialty Care BU, New London, CT, USA, [2]Pfizer Primary Care BU, New York, NY, USA, [3]Pfizer Clinical Research Statistics, Groton, CT, USA, [4]Moscow, St-Petersburg, Smolensk, Russia, [5] Novartis Institutes, Cambridge, MA, USA, [6]Tallinn, Estonia

Background: Co-administration of Sertraline and Elzasonan (SEC) may be more rapid-acting. Elzasonan is a 5-HT1B receptor antagonist. Combination of an SSRI and a 5-HT1B antagonist increases Serotonin release (Rollema H, et al., 1996).

Methods: Time to onset of response for SEC was compared to Sertraline with a mixture survival model, also referred to as the Laska-Segal conditional method (Laska, EM, et al, 1995) in an 8 wk DB, PBO-controlled study (n=262). Both response rate and time to onset of response (MADRS) were analyzed. Statistical design was a fixed-information group-sequential with the sample size re-estimation at interim.

Results: The Kaplan-Meier estimates of time to response, conditional on response, and the fitted log-logistic mixture-model estimates (Farewell, VT, 1982) were obtained (see the figure). The log-logistic estimates of the conditional time-to-response distributions indicated that among responders the estimated time to response for SEC is 1.2 times greater than that of Sertraline (p=.015 based on a likelihood ratio test). Observed cases analysis showed that at Weeks 2 and 3, 14.9% and 15.7% of the subjects of the SEC group were responders, while only 3.3% and 7.0% of the Sertraline group were responders.



Conclusions: Laska-Segal paradigm with Kaplan-Meyer approach in the proof-of-concept study provided evidence that a combination drug has a more rapid response than one of the components.

References: 1.Rollema H., Clarke T, Sprouse JS, et al. Combined administration of a 5-hydroxytryptamine (5-HT)1D antagonist and a 5-HT reuptake inhibitor synergistically increases 5-HT release in guinea pig hypothalamus in vivo. J Neurochem 1996; 67(5):2204-7. 2. Laska, EM, Siegel C. Characterizing onset in psychopharmacological clinical trials Psychopharmacol Bull 1995; 31(1):29-35. 3. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 1982; 38:1041-1046.

11 ## Placebo Response in Trials of Antidepressants in Patients With Major Depressive Disorder

Christine Guico-Pabia[1] Jeff Musgnung[1] Ron Pedersen[1] Philip Ninan[1]

[1]*Wyeth Research, Collegeville, Pennsylvania*

Introduction: The degree of placebo response provides a pivotal context for defining antidepressant efficacy in clinical trials. Evidence suggests an increase in the placebo response in the past decades (Papakostas & Fava 2008). A variety of factors such as baseline severity of symptoms and attention paid to subjects during clinical trials have been proposed as explanations. The objective of this analysis is to examine the placebo response in Wyeth-sponsored antidepressant trials and was to explore factors that may be associated with changes in placebo response.

Methods: The analysis included data from all Wyeth-sponsored, randomized, double-blind, placebo-controlled studies of venlafaxine and desvenlafaxine completed as of August 2009 involving adult patients with Diagnostic and Statistical Manual of Mental Disorders (DSM) defined MDD. Data from 22 studies in which patients received venlafaxine (75-300 mg/d) or placebo for up to 8 weeks and from 9 studies in

which patients received desvenlafaxine (50-400 mg/d) or placebo for up to 8 weeks were summarized. Effect sizes (using Cohen's d) were calculated based on the 17-item Hamilton Rating Scale for Depression (HAM-D17) total score. Effect sizes were plotted against study start date, mean baseline HAM-D score, minimum baseline HAM-D score, and mean number of assessments per visit.

Results: In venlafaxine and desvenlafaxine studies, effect sizes generally decreased over time, suggesting an increase in placebo response. Mean baseline HAM-D scores did not appear to substantially influence effect size, although effect sizes tended to be lower as minimum baseline HAM-D scores increased. Effect sizes tended to decrease as the number of assessments per visit increased.

Conclusions: Further investigation and analysis of these and other factors that may influence placebo response in antidepressant studies is necessary. Future studies of antidepressants should be designed with such factors in mind. The demonstration of the efficacy of novel medications with potential antidepressant efficacy is jeopardized by the increase in placebo response in clinical trials.

Research supported by Wyeth Research

12 Examining Patient Validity for Clinical Trials: A Post-hoc Analysis of Placebo Responders

Howard Hassman D.O. and Sean Haley Ph.D.

Introduction: In an effort to identify how early screening processes and related patient selection might be improved to better specify “true” study candidates, the authors have endeavored to identify shared characteristics among patients that experience a positive placebo response while providing empirically driven guidance to assist investigators seeking to select the highest quality subjects for inclusion in the trial.

Methods: The authors identified twelve (12) studies from four (4) pharmaceutical sponsors representing four (4) disease states (General Anxiety Disorder, Major Depressive Disorder, Schizophrenia and Bi-polar Depression) to investigate patient factors related to placebo response.

Several variables suspected of being related to a placebo response were selected for analysis including: age, sex, race, any employment, referral source, years of disease illness for the index disease, adverse events during the trial, prior exposures to specific medication screening, diagnosis of a co-morbid anxiety disorder at screening, the number of prior clinical trials, co-morbid-psychiatric diagnosis, report of a co-morbid medical condition prior to screening, specifically: gastrointenstinal disease/disorders, cardio-vascular disease/disorders, and surgery.

For all analyses, placebo responders were defined as subjects assigned to placebo whose last observed HAMD score demonstrated a 50% or greater reduction (improvement) over their baseline HAMD score.

Independent t-tests comparing placebo responders to non responders were conducted on all continuous variables (age, Ham-D score at baseline and years of illness). Pearson’s Chi-Square tests were used for all remaining non-parametric variables.

Results: Approximately 33% of the sample (N=23) had CGI-severity scores indicating that they were markedly ill or worse.

In addition, Pearson’s Chi Square comparisons found no significant difference for sex, race, employment, referral source, severity of illness as measured by CGI at baseline, any adverse event, presence of an anxiety disorder, or reports of any gastro-intestinal or cardio vascular disease between placebo responders and non-responders.

Recommendations/Conclusions: Placebo patients make up a small portion of all patients and placebo responders comprise even a smaller portion of the set - even when conducting a post hoc analysis. Sponsors do not require, nor do most clinical sites voluntarily obtain, valuable data regarding the potential subjects’ motivation for participation in the trial. As a result, this data is extremely hard for the site and sponsor to analyze. It is extremely difficult for the study site to obtain the unblinding codes and relevant treatment groups for trials conducted.

It is the recommendation of the authors that sponsors and study sites collaborate on a project (or series of projects) that would help to identify specific criteria or characteristics that would assist investigators in identifying outliers (placebo responders) prior to admission in a clinical trial. Such a project would test the hypothesis of whether using patients who frequently participate in clinical trials has an impact upon their clinical response and would capture demographic, medical, and psychiatric information on participants in order to help sponsors and investigators alike further understand the phenomenon of limited clinical

efficacy in US subjects.

Please direct all inquires to Dr. Howard Hassman, CRI Worldwide, Phone 856-533-5011

13 Sensitivity of a Comprehensive Rating Scale for Bipolar Disorders to Assess Improvements with Treatment

Charles L. Bowden, Jodi Gonzalez, Vivek Singh, Jim Mintz, Peter Thompson, Carmina Bernardo

The University of Texas Health Science Center at San Antonio

Aim: To establish the sensitivity to change with treatment of the Bipolar Inventory of Signs and Symptoms Scale (BISS) which assesses the full spectrum of symptom manifestations in Bipolar Disorders (BD).

Method: BD I or II patients were assessed with the BISS and concurrently the YMRS or the Mania Rating Scale (MRS), MADRS and CGI. A second assessment was performed at least 4 weeks after the initial assessment. The 114 subjects assessed for sensitivity to improvement in depression were required to have baseline CGI-D scores ≥ 2 and to be in studies which provided active treatments for depression. Analogous criteria were required for the 68 subjects studied for sensitivity to improvement in manic symptomatology.

Results: Improvement effect size by the BISS-Depression scale was1.01 and 0.95 by the MADRS. Improvement effect size by the BISS Mania scale was 0.61and 0.56 by the YMRS. We assessed sensitivity to improvement in relationship to baseline severity, based on CGI-D and M scores at baseline. The overall 2 way ANOVA indicated no significant difference between BISS and MADRS effect sizes, however the interaction between change effect size and baseline CGI-D score indicated a significant increase in change effect sizes as CGI-D scores increased from mild-moderate to marked-severe. BISS and MADRS effect size improvement in depression was in the low range for patients with CGI-D scores of 2. Patients with CGI-D scores of 3 had medium ES change with the BISS, but small with the MADRS. Within cell t tests for subjects with CGI-D baseline scores of 3 indicated significant change effect size for the BISS –D vs non-significant change for the MADRS. Patients with CGI-D scores ≥ 4 had large ES change with both scales.

For improvement in manic symptomatology, the ANOVA indicated a significant interaction between scale and severity. Patients with baseline CGI-M scores of 2 had low ES changes with both scales. Patients with baseline CGI-M scores of 3 had large ES changes with BISS-M vs medium effect size change with YMRS/MRS assessment. Patients with baseline CGI-M ≥ 4 had large ES changes with both BISS-M and YMRS/MRS assessments.

Conclusion: BISS-D and BISS-M scores yielded moderate or large effect size changes for improvement which were similar to MADRS and YMRS/MRS effect size scores for depression and mania respectively. The BISS provides somewhat larger ES improvement differences than the MADRS for patients with mild to moderate depressive symptoms; BISS and MADRS provided equivalent effect sizes at the highest level of severity of depression. BISS-M effect size improvement was somewhat more sensitive than YMRS/MRS for patients with moderate initial severity and somewhat less sensitive with high initial manic severity.

14 A Comparison of Asia to Rest of World in Use of Doctorate vs. Non-Doctorate Level Investigators for Symptom Rating and Diagnosis

David G. Daniel, MD[1] Gwyneth M. Moya, MPH[1] John Bartko, PhD[2]

(1) United BioSource Corporation, Wayne, PA, (2) Private Statistical Consultant, Newville, PA

Background: Asia is one of the most rapidly growing regions for conduct of global CNS clinical trials. Regional differences in rater credentials and interview skills in evaluating psychopathology are potential confounds in interpretation of clinical trials data and represent a challenge in training of raters for international multi-site studies. Relatively little comparative data exists with respect to Asian and non-Asian raters

Method: Training data from industry sponsored multi-center clinical trials utilizing the Young Mania Rating Scale (YMRS), Montgomery and Asberg Depression Rating Scale (MADRS) Positive and Negative Symptom Rating Scale (PANSS) (for all three scales combined, n=9294), Structured Clinical Interview for Diagnosis (SCID) (n=3168) and Research Interview Skills Assessment (RISA) (n=118) was collected from the years 2004-2008, inclusive. This data was retrospectively analyzed to assess for differences between Asia and the rest of the world (ROW) in use of doctorate vs. non-doctorate investigators as well as proficiency in interview skills.

Discussion: Regional variation in credentials of raters are recognized but poorly characterized. In this retrospective analysis the proportion of doctorate to non-doctorate investigators administering rating scales and structured diagnostic instruments was higher in Asia than in North America but lower than in European and South American raters. In a subset of raters who were administered the RISA levels of proficiency in interview skills were similar between Asian and North America raters but lower than in eastern European raters. With the rapid growth of CNS clinical trials in Asia, it will be important to characterize potential differences in rater qualifications, interview skills and perceptions of psychopathology that could lead to regional and country effects in clinical trials ratings. Future analyses will address Asian vs, ROW differences in measurement of symptom severity as well as regional variation within Asia.

References: Rubio-Stipec M, Canino I, Hsiao-Rei Hicks, M and Tsuang MT:Cultural Factors Influencing the Selection, Use, and Interpretation of Psychiatric Measures. In Handbook of Psychiatric Measures, Second Edition, American Psychiatric Publishing, Inc 2008.

15 Cross-Cultural Comparisons of American and Japanese Clinical Raters on Patients with Major Depressive Disorder using the Hamilton-Depression Scale-17 (HAM-D_{17})

Graciete Lo, MA[1,2] Christian Yavorsky, PhD[1] Karen A.Tourian, MD[3]; Bruno Pitrosky, PhD[3] Linda Mele, MS[3] Ashleigh DeFries, BA[1,4,5] Mark Opler, PhD, MPH[1, 6]

[1]*ProPhase LLC,* [2]*Fordham University,* [3]Wyeth Research, [4]Teachers College, [5]Columbia University, [6]*New York University School of Medicine*

Introduction: Where clinical trials are conducted internationally, it is imperative to attend to cultural influences on clinical assessment tools. Cross-cultural literature has consistently shown that the expression of distress differs across cultures (Kirmayer, 2001). In a study looking at differences in HAM-D scores in Japan, United States, and Europe, Ohishi & Kanijina (2002) indicated that the items 'depressed mood' and 'feelings of guilt' were rated as less severe in the Japanese cohort. This suggests that expressed emotionality may be assessed differently across cultures and in this study we sought to further explore this dimension.

Methods: We examined the results from 54 American and 106 Japanese raters in a global trial for desvenlafaxine (Wyeth Research). For training, raters watched and rated two videos of HAM-D interviews conducted in English, each with a depressed woman of Asian descent (Japanese subtitles were provided). The first video depicted a more severely depressed patient (Video 1); the second video was conducted with a moderately depressed patient (Video 2).

Results: In Video 1 we found no significant difference in the overall HAM-D score between Japanese (M=22.88, SD=3.26) and American raters (M=23.35, SD=3.17), $t(133) = 0.816$, p=.416 (two-tailed). However, in Video 2, we found a significant difference in the overall HAM-D scores for Japanese raters (M=17.17, SD=3.08) and American raters (M=18.13, SD=2.35), $t(154) = 2.00$, $p < .05$, though the magnitude of the differences in the means was small ($\eta 2 = .025$). On Insomnia-Late, Anxiety psychic, and Insight items, Japanese clinicians rated *both* patients as more severe than American raters.

Conclusion: Our analyses suggest that there are differences in how American and Japanese raters evaluate symptom severities of identical patients. While American and Japanese raters may not rate severely depressed patients (Video 1) that differently from each other, there seems to be cultural influences on how they rate patients with moderate depression (Video 2). The assessment of severe symptomatology does not seem to be impacted by cultural differences. However, rating mild to moderate level of depression may pose more of a challenge and has implications for training as well as interpretation of results from trial conducted in these regions.

16 A Meta-Analysis of Computerized Assessment Batteries in Schizophrenia Medication Trials

Henry J. Riordan, Ph.D[1] Luke Eastman[2] Lauren Hoffman[2] Neal Cutler, MD[1] & Paul J. Moberg, Ph.D.[2]

[1]*Worldwide Clinical Trials, King of Prussia, PA 19406;* [2]*University of Pennsylvania, Dept. of Psychiatry, Philadelphia, PA 19104*

Introduction: A variety of computerized assessment batteries (CABs) have been utilized to assess cognitive impairment in schizophrenia; however, there is no consensus regarding CABs sensitivity to medication effects. This meta-analysis provides a quantitative overview of CABs used in schizophrenia research by examining medication trials with at least one pre and post cognitive assessment.

Methods: A structured search of the CAB literature using the PsycInfo, MEDLINE, PubMed, and Google Scholar databases yielded 15 suitable publications that met inclusion criteria for meta-analytic review. Each CAB website was also examined for relevant publications, resulting in a total of 81 separate pre-post effects. CABs reviewed included CANTAB, ANAM, CogState, CogLab, and MINDSTREAMS. Specific tests from each CAB were extracted and grouped into cognitive domains reflecting executive function, working memory, verbal and non-verbal memory, visuospatial reasoning and motor functioning. Effect sizes (ES) (Cohen's d) were then calculated for each CAB, their component subtests, and for each cognitive domain.

Results: Analysis of medication effects on cognitive functioning, across different medication types, revealed an overall moderate effect size ($d = 0.523$) for all CABs that was significantly heterogeneous ($p < 0.001$). Of the five CABs, CogLab yielded the largest effect size ($d = 0.79$) followed by ANAM and then CogState. Effect sizes were largely driven by battery composition with measures of attention ($d = 0.809$) and visuospatial reasoning ($d = 0.702$) yielding relatively higher ESs than non-verbal memory ($d = 0.459$) and executive functioning ($d = 0.403$); although these four domains did not differ significantly from each other. Type of treatment intervention also impacted ES with combination treatment (Haloperidol plus nicotine) yielding the largest ES ($d = 1.05$) followed by Haloperidol alone, and then various antipsychotics and nooptropics. Important moderator variables included previous medication type, inpatient/outpatient status, number of follow-up cognitive assessments, PANSS negative symptomatology score, and patient age.

Conclusion: This meta-analysis suggests that it is possible to more confidently select CABs, their component subtests, and cognitive domains that are more likely to be sensitive in treatment trials; and that this sensitivity is moderated by medication type and important disease-related and demographic variables.

17

The InterSePT Scale for Suicidal Thinking-Plus (ISST-Plus): A Modified Instrument for the Comprehensive Assessment of Suicidality in Clinical Trials of Patients with Schizophrenia or Schizoaffective Disorder

Larry Alphs[1] Jean-Pierre Lindenmayer[2]

1. Ortho-McNeil Janssen Scientific Affairs, Titusville, NJ.; 2. NYU Langone Medical Center, New York, NY

Introduction: Approximately 50% of individuals with schizophrenia will attempt suicide during their lifetime. [1-3] The FDA's Division of Psychiatry Products (DPP) recently implemented a policy requiring a prospective suicidality assessment in protocols developed under the Division's review. An acceptable assessment instrument should map to the Columbia Classification Algorithm for Suicide Assessment (C-CASA). The 4-part ISST-Plus instrument was developed to meet the FDA requirement and offers advantages over other instruments in the assessment of suicidal thinking severity and behavior and to facilitate comparisons across data points and patients within clinical trials.

Methods: The ISST-Plus Part I rates suicidal thinking during the past 7 days using the 12-items in the original ISST at 3 severity levels. Part II assesses suicidal behavior (injuries, accidents, overdose, suicide preparation) since the last visit. Rater's global assessment of the patient's suicidality at the interview is recorded in Part III; death by suicide during the study is recorded in Part IV. Reported suicidal behavior or completed suicide information can be entered in the structured Suicide Attempt Narrative section.

Results: Part I of the ISST-Plus assesses severity of suicidal thinking within a defined time interval, allowing comparisons over time and across patients. Part II permits rating of suicidal behavior during a trial. The Suicide Attempt Narrative permits efficient collection of suicidal behavior information that can be used for safety reporting. Global assessment of suicidality (Part III) facilitates risk assessment. The ISST-Plus takes 15 to 20 minutes to complete and maps to the C-CASA categories of completed suicide, suicide attempts, preparatory acts toward imminent suicidal behavior, suicidal ideation, self-injurious behavior but no suicidal intent, no deliberate self-harm, and self-injurious behavior but suicidal intent unknown.

Conclusion: Although not yet fully validated, the ISST-Plus permits ready estimation of risk of suicidality and tracks changes in severity of suicidality over time. The authors consider the ISST-Plus, with its recent modifications (separate assessment of suicidal thinking and behavior, mapping to the C-CASA and measurement of global severity of suicidality) to be a useful tool for assessing suicidality in clinical trials and an instrument that may meet the new policy requirements by the FDA's DPP.

References: 1. Palmer BA, Pankratz VS, Bostwick JM. The lifetime risk of suicide in schizophrenia: a Reexamination. Arch Gen Psychiatry 2005;62:247-253. 2. Fenton WS, McGlashan TH, Victor BJ, Blyler CR. Symptoms, subtype and suicidality in patients with schizophrenia spectrum disorders. Am J Psychiatry 1997;154:199-204. 3. Radomsky ED, Haas GL, Mann JJ, Sweeney JA. Suicidal behavior in patients with schizophrenia and other psychotic disorders. Am J Psychiatry 1999;156:1590-1595.

18 **Three Levels of Rater Performance in Standardized Positive and Negative Syndrome Scale (PANSS) Training**

Graciete Lo, MA;[1,2] Christian Yavorsky, PhD;[1] Mark Opler, PhD;[1,3] Ashleigh DeFries, BA;[1,4,5]

[1]ProPhase LLC; [2]Fordham University; [3]New York University; [4]Teachers College; [5]Columbia University

Introduction: The importance of rater training for reliability and validity in clinical trials is well documented (Ventura, Green, Shaner & Lieberman, 1993; Ivancevich, 1979; Muller & Szegedi, 2003; Muller & Wetzel, 1998). There is less agreement about what constitutes adequate training and what defines a successful outcome. Equally, the assumption that "training" and "raters" are a unitary construction is problematic. In this study the authors examined data from several large training events that used standardized training procedures and raters with similar levels of education and experience to determine if there were any differences between rater performance at baseline (before training) and endpoint (after training) that emerge independent of these factors.

Methods: Results from multiple training events held internationally were analyzed to determine if differences between baseline and endpoint scores were significant. 308 raters scored videotaped interviews of the Positive and Negative Syndrome Scale (PANSS) in training events. These results were then grouped into three categories based on concordance with gold-standard scores and change from baseline scores.

Results: Three subgroups of raters emerge based on concordance with gold-standard ratings:raters that score high at baseline and endpoint; raters that score fair at baseline and good at endpoint; raters that do not appear to improve. For the stronger initial raters that continued to perform well after training there was a less substantial change ($t(96)=2.953$, $p<.005$) than those raters that did less well at baseline but improved after training ($t(160)=4.037$, $p<.001$). The smaller group (n=52) that did not appear to show improvement after training had lower ICCs for negative (range .60-.72) and general subscale (range .67-.85) items.

Conclusion: Within training groups there appear to be three groups of raters that emerge independently of rater qualification and training received:those that performed well initially and well at endpoint; another group that performed marginally at the beginning of training and showed improvement by the end; and a third group that did not improve as a result of training. Analysis of ICCs suggests targeted training for individuals that perform less well at baseline could be beneficial.

19 **Same Versus Different Raters and Rater Quality in a Clinical Trial of Major Depressive Disorder: Impact on Placebo Response and Signal Detection**

Kenneth A. Kobak, PhD.,[1] Joshua D. Lipsitz, Ph.D.,[2] Alan D. Feiger, M.D.,[3] Michael J. Detke, M.D., Ph.D.[1]

[1]MedAvante Research Institute, [2]Columbia University, [3]University of Colorado Depression Center

Introduction: Minimizing error variance in clinical trials has traditionally been accomplished by using same rater at each visit. However, clinicians (and patients) typically expect to see improvement over time rather than no change or worsening. Thus, knowing the study visit may subtly bias clinicians' ratings. In addition, seeing the same patient week after week can result in less thorough or independent probing. Using different raters may tend to mitigate this, as well as the potentially confounding impact of the 'therapeutic alliance' (vs. change due to study drug).

Method: We retrospectively compared patients who had the same rater at baseline and endpoint (n=163) to patients who had a different rater at baseline and endpoint (N=53) in a multisite (N=20) depression study. The study sponsor provided data from the active comparator (paroxetine; n=109) and placebo (n=107) cells.

Results: Subjects with the same interviewer at baseline and endpoint had a mean HAMD change (drug minus placebo) of +0.56 (mean change of 9.1 and 9.7 for paroxetine and placebo respectively), p=.625. Those with different raters had a mean HAMD change of -3.76 (mean change of 11.5 and 7.7 on paroxetine and placebo respectively), p=.065. The difference in change for same raters and different raters was not statistically significant, p = .062. While the small cell size for different raters may have precluded statistical significance, results indicate a trend in this direction.

The greatest paroxetine-placebo difference was found with good interview quality1 and different raters at baseline and endpoint (mean HAMD change = -15.5, N=5, p= .008) (none of the other comparisons were significant). The smallest difference was same raters and poor interview quality (mean HAMD change = 2.87, N=25). The small cell sizes make the risk for Type I error great, and thus this result should be interpreted with caution.

Conclusion: In the present study, different raters at baseline and endpoint was associated with larger drug-placebo differences. Limitations include the use of retrospective data, and a small number of subjects per cell. It is unclear if these findings generalize to other disorders.

1. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. Am J Psychiatry. Mar 2005;162(3):628.

20 Signal Enhancement System (SES) Identifies Inaccurate Ratings and Quality of Clinical Interview Conducted in a CNS Trial.

Richa Gaur PhD[1], Martha Sajatovic MD[2], Susan De Santi PhD[3],

[1]TCG Newark, Delaware, USA; [2]Case Western University, Ohio, USA; [3]New York University, New York, USA

Introduction: The Signal Enhancement System (SES) has been developed for clinical interview quality control, overall study assessment standardization, and scale rating integrity at the investigating sites. The present study evaluates the performance of SES implemented in a depression study.

Method: This study of major depressive disorders included three US sites and six SRs trained and certified on the Montgomery Asberg Depression Rating Scale (MADRS), prior to the study start. During the study, MADRS patient interviews and ratings were monitored with the help of SES. Audio and video recording of the interview session is transmitted along with the ratings and collateral information over the web to a secure server. IR (blind to SR ratings) reviews the recorded video via the Internet, evaluates interview quality by completion of the Rater Applied Performance Scale (RAPS). The IR is identified as the local "expert" and is strongly encouraged to review the recorded interview, consider the SR narrative for score justification, and provide the score that best captures the patients severity at the time of the interview. 29 patient interviews were rated separately by the SR and IR, and discrepancies in ratings and interview quality were analyzed for the baseline visit.

Results: Of the 29 patient interviews rated by the SR and IR, 62% were consistent and required no further discussion. 38% of the patient interviews received discrepant ratings between the SR and IR and required further attention. 82% discrepancies observed between the SR and IR was due to overrating of patient symptoms by the SR. As per the RAPS feedback, 21% interviews were rated as unsatisfactory by the IRs as raters did not adhere to MADRS structured interview guide, which was the prime cause of discrepancies between the SR and IR.

Conclusions: SES correctly indentified inaccurate ratings and poor interview quality which was due to non adherence to structured interview guide (crucial for the maintenance of reliability of ratings). Thus SES is an efficient tool to identify ratings that do not conform to pre-set rating standards and poor interview quality.

21 A Survey of Rater Perceptions and Expectations of a Rater Training Program in a CNS Drug Trial

Richa Gaur PhD[1], Martha Sajatovic MD[2], Susan De Santi PhD[3],

[1]TCG Newark, Delaware, USA; [2]Case Western University, Ohio, USA; [3]New York University, New York, USA

Introduction: Rater training programs have been developed to provide effective training for raters for accurate and consistent ratings throughout the trial. While training programs focus on improving theoretical and applied rating skills of raters, the training needs of individual raters may not be incorporated into these programs. Moreover, there is a lack of standardized training methods. This is a cross-sectional survey of perceptions and expectations of rater training programs among experienced CNS trials investigators.

Method: The 19 item survey was developed utilizing input from rater training experts and clinical researchers. The format consisted of closed-ended and open-ended questions that queried perceived strengths and weaknesses of rater training programs, training needs of individual raters, and suggestions for future programs. The survey was emailed to 742 CNS trials investigators. The respondents had to click on the link in the email to go to the survey site. The 34 respondents, (5% of queried sample) were CNS trials investigators from 6 continents with an average experience of 11 years in clinical rating scales.

Results: Most respondents (77%, N=26) identified "understanding of the items on a scale" as the most important component of a rater training curriculum, whilst 68%, (N=23) felt "discussion on common errors committed while rating" followed by "interviewing and observation skill training" (65%, N=22), "cultural specific training on rating scales"," reducing bias while rating" and "rating complex patients " (56%,

N=19). Over two-thirds (68%, N= 23) of raters reported a "lack of interviewing competency" and " accurate observation and interpretation" as the most challenging obstacle to an accurate rating followed by 53%, (N=18) reported as "clinical research interviewing skills", "rating acutely psychotic patients" and " lack of time to interview patient at the site".

Conclusions: Experienced CNS study investigators perceive typical rater training programs as inadequate for the provision of reliable and accurate outcomes ratings. Detailed assessment of interview skills prior to the start of the study as well as continuous centralized monitoring of interviews by independent raters is likely to reduce the variance in CNS clinical trials.

22 The Challenge of Patient Ascertainment in Clinical Trials – New Data

Michael Detke[1, 2], Janet Williams[1, 3], Kenneth Kobak[1], Amy Ellis[1], Earl Giller[1], Andrew Leon[4], Scott Reines[1], John Kane[5]

[1]MedAvante Research Institute, [2]Indiana University School of Medicine & Harvard Medical School, [3]Columbia University College of Physicians and Surgeons, [4]Weill Cornell Medical College, [5]Zucker Hillside Hospital & Albert Einstein College of Medicine

Introduction: Clinical trials fail too frequently (up to 50% failures in trials powered at 80-90%). Signal detection might be enhanced with more reliable scales, greater rater reliability, or the use of independent assessments; here we focus on the last of these. Previous studies showed that 1/3 to 1/2 of the patients enrolled in two MDD trials by site raters would be excluded based on the patient's self-rating or remote blinded clinicians' ratings of initial severity. New data on the extent and characteristics of patient ascertainment discrepancies and various methods to mediate it will be presented.

Methods: Inter-rater reliability and internal consistency reliability were assessed in one MDD study. Two doses of an experimental compound were compared to placebo in a GAD study in which remote blinded clinicians and site raters assessed patients on the HAMA. In ongoing studies (of MDD, GAD & SZ) patients were assessed by both site raters and by remote blinded clinicians. In two of these studies, accuracy of diagnosis was examined.

Results: Internal consistency reliability (Cronbach's alpha) was strong for remote blinded clinicians at screening and endpoint and for site raters at endpoint (.67 - .83) but much lower for site raters at screening (.38). In the completed and ongoing studies of MDD, GAD & SZ, 34% (range: 5-56%) of patients included by site raters would have been excluded based on remote blinded clinicians' ratings of initial severity. SCID-CT assessments by remote blinded clinicians also revealed potential diagnostic errors in patients previously screened for study entry by site-based raters. In one study of GAD, patient ascertainment by remote blinded clinicians increased the drug effect size from .43 to .74.

Conclusion: Patient ascertainment issues are pervasive and substantial; on symptom severity alone over 1/3 of patients enrolled in clinical trials may not meet protocol-specified inclusion/exclusion criteria. Diagnosis is an additional source of potential error. Independent assessment of symptom severity by patients appears to have potential benefit in one MDD study. Remote blinded clinicians may be beneficial for diagnosis and symptom severity assessment across several diagnoses. Accurate patient ascertainment may substantially increase effect size.

23 Does Placebo Response Differ Between Objective and Subjective ADHD Measures?

Calvin R. Sumner, MD,[1] Virginia K. Sutton, PhD,[2] and Jeffrey H. Newcorn, MD,[3]

[1]Biobehavioral Diagnostics; [2]i3 Global; [3]Mount Sinai School of Medicine

Introduction: Placebo response is a challenge in conducting ADHD trials. A pilot study compared the QuotientTM ADHD Test, a computerized assessment of hyperactivity, inattention, and impulsivity, and ADHD rating scales; lack of strong agreement between these measurements led to an examination of placebo response associated with each.

Methods: Children aged 6 to 14 with a DSM-IV ADHD diagnosis based on clinician Kiddie-Sads-Present and Lifetime Version interview were randomized in one of two sequence groups (placebo, low dose, and medium dose or low dose, medium dose, and placebo) to atomoxetine or extended-release methylphenidate for a 3-week, double-blind trial. Placebo response was defined using three thresholds – any improvement, >25% improvement, or >40% improvement from baseline on the Quotient Global and ADHD-Rating Scale (ADHD-RS) Total scores. Lin's concordance coefficient measured baseline and placebo score agreement.

Results: Of the 30 subjects with placebo and baseline scores, 90%, 47%, and 27% met the three response thresholds (any, >25%, or >40%, respectively) on the ADHD-RS Total score compared with 27%, 7% and 0% of subjects on the Quotient Global Score. Lin's concordance correlation coefficient was 0.78 and 0.38

for the Quotient Global and the ADHD-RS Total scores, respectively.

Conclusion: Although larger trials are warranted, we tentatively conclude that using objective measures and higher response thresholds may enhance assay sensitivity in clinical trials.

24 Test-retest Characteristics of the MATRICS Consensus Cognitive Battery in a 29-site Schizophrenia Clinical Trial of Lurasidone Versus Risperidone

Kolleen Hurley Fox[1] Richard S.E. Keefe[2] Philip D. Harvey[3] Josephine Cucchiaro[4] Cynthia Siu[5] Antony Loebel[4]

[1]NeuroCog Trials, Inc., Durham, NC, USA, [2]Duke University Medical Center, Durham, NC, USA, [3]Emory University, Atlanta, NC, USA, [4]Dainippon Sumitomo Pharma USA, Fort Lee, NJ, USA, [5]Data Power (DP), Inc., Ringoes, NJ, 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. It has been an empirical question whether the MCCB would demonstrate these favorable characteristics when administered in the context of the type of large multi-site industry trial for which it was designed.

Methods: 323 clinically-stable outpatients with schizophrenia were randomized 2:1 to flexibly-dosed lurasidone and flexibly-dosed risperidone, respectively. Testers from 29 sites were trained and certified, and all MCCB data were reviewed and re-scored centrally. The MCCB was administered at screening and 7-21 days later at baseline. A measure of functional capacity, the UCSD Performance-based Skills Assessment – Brief Version (UPSA-B) was also measured at baseline. The MCCB generates a composite score and cognitive domain scores standardized to a normative population with mean (T) = 50 and SD = 10.

Results: Baseline T-scores for the 7 MCCB cognitive domains and a composite score were determined for 231 male and 92 female subjects, mean age 43.1 years (SD=10.4), mean PANSS total score 67.5 (SD=11.7) and mean UPSA-B total score 70.0 (SD=16.2). Only 14 test scores were missing out of a total of 6460 test assessments for the 10 MCCB tests performed in 323 subjects at 2 occasions (99.8% complete). At baseline, all 323 (100%) patients had sufficient data for computing a composite score according to the MCCB criteria. The mean (SD) MCCB composite score was 24.7 (12.1) at screening and 26.9 (12.4) at baseline. The test-retest reliability for the MCCB composite score was very high (ICC=0.88). Construct validity was also strong, as the MCCB composite score demonstrated a large correlation with the UPSA-B composite score (r=.61, df=304, P<.001). The practice effect on the composite score was small (z=0.18). Discussion: In the context of a 29-site clinical trial in stable outpatients with schizophrenia, the MCCB is sensitive to cognitive deficits in all domains, demonstrates excellent test-retest reliability and construct validity, and small practice effects.

25 Preliminary Psychometric Comparisons of Remote-Televideo and Face-To-Face Administration of a Commonly Applied Neurocognitive Assessment Battery

Aaron S. Kemp[1,2] Kirstan N. Gooch[2] Joisabel L. Goldberg[2] Sitha Bun[2] Dimitrios Gripeos[1,3] Christopher Reist[1,3] Barton W. Palmer[4,5] Mark Bondi[4,5] James P. O'Halloran[2]

[1]Psychiatry and Human Behavior, University of California, Irvine School of Medicine; [2]NeuroComp Systems, Inc., Irvine, CA; [3]Veterans Administration Healthcare System, Long Beach, CA; [4]Psychiatry, University of California, San Diego; [5]Veterans Medical Research Foundation, San Diego, CA

Introduction: Neurocognitive assessment has become an integral component of controlled clinical trials of candidate "cognitive-enhancing" treatments for a wide range of CNS disorders. However, manual administration of large, paper-based, neurocognitive assessment batteries is often inefficient, error-prone, and inconsistent across multiple sites. Existing computerized testing systems are also limited in both the assessment instruments available and the range of impairments that can be accommodated with the subject sitting alone at a single display. Therefore, a unique, dual-display computerized testing system was developed, with funding from the National Institutes of Health, that integrates (rather than replaces) the examiner for computerized administration of standard neurocognitive assessment batteries (O'Halloran et al., 2008 & Kemp et al., 2008). While originally developed to be administered with the two displays positioned side-by-side for face-to-face administration, the dual-display configuration has now been extended to support two-way, wireless televideo communications for remote administration by an examiner at a distal location. The purpose of the current study was to evaluate the psychometric feasibility of this novel configuration by comparing the concurrent validity and test-retest reliability of computerized,

remote-televideo (RT) administration of a representative battery of common neurocognitive assessment instruments with traditional, face-to-face (FF) administration of the same battery on paper.

Methods: The neurocognitive battery was administered to 25 healthy subjects with no history of psychiatric diagnoses, 5 patients with schizophrenia, and 10 patients with mild cognitive impairment, via both methods ~14 days apart with the order of administration counterbalanced across participants.

Results: Intraclass Correlation Coefficient (ICC) comparisons of concurrent validity between RT and FF batteries yielded highly significant measures of agreement for all tests and only one significant mean difference was found between the methods using paired-samples t-test comparisons. The ICCs for test-retest reliability were also highly significant for all tests compared.

Conclusions: While further validation within the specific patient populations for which application is intended is still on-going, the current, preliminary results support the psychometric feasibility of administering a computerized battery of commonly applied neurocognitive assessment instruments via remote, televideo interactions with an expert examiner at a distal location.

References: O'Halloran JP, Kemp AS, Gooch K, Harvey P, Palmer B, Reist C, Schneider L (2008). Psychometric Comparison of Computerized and Standard Administration of the Neurocognitive Assessment Instruments Selected by the CATIE and MATRICS Consortia among Patients with Schizophrenia. Schizophrenia Research; 106(1):33-41.

Kemp AS, Mohs R, Salmon D, Tariot P, Ismail MS, Sano M, Schneider L, O'Halloran JP (2008). Psychometric Comparison of Computerized and Standard Administration of the Alzheimer's Disease Assessment Scale (ADAS) among Patients with Mild to Moderate Alzheimer's Disease. International Society for CNS Clinical Trials and Methodologies (ISCTM) 2008 Autumn Conference; Toronto, ON; October 6-8, 2008. Recipient of the "Distinguished Poster Award".

Source of Funding: NIH Small Business Innovation Research (SBIR) Program (R43AG032629-01

26 A Comparison of the Conners' Continuous Perfomance Test to the CDR Computerised Cognitive Assessment System

Keith A. Wesnes,[1] Helen Brooker,[1] Jeffrey D. Baker,[2] Robert A. Lenz,[2]

[1]United BioSource Corporation, Goring on Thames, RG8 9RD, UK, [2]Abbott Laboratories, Abbott Park, IL 60064-6084, USA

Introduction: The Conners' Continuous Performance Test (CPT) and the CDR Computerised Cognitive Assessment System have both been extensively validated and are both widely used in worldwide clinical trials. To date, no evaluation of the relative aspects of cognitive function assessed by the two systems has been performed.

Methods: Both the CPT and the CDR System were administered to healthy young volunteers in a Phase I clinical trial. In this paper the various measures from the CPT and the CDR were inter-correlated, and the various associations and dissociations were evaluated using correlation techniques across 14 Connor's parameters and more than 30 from the CDR System.

Results and Conclusions: The data provide good insight into the aspects of function assessed by each system, and confirm that both are useful tools for clinical research.

27 The Profile of Cognitive Impairment in MCI

Keith Wesnes[1]; Helen Brooker[1]; Paul Newhouse[2]; Edward Levin[3]; Heidi White[3]; Emily Coderre[2]; Heather Wilkins[2]; Ken Kellar[4]; Paul Aisen[4]

[1]United BioSource Corporation, Goring, UK, [2]Clinical Neuroscience Research Unit, University of Vermont College of Medicine, [3]Duke University Medical Center, [4]Georgetown University

Introduction: Two studies in which the CDR System was administered to MCI patients are reviewed to evaluate the profile of impairments in the patients on tests of attention, working and episodic memory.

Methods: The first study was run at the Hammersmith Memory Clinic where the CDR System was used as a routine part of the evaluations for over 10 years (Nicholl et al, 1995). In the second study (Newhouse et al, 2009), the CDR System was part of the evaluation profile in a randomized clinical trial of the effects of nicotine on cognitive function in MCI. 74 non-smoking subjects from 3 US sites who met criteria for amnestic MCI were randomized to receive either double-blind transdermal nicotine (NIC) or placebo patch for the first 6 months of the study

Results: In the memory clinic, compared to 'worried but well' patients, attention was impaired in MCI patients, but accuracy scores on a working memory task as well as verbal and pictorial episodic recognition tasks were not notably poorer than the controls. However, the speed of responses on the working and episodic memory tasks was slowed to the degree of impairment seen in demented patients.

In the nicotine trial, the baseline CDR System data were compared to the CDR System database for age and gender matched controls. The profile of impairment showed a remarkable similarity to the 1995 paper, attention being impaired, as well as the speed of recognition in working and episodic memory tasks. Although there was a small and significant decline on the ability to recognise pictures, there was no decline on word recognition or working memory supporting the profile seen in the 1995 study. Word recall tests were used in the present study but not the memory clinic, and delayed word recall was not surprisingly found to be at near floor levels, though immediate word recall was not notably impaired. Over the six months of the study, nicotine produced significant ($p<.05$) improvements in delayed word recall accuracy, speed of memory and choice reaction time accuracy.

Conclusions: The cognitive profile of amnestic MCI, besides being characterized by markedly poor delayed verbal recall, is accompanied to disruptions to attention and the time taken to retrieve information from working memory. The nicotine study showed that these impairments can respond to therapeutic intervention. However, there is growing interest in charting the progress of cognitive decline in MCI until early signs of dementia appear, but delayed recall measures which are already at floor levels will not be useful in this respect. However, many other aspects of function are not at floor levels and could be useful outcome measures for longer term intervention strategies. In fact the placebo subjects in the nicotine study showed significant declines to attention over the six months of the study, Nicholl CG, Lynch S, Kelly CA, White L, Simpson L, Simpson PM, Wesnes K, Pitt BMN (1995). The Cognitive Drug Research computerised assessment system in the evaluation of early dementia - is speed of the essence? International Journal of Geriatric Psychiatry, 10, 199-206.

Newhouse P, Levin E, White H, Coderre E, Wilkins H, Kellar K, Aisen P, Wesnes K (2009) Transdermal Nicotine Treatment of Mild Cognitive Impairment (MCI): A Multi-Center Pilot Study. American Association for Geriatric Psychiatry meeting, March 2-4, 2009, Honolulu Hawaii

28 **Placebo Arm and Test-Retest Data for the MATRICS Consensus Cognitive Battery in a 20-Site Schizophrenia Clinical Trial of R3487/MEM3454**

R. Keefe[1]; C. Siu[2]; K. Fox[3]; D. A. Lowe[4]; G.Garibaldi[5] L. Santarelli[5] D. Umbricht[5] S. Murray[6]

[1]Psychiatry, Duke University Medical Center, Durham, NC, USA; [2]Data Power (DP), Inc., Ringoes, NJ, USA; [3]NeuroCog Trials, Inc., Durham, NC, USA; [4]Psychogenics. Inc., Tarrytown, NY, USA; [5]Roche Pharmaceuticals, Nutley, NJ, USA; [6]Omeros Corporation, Seattle, WA, USA.

Introduction: 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. It has been an empirical question whether the MCCB would demonstrate these favorable characteristics when administered in the context of the type of large multi-site industry trial for which it was designed. Further, very few data have been presented on the characteristics of the MCCB in the placebo arm of a cognitive enhancement study.

Methods: Patients with schizophrenia maintained on a stable dose of a second generation antipsychotic therapy were enrolled into a randomized, double-blind, placebo controlled trial of R3487/MEM3454. Testers from 20 sites were trained and certified, and all MCCB data were reviewed and re-scored centrally. Cognitive functioning was assessed at screening, baseline and weeks 4, 8 and 10 with the MCCB. The interval between screening and baseline was 1-2 weeks. The MCCB generates a composite score and cognitive domain scores standardized to a normative population with mean (T) = 50 and SD = 10. Functional capacity was measured with the UCSD Performance-based Skills Assessment, 2nd edition (UPSA-2).

Results: The MCCB composite scores had excellent test-retest reliability (ICC = 0.88) and sensitivity to impairment (T-score = 25.1+/-11.6 at screening; 27.6+/-SD=12.1 at baseline). The Pearson correlation with the UPSA-2 scores was 0.56 ($P<.001$). In the placebo group, the effect size of the improvement on the MCCB composite score from baseline to week 8, encompassing three assessments, was small ($d=0.2$). The reliability between assessments in the placebo group was high (all ICC's >.90). The MCCB had a very low rate of missing data with none of the 215 patients missing a composite score. Conclusion:In the context of a 20-site clinical trial in stable patients with schizophrenia, the MCCB is sensitive to cognitive deficits in all domains, demonstrates excellent test-retest reliability and construct validity, and small practice and placebo effects.

29 **Computerized Cognitive Testing in Schizophrenia: Patient Experiences.**

Smita Pandey Bhat[1] PhD, Geetika Nath[1], Amit Sharma[1], Susan De Santi[2] PhD

[1]The Cognition Group, Newark Delaware, U.S.A; [2]NYU School of Medicine, Centre for Brain Health, New York, USA.

Introduction: Cognitive deficit is a core characteristic of schizophrenia. Recent research reports the use of Computerized Cognitive batteries with Schizophrenia patients. However, the ability of Schizophrenia patients to use these tests has not yet been formally examined.

Aim: We examined the ability of Schizophrenia patients to perform computerized cognitive testing. We measured the percent of schizophrenia patients failing to complete the battery, specific tests that were difficult to perform and reasons for non-completion of tests.

Method: 469 schizophrenia patients participated in a large multicenter, multinational clinical trial, conducted for twelve months. Seven tests from the Cogtest battery were administered including Strategic Target Detection Tests (STDT), Continuous Performance Task – Identical Pairs (CPT-IP), Word List Memory, (WLM), Word List Memory Delayed (WLMD), Tapping Speed Test (TST), Penn's Emotion Acuity Test (PEAT) and Workstation Orientation (WO). Six assessments took place (Screening, baseline, Week 6, Month 3, Month 6 and Month 12). Tests not completed were coded electronically with explanations.

Results: On an average 12% failed to complete testing; 10% at screening, 10% at baseline, 12% at week- 6, 7% at month-3, 9% at month-6 and 26% at month-12.The major tests not completed were STDT (0.77%) followed by CPT-IP (0.73%), the WLMD (0.73%) and PEAT (0.50%). The important reasons for non-completion were refusal of the patients (3.36%), technical reasons (2.56%) and patients being too tired (0.87%).

Conclusion: 88% of Schizophrenia patients successfully completed the computerized cognitive testing (Cogtest), suggesting that this battery is an efficient measure of cognitive functioning in multicenter multinational clinical trials