A Review of 175 Trials Identifying Cognition Enhancement Using a Single Computerised System Designed Specifically for Clinical Neuroscience

Keith A. Wesnes, BSc, PhD, FSS, CPsychol, FBPsS
Practice Leader, Bracket Global, Goring-on-Thames, RG8 0EN, UK Visiting Professor, Division of Psychology, Northumbria University, Newcastle, UK Adjunct Professor, Centre for Human Psychopharmacology, Swinburne University, Melbourne, Australia

Methodological Question Being Addressed: Could a single test system be used in a wide variety of clinical indications and different study designs and yet successfully and consistently identify cognition enhancement?

Introduction: As human longevity steadily grows in more fortunate world regions; cognition enhancement has become a ‘hot topic’ and is currently arousing intense public and scientific debate. While definitions vary, cognition enhancement can be defined as improved ability to perform tasks involving mental ability; either by counteracting impairment, or by producing improvement above existing levels. The principal and arguably only direct and objective measure of cognitive ability involves the use of tasks that demand mental efficiency. This paper concerns a computerised test system designed by the author which had its roots in a PhD project at Reading University (UK) which began in 1972 to determine whether the brain reticular cholinergic systems were involved in the control of human attention. It rapidly became evident that to detect subtle cognitive improvements in healthy students, automated procedures that captured cognitive reaction times as well as accuracy were essential. The early laboratory computers of the 1970s offered the first solutions, while portable laboratory microcomputers allowed cognitive testing to migrate from the laboratory to diverse clinical settings and even patients home in the early 1980s.

Methods: The system which emerged from this early research and upon which this presentation is based has been used in over 1300 trials worldwide. The core tests cover attention, information processing, working memory, executive control, and various aspects of verbal and non-verbal episodic memory. The core tests have remained constant over the last 29 years and have been supplemented with others. The public domain studies in which the system has identified cognition enhancement are reviewed in this paper. Three major domains are evaluated (1) Attention/information processing, (2) Working memory/executive control, and (3) Episodic memory. For each study the Tables indicate whether or not the domain was assessed, and if so whether significant improvement was identified, or significant impairment, or no reliable change.

Results: 175 studies were identified published in 125 peer-reviewed journals, 7 chapters & 1 review; as well as 39 other studies published as abstracts from conference presentations and 3 unpublished conference presentations. The 5 peer-reviewed negative studies are additionally included for completeness, including one published by JAMA who considered it a well conducted negative trial. The trials which were conducted from 1975 to the present involve data from over 11,000 healthy volunteers aged 5 to 87 years & 26 different patient populations. The interventions include pharmaceuticals, nutraceuticals, natural products, everyday drugs including those of abuse, various surgical procedures and even classroom ventilation.

Conclusions: To the author’s knowledge, this is the largest database ever assembled of cognitive enhancement identified with a single test or test system. The answer to the question of whether a single system can be used in a diverse range of conditions to detect cognition enhancement appears to be a positive one. Further work will include the identification of the relative effect sizes of the various improvements in all of the studies.

Key Words: Cognition enhancement; attention; memory; computerised cognitive testing

Disclosures: The author is employed by Bracket which provides clinical trial services to the pharmaceutical and nutraceutical industries including the test system described in this paper.
2 An Instrument for Conducting Cognitive Testing Via the Internet
Keith A. Wesnes, BSc, PhD, FSS, CPsychol, FBPsS

Practice Leader, Bracket Global, Goring-on-Thames, RG8 0EN, UK Visiting Professor, Division of Psychology, Northumbria University, Newcastle, UK Adjunct Professor, Centre for Human Psychopharmacology, Swinburne University, Melbourne, Australia

Methodological Question Being Addressed: Can cognitive testing be conducted reliably over the internet?

Introduction: Late 2012 saw the start of a new frontier for clinical trials when the FDA approved a yearlong study in which all assessments are made remotely apart from 2 clinician visits. Such trials in which cognitive function was the endpoint would benefit from assessment of mental function which could be conducted via the internet from the home or workplace. The author has just published a remote trial on the effects of breakfast in schoolchildren conducted in 2004 (Wesnes et al. Breakfast is associated with enhanced cognitive function in schoolchildren: An internet based study. Appetite 2012 59: 646-9), although the same journal rejected the same manuscript 6 years earlier! The present study compares data from a large internet study of cognitive function with data from the same tests administered under laboratory conditions.

Methods: A website offered feedback on cognitive function. Individuals clicked on the link entered their age and gender, and could perform 4 tests lasting 10 minutes (a 3-minute vigilance task, simple & choice reaction time and picture recognition). Five language versions could be selected, English, Greek, Hungarian, Portuguese & Spanish.

Results: In the 18 months following August 2010, 120,171 individuals logged on and entered demographics, 111,203 completed the first task and 97,171 all tasks; this latter cohort then receiving graphical feedback on the degree to which they may have favourably exceeded their age-norms. The age range of participants was 4 to 105 yrs, the male:female ratio was 41:59, and 84% were aged 18 to 60. Within this cohort the numbers were fairly evenly distributed over 5 year age-bands. Compared to laboratory data the patterns over the age-range on all task measures were directly comparable, young children showing the poorest performance, subsequently peaking during the late teenage years and declining steadily thereafter. Gender differences were also consistent with laboratory data.

Conclusions: This study demonstrates large-scale remote cognitive testing to be feasible in virtually countless clinical trial applications, a few examples including post-marketing safety (or efficacy) evaluations of novel medicines, remote studies of nutritional products, long-term follow-up studies in childhood cancer survivor cohorts, and the long duration trials which are now starting in the new indication of preclinical Alzheimer’s disease.

Key Words: Remote clinical trials; Internet based cognitive testing; attention; memory

Disclosures: The author is employed by Bracket which provides clinical trial services to the pharmaceutical and nutraceutical industries including the test system described in this paper.

3 A Rasch Model Analysis to Assess Cross-Cultural Differences in Negative Symptoms in Schizophrenia
Anzalee Khan, PhD 1,2, Mark Opler, MPH, PhD 2,3, Stacy Liechti, PhD 4, Christian Yavorsky, PhD 5, Luka Lucic, PhD 2,6, Brian Rothman, PhD 5, Sofija Jovic, PhD 2,7

1 Nathan S. Kline Institute for Psychiatric Research, 2 ProPhase LLC, 3 New York University School of Medicine 4 The PANSS Institute, 5 CROnos CCS, 6 Pratt Institute, 7 Massachusetts Institute of Technology

Methodological Question Being Addressed: Treatment of negative symptoms has been identified as an unmet need for individuals with schizophrenia. Targeting negative symptoms in treatment may have significant functional and ecological benefits. Difficulties with signal detection with negative symptoms may be related to measurement, the specific symptom being examined, or to the interpretation of these symptoms across cultures. The objective of this study was to examine the cross-cultural differences of the Negative symptom domain (Marder et al., 1997) of the Positive and Negative Syndrome Scale (PANSS) across eight selected geographical regions using Rasch analysis. The specific aims are: (1) To examine measurement properties of negative symptoms, namely, dimensionality and response structure; (2) To examine how each of the domain items function across the eight geographic regions; and (3) To examine possibilities to enhance the measurement properties by providing enhanced training and guidance for raters on items deemed problematic for a particular geographical region.

Participants: The data was obtained for 6,360 patients from eight different geographical regions: USA and Canada (n = 3364); India (n = 314); Russia and Ukraine (n = 274); Northern Europe (n = 497); Southern European (n = 434); Spain and Italy (n = 245), South Africa (n = 311) and Northern Europe (Germany, Austria, Switzerland, Netherlands, France
and Great Britain) (n = 921).

Methods: We first performed a principle components analysis to assess unidimensionality of the Negative Marder Domain and then applied Rasch analysis to examine cross-cultural differences among each of the 6 items (Emotion Withdrawal, Lack of Spontaneity, Poor Rapport, Blunted Affect, Active Social Avoidance, and Motor Retardation). A Rasch rating scale model was used to identify invariance of item calibrations for the 8 geographic regions.

Results: Lower item calibration reflects items easy to endorse, in which raters offered less difficulty scoring. Alternatively, higher item calibration reflects items more difficulty scoring. The most difficult item to score for all regions is Lack of Spontaneity with India showing the most difficulty $\Delta = 0.71$, and Northern Europe and the United States of America showing the least difficulty $\Delta = 0.26$, each. The second most difficult item for raters to score was Blunted Affect for most countries including Northern Europe ($\Delta = 0.29$), India ($\Delta = 0.27$), and Russia and Ukraine ($\Delta = 0.24$). Russia and Ukraine, and Northern Europe had difficulties scoring Active Social Avoidance $\Delta = 0.19$. Raters from South Africa had difficulties scoring item Motor Retardation, $\Delta = 0.29$.

Conclusions: This study looks at a cross-cultural comparison of the psychometric performance, mean scale scores, and item and scale-summary for the Negative Symptom domain items, using a representative population. There were significant differences in response to Negative symptom items, possibly caused by a lack of equivalence between the original English and translated versions, cultural differences among interpretations of items, culturally-specific symptom presentations, or scoring parameters. Knowing which items are problematic for various cultures can help guide training for specific geographical regions, which may optimize signal detection in multicenter clinical trials.

4 Validation and normalization of the Russian Version of the Positive and Negative Syndrome Scale (PANSS-Ru) in Schizophrenia: Preliminary Findings

Evgenia Ivanova, PhD 1, Alexander Reznik, PhD, MD 2, Mark Opler, MPH, PhD 1,3, Christian Yavorsky, PhD 1,4, Anzalee Khan, PhD 1,5,6, Sofija Jovic, PhD 1,8, Brian Rothman, PhD 1, Luka Lucic, PhD 1,9

1 ProPhase LLC, New York, NY; Healthy Russia Inc, Moscow, 2 Moscow Regional Psychiatric Hospital, Moscow, 3 New York University School of Medicine, New York, NY, 4 CROnos CCS, Hamilton, NJ, 5 Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY, 6 Manhattan Psychiatric Center, Wards Island, NY 7 Manhattan Psychiatric Center, Wards Island, NY, 8 Massachusetts Institute of Technology, Boston, MA, 9 Pratt Institute, Brooklyn, NY

Methodological Question Being Addressed: The Positive and Negative Syndrome Scale (PANSS) assesses treatment response or clinical severity in schizophrenia and psychotic disorders. It has become the ‘gold standard’ for clinical trials for the assessment of psychopathology. Psychopharmacology is facing a major challenge in the signal to background ratio for the accurate detection of change in antipsychotic trials, while evidence based practice in psychiatry is in need of objective benchmarks. In order to determine the status of a patient or group of patients, comparisons with the general population are necessary. Even if several groups of patients are to be compared, gender, age, co-morbid diagnoses, and other factors affecting clinical status are not identically distributed. The methodological question is whether a Russian translation of the PANSS and normed-reference data from a representative Russian population can be psychometrically sound and allow for clinical and research comparisons?

Methods: This study is designed in two parts: Phase 1: 40 in- and out-patients with schizophrenia and other psychotic disorders. The purpose of Phase 1 is to establish the initial psychometric properties of the PANSS-Ru by demonstrating the reliability and validity of the translated version. Phase 2 seeks to establish normed-reference data for the PANSS-Ru by administering the scale to a representative Russian sample (including individuals with schizophrenia and psychotic disorders and those without (i.e., controls)). 375 individuals (n = 250 in/outpatients with schizophrenia and other psychotic disorders; n = 125 controls). The participants' responses were statistically analyzed to verify the adapted instrument's internal consistency, stability reliability, discriminative validity, and construct validity. We calculated fifth percentile norms and presented them as step functions. Data were compared to North American norms.

Results: From communities across Russia, we selected a proportionate random sample of residents' age 18 to 71 years, to evaluate their health status with respect to schizophrenia. To date, a total of 40 eligible respondents have been enrolled. Results substantiated the psychometric qualities of the adapted instrument. An alpha coefficient of .88 verified the adapted instrument's internal consistency. Results from test-retest comparisons verified that the instrument's time stability (range = 0.67 to 0.92). The subscale scores of PANSS were normally distributed. Correlation between the subscale scores and the total score ranged from 0.76 to 0.86. Correlation with corresponding subscales reported by Kay et al., was above 0.83. Internal consistencies met the minimum criteria ($\alpha > 0.745$). Mean subscale and total score were equivalent to United States general population norms within 13%. However, there was a difference of more than five norm-based scoring points for mean general psychopathology scores. In screening for schizophrenia, the norms had a sensitivity of
85% and a specificity of 89%.

**Conclusions:** The PANSS scores were normally distributed, resulting in compatible percentile values reported by the original author. Preliminary results of the study show that different dimensions of symptom presentations in the Russian population may help to improve symptom-specific treatments, and will also provide comparison data for a Russian population. The instrument's discriminative power and factor analysis to verify construct validity are underway.

### 5 How Do North American Sites Compare with Rest of World in PANSS Interview and Ratings Quality?

David G. Daniel, MD, Alan Kott, MD

*Bracket Global, LLC*

**Methodological Question Being Addressed:** Is the quality of psychopathology assessment in schizophrenia clinical trials in North America similar or different than in the rest of the world?

**Introduction:** As drug-placebo differences have diminished over time clinical trials sites in North America have come under increasing scrutiny (1,2). We compare the quality of interview and ratings procedures delivered by North American (NA) vs. rest of world (ROW) sites using pooled results of audio/video surveillance measures from ten international schizophrenia clinical trials.

**Method:** A proprietary video/audio recording system was utilized to record PANSS rating procedures for review by calibrated external reviewers. Instruments used to assess ratings and interview quality included the Research Interview Assessment Scale (RISA) and the Rater Quality Questionnaire (RQQ). The RISA is a four domain, 16 item scale with higher scores reflecting better quality (3). The RQQ is a 2 item global measure addressing: 1) the quality of patient and/or informant interview data; and 2) proper application of the rules and anchor points of the rating scale or structured interview. Each RQQ domain is evaluated on a Likert-like scale ranging from one to three with lower scores representing higher quality (4).

**Results:** Mean total RISA scores NA (27.9 +/- 3.00) were modestly lower than ROW (28.3 +/- 2.54) (t(1899)=2.94, n=1901, p<0.01). However, NA raters scored lower (indicating higher quality) than ROW on the RQQ global interview quality axis (1.27 +/- 0.49 vs. 1.34 +/-0.50 t(1614)=2.54, (n=1616), p<0.05. There were no significant differences between NA and ROW scores on the RQQ ratings quality axis. (1.27 +/- 0.47 vs. 1.30 +/-0.49 t(1614)=1.26, (n=1616)p=NS.

**Discussion:** North American clinical trialists scored as well or better than their ROW counterparts on the RQQ, which evaluates the quality of ratings information collected by interview and adherence to rating scale rules. On the RISA, which evaluates a broad range of interview behaviors, including promotion of placebo response, NA raters scored modestly worse than their ROW counterparts. This report is preliminary as data continues to be collected in ongoing clinical trials.

**References:**


### 6 Could a Simplified PANSS be More Efficient?

Gary Sachs, Jean Dries, David Daniel

*Bracket Global*

**Methodological Question Being Addressed:** Is it possible to simplify the PANSS and reduce respondent burden in schizophrenia studies? As an initial exploratory step in a process aiming to improve the efficiency of the PANSS, we examined the frequency of item ratings beneath the level of clinically significant severity at baseline.

**Introduction:** In clinical trials for mood disorders, multi-item outcome measures with more items often perform worse
than the same or similar scales with fewer items. Hence, despite the availability 21, 25, 28, and 31 item versions of the Hamilton Depression Rating Scale, most RCTs use the 17 item version or the 10 item Montgomery Asberg Depression Rating Scale.

The PANSS is the most widely used scale in global RCTs investigating treatments for schizophrenia. This 30 item instrument assesses many complex domains, which makes it challenging as well as costly when implemented in global clinical trials. Certification as a PANSS rater requires considerable training and is typically administered at most or all study visits. Each administration requires about 45 minutes of interaction between subject and site staff. A shorter version of the PANSS could lessen respondent burden as well as the challenges associated with translating scales, and training raters in global studies.

Methods: We reviewed the Bracket clinical trial database and selected the two most recent schizophrenia trials for which baseline data was available. Scores for each of the PANSS items was classified as “Below clinical significance” or “Clinically significant”. “Below clinical significance” was operationally defined as item severity ratings of “1”(Absent) or “2”( Minimal - Questionable pathology; may be at the upper extreme of normal limits). “Clinically significant” was operationally defined as the item severity ratings of 3 (mild) or higher. The frequency of scores “Below clinical significance” was computed for each study. For each study, a descriptive analysis was conducted to determine the frequency that PANSS items were rated “Below clinical significance” at baseline.

Results: Two multicentre acute schizophrenia trials were identified. Study 1 included 463 subjects and was conducted at 71 sites in 12 countries. Study 2 included 815 subjects and was conducted at 110 sites in 16 countries.

<table>
<thead>
<tr>
<th></th>
<th>Total Mean (s.d)</th>
<th>P Scale Mean (s.d)</th>
<th>N Scale Mean (s.d)</th>
<th>G Scale Mean (s.d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>463</td>
<td>97.5 (10.2)</td>
<td>25.5 (3.8)</td>
<td>24.9 (4.6)</td>
</tr>
<tr>
<td>Study 2</td>
<td>815</td>
<td>95.1 (12.8)</td>
<td>25.3 (4.4)</td>
<td>23.3 (5.1)</td>
</tr>
</tbody>
</table>

In both studies, the ten items most frequently rated “Below clinical significance” were: G14 Poor Impulse control, P7 Hostility, G5 Mannerisms and Posturing, G8 Uncooperativeness, G10 Disorientation, G6 Depression, G7 Motor Retardation, P5 Grandiosity, G1 Somatic Concern, and G3 Guilt Feelings.

Conclusions: Items frequently rated absent at baseline may be a source of noise in RCTs. We identified 10 PANSS items that were often absent at baseline in two studies. An abbreviated PANSS might eliminate such items without loss of sensitivity. Further studies are required to determine the impact of a shorter simpler PANSS.

Disclosures: The authors are fulltime employees of Bracket.

7 Review of Methods Used to Adapt Neurocognitive Assessments in Other Languages and Cultures

Christelle Giroudet, Caroline Anfray, Dorothee S. Fuller

MAPI Institute, Lyon, France

Methodological Question Being Addressed: This abstract intends to address the methodological question of how to adapt neurocognitive assessments in other languages and cultures.

Introduction: Neurocognitive assessments explore cognitive functioning which typically include orientation, new-learning/memory, intelligence, language, visuoperception, and executive function. They are usually composed of three elements: the stimulus (e.g., images, digits, letters, words, story, objects from daily life), the instructions to the patients (read by the rater), and the instructions to the rater on how to administer and score the test. The objective of this study is to review the methodologies used to adapt neurocognitive assessments in other languages and cultures.

Methods: A review of projects involving the translation of neurocognitive assessments was conducted focusing on the methodologies used to cross-culturally adapt the instructions and the stimulus.

Results: Sixteen projects were reviewed. The process used to adapt the instructions involved the following steps: definition of concepts, two forward translations (from source to target language), reconciliation of forward translations, a backward translation (from target to source language), resolution of the back and forward translations, a clinician’s review, and proofreading. The process used to adapt the stimulus was similar to the process used for the instructions with the addition of one step: involvement of other experts to help establish and prepare stimuli (e.g., speech therapists and/or neuropsychologists) prior to the clinician’s review (i.e., at the time of the conceptual definition). Examples of challenges
encountered during the translation process of neurocognitive assessments, such as the National Institutes of Health Stroke Scale (NIHSS) List of Words/Sentences, the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), or the Rey Auditory Verbal Learning Test (RAVLT) will be presented.

**Conclusion:** The translation of neurocognitive assessments follows a methodology similar to the one used for patient-reported outcome measures which ensures that content validity is adequately similar between the translations and the original. The involvement of experts in the field, such as speech therapists and/or neuropsychologists, early in the process is paramount to guarantee the correct adaptation of the stimulus to the target languages and cultures.

8 **Translation of the National Institutes of Health Stroke Scale (NIHSS) List of Sentences: A Challenging Task**

Boucher F\(^1\), Cooke C\(^2\), Lyden P\(^3\), Acquadro C\(^1\), Fuller, DS\(^1\)

\(^1\)Mapi Institute, Lyon, France; \(^2\)Pfizer, Sandwich, UK; \(^3\)Department of Neurology, Cedars-Sinai Medical Center Los Angeles, USA

**Methodological Question Being Addressed:** This abstract intends to address the methodological question of translating the list of sentences of the National Institutes of Health Stroke Scale (NIHSS). The main difficulty was in developing the best methodology to create translations covering the same level of concretion/abstraction as the original US.

**Introduction:** The NIHSS was developed in the US to assess stroke severity across 11 categories. Aphasia is evaluated with a battery of items, including a list of five sentences with different levels of abstraction: “You know how”; “Down to earth”; “I got home from work”; “Near the table in the dining room”; “They heard him speak on the radio last night”. The objective of our study was to present the methodology used to translate this list into Canadian French, Bulgarian, Korean and US Spanish, and the resulting outcomes. The aim of the translation was not to find conceptual equivalents but sentences with the same level of concretion/abstraction and of similar length as the original.

**Methods:** The traditional translation process had to be adapted. In each country, instead of a classical linguistic validation (i.e., forward/backward translation, clinician review and/or cognitive interviews), a thorough forward translation was performed with a neurologist and a speech therapist, native speakers of the target languages. Candidate sentences were then reviewed with an expert panel to assure fulfillment of the original test of expression.

**Results:** All sentences had challenges to consider. Two sentences proved especially difficult to translate. For the expression "down to earth" an expression using the same level of abstraction had to be found. After discussion with the speech therapist and neurologist, idiomatic expressions in all languages were found: “golden hands” in Bulgarian, “love at first sight” in Canadian French, "the worst wheel of a cart always creeks most" in Korean and "she is in the clouds" in US Spanish. The other difficult sentence, particularly in Bulgarian, was “they heard him speak on the radio last night.” The preposition used (i.e., on) should have a meaning on its own and a figurative meaning when used in the entire expression. In Bulgarian, this was not the case. Another idiomatic expression was found that rendered this confusion conveyed by the preposition on: "they heard him speak before the television last night."

**Conclusion**. This study showed that finding sentences equivalent to the level of abstraction of the original ones used to assess aphasia is challenging and requires the collaboration of specialists and the developer in each target language.

9 **Translation of the Columbia Suicide Severity Rating Scale (C-SSRS) for Use in 33 Countries**

Géraldine Grataloup\(^1\), Nathalie Fernandez\(^1\), Dorothee S. Fuller \(^1\), Kelly Posner \(^2\)

\(^1\)MAPI Institute, Lyon, France; \(^2\)New York State Psychiatric Institute, New York, NY, USA

**Methodological Question Being Addressed:** This abstract intends to address the methodological question of how to adapt the Columbia Suicide Severity Rating Scale (C-SSRS) in other languages and cultures.

**Introduction:** To help clinicians determine the presence of suicidal ideation and behaviour, the C-SSRS was developed in US English and contains clear definitions and suggested probes to assess severity and intensity of suicidal ideation, the different types of suicidal behaviours and the lethality of suicide attempts. Prior to use in an international study to investigate suicidal ideation and behaviour, the clinician-rated C-SSRS had to be translated into 45 languages for 33 countries. A rigorous methodology was required to ensure conceptual equivalence and cultural relevance across languages.

**Methods:** The process was conducted by specialists in each target country, following a standardized methodology: 1) two forward translations by native target language speakers; 2) comparison and reconciliation of the translations; 3) back translation by a native English speaker; 4) comparison of original and back translation; and 5) review by a clinician.
**Results:** Cultural and linguistic challenges emerged during the process. On the cultural level, the differences in the approach to suicide and its methods based on differences in tradition and availability of means required finding suitable alternatives in the target languages. For instance in Taiwan; the original sentence "If a person pulls trigger while gun is in mouth but gun is broken no injury results, this is considered an attempt." was replaced by the following alternative: "If a person tried to hang himself/herself but the rope broke no injury results, this is considered an attempt." On the linguistic level, it was important to differentiate between medical and psychiatric hospitalisation after a suicide attempt and appropriate solutions across languages had to be found. Examples of challenges and their solutions will be discussed in the poster.

**Conclusion:** The 45 language versions, of the C-SSRS (a total of over 100 translations now exist), were established according to a rigorous methodology to ensure conceptual equivalence and cultural relevance across languages. The translations may now be used in international studies to assess suicidal ideation and behaviour and facilitate the comparison and pooling of data. The analysis of the psychometric results will be necessary to see if and how suicidal ideation and behaviour compare across countries and cultures.

10  **The Comparison of Rater Performance With and Without a Data-monitoring Program: What is the Impact of In-study Remediation on the Quality of Montgomery-Åsberg Depression Rating Scale Scores?**

Yavorsky C1, DiClemente G1, Wolanski K1, Brown B2

*Cronos CCS, Kashima Consulting*

**Methodological Question Being Addressed:** How do we maintain rater reliability for the duration of a clinical trial given the nature of the measurement tools and is data-monitoring an adequate strategy to mitigate risk to reliability and thus sample size and power calculations?

**Introduction:** Despite the emphasis on rater training at the outset of CNS clinical trials and in many cases re-current training, there are still significant concerns about the reliability of in-study data. The well-documented phenomenon of rater drift suggests that degradation of training impact over time. This drift can cause problems for both reliability and validity in clinical trials. When multiple raters are used in a clinical trial, differences between raters in terms of interviewing technique and scoring criteria introduce variability that can distort the outcome measures. Idiosyncratic rating practice can emerge, and if not dealt with, can introduce systematic error into study results. Maintenance of high reliability of assessment during a clinical trial would greatly enhance the sensitivity of the data quality and the treatment effects. 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:** 22 raters, trained on the use of the SIGMA through typical IM-based training (didactic training followed by scoring a patient video), assessed the same standardized subject, using the Montgomery-Åsberg Depression Rating Scale (MADRS), over five unique visits. All interviews were conducted independently, and the standardized subject adhered to a script for each visit. Each standardized visit was associated with pre-determined gold-standard scores. Raters were randomly assigned to one of two groups: active data monitoring (data monitoring with intervention where required; receipt of feedback based on flags at all applicable visits) and inactive data monitoring (no feedback provided). Raters were aware of the study design and purpose of the study and to which intervention group they were assigned to. No IRB approval was obtained because this was an internal, ad-hoc, training exercise and covered under 45 CFR 46.101b as an exempted project. Raters were additionally classified as clinically-naïve or clinically-trained based on their background.

**Results:** All participants completed the study and there were no blank data fields. There were no significant differences between individuals with rating experience and no rating experience even when years of total education were controlled for in terms of agreement with gold standard at the baseline visit. Differences emerged as a function of time with the un-remediated group showing progressively poorer absolute correlation with gold standard scores. There were no significant changes in RAPS scores in the un-remediated group while the group that did receive remediation improved by 34% (range -5% to 58%).

**Conclusions:** Active data-monitoring provides clear benefit in rating technique as indicated by the Rater Applied Performance Scale. This study has a small sample size so generalizability is limited though trends emerged that suggested that the no-intervention group did not perform as well as the active data monitoring group. Although small scale, the results here are generally confirmatory of those found in large clinical trials where experience has not been controlled for.
Cognitive Remediation Effects: Meta-analysis and New Data from Psychiatric Inpatients

Anthony O. Ahmed, Nancy J. Doane, Barbara Jenkins, Daniel J. Fridberg

Department of Psychiatry and Health Behavior, Georgia Health Sciences University

Methodological Question Being Addressed: Does the extensive use of exclusion criteria limit the generalizability of cognitive remediation studies conducted in laboratory settings and outpatient/day programs to inpatient settings?

Introduction: Studies suggest that cognitive remediation reduces neurocognitive and concomitant functional deficits in people with schizophrenia. There are however few evaluations of CRT benefits in psychiatric inpatient and forensic settings. We evaluated the generalizability of studies conducted in laboratories and outpatient/day programs by examining whether exclusion criteria inflated effect sizes by decreasing sample heterogeneity. We also evaluated the benefit of cognitive remediation for state hospital patients.

Methods: Study 1: We identified cognitive remediation trials published between 1990 and 2013 by searching MEDLINE, Science Direct, PsychINFO, and EMBASE. We also searched the reference list of identified studies and previous meta-analytic reviews. The search produced 74 articles for inclusion. For each study, we computed mean separations, pooled SDs, and effect sizes on global cognition, symptoms, and functioning and classified and counted the number of exclusion criteria by studies.

Study 2: Forty-four patients with schizophrenia and schizoaffective disorder most of whom had been adjudged as “Incompetent to Stand Trial (IST)” or “Not Guilty by Reason of Insanity (NGRI)” participated in the study. Patients were randomized to a 24-week cognitive remediation group (CRT) versus a wait list-control (WLC) condition. The CRT group completed three 50-minute sessions weekly that comprised 40 minutes of computer activity and a 10-minute bridging group. Participants completed the MATRICS Consensus Cognitive Battery; Positive and Negative Syndrome Scale; UCSD Performance-Based Skills Assessment; Social Adaptive Functioning Evaluation; Readiness for Discharge Questionnaire; and Maryland Assessment of Recovery in Serious Mental Illness.

Results: Study 1: Effect sizes for cognitive remediation trials ranged from minimal to moderate with an average effect size of .48 for global cognition. There were no effect size differences between studies that included inpatient samples and those that did not across outcome measures. Although baseline pooled SD accounted for study effect size as much as actual treatment-versus-control group mean separation; the number of study inclusion/exclusion criteria (range: 3-11) was uncorrelated with pooled SD.

Study 2: There were no significant differences between the CRT and WLC patients on any outcome measures at baseline. Forensic patients had greater deficits in working memory, verbal, visual learning, and global cognition relative to mental health patients. Post-treatment, the CRT group demonstrated greater improvements in working memory, processing speed, and overall cognition relative to the WLC group. The RM ANCOVA group-by-time interaction achieved statistical significance. The CRT group also demonstrated greater improvements in positive symptoms, functional capacity, but not functional performance or discharge readiness. The effect sizes were generally moderate to large.

Conclusions: The meta-analysis suggests that cognitive remediation produces small to moderate improvements in neurocognition, symptoms, and psychosocial functioning. The effect sizes obtained across studies are not an artifact of highly homogenous samples caused by the use of numerous exclusion criteria or study setting. The benefits of cognitive remediation for patients with schizophrenia receiving care in a forensic setting suggest that it may be a useful adjunct to treatment efforts at restoring functional and legal competence. It may also improve clinical symptoms by contributing to treatment engagement such as participation in psycho-pharmacological and other rehabilitative interventions.

Disclosures: The authors report no conflict of interest for this work

The Unreliability of Reliability Statistics: Calculating Interrater Reliability in CNS Clinical Trials

Popp D¹, Mallinckrodt CH², Williams JBW¹, ³, Detke MJ¹, ⁴

¹MedAvante, Inc., ²Eli Lilly Company, ³Department of Psychiatry, College of Physicians and Surgeons, Columbia University, ⁴Indiana University School of Medicine

Methodological Question Being Addressed: To define appropriate standards for estimating and reporting interrater reliability (IRR) in CNS clinical trials.

Introduction: Clinical trial reporting seldom includes estimates of interrater reliability and when reported, selection of
reliability statistics is often inconsistent or inappropriate. Power calculations rarely consider the imperfect reliability of subjective assessments in CNS trials. Inaccurate or inflated reliability estimates can impact study power, sample size and signal detection. A drop in reliability from .90 to .70 decreases study power from 76 percent to 65 percent. A 30 percent increase in sample size is required to maintain power with this decrease in reliability. Poor or inaccurate reliability can have significant consequences ranging from increased R&D costs to significant delays in getting effective drugs to patients who need them. Guidelines are proposed for selection of appropriate reliability measures for CNS clinical trials.

Methods: We evaluate the appropriateness of commonly selected reliability statistics and how they are used with various types of data and methodologies, illustrating frequently misused reliability statistics that demonstrate the impact of inappropriate analytic selection on estimates of reliability.

Results: In CNS clinical trial research, IRR can be measured for both diagnosis and outcome variables (e.g., severity scales). IRR is typically measured using one or both of the following methodologies: at an investigator meeting when a large group of raters independently score one or more subject videotapes prior to study start and/or during a trial via in-study surveillance when an expert clinician reviews and independently scores audio/video taped in-study assessments providing feedback to raters when indicated. We propose a statistical decision tree that can be used to determine the appropriate IRR measure for various methodologies based on the type of variable, the number of raters and the number of subjects or observations. For example, to estimate IRR for a continuous outcome measure such as the MADRS, using data commonly obtained during an investigator meeting where a large group of raters rate two or more videotaped subjects, an intra class correlation (ICC) should be used. While reliability statistics require observations of more than one subject, it is not always possible to obtain multiple observations. We present analytic strategies that are appropriate when only one subject is rated. Common misuses of reliability statistics include using Kappa for continuous variables and treating individual scale items as independent observations when data is available from only one subject. Comparisons of ICC to Kappa using a 5, 10, and 20 percent agreement criterion demonstrate that selecting a broader criterion range for agreement can artificially inflate IRR using the same data. Finally, results suggest that ICCs calculated from a single observation by treating items as subjects may be inversely related to scale reliability. Appropriate statistics for single observations can reveal IRR issues masked by this approach.

Discussion: Reliability can have significant impact on clinical trial outcomes. When selecting reliability statistics, researchers must take into account the type of variable (e.g., binary, nominal or interval), the number of raters, composition of the rater pool (i.e., same raters rate all subjects vs. raters selected from a larger pool) and the number of observations using the guidelines presented for various methodologies.

13 A Phase II Randomized, Double-Blind, Placebo-Controlled, Trial of GLYX-13 for the Rapid Treatment of Major Depressive Disorder using Central Ratings

Popp D1, Burch R2, Williams JBW1,3, Price LM1, and Detke MJ1,4

1MedAvante, Inc., 2Naurex, Inc., 3Department of Psychiatry, Columbia University, 4Department of Psychiatry, Indiana University

Methodological Question Being Addressed: To assess the ability of blinded, independent central raters via telephone to detect drug-placebo separation of a fast acting single intravenous dose antidepressant in subjects with inadequate response to previous treatment for MDD.

Introduction: NMDA receptor ligands have been shown to rapidly treat depression but are associated with psychotomimetic effects. GLYX-13 is an NMDA receptor glycine site functional partial agonist with ~ 25% of the agonist activity of glycine or D-serine. Animal models suggest a single intravenous dose may produce long-term efficacy without psychotomimetic effects. A phase II randomized, double-blind, placebo-controlled trial was conducted to assess the efficacy of GLYX-13 with central raters. The current study examines the effects of a single dose of GLYX-13 in subjects with inadequate response to previous treatment for MDD.

Methods: 48 male and 68 female subjects received a single dose of GLYX-13 (1/5-10/30-mg/kg) or placebo. Central raters assessed subjects via telephone using the HDRS-17 at Screening, Baseline, Days 1, 3, 7, 14, 21 and 28.

Results: The a priori primary efficacy ANCOVA on pooled drug dose versus placebo was not significant for change from baseline to Day 1 on HDRS-17 total score. MMRM revealed a statistically significant reduction in HDRS-17 total score versus placebo at Day 3 for 5-mg/kg (-4.4; p<.05) and a trend at Day 1 for 5-mg/kg (-3.5; p=.068) and at Day 7 for 5 and 10-mg/kg (-4.0 for both; p’s=.059 and .073). GLYX-13 did not cause psychotomimetic side effects at any dose studied.

Conclusion: This study suggests that GLYX-13, an NMDA receptor glycine site functional partial agonist, rapidly
A Randomized Trial Administering Aspirin, Minocycline or Pramipexole vs Placebo as Add-on to Antipsychotics in Patients with Schizophrenia or Schizoaffective Disorder

Mark Weiser1,2, Shimon Burstein1, Liliana Fodoreanu3, Roxana Chiriță4, Ghiorghe Talău5, Diana Cirjaliu6, John Davis7, Michael Davidson1,2

1 Department of Psychiatry, Sheba Medical Center, Tel Hashomer 52621, Israel, 2Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel, 3Spitalul Clinic Judetean Cluj. Victor Babes 43, Cluj, 4Spitalul Clinic de Psychiatrie Socola, Iași. Șos. Bucium 36, Iași, 5Spitalul Clinic de Psychiatrie Gh. Preda Sibiu. Bagdazar 12, Sibiu, 6Clinica de Psihiatrie Palazu Mare, Constanta. Tomis 145, Constanta, 7Department of Psychiatry, University of Illinois, 1601 W. Taylor Street, PI 324W, MC 912, Chicago, IL 60607

Introduction: This is a randomized placebo controlled trial (RCT) testing the inflammatory hypothesis and elaborating on the dopamine hypothesis in schizophrenia. Aspirin, an inhibitor of both COX-1 and COX-2, and minocycline a tetracycline with anti-inflammatory effects (both of which showed in smaller trials some benefit in schizophrenia) were selected to test the anti-inflammatory hypothesis. Pramipexole, a pre-synaptic dopamine auto-receptor agonist, was chosen to elaborate on the dopamine hypothesis.

Methods: This multi-center, N=400 trial was designed with one placebo arm to be employed as a comparator for 3 active arms. Inclusion criteria were 4 (moderate) or above on CGI-S and 3 (moderate) score on two of the following four PANSS items: delusions, hallucinatory behaviors, conceptual disorganization or suspiciousness/ persecution, and/or a total PANSS negative symptoms score above 18. Before entering the trial and throughout the trial all subjects received anti-psychotics at doses within PORT recommendations. Upon entering the trial they were randomized to aspirin 1000 mg/d + pantoprazole 40 mg/d, minocycline 200 mg/d, pramipexole 1.5 mg, or placebo. Duration of the study was 16 weeks. Primary outcome measure was changes in total PANSS scores, secondary outcome measures included PANSS subscales and CGI.

Results: Mean age of patients was 42, 50% were females, mean duration of illness was 13 years, mean PANSS total score at baseline was 92, mean CGI at baseline was 4.7. The ANOVA for overall change for all comparison of 3 drugs and placebo for the primary outcome of the total PANSS scores was significant, p=0.0343. Individual comparisons between each drug and placebo showed trends for significance (Effect size, ES=0.26, p= 0.0561) for aspirin, and were non-significant for minocycline (ES=-0.14, p=0.328) and for pramipexole (ES=0.01, p=0.952 ) For positive symptoms the overall ANOVA was not significant, p=0.0842. Individual comparisons between each drug and placebo showed a trend for significance for aspirin (ES=0.24, p= 0.079), and were non-significant for minocycline (ES=0.04, p=0.771) and pramipexole (ES=0.11, p=0.451). For negative symptoms the overall ANOVA was not significant, p=0.0995, as were individual comparisons between each drug and placebo: aspirin (ES=0.07, p= 0.586), minocycline (ES=-0.93, p=0.095) and pramipexole (ES=0.06, p=0.451). For general psychopathology the overall ANOVA was significant, p=0.0214. Individual comparisons between each drug and placebo were significant for aspirin (ES=0.31, p= 0.040), non-significant for minocycline (ES=-0.16, p=0.307) and pramipexole (ES=0.05, p=0.717). For CGI the overall ANOVA was significant, p=0. 0.0140. Individual comparisons between each drug and placebo were non-significant: aspirin (ES=0.23, p= 0.102), minocycline (ES=-0.24, p=0.107) and pramipexole (ES=0.04, p=0.799).

Discussion: This relatively large RCT was intended to provide a more definitive answer regarding the benefits of aspirin, minocycline and pramipexole reported in previous smaller trials. Although the overall ANOVA for PANSS total was significant, post-hoc analyses were only significant at trend level for aspirin, uncorrected for multiple comparisons, provided only a equivocal small advantage over placebo in total, positive and general psychopathology PANSS scores, and CGI. Based on these the failure to obtain the conventional .05 level, it is not possible to confirm the benefit of aspirin, nor to rule out the possibility that inhibition of COX-1 or COX-2, or other biological effects, both inflammatory and non-inflammatory) of aspirin are implicated in the symptomatology of schizophrenia. It is not unusual to have pilot trials done by enthusiasts not confirmed. Although consistently replicated, the effect of aspirin in schizophrenia is too elusive and small to be of clinical significance. On the other hand these findings call for a renewed basic science effort to investigate what in the biological activity of aspirin is related to amelioration of the symptoms of schizophrenia.

Disclosure: USA ClinicalTrials.gov identifier: NCT01320982, funded by the Stanley Medical Research Institute.
Long Acting Injectable vs. Oral Antipsychotics for Schizophrenia: A Meta-Analytic Examination of Study Design Impacting on Results

Taishiro Kishimoto1,2,3, Masahiro Nitta1, Michael Borenstein4, John M. Kane1,3,5,6, Christoph U. Correll1,3,5,6

1The Zucker Hillside Hospital, Psychiatry Research, North Shore - Long Island Jewish Health System, Glen Oaks, New York, USA, 2Keio University School of Medicine, Neuropsychiatry, Shinjuku-ku, Tokyo Japan, 3The Feinstein Institute for Medical Research, Manhasset, New York, USA, 4Biostat, Inc, New Jersey, USA, 5Hofstra North Shore LIJ School of Medicine, Hempstead, New York, USA, 6Albert Einstein College of Medicine, Bronx, New York, USA

Methodological Question Being Addressed: How much impact do the study designs have on the results of long acting injectable (LAI) vs. oral antipsychotics comparisons? How can we integrate the results of different study designs? What is the ideal design to understand the effectiveness of LAIs?

Introduction: High non-adherence rates in schizophrenia can limit the efficacy of pharmacotherapy, therefore, the use of long-acting injectable antipsychotics (LAI) is considered to be an important treatment option. However, new, large, randomized controlled trials (RCTs) showed no benefit of LAIs over oral antipsychotics (OAPs), and our recent meta-analysis on RCTs showed non-superiority of LAI over OAPs [1]. On the other hand, multiple non-RCTs such as mirror-image studies (where the data before and after initiation of LAIs are compared), and naturalistic cohort studies have shown the superiority of LAIs.

By conducting meta-analyses on RCTs and non-RCTs, we compare the effect sizes of LAIs in comparison to OAPs within and across each study design; i.e. RCTs, mirror-image and naturalistic cohort studies. We will then discuss the potential mediating/moderating factors influencing the differences in results between study designs, and make suggestions regarding the ideal design for LAI trials in the future.

Methods: Systematic review/meta-analysis was conducted on RCTs and non-RCTs lasting ≥6 months comparing LAIs and OAPs. RCTs, mirror-image studies (where the data before and after initiation of LAIs were compared), and naturalistic cohort studies were analyzed separately. Primary outcome was hospitalization risk (proportion of patients experiencing ≥1 hospitalization). When relapse, but not hospitalization was reported, we used relapse instead of hospitalization. Pooled relative risks (RR) together with their 95% confidence intervals (CIs) were calculated, using random-effects model.

Results: A total of 55 studies were included in the analysis (RCTs: 21 studies, n=5,176; mirror-image: 16 studies, n=4,066; naturalistic cohort: 19 studies, n=6,020). Across RCTs and naturalistic cohort studies, LAIs were not significantly superior to OAPs in preventing hospitalization (RR=0.917, 95%CI: 0.805-1.044, p=0.190; RR=0.915, 95%CI: 0.773-1.082, p=0.299 respectively). On the other hand, across mirror-image studies, LAIs showed strong superiority over OAPs (RR=0.428, 95%CI: 0.349-0.525, p<0.001). The difference between study groups were significant (Q-value=42.25, df=2, p<0.001). Although information was limited, there seemed to be no relevant differences in terms of treatment (e.g. medication, dosage), but some patient characteristics as well as procedural differences existed (e.g. age, informed consent, assessments).

Conclusion: The results of the three study designs were in strong contrast with each other. Substantial cohort differences may account for the different results in that 1) patients in RCTs are likely to be more adherent, 2) patients in non-RCTs are mostly LAI-targeted populations treated in clinical practice that are likely to be non-adherent, 3) patients in the oral arm in naturalistic studies are likely to be less severe than patients started on LAIs. Although mirror-image studies showed strong LAIs superiority, one needs to take into consideration the biases in mirror-image studies such as expectation bias and time effect. Future, simple RCTs may benefit from more closely replicating routine clinical circumstances. Two-way-mirror-image (oral to LAI, LAI to oral) might be another option.

Reference: [1] Kishimoto et al. Long-acting injectable vs. oral antipsychotics for relapse prevention in schizophrenia: A meta-analysis of randomized trials. Schizophrenia Bulletin (in press) Disclosures: Dr. Kishimoto has received consultant fees from Otsuka, Pfizer, Dainippon Sumitomo, and speaker’s honoraria from Banyu, Eli Lilly, Dainippon Sumitomo, Janssen, Novartis, Otsuka and Pfizer. He has received grant support from the Byoutaitaisyakenkyukai Fellowship (Fellowship of Astellas Foundation of Research on Metabolic Disorders) and Eli Lilly Fellowship for Clinical Psychopharmacology. Dr. Nitta is an employee of Dainippon-Sumitomo Pharma, Japan. Dr. Borenstein is founder of Biostat, Inc, New Jersey, USA. He has received grants from NIH and IES to develop software for meta analysis, and has a commercial interest in the software Comprehensive Meta-Analysis. Dr. Kane has been a consultant to Alkermes, Amgen, Astra-Zeneca, Janssen, Pfizer, Eli Lilly, Bristol-Myers Squibb, Dainippon Sumitomo/Sepracor/Sunovion, Johnson & Johnson, Otsuka, Pierre Fabre. Vanda, Proteus, Takeda, Targacept, IntraCellular Therapies, Merck,
Lundbeck, Novartis, Roche, Rules Based Medicine, Sunovion and has received honoraria for lectures from Otsuka, Eli Lilly, Esai, Boehringer-Ingelheim, Bristol-Myers Squibb, Merck and Janssen. He is a shareholder of MedAvante. He has received grant support from The National Institute of Mental Health. Dr. Correll has been a consultant and/or advisor to or has received honoraria from: Actelion, Alexza; American Academy of Child and Adolescent Psychiatry, AstraZeneca, Biotis, Bristol-Myers Squibb, Cephalon, Desitin, Eli Lilly, Gerson Lehrman Group, GSK, IntraCellular Therapies, Lundbeck, Medavante, Medscape, Merck, National Institute of Mental Health, Novartis, Ortho-McNeill/Janssen/J&J, Otsuka, Pfizer, ProPhase, Sunovion, Takeda and Teva. He has received grant support from BMS, Feinstein Institute for Medical Research, Janssen/J&J, National Institute of Mental Health (NIMH), National Alliance for Research in Schizophrenia and Depression (NARSAD), and Otsuka.

16 Neurocognitive Profile of Chronic Cannabis Users

Thompson, T\textsuperscript{1}, Boyd, AF\textsuperscript{2}, Morgan, DW\textsuperscript{2}, Gray, KM\textsuperscript{3}

\textsuperscript{1}UNC Eshelman School of Pharmacy, \textsuperscript{2}CNS Vital Signs, LLC, \textsuperscript{3}Medical University of South Carolina

Methodological Question Being Addressed: Does chronic cannabis smoking result in neurocognitive deficits?

Introduction: Prohibition of cannabis began in the late 1930s. Currently, 18 states have legalized medical cannabis; 12 of these have decriminalized-possession laws. In 2012, the states of Washington and Colorado passed legislation legalizing recreational use of cannabis. Federally, use, transportation, cultivation, and possession of cannabis remains illegal. While increased use may emerge, studies indicate that use of cannabis impairs performance and memory.\textsuperscript{2-4} Guidelines for use may need to be expanded to consider the deleterious effects of chronic cannabis use.

Methods: Neurocognitive data was pooled from three studies for subjects 18 to 23 years old at the time of the study. Per protocol, CNS Vital Signs\textsuperscript{8} a computerized neurocognitive battery,\textsuperscript{8} was administered to all test subjects. CNS Vital Signs produces age-matched standard scores across several neurocognitive domains as well as a neurocognitive index (NCI), which is a single-value summary score. “Normal” subjects will record a mean domain or NCI score of 100 (SD=15). Higher scores indicate superior neurocognitive performance. Per protocol, all subjects were administered a non-scored practice test at a screening visit, allowing the subjects to become familiar with the test before the testing at the subject’s randomization visit occurred. The number of patients and design of the three studies was as follows:

1. Thirteen patients from a randomized, double-blinded placebo-controlled study of subjects administered lorazepam as a 2 mg oral dose.\textsuperscript{1}
2. Thirteen patients from an open-label acute alcohol intoxication study of subjects with a blood alcohol content target of 0.08 to 0.12.\textsuperscript{6}
3. Eighty-six patients from a randomized eight-week, double-blinded placebo-controlled study of treatment-seeking cannabis-dependent participants.\textsuperscript{7}

Descriptive statistics of age-matched standard scores are presented here.

Results: Neurocognitive impairment was demonstrated in each study group. In the lorazepam group, subjects recorded a baseline neurocognitive index (NCI) recorded a score of 98. The NCI after lorazepam administration was 88 (SD=12), indicating a decrease from baseline of 10 points. Similarly, the acute alcohol intoxication group recorded a baseline NCI of 110. Intoxicated subjects recorded an NCI of 94 (SD=14.2), indicating a decline of 16 points from baseline. Cannabis-dependent subjects at randomization recorded a mean NCI of 86 (SD=34). Comparison to baseline for the cannabis-dependent subjects was not available as participants were chronic users.

Conclusions: Neurocognitive testing across three groups indicate that long-term cannabis may be cognitively detrimental.\textsuperscript{2-5,7} The degree of neurocognitive impairment possibly due to chronic cannabis smoking described in this data reveals cognitive impairments greater than the effects of acute lorazepam administration as well as acute alcohol intoxication.

17 fMRI Imaging Biomarkers for Predicting Treatment Response in Drug Trials

Ariana Anderson

Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, UCLA

The Methodological Question Being Addressed: Can fMRI biomarkers predict treatment outcome for addiction, and infer the interaction between cognitive activity and the medication? Which biomarkers are prospectively and retrospectively translatable, independent of the stimulus or experimental design used?
**Introduction:** We demonstrate using fMRI networks as biomarkers to predict post-treatment smoking levels in a drug trial of bupropion for nicotine addiction, proposing the specific brain networks with which bupropion interacts. Previous work in pharmaco-fMRI analyzes brain activity relative to a specific stimulus [1][2][3]. In contrast, we predict smoking outcomes using baseline fMRI networks as biomarkers, in a study comparing treatments of nicotine addiction (bupropion, counseling, or placebo). Instead of correlating activation with specific stimuli, we instead model the brain networks observed during baseline scan-time. This approach allows us to answer two fundamental questions: 1) Can baseline fMRI networks predict a patient’s treatment outcome with/without bupropion, and 2.) Which brain networks will interact with the pharmacological intervention?

**Methods:** 48 healthy patients who smoked ≥10 cigarettes/day and met the DSM-IV criterion for tobacco dependence were scanned up to 3 times pre-treatment using a video-craving stimulus[3]. These 143 total scans were summarized by the categorical brain networks found during scan-time, obtained by single-scan ICA results being dimension-reduced and spatially clustered into 25 total networks across a larger database of scans [4]. The ICA method is blind to the task being performed, seeking instead brain networks that are statistically (and spatially) independent. Each scan was summarized by its categorical membership; for example, Scan 1 for Subject 1 could contain brain networks from clusters 1,3,4,8,15,15,25. The total baseline occurrence of the 25 possible networks in the subject’s 3 scans is used to predict the post-treatment smoking level along with an indicator for treatment group (medication vs. placebo/counseling), using a Poisson link-function in a general linear model. A second model included a subset of treatment:network interaction effects. Each predictor’s contribution is evaluated using nested models with a chi-squared distribution.

**Results:** Our results show that dropping the main effect for the treatment does relatively little to change the model fit (p>.05), but the medication:network interaction effects are all highly significant. A total of 8 of the 25 networks, as main-effects, are significant in predicting treatment outcome at the p<.001 level. 6/10 possible network/treatment interaction effects are significant at the p<.0001 level.

**Conclusion:** This conceptually models what we know of addiction; it is a disorder that is reflected in neural activity. We can prospectively model who is likely to quit, and identify which networks predict smoking outcomes and which networks interact with bupropion. This suggests that bupropion does not treat addiction directly but instead modulates brain networks. These fMRI biomarkers are independent of the stimulus applied, which makes them directly translatable to studies with different stimuli and experimental design.

This work is not without limitations. Although this study is quite large for fMRI analysis (143 sessions, 48 patients), by clinical trial standards it is miniscule. Our results are with respect to final smoking levels and are therefore silent about brain networks altered by bupropion that are independent of smoking reduction, although we have already modeled this network in a separate analysis. Finally, our regression model doesn’t evaluate whether bupropion suppresses craving directly or instead inhibits the dopamine response to nicotine, which our future work will test using directed cyclic graphs.


18 APOE Genotype Modulates 1H-MRS Metabolites in the Aging Brain

Jesus J. Gomar, Marc L. Gordon, Peter B. Kingsley, Ulug M. Aziz, Jeremy Koppel, Concepcion Conejero-Goldberg, Peter Davies, Terry E. Goldberg

*The Litwin-Zucker Research Center, Feinstein Institute for Medical Research, Manhasset, NY, USA*

**Methodological Question Being Addressed:** Have healthy older subject’s carriers of the APOE E4 allele higher levels of choline (Cho) and myoinositol (mI) and lower levels of N-acetylaspartate (NAA)? Is there a mediator effect through 1H-MRS metabolites on cognitive performance in older subject’s carriers of the APOE E4 allele?

**Introduction:** Alzheimer’s disease (AD) and Mild Cognitive Impairment (MCI) patients show evidence of reduction of
NAA/Creatine (Cr) ratio and increase of Cho/Cr and mI/Cr ratios on proton magnetic resonance spectroscopy (1H-MRS). 1H-MRS studies on healthy aging have reported inconsistent findings, and only one small study investigated the role of APOE in the findings. Therefore, we sought to determine relations between APOE, age, and 1H-MRS metabolite ratios. We tested the predictions that APOE E4 status would 1) interact with age, and 2) impact the level 1H-MRS metabolites, and that these metabolites would mediate the relationship between APOE and cognition.

**Methods:** 112 subjects between 50 and 86 years were recruited. They had MMSE scores \( \geq 24 \) and did not meet Petersen’s criteria for MCI. Participants underwent 1H-MRS, genotyping and neuropsychological testing. A 10.8 cm³ (2 x 2 x 2.7 cm) mid-sagittal oblique posterior cingulate region 1H-MRS voxel was prescribed with the long axis parallel to the parieto-occipital sulcus. Double-spin-echo (PRESS) spectra were acquired on a GE Twinspeed 3T MRI scanner, with 128 excitations acquired at TR = 1600 ms, TE = 30 ms. For 1H-MRS analysis, the relative metabolite levels and ratios were determined for NAA, Cr, Cho, and mI. General Linear Models (GLM) were used to examine the effect of APOE, age, and their interaction on individual 1H-MRS metabolites, and Structural Equation Modeling (SEM) was performed in order to determine causal relationships between those variables and a cognitive composite measure.

**Results:** GLM analysis demonstrated that older APOE E4 carriers had higher Cho/Cr, and APOE E4 carriers had higher levels of ml/Cr than APOE3 homozygotes. No effects were found for NAA/Cr. The effect sizes for the comparison between older carriers of APOE E4 and younger APOE E3 homozygotes were in the high medium to large range. SEM resulted in a model with an excellent goodness of fit and in which the APOE x age interaction had a significant effect on 1H-MRS metabolites (Cho/Cr and ml/Cr), such that higher Cho/Cr and ml/Cr ratios were associated with worse cognition (as predicted). Furthermore, the APOE x age variable’s modulation of cognition was mediated by 1H-MRS metabolites. As expected, age also had a direct and negative effect on cognition.

**Conclusions:** In a healthy aging normal population, Cho/Cr and ml/Cr were significantly increased in older carriers of the E4 APOE allele. Additionally, SEM analysis suggested 1) higher Cho/Cr and ml/Cr were influenced by APOE x age interaction, and 2) the influence of APOE E4 status and older age on cognition was mediated by 1H-MRS metabolites. In concert, Cho and ml may increase the risk of APOE E4-driven neurodegeneration.

**Disclosures:** This study was supported by the Litwin-Zucker Research Center and NIH RO1 AG038734 (Goldberg TE, PI). Terry E. Goldberg receives royalties for the use of the Brief Assessment of Cognition in Schizophrenia (BACS) in clinical trials.

19 **The Effects of Human Demographic Characteristics on Translated Neurocognitive Measures Specific for Electroconvulsive Therapy**

Shawn M. McClintock¹ ², Mustafa M. Husain², Matthieu Chansard³, Ira Bernstein², Herbert Terrace³, Sarah H. Lisanby² ³, Munro Cullum²

¹Duke University Medical Center, Durham, NC, USA; ²UT Southwestern Medical Center, Dallas, TX, USA; ³Columbia University, New York State Psychiatrist Institute, New York, NY, USA

**Methodological Question Being Addressed:** Preclinical neurocognitive measures are beneficial as they allow for sensitivity and specificity when addressing a critical question within a finite controlled environment. However, when such measures are placed in clinical science without adequate translation, the information provided could be limited by confounding factors in the clinical model, such as demographic characteristics. Thus, the methodological question to be addressed is whether or not demographic characteristics affect performance on neurocognitive measures designed for preclinical models.

**Introduction:** Currently, there are no psychometrically sound neurocognitive measures specifically designed to measure the effects of electroconvulsive therapy (ECT). We previously reported our translation of touch-screen, computerized neurocognitive measures from a preclinical model in order to be used in clinical research. These measures comprehensively assess those neurocognitive domains affected by ECT including attention, processing speed, short and long term memory, working memory, and executive function. The purpose of this study was to determine if human demographic characteristics affected performance on these translated neurocognitive measures.

**Methods:** We enrolled 89 healthy participants from an academic medical center who were free of medical and psychiatric illness. The translated neurocognitive measures were placed on custom-modified, touch screen computer. The participants completed the neurocognitive measures including Target Identification, Target Sequencing, Spatial Configuration, Serial Target Recognition, and MetaCognition. Pearson correlation coefficients were used to examine associations between demographic characteristics and performance on neurocognitive measures at baseline.
Results: The study cohort had a mean age of 39.1 (SD=14.0), education of 16.0 (SD=2.5), and estimated IQ of 107.0 (SD=7.7). Age was associated with accuracy on Target Sequencing (r=-.39, p=0.0002) and Serial Target Recognition (r=-.32, p=0.005), and reaction time on Spatial Configuration (r=.34, p=0.001), Serial Target Recognition (r=.51, p=0.0001), and MetaCognition (r=.50, p=0.04). Education was associated with accuracy on Target Sequencing (r=.21, p=0.05) and Serial Target Recognition (r=.34, p=0.003). Estimated IQ was associated with accuracy on Target Sequencing (r=.31, p=0.004) and Serial Target Recognition (r=.45, p=0.0001).

Conclusions: Human demographic variables of age, education, and estimated IQ were associated with performance on the translated neurocognitive measures. These findings are consistent with prior literature that has consistently demonstrated that demographic variables attenuate performance on standard clinical neuropsychological measures. These results are encouraging and support the continued translation of these measures from the preclinical model for use in clinical research and practice. Further, they provide information regarding the need to interpret performance with respect to the specific demographic.

20 The Waiting Room Task: A Measure of Oxytocin-related Social Cognition

David L. Roberts, Consuelo Walss-Bass, João M. Fernandes, Dawn I. Velligan

University of Texas Health Science Center, San Antonio

Methodological Question Being Addressed: There is a need for practical measures that link functionally meaningful behavioral domains with biological markers.

Introduction: The biological bases of social dysfunction are not well understood. Emerging evidence suggests that the peptide oxytocin influences social functioning, and that this relationship may be mediated by social cognition. There is evidence that oxytocin administration may improve deficient social cognition in people with schizophrenia (SCZ). However, there is also evidence that whereas oxytocin is associated with prosocial judgments of in-group members it is also associated with derogatory judgments of out-group members. The aim of the current project is to develop a behavioral measure that is practical for use in clinical trials, and that is sensitive to the complex relationship between oxytocin and social cognition.

Method: We reasoned that oxytocin may be a marker for feelings of social self-referentiality that at high levels may manifest as feelings of vulnerability to threat or intrusion. We developed the Waiting Room Task (WRT) to be sensitive to both adaptive and pathological levels of self-referential bias. WRT participants view a series of brief video clips in which actors look at or away from the camera with different facial expressions. For each clip, participants judge whether it seemed that the actor looked at them and/or had a thought about them. The WRT takes approximately ten minutes to administer, and can be objectively scored by computer. The WRT produces ratings of bias toward inferring self-referential gaze and self-referential thought, and has shown initial evidence of validity in healthy and SCZ samples. The present study assessed the WRT’s association with blood plasma levels of oxytocin and measures of symptoms and neurocognition among 60 adults with DSM-IV SCZ and 20 demographically matched healthy controls.

Results: Controls had significantly lower blood plasma levels of oxytocin than did non-delusional SCZ patients (p=0.008) who, in turn, had significantly lower oxytocin than did delusional SCZ patients (p=0.037). Among controls and SCZ patients with delusions, oxytocin level was significantly correlated with bias toward self-referential gaze (controls: r=0.472, p=0.009) and thought (controls: r=0.434, p=0.015; SCZ: r=0.548, p=0.001). This bias was not present in patients without delusions, and oxytocin was unrelated to memory or executive function. Overall, controls exhibited significantly lower bias than SCZ participants in self-referential gaze and thought (p=0.023 and p=0.018, respectively).

Conclusions: The WRT is sensitive to self-referential bias and oxytocin level independent of traditional neurocognitive domains in both healthy and delusional participants. Oxytocin may be a marker for both normal and pathological levels of self-referential bias, and the WRT may be a useful and practical measure for use in clinical trials involving oxytocin, social cognition, and social functioning.

21 The Presence of a Placebo Comparator Arm and the Number of Active Treatment Arms Impact Results in Medication Trials for Generalized Anxiety Disorder and Panic Disorder

Mark Sinyor¹, Nadine Zahan², Ayal Schaffer³, Anthony J. Levitt⁴, Krista L. Lancot⁵, Tetsuhiro Harimoto⁶

¹Assistant Professor, Department of Psychiatry, University of Toronto. Psychiatrist, Sunnybrook Health Sciences Centre, Toronto, Canada. ²Project Manager, Sunnybrook Health Sciences Centre, Toronto, Canada. ³Associate Professor, Department of Psychiatry, University of Toronto. Head, Mood Disorders Program, Sunnybrook Health Sciences Centre, Toronto, Canada. ⁴Professor, Department of Psychiatry, University of Toronto. Psychiatrist, Sunnybrook Health Sciences Centre, Toronto, Canada. ⁵Head, Neuropsychopharmacology Research Program and Senior Scientist, Brain
Methodological Question Being Addressed: Do the inclusion of a placebo group and/or the number of treatment arms in trials for generalized anxiety disorder (GAD) and panic disorder (PD) influence outcomes?

Introduction: The placebo effect is thought to be mediated through expectations. Expectations are particularly important in psychiatric illness and they can change depending on the type of randomized controlled trial (RCT) a subject is enrolled in. A small but growing literature has demonstrated that in major depressive disorder (MDD), an increased probability of receiving a placebo results in poorer remission rates and response rates to antidepressants and placebos.

Results: Remission rates to active medication in GAD were 43%, 34% and 33% for DRUG-DRUG, DRUG-DRUG-PLACEBO and DRUG-PLACEBO trials respectively (DvD vs. DvDvP, p=0.001; DvD vs. DvP, p<0.001; DvDvP vs. DvP, p=0.003). Remission rates to active medication in PD were 61%, 55%, 53% for DRUG-DRUG, DRUG-DRUG-PLACEBO and DRUG-PLACEBO trials respectively (DvD vs. DvDvP, p=0.08; DvD vs. DvP, p<0.001; DvDvP vs. DvP, p<0.001). Remission rates to placebo in GAD were 23% and 21% (p<0.001) and in PD were 40% and 39% (p=0.03) for DRUG-DRUG-PLACEBO and DRUG-PLACEBO trials respectively.

Conclusions: These results are similar to what has been demonstrated in MDD in that trials with a placebo comparator demonstrate poorer response to the active medication. This effect appears to increase with increasing odds of receiving a placebo as we previously showed for MDD. This demonstrates that the lessebo effect seems to occur regardless of whether we are looking at trials for MDD, GAD or PD. Given that it has now been observed in multiple psychiatric disorders, further study of the lessebo effect is warranted.

Disclosure: I previously received grant funding from the Physicians' Services Inc. Foundation in Ontario, Canada.

22 Risk of Prospective Suicidal Behavior Reports among Psychiatric and non-Psychiatric Patients Using Lifetime Reports at Baseline

JC Mundt¹, JH Greist², M Federico³

¹Center for Psychological Consultation, ²Healthcare Technology Systems, ³ERT

Methodological Question Being Addressed: Previous analyses have shown that baseline reports of prior lifetime suicidal ideation and behavior (SIB) increase the likelihood of prospectively reporting suicidal behavior during study participation. The prior findings were found in patients predominately being treated for psychiatric disorders (MDD and PTSD) which have higher rates of SIB than non-psychiatric disorders. No prior analyses have investigated whether similar relationships between lifetime SIB and prospective reports of suicidal behavior hold for non-psychiatric disorders.

Introduction: (1) To replicate prior analyses that found increased risk of prospective reports of suicidal behavior among psychiatric patients reporting lifetime suicidal behavior and/or suicidal ideation with an intention to act at baseline. (2) To evaluate the generalizability of such findings among patients being treated for non-psychiatric conditions.

Methods: Data from 74,406 eC-SSRS assessments of SIB from 15,801 patients participating in clinical research studies were extracted for analysis. Evaluable patients with a lifetime assessment of SIB at baseline and one or more prospective follow-up assessments were identified and categorized as being diagnosed with a psychiatric or non-psychiatric disorder. Baseline reports of lifetime SIB were used to assign each patient a safety concern code (SCC) of ‘None’, ‘Ideation Only’, ‘Behavior Only’, or ‘Both’. Patients with a SCC of ‘None’ were used as the reference group to determine whether or not the patients assigned each of the other SCCs were more likely to prospectively report suicidal behavior during subsequent study participation.

Results: A total of 6,760 psychiatric (Major Depression, Generalized Anxiety, Post-traumatic stress, Opioid dependency) and 2,077 non-psychiatric patients were analyzed. Of the 4,967 psychiatric patients (73.5%) with a SCC of ‘None’, 103 (2.1%) prospectively reported a suicidal behavior during study participation. In comparison, 14 (9.0%) of the 156 psychiatric patients with ‘Ideation Only’ (OR = 4.7; 95%CI: 2.6-8.3; p < .001), 92 (10.7%) of the 861 ‘Behavior Only’ psychiatric patients (OR = 5.7; 95%CI: 4.2-7.6; p < .001), and 128 (16.5%) of the 776 psychiatric patients with a SCC of ‘Both’ (OR = 9.3; 95%CI: 7.1-12.2; p < .001) did so. Among the 2,077 non-psychiatric patients participating in non-psychiatric research studies, 1,983 (95.5%) had a SCC of ‘None’ and 7 (0.4%) prospectively reported a suicidal behavior. None of the 12 non-psychiatric patients with ‘Ideation Only’, 2 (4.3%) of the 47 ‘Behavior Only’ non-psychiatric patients (OR = 12.5; 95%CI: 2.5-62.1; p = .002), and 2 (5.7%) of the 35 non-psychiatric patients with ‘Both’
(OR = 17.1; 95%CI: 3.4-85.5; p = .001) did so during study participation.

**Conclusions:** Baseline assessments of most severe lifetime ideation that includes an intention to act, prior suicidal behaviors, or both, significantly increases the risk of patients prospectively reporting suicidal behavior during participation in clinical research studies. While such reports are substantially higher among psychiatric patients, the increased risk conveyed by the lifetime assessments does generalize to non-psychiatric patients also.

**Disclosures:** All authors have a financial interest in the eC-SSRS.

---

**Methodological Issues on the Impact of Central Nervous Disorders or Psychotropic Drugs on Driving Performance as Encountered by an Alcohol Reference Study in Driving Simulation**

Kaußner Y¹, Kenntner-Mabiala R¹, Hoffmann S¹, Neukum A¹, Kohnen R²

¹Wueerzburg Institute for Traffic Sciences, Wurzburg, Germany; ²ReSearch Pharmaceutical Services, Inc., Fort Washington, PA, USA

**Methodological Question Being Addressed:** This poster investigates which driving test scenarios and which driving performance endpoints are most adequate for the evaluation of driving-related deficits due to drug treatments using driving simulation in clinical studies.

**Introduction:** Experimental settings are recommended to investigate the impact of central nervous disorders and psychoactive drugs on driver fitness. Epidemiological studies are less promising because prevalences of these conditions are still relatively rare in the total population and amongst drivers involved in accidents. In experimental studies, the operationalization of driving performance represents a major methodological issue that has to be carefully considered. Here, modern driving simulators are increasingly favored: Test scenarios can be systematically designed, presented and reproduced without risk. Furthermore, as in real traffic, subjects can practice compensatory strategies. Thus, simulation combines the advantages of classical psychometric test batteries and on road trials and solves the shortcomings. However, some methodological issues have to be specifically addressed when using simulators. Firstly, the test scenarios should be both representative for real driving and sensitive to the condition to be investigated. Furthermore, driving performance should be assessed as a whole by various endpoints that refer to the operational and tactical level of the driving task. Finally, the simulated driving test has to be validated. The aim of the present study was to investigate these issues by collecting alcohol reference data for a representative simulated test course and various driving performance parameters.

**Methods:** The driving performance of 24 healthy volunteers under the influence of a placebo, 0.05%, and 0.08% blood alcohol concentration (BAC), was measured in our institute’s motion-based simulator with a double blind, randomized, crossover design. The test course included a representative set of driving scenarios on rural roads, highways and in urban traffic. The selection and design of the scenarios ensured that the operational and tactical levels of the driving task were addressed. Most of the scenarios were already proven sensitive to the influence of neurological disorders and psychotropic drugs in previous studies. Various well established driving parameters on lateral and longitudinal control recorded by the simulation were analyzed. In addition, specifically trained raters assessed subjects’ driving performance on a whole and registered and classified their driving errors as well.

**Results:** At large, the completion of the test course was worse under the influence of alcohol (p<.05 for raters’ assessment, p<.001 for total number of errors for both alcohol conditions compared to placebo). However, the parameters differentiated distinctly between the alcohol conditions depending on the underlying scenarios. For example, lane keeping performance as part of the operational level and measured by the standard deviation of lane position deteriorated in easy tracking scenarios (p<.05). In contrast, for complex tracking scenarios, the frequency of lane departure as an indicator of very poor tracking was sensitive (p<.005). For cognitively demanding driving tasks representing the tactical level, the number of related driving errors was found to be selective (p<.005).

**Conclusions:** The present study clearly confirms that the selection of driving scenarios and driving performance parameters is of crucial importance for an assessment of driving-related deficits due to central nervous disorders or psychotropic drugs. We recommend to evaluate driving performance as a whole by a profile-like analysis of various parameters at different levels of the driving task in future studies.
Development of the Readiness for Work Questionnaire in Schizophrenia

Potkin, SG¹, Bugarski-Kirola, D², Edgar, E³, Le Scouller, S⁴, Kunovac, J⁴, Velasco, EM⁵, Garibaldi, GM²

¹University of California, Irvine, Department of Psychiatry and Human Behavior, 5251 California, Suite 240, CA, USA 92617, ²F. Hoffmann-La Roche LTD, Basel, Switzerland, ³Roche Products LTD, Welwyn Garden City, UK, ⁴Excell Research Inc., Oceanside, California, USA, ⁵Resolution Psychopharmacology Research Institute, Mendoza, Argentina

The Methodological Question Being Addressed: Unemployment rates are high among patients with schizophrenia and unemployment has a negative impact upon patients’ quality of life. There is also a significant societal and economic burden associated with unemployment. The ability to work should be considered as a target for treatment with potentially wide ranging benefits. However, employment status is dependent upon several factors beyond ability to work, including the availability of work, stigma of mental illness and cultural influences, and patients’ work history and educational background. Thus a clinician rated scale evaluating readiness for work independent of current work availability would be a potentially valuable assessment tool.

Introduction: The aim was to conduct a series of studies to create and validate a work readiness questionnaire, which would allow clinicians to assess and rate patient function with respect to the ability to engage in socially useful activity.

Methods: Three separate studies were conducted to evaluate: content validity using qualitative analyses of interviews with expert respondents; test-retest and inter-rater reliability using ratings of videotaped patient and caregiver interviews; and construct validity in a cross-sectional, observational study of 200 patients with schizophrenia, 25% of whom were working.

Results: Content validity was supported, with practicing clinicians endorsing the importance of concepts in the questionnaire, including adherence to treatment, physical appearance, social competence and symptom control. Reliability was adequate, with the final readiness decision demonstrating good test-retest reliability (tetrachoric correlation 0.73) and moderate inter-rater reliability (Kappa statistic, 0.43; tetrachoric correlation, 0.69). Readiness for work was associated with statistically significantly higher levels of functioning and lower levels of negative symptoms (P<0.0001). For the prediction of work status, a low Positive Predictive Value (32%) and a high Negative Predictive Value (89%) were obtained, consistent with the interpretation that the scale clearly identifies patients unable to work and not working, but amongst those ready to work, there may be a proportion unable to work due to demographic and socioeconomic factors.

Conclusions: The WoRQ has been shown to possess suitable psychometric properties for use in a clinical trial setting. This was established in patients with a broad range of symptom severity. However, potential sensitivity to therapeutic intervention has yet to be evaluated.

Disclosures: This research was supported by F. Hoffmann-La Roche LTD.

Examining the Reliability and Usefulness of Interim Analysis Data: Case of Antipsychotic Drug Trials

Jonathan Rabinowitz¹, Nomi Werbeloff³, Francine Mandel², Stephen Levine³, Ivo Caers⁴, François Menard⁵, Shitij Kapur⁶

¹Bar Ilan University, Ramat Gan, Israel, ²Pfizer, New York, ³Haifa University, Haifa, Israel, ⁴Janssen Research & Development, Beerse, Belgium, ⁵Eli Lilly, Indianapolis, ⁶Institute of Psychiatry, Kings College, London

Methodological Question Being Addressed: Interim analyses yield effect size information that assist in deciding about potential sample size adjustments. This is predicated on the assumption that the results of the first cohort to complete the study will be representative of the patients to follow. The extent to which this is true is unknown.

Introduction: (1) To examine the similarity of efficacy results of the first and second half of patients to complete trials; (2) To compare the sample size needed based on interim results using repowering, conditional probability and bootstrapping.

Methods: Using data from a large repository of placebo controlled randomized trials of antipsychotic medications we examined the similarity of efficacy results of the first and second half of patients to complete trials. We also compared the sample size needed based on interim results using repowering, conditional probability and bootstrapping. NEWMEDS repository includes anonymized individual data from 29 placebo-controlled trials of second-generation antipsychotics (drug, n=6971, placebo, n=2200) conducted by AstraZeneca, Eli Lilly, Janssen, Lundbeck, and Pfizer. Patients in each study were divided by median of when they entered the study, which was available for 20 studies. The
effect size of the two cohorts was compared. First cohort effect sizes were used to re-estimate sample sizes by repowering the study, using conditional probability and bootstrapping.

**Results:** The mean difference in placebo vs. active treatment difference in effect size (weighted by sample size in the study) between cohorts was -0.01 and the median -0.03, indicating that overall the first and second cohorts yielded similar results. In nine trials the differences were positive, that is that the second cohort showed more placebo vs. active difference, in four the differences were negative, where the first cohort showed more difference these differences were small (less than -0.08). These would not likely invalidate extrapolation of results from the first to the second cohort. In 7 of the 20 studies the first cohort well outperformed the second cohort (from -12 to -44). Baseline to endpoint change was not correlated to time of recruitment (placebo r=.014, active treatment, r=-.01). Repowering using conditional probability together with bootstrapping was the most efficient as it required adding the fewest number of patients.

**Conclusions:** Results suggest that earlier cohorts in studies of antipsychotic medications may reliably serve to estimate trial effects in interim analyses and to efficiently adjust sample size calculations using conditional probability.

**Disclosure:** This research was supported by the Innovative Medicine Initiative Joint Undertaking under grant agreement n° 115008 of which resources are composed of EFPIA in-kind contribution and financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013). The authors report no conflicts of interest for this work.

26 Psychopathological Characteristics of First Episode, Chronic Inpatients and Ambulatory Patients with Schizophrenia: A Non-parametric Item Response Analysis

Harvey PD¹, Khan A²,³, Lindenmayer JP²,³, Opler M³,⁴, Yavorsky C⁴,⁵

¹University of Miami Miller School of Medicine, Miami, FL, ²Nathan S. Kline Institute for Psychiatric Research, Orangeburg, ³New York University School of Medicine, New York, NY, NY, ⁴ProPhase, LLC New York, NY, ⁵CROnos CCS, Hamilton, NJ

**Methodological Question Being Addressed:** While the Positive and Negative Syndrome Scale (PANSS) assesses multiple dimensions of schizophrenia, evaluation of each unidimensional domain has not been conducted to assess differences in the characteristics of psychopathology in subpopulations. There is considerable interest in identifying the course and treatment needs across different stages of illness. Epidemiological studies show that symptoms develop over time and can vary across subpopulations. For instance, first episode patients come to attention for recent-onset adjustment problems, while patients who cannot be discharged from long-stay psychiatric care may have substantial hostility or disorganization. Patients living successfully in the community are unlikely to be highly psychotic or disorganized. Can knowing the specific symptom or domain most characteristic of a subpopulation aid in identifying individually selected efficacy endpoints? The aim is to explore the clinical symptoms in three populations of schizophrenia (first episode, long-stay inpatients, and ambulatory patients).

**Methods:** Nonparametric IRT was used to examine PANSS items and severity anchors, and the 5-factor domains to discriminate symptom severity among Ambulatory (n = 831), Chronic Inpatient (n = 694) and First Episodes (n = 305). Option (OCCs) and Item Characteristic Curves (ICCs) examined the probability of rating each anchor as a function of 5-factor domains. Items were judged as Very Good, Good and Weak based on OCCs, ICCs and slopes of the IRT curves. The average item information function (IIF) assessed the amount of information in each group about symptom severity. First cohort effect sizes were used to re-estimate sample sizes by repowering the study, using conditional probability and bootstrapping.

**Results:** First Episode: The IIF for the anxiety domain was highly discriminating (showing greater range) for First Episodes (0.17 to 0.45), compared to Ambulatory (range: 0.16 to 0.32) and Chronic inpatients (range: 0.21 to 0.41). Disorganized symptoms showing highly discriminating items included Conceptual Disorganization, Difficulty in Abstract Thinking, Poor Attention and Preoccupation. The Hostility domain had highly discriminating IIFs for First Episodes (0.18 to 0.30) compared to the Ambulatory patients (0.16 to 0.24). Negative domain items (Poor Rapport, Passive/Apathetic Social Withdrawal) were scored Weak, whereas most Positive domain items were also scored as Weak. Chronic Inpatient: All items of the Anxiety domain were Weak, along with some items of the Positive domain (Suspiciousness/Persecution, Stereotyped Thinking, Somatic Concerns); for all other domains, items were Very Good or Good, with the Negative domain showing the highest IIF (0.15 to 0.28). Ambulatory: All items in the Anxiety and Hostility domain were scores as Weak, with the exception of Excitement. For the Disorganized domain, Conceptual Disorganization, Disorientation and Disturbance of Volition were Good, while other items were considered Weak. Suspiciousness/Persecution and Unusual Thought Content were not highly discriminating items for the Ambulatory group. Negative Symptoms were highly discriminating, IIFs = 0.09 to 0.17.

**Conclusions:** Anxiety, Hostility, Negative and some Positive items are better represented in First Episode patients. Disorganization, Hostility, Negative and some Positive items are better represented in Chronic Inpatients. Negative and some items of Disorganized and Positive symptoms are better represented in Ambulatory patients.
Identifying symptom domains specifically characteristic of subpopulations may be more useful in assessing efficacy endpoints than total or subscale scores.

27 A Comparison of Imputation and Analysis Methods for Examining Informatively Missing Data: Lessons from Chronic Pain Data
Karen Kesler, PhD, Herbert Harris, MD, PhD

Methodological Question Being Addressed: What is the best imputation and/or analysis method to offset bias due to subject dropout in CNS trials

Introduction: To evaluate therapies for CNS disorders, we rely on longitudinal data in order to assess both short and long term effects. Unfortunately, this type of study design is susceptible to bias and incorrect conclusions due to missing data. The most problematic missing data is due to subjects dropping out of the study often because of lack of efficacy or side effects. There are a host of imputation and analysis methods to combat this problem, but much controversy exists as to which methods give the most valid results. Quantifying the effects of these various potential solutions is intrinsically difficult due to the problem itself—without the knowing what the true results are for the missing subjects, we cannot directly measure the bias or make meaningful comparisons between imputation and analysis methods.

Methods: Our approach to this problem involves use of real data from a chronic pain clinical trial to examine the bias induced in longitudinal data by various imputation and analysis methods. We compare these results to determine the best practical strategy for managing this common problem. The trial examines a control and experimental groups with chronic pain over 4 weeks using a Numeric Pain Rating scale every day. To accurately measure bias, one needs complete data on all subjects as well as the dropout pattern. We have taken the actual data with informative dropouts and carefully created realistic, full data trajectories for all subjects. Together with the known dropout time, we are then able to conduct analyses with a variety of imputation and analysis methods and calculate the actual bias for each one. We chose several methods to compare:

- “Observation Carried Forward” Imputation: Imputing both the “Last Observation Carried Forward” (LOCF) and the mixture of “Worst Observation (for experimental subjects) and Last Observation (for control subjects) Carried Forward”.
- Population Mean Imputation.
- Random Effects Analysis: No imputation, just using a Mixed Effects Longitudinal model with a random subject effect, with and without controlling for the reason for dropout.
- Pattern Mixture Model

Since our estimate of interest was the difference in pain scores between the two treatment groups, we compared the bias and standard error of that difference, along with the conclusions drawn for testing a hypothesis of no difference between the groups.

Results: We found that the imputation methods examined provided less bias than only using completers, but underestimated the variability of the estimate. Random effects models reduced this bias and controlling for the reason for dropout inflated the variance estimate. Pattern mixture models were associated with the lowest bias and largest estimate of variance.

Conclusions: We conclude that Pattern Mixture models provide an excellent sensitivity analysis, but for practical purposes are not ideal for a primary analysis. We recommend the random effects models on the longitudinal data. We discuss the generalization of these findings to clinical trial data from other CNS indications.

28 Electronic Assessment of Suicidal Ideation and Behavior for Meta-Analyses across Multiple Trials and Treatment Indications
Greist, JH1, Mundt, JC2, Federico, M3

Methodological Question Being Addressed: The 2012 FDA guidance regarding suicidal ideation and behavior (SIB) recommends prospective assessment of SIB at baseline and all planned visits when other clinical outcomes are to be collected in all drug development trials for psychiatric indications, as well as for antiepileptic or other drugs with CNS activity. A primary reason for this recommendation is to guarantee more complete SIB assessments concurrent with administered treatments and placebo. Signals indicative of increased risk would be easier to detect in individual trials,
and aggregation of data across multiple trials for meta-analyses could more readily confirm true signals or provide counter-evidence of spurious, false-positive events. The FDA guidance states “The full assessment of suicidal ideation and behavior generally should involve a pooled analysis of all controlled trials, so that it will not be possible to conclude that a drug has no effect on suicidal ideation and behavior until a substantial database is available for this analysis.”

**Introduction:** Databases created from use of the C-SSRS (or other acceptable instruments, such as the electronic self-report eC-SSRS) that directly classify SIB into preferred categories can provide the basis for such analyses. Aggregation and analysis of the database described below was designed to meet the FDA recommendations, and to provide clinical insight regarding differences in SIB prevalence rates among patients participating in clinical studies investigating treatments for different therapeutic areas and indications.

**Methods:** Data from 74,406 eC-SSRS assessments of SIB were extracted from 35 clinical trials. Patient-reported ideation and behavior were compared at baseline (lifetime) and prospectively during trial participation across multiple therapeutic areas and 17 treatment indications. Trials were categorized as Psychiatric (MDD, PTSD, opioid dependency, GAD), Neurologic (pain, epilepsy, insomnia, multiple sclerosis, Parkinson’s, restless leg syndrome), or non-CNS studies (pulmonary, fibromyalgia, analgesia, anti-viral).

**Results:** The number and type of eC-SSRS assessments administered within each therapeutic category and the percentage of observations with each SIB are provided below.

<table>
<thead>
<tr>
<th>Therapeutic Category</th>
<th>eC-SSRS Assessments</th>
<th>Most Severe Suicidal Ideation (%)</th>
<th>Patient-reported Suicidal Behavior (%)</th>
<th>Positive Reports (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Psychiatric (16 studies)</td>
<td>Baseline = 10,233</td>
<td>SLC = 51,819</td>
<td>46.40</td>
<td>19.83</td>
</tr>
<tr>
<td></td>
<td>Baseline = 1,912</td>
<td>SLC = 4,352</td>
<td>82.01</td>
<td>8.71</td>
</tr>
<tr>
<td>Neurologic (13 studies)</td>
<td>Baseline = 1,991</td>
<td>SLC = 4,099</td>
<td>82.97</td>
<td>7.73</td>
</tr>
<tr>
<td>Non-CNS (6 studies)</td>
<td>Baseline = 1,991</td>
<td>SLC = 4,099</td>
<td>98.88</td>
<td>0.93</td>
</tr>
</tbody>
</table>

**Conclusions:** Neurologic and non-CNS studies found similar prevalence rates of SIB using the eC-SSRS at baseline and prospectively during study participation. As anticipated, these rates were substantially less than the suicidal ideation and behavior rates self-reported by psychiatric patients. Although less frequent, these positive findings remain a safety concern in non-psychiatric trials.

**Disclosures:** All authors have financial interests in the eC-SSRS.

**29 Working Group Dinner Updates**