Psychopharmacology (2006) 186: 7–17
DOI 10.1007/s00213-006-0331-5

ORIGINAL INVESTIGATION

Paul Maruff · John Werth · Bruno Giordani · Angela F. Caveney · Douglas Feltner · Peter J. Snyder

A statistical approach for classifying change in cognitive function in individuals following pharmacologic challenge: an example with alprazolam

Received: 8 June 2005 / Accepted: 16 January 2006 / Published online: 28 March 2006

Abstract *Introduction:* The effects of any drug treatment on cognitive function are typically studied in groups of subjects. Observations made about the behavior of the drug, in the study sample, are then generalized to the population from which the sample was drawn. However, the magnitude and pharmacodynamic qualities of the response to many central nervous system-active drugs are known to vary in the population. Therefore, it is useful to consider statistical models for the detection of cognitive change in response to a drug treatment in individual subjects. *Materials and methods:* In this report, we first outline the statistical assumptions and requirements for the reliable estimation of clinically relevant individual change in cognition. We then used the sedative benzodiazepine, alprazolam, as a pharmacologic challenge in healthy volunteer subjects to test our statistical model, using a parallel groups placebo-controlled study design. After treatment, the nature and severity of alprazolam-induced cognitive change was determined for each individual. *Results:* Our proposed method and analysis showed an excellent sensitivity and specificity for alprazolam-related cognitive deterioration in individuals. *Discussion and conclusions:* These findings, although preliminary, suggest that statistically reliable decisions about the effects of sedative drugs on cognition can be made for individuals.

P. Maruff (✉)
CogState Ltd,
Melbourne, Australia
e-mail: pmaruff@cogstate.com

J. Werth · D. Feltner
CNS Early Development,
Pfizer Global Research & Development,
Ann Arbor, MI, USA

B. Giordani · A. F. Caveney
Section of Neuropsychology, Department of Psychiatry,
University of Michigan School of Medicine,
Ann Arbor, MI, USA

P. J. Snyder
Department of Psychology,
University of Connecticut (Storrs, CT)
& CNS Early Development,
Pfizer Global Research & Development,
Groton, CT, USA

Introduction

Most conclusions about the psychoactive effects of licensed or novel drugs are based on the comparison of cognitive measures between or within groups of subjects. For example, the effects of a drug on cognitive function are determined optimally by comparing the average cognitive performance of a group of subjects, after drug treatment to that of the same group, or a matched group, who have been treated with a placebo (e.g., Hinkle et al. 1996). Observations made about the behavior of the drug in the study sample are generalized to the population from which the sample was drawn (e.g., healthy adults, patients with psychiatric disorder). However, there is a growing awareness that response to psychoactive drugs may vary in the population and that some of this variation can be explained by subject-specific factors such as physical characteristics, disease severity, or genetic status (e.g., Macher and Corcq 2004; Calabrese et al. 2004; Lesch and Gutknecht 2004). Therefore, it is appropriate to develop statistical models for the detection of cognitive change in response to a drug treatment in individual subjects. For example, it is well known that not all study subjects show the expected cognitive response to drug challenges (e.g., treatment nonresponders Redelmeier and Tversky 1990). The identification and study of "treatment nonresponders" or "treatment responders" could provide critical information for understanding population variability, for design of future studies, for fine-tuning of subject inclusion criteria, and eventually for regulatory labelling. Second, the ability to measure cognitive effects reliably in individuals may increase the statistical power of small phase I and II studies. With this approach, individuals who do not respond to a treatment based on prospectively defined criteria might be removed or replaced with "treatment responders". Third, the ability to measure a drug effect in individual patients may

assist clinicians in determining whether their patients respond to a treatment as expected from published clinical trial data.

The definition of a cognitive response or nonresponse to treatment requires that some criterion for either be defined on a specific outcome measure (Guyatt et al. 1997; Macher and Corcq 2004). The definition of such a criterion for change in cognitive performance in individuals requires that at least three issues be considered. First, any criterion value for classifying change will contain error and this error is greater for individuals than groups (Jenkinson 1995; Williams and Naylor 1992). Therefore, the magnitude of change in a cognitive outcome measure necessary to be classified as statistically significant or clinically meaningful in an individual may need to be greater than that required for a group of subjects. Second, if a criterion value for classifying cognitive change in individuals is estimated from a group study, then this criterion will also contain error. Without appreciating the magnitude of this error, individuals whose performance has merely exceeded the group-derived criterion may be considered to have truly changed, even though the change value is within the range of error associated with the criterion (Redelmeier and Tversky 1990). Third, while these two issues are relevant when cognitive change is defined using a single cognitive performance measure, most studies of the cognitive change in response to drug treatment actually use multiple performance measures. This is because qualitative differences in the patterns of treatment-related change across measures can be used to characterize the effects of different drugs or different doses of the same drug. However, as the number of performance measures used to define cognitive impairment in individuals increases, the probability of false positive classification also increases. For group studies, this risk is well understood and corrections in Type I error for multiple comparisons are generally applied (e.g., Bonferroni correction; Hinkle et al. 1996). By contrast, methods for the application of corrections for Type I error to individuals are not as well known and therefore used rarely.

It may be possible to provide a statistically reliable basis for inferences about drug treatment-related cognitive change in individuals. First, we suggest that performance data for the individual should not be considered representative of the population from which the individual was selected, but rather as representative of the population of data for that individual. Hence, statistical inferences about the presence of cognitive change pertain only to the individual and not the population. Importantly, the assumptions that operate for inferential statistics in groups must also be satisfied for individuals. Thus, distributions of the data for individuals should be normal and possess homogeneity of variance between treatment conditions. Every effort should be made to meet these assumptions by utilizing strategies such as increasing the number of observations, improving the metric properties of tests and minimizing the effects of extraneous factors such as practice and motivation (Anastasi and Urbina 1997; Bland and Altman 1995). In a series of recent studies, we identified a series of outcome measures from the CogState battery http://www.cogstate.com) that satisfy these statistical requirements (Mollica et al. 2005; Mollica et al. 2004; Collie et al. 2003a,b).

Statistical questions about whether the cognitive change observed in an individual is statistically reliable also require that any observed change be compared to some estimate of the normal variability in performance over time. Recent considerations of this issue have shown how variability estimated in groups of individuals measured over time can be used to provide a denominator against which change in performance in individuals can be compared (Bland and Altman 1996; Collie et al. 2002; Jacobson and Truax 1991). Importantly though, the most powerful of these techniques are parametric and once again require that data have the appropriate characteristics (Anastasi and Urbina 1997; Bland and Altman 1996). Second, the definition of a criterion for change on any individual cognitive measure should be a statistical difference rather than an absolute difference in performance. Statistical criteria appreciate that any specific value nominated as a criterion will contain error and consequently place confidence intervals around this criterion value. Third, where multiple criteria for change are applied to the same individual, the potential for a Type I error can be controlled by considering simultaneously the number of outcome measures, the criterion for abnormality on any one measure, and the number of abnormal tests results necessary for a classification of change and whether the hypothesis is one- or two-tailed. For example, a series of simulations based on the binomial probability distributions show that to contain Type I error at less than 5% (with a two-tailed hypothesis) in a test battery that yields nine outcome measures, change should be classified as reliable only when there is a decline in performance of 1.96 SD or greater on at least three of the outcome measures. If the criterion for change on any one single measure is defined as a decrease of one standard deviation, then a classification of reliable change would require this to occur for all nine measures to preserve the false positive classification rate at 5% (see Ingraham and Aiken 1996 for further discussion of this point). However, the Ingraham and Aiken (1996) simulations assume that the outcome measures used are independent. Therefore, these simulations may be too conservative for classification criteria applied to questions of cognitive change. It would be useful to compute the actual false positive classification rate associated with any criteria for cognitive change. Another method for minimizing the risk of Type I error when multiple performance measures are used is to summarize all outcome measures into a single composite change score (see Rasmussen et al. 2001, Mollica et al. 2004). In fact, composite change scores may be more appropriate for identifying treatment-related cognitive impairment in individuals as it is possible that a psychoactive drugs will give rise to a generalized but subtle change over a range of cognitive functions rather than to large impairments or improvements in measures of specific cognitive functions (as would be required to see a deterioration of greater than 1.96 SD).

Benzodiazepines are a class of drug for which the effects on cognitive function in groups of healthy individuals are well understood (de Visser et al. 2003; Vester and Volkerts 2004; Stewart 2005). The benzodiazepine alprazolam is

known to induce marked impairment in cognitive function after a single dose as low as 0.5 mg and this impairment has been observed most reliably in psychomotor, attentional, and memory functions (Vester et al. 2002; Vester and Volkerts 2004; Snyder et al. 2005). Alprazolam reaches maximum concentration in plasma (e.g., C_{max}) in approximately 1 h and has few side effects (D'Souza et al. 2001; Wong et al. 1998). Given these features, alprazolam treatment provides a good challenge for measuring the magnitude of impairment in individual subjects according to the statistical and methodological principles outlined above. Because all of the studies of alprazolam related cognitive impairment have been conducted in groups, the extent to which alprazolam-related cognitive impairment can be identified reliably in the cognitive performance of individuals has not been established. Recently, we reported that the CogState test battery and the Groton Maze Learning Test (GMLT) developed by us specifically to assess cognitive change were sensitive to the effects of 1 and 0.5 mg of alprazolam on cognitive function (Snyder et al. 2005). We found that, when analyzed at a group level, an acute dose of 1 mg gave rise to significant impairment across all aspects of cognitive function measured. Although the magnitude of impairment was greater for tests of attention and psychomotor speed (e.g., effect sizes for the difference between baseline and posttreatment approximately 1) compared to the deterioration detected on measures of memory and executive (e.g., effect sizes for the difference between baseline and posttreatment approximately 0.5).

The cognitive measures used in the Snyder et al. (2005) study had been developed specifically so that they yielded data with optimal metric properties for the detection of change (Mollica et al. 2004; Mollica et al. 2005) and these measures had been challenged in a group of subjects who were given a moderate dose of alprazolam in a highly controlled clinical trial environment. Therefore, the data from the Snyder et al. (2005) study provided a strong basis for challenging empirically the statistical and methodological principles argued to be important for the classification of treatment-related cognitive change in individuals. The aim of the current study was to reanalyze some of the data from our earlier study in accord with these recommendations to determine the extent to which cognitive change could be classified in individual subjects. To achieve this, the stability of performance of each outcome measure was estimated from the multiple pretreatment assessments. These estimates were then used as the reference against which the magnitude of any treatment-related cognitive change could be determined. We then compared the incidence of treatment-related cognitive change in the group who had been challenged with a single acute 1-mg dose of alprazolam and those who received placebo. We hypothesized that the statistical and methodological approach developed in this study would allow classification of cognitive change in a large proportion of subjects who had received an acute dose of alprazolam, while at the same time maintaining a low rate of false positive classification in individuals who had received placebo.

Materials and methods

Subjects

The data from 26 healthy individuals was analyzed in the current study (14 had received alprazolam and 12 had received placebo). All were in good health as determined by medical history, physical examination, vital signs, electrocardiogram, and clinical laboratory measurements, The exclusion criteria for the selection of the study subjects were if any of the following conditions applied: routine use of sedative or other psychoactive medication, participation in any investigational or marketed drugs during the 30-day period before the start of the study (baseline), pregnancy or lactation, a history of significant adverse reaction to gabapentin, pregabalin, or benzodiazepine anxiolytics, significant urine concentration of any illicit drug, a history of glaucoma or red–green color blindness, or a history of any central nervous system condition such as Parkinson's disease, significant head trauma (head injury with loss of consciousness within the past 5 years), brain tumors, major depression, alcohol or illicit drug abuse. All subjects provided written informed consent before any study procedures. This study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice and the protocol was approved by institutional ethics committees. As this study as conducted at the level of individual subjects, only those who provided baseline and postdrug cognitive data for all performance measures were included. Using this additional criterion, two patients who had been randomized to the placebo condition in the initial group study (e.g Snyder et al. 2005) were not analyzed in the current study. There was no difference between the groups of subjects randomized to placebo or alprazolam for weight (alprazolam mean=78.4 kg, SD=17.1, placebo=82.3, SD=18.1, $t<1$), height (alprazolam=172.1 cm SD=8.1; placebo mean=171.3 cm, SD=9.2, $t<1$), age (alprazolam mean=31.2 years SD=8.2; placebo mean=32.4; SD=9.6; $t<1$), education level (alprazolam mean=12 years SD=2.3; placebo mean=12 years; SD=3.1; $t<1$), or gender (alprazolam number of males=14/14; placebo number of males=11/12 $\chi 2=0.1$).

Study design

Subjects were assigned randomly to receive placebo TID or 1 mg alprazolam TID in a double blind parallel groups design. The present analyses utilized performance data at baseline and after the first dose of study medication. The study blind was maintained by administering the same number of identical capsules to each subject within each cohort. The administration and ingestion of study medication was supervised by clinical research personnel. Cognitive testing required approximately 30 min in total (20 min for the CogState battery and 10 min for the GMLT). These cognitive tests were administered three times for more than three days before baseline (day 3, day 2, day 1), then at baseline and then again at C_{max} for

alprazolam. Because the time to maximum concentration (T_{max}) for alprazolam is approximately 1 h post dosing (Wong et al. 1998; D'Souza et al. 2001) and the test battery required 30 min to complete testing always began 45 min after dosing so that we bracketed around T_{max} for this drug. The order of administration for the CogState and GMLT was pseudo-randomized across subjects and test sessions to control for test order effects, although the tests in the CogState battery were always given in the same order. Data from the Day 3 assessment was classified as a familiarization assessment and was not considered in the current analysis.

Apparatus

The cognitive test battery consisted of eight different tests of psychomotor function, attention, executive function, memory, and problem solving/strategy use. All were presented on lap top computers. The tests were selected because of their brevity, demonstrated utility for within-subjects experimental designs, and for the parametric properties of their outcome measures (cf., Falleti et al. 2003; Schroder et al. 2004). The CogState battery consisted of seven tasks and has been described elsewhere in detail (Falleti et al. 2003; Mollica et al. 2004; Snyder et al. 2005). The tests used the outcome measure they generate and the cognitive functions that they measure are summarized in Table 1. These tasks are in the form of card games that were presented in succession on a green background. At the beginning of each task, written instructions were presented to the left of the screen to indicate the rule. Each subject was then given an interactive demonstration and, once they had successfully given a sufficient number of correct responses to demonstrate their awareness of the rules, the task began. A grey keyboard resembling a computer keyboard appeared in the lower half of the computer screen and the cards associated with each task were presented in the upper half. The subject was required to respond with two keys throughout the entire battery, either the "D" or "K" key. For right hand dominant subjects "Yes" responses were indicated by pressing the "K" key and "No" responses by pressing the "D". This mapping was reversed for left-hand dominant subjects. The beginning of each new task was indicated with a shuffling of the cards. An error beep sounded if a subject pressed an incorrect key at any time. Each trial was time-limited and the same error beep sounded if a response was not made within the required time. A subject could pause the test at any stage using the escape key. The performance measures for each task were the speed of the response (i.e., reaction time recorded in milliseconds) and the accuracy of response (i.e., correct or incorrect). The seven tasks included in the battery were as follows.

1. *Detect* A single card was presented at in the middle of the monitor and the subject was required to respond as soon as the card turned face up. The instructions provided by the "on-screen" helper were: "Has the card turned face up? If yes, press the 'K' key". This was completed three times throughout the battery: at the beginning, in the middle (before the working memory task), and at the end. On each occasion, this task continued until 25 correct responses were recorded or the maximum allowed time elapsed (2 min). For the current study, performance was averaged across the three trials.
2. *Monitor* Five cards were presented face-up in the middle of the screen. Horizontal white lines were marked above and below these cards. The five cards move randomly in either an upwards or downwards direction. The direction of each card was independent of the others. The instructions provided by the "on-screen" helper were: "Press 'yes' as soon as a moving card touches a white line". Following this, the card returned to the middle of the screen and began to move again. A response was classified as being correct if the subject pressed the "K" key after one of the cards had touched the upper or lower white line. If the "K" key was not pressed when a card touched either line, an error buzzer sounded until a response was made. This task continued until 25 correct responses were recorded or until the maximum allowed time elapsed (2 min).
3. *Identify* A white card with either a black or red square in its center was presented in the middle of the computer screen. The instructions provided by the "on-screen" helper were: "Is the face-up card red? If the answer is yes press the 'K' key, if the answer is no press the 'D' key". Once the subject demonstrated that they understood the rules, the cards reverted to conventional playing cards and the participants responded to the color of the cards' suit. This task continued until 25 correct responses were recorded or the maximum allowed time elapsed (2 min).
4. *Match* Two white cards, aligned horizontally, with either a black or red square on each were presented in the middle of the screen. The instructions provided by the "on-screen" helper were: "Are the two face-up cards the same color? If the answer is yes then press the 'K' key, if the answer is no, press the 'D' key". Once the subject demonstrated that they understood the rules, the cards reverted to conventional playing cards. This task continued until 25 correct responses were recorded or the maximum allowed time elapsed (2 min).
5. *Remember one back* Two playing cards were presented in the middle of the screen; one was face-up and the other face-down. The instructions provided by the "on-screen" helper were: "Does the face-up card exactly match the one before? If the answer is yes, press the 'K' key, if the answer is no, press the 'D' key". The correct answer for the first card was always "no" as no card had previously been shown. After the subject had pressed the "D" key, the card turned over and a new card appeared face-up. If this card was the same as the previous card, the subject was required to press "K" (yes), if not "D" (no) should have been pressed. This task continued until 25 correct responses were

recorded or the maximum allowed time elapsed (2 min).

6. *Sort* Six pairs of cards (each pair aligned vertically) were presented in a display at the top of the computer screen. A single pair of cards (also aligned vertically) was presented at the bottom of the screen, which may or may not have matched one of the pairs above. The instructions provided by the "on-screen" helper were: "Do the two face-up cards match one of the pairs above? If yes, press the 'K' key, if no, press the 'D' key". Once the subject responded, a new pair of cards was presented at the bottom of the screen. This task continued until 25 correct responses were recorded or the maximum allowed time elapsed.
7. *Learn* Five card pairs (aligned vertically) were presented at the top of the screen. A single card pair (aligned vertically) was also presented at the bottom of the screen. The instructions provided by the "on-screen" helper were "Do the two face-up cards match one of the pairs above? If yes, then press the 'K' key, if no, press the 'D' key". As soon as the subject pressed the "K" or "D" key, a new pair of cards was shown at the bottom of the screen. If the single pair of cards matched one of the pairs of cards in the display, the pair at the top of the screen turned face down. This continued until all of the card pairs at the top of the screen were turned face down, with the exception of the central pair, which remained face up as a control task. The subject was required to remember the card pairs at the top of the screen and to determine whether the pair at the bottom of the screen matched one of these. The pairs of cards presented at the bottom of the screen could be correct pairs or foils (i.e., card pairs that were incorrect). Each incorrect pair was presented only once during the test, whereas each correct pair was presented five times. The probability that the pair of cards presented was correct was 25%. When the subject responded correctly, the card pair shown moved to the top of the screen (indicating that it was one of the hidden pairs). After an incorrect response, the card pair remained in the same location but was shuffled to the bottom of the deck while the error buzzer sounded. This task continued until 25 correct responses were recorded or the maximum allowed time elapsed
   - The Groton Maze Learning Task (GMLT) was presented on a tablet computer (i.e., a portable computer with a touchscreen) where the subject used an electronic stylus to respond. Once again feedback was given by the computer.
8. *Maze Learning* The Groton Maze Learning Test (GMLT) was developed by one of the authors (P.J. S.), and it is based on an earlier hidden maze task developed by Milner (1965). The GMLT has been used to examine spatial memory and spatial working memory (see Darby and Walsh 2005, for review), and it consists of a 10×10 grid of tiles presented on a computer touchscreen in which 28-step pathway was hidden. The tile at the start and the finish were indicated in locations at the top left and bottom right locations of the grid. The subject was instructed to move one tile from the start location and then to continue one tile at a time toward the end. While moving through the hidden maze, subjects were required to adhere to two rules; first, they could not move diagonally and second, they could not move back to a location where they had previously been. After each move made by the subject, the computer indicated whether this was correct (i.e., was the next step in the pathway) or incorrect (was not the next step in the pathway, or that the subject had broken one of the rules). If the choice was incorrect, the subject was required to touch the last correct tile and then choose a different tile to advance toward. When the subject completed the pathway, he or she repeated this process for a total of five successive learning trials. On each trial, the number and type of errors (first errors vs perseverative errors) made as well as the time (in ms) to complete each trial was recorded. This current paper reports results for only a measure of problem solving and spatial working memory on this task (the number of correct moves per second, averaged over the five learning trials). Twenty well-matched alternate forms for this test were selected in pseudo-random order, to

Table 1 Cognitive functions assessed in the current study and their performance measures

Cognitive task	Cognitive function measured	Measure type	Metric
Detect	Psychomotor	Speed	Log10RT
Monitor	Vigilance	Speed	Log10RT
Identify	Simple attention	Speed	Log10RT
Match	Simple attention	Speed	Log10RT
Remember one back	Executive function (working memory)	Speed	Log10RT
Remember one back	Executive function (working memory)	Accuracy	Arcsine percentage accuracy
Sort	Executive function and complex attention (sorting)	Accuracy	Arcsine percentage accuracy
Maze learn	Executive function (problem solving/spatial working memory)	Efficiency	Moves/second
Learn	Associate learning and memory	Accuracy	Arcsine percentage accuracy

ensure that no subject completed the same hidden path form more than once throughout the study.

Data analysis

Data cleaning and preprocessing

Data from all practice trials was removed from data bases. All data from the CogState tasks were then inspected and anticipatory responses (RTs <100 ms) and abnormally slow responses (RTs >3,500 ms) were considered errors and excluded from further analysis. The number of trials on which correct responses occurred was also computed and this was expressed as a proportion of the total trials. As distributions of percent correct from the participant group were characterized by positive skew, the percent correct value for each individual for each measure was normalized using an arcsine transformation (Anastasi and Urbina 1997). The speed of decisions was computed by taking the average reaction time of correct responses for each task. Inspection of the distributions of reaction times for correct responses in each individual indicated a negative skew on all of the outcome measures. Therefore, before computing the average reaction time, the response time for each trial was normalized using a logarithmic base 10 transformation (Anastasi and Urbina 1997). For each assessment of the GMLT, the time required to complete the pathway on each of the five trials was measured and a total time to complete the five trials computed. Because there are 28 correct steps required to complete each trial of the pathway, the total time was divided by 140 (i.e., 5×28) to yield a measure of correct moves/second. Because this measure was distributed normally, it was not transformed.

Stability of cognitive performance over time

Although the stability of the performance measures used in this study has been established performance across the three baseline assessments (Day 2, Day 1, and baseline) was analyzed in the 26 subjects to determine whether any learning effects were present and to compute stability estimates for each test. Baseline data from each subject for each task was submitted to a series of one-way analyses of variance (ANOVA). The within-subject standard deviation (WSD) was derived from the ANOVA output by computing the square root of the mean square residual term (Bland and Altman 1996). The coefficient of variation was then calculated for each measure by expressing the WSD as a function of the performance on each test averaged over the three baseline assessments (i.e., grand mean).

Effect of treatment on cognitive performance of individuals

Computation of reliable change for individual performance measures To quantify the magnitude of cognitive change associated with each performance measure under

Table 2 Group mean (SD) performance across the three baseline assessments, stability (within subject standard deviation, WSD) and coefficient of variation (CoVar)

Performance measure	B1 (*n*=26)	B2 (*n*=26)	B3 (*n*=26)	*F*	Mean B	WSD	CoVar
Detect (speed)	2.51 (0.1)	2.49 (0.08)	2.51 (0.08)	1.3	2.5 (0.07)	0.04	0.02
Monitor (speed)	2.53 (0.11)	2.53 (0.11)	2.53 (0.13)	0	2.53 (0.09)	0.08	0.03
Identify (speed)	2.69 (0.15)	2.71 (0.08)	2.74 (0.08)	1.8	2.71 (0.08)	0.09	0.03
Match (speed)	2.84 (0.07)	2.85 (0.1)	2.84 (0.08)	0.1	2.84 (0.07)	0.04	0.01
Remember one back (speed)	2.87 (0.13)	2.85 (0.11)	2.84 (0.08)	1.3	2.85 (0.09)	0.07	0.02
Remember one back (acc)	1.16 (0.29)	1.26 (0.3)	1.33 (0.26)	1.2	1.25 (0.22)	0.13	0.10
Maze learn (efficiency)	1.14 (0.21)	1.12 (0.3)	1.15 (0.22)	0.8	1.14 (0.18)	0.10	0.09
Sort (acc)	1.11 (0.34)	1.08 (0.24)	1.15 (0.26)	0.8	1.11 (0.23)	0.11	0.10
Learn (acc)	0.83 (0.22)	0.85 (0.19)	0.9 (0.18)	1.6	0.86 (0.18)	0.11	0.13

The *F* value given here is the effect of time derived from an analysis of variance comparing performance between the three baseline assessments

the placebo and alprazolam conditions in each individual subject, a cognitive change score was computed for each outcome measure using the following formula:

$$\text{cognitive change score}_{(\text{measure a})} = \frac{\text{baseline}_{(\text{measure a})} - \text{post baseline}_{(\text{measure a})}}{\text{WSD}_{(\text{measure a})} \text{ of the group}}$$

In this equation, "measure a" is any one of the nine performance measures (detailed in Table 1), the WSD is the within-subject standard deviation derived from the three baseline conditions for the same measure, baseline is the mean of the three baseline scores for "measure a", and postbaseline (both shown on Table 2) is the score on "measure a" at C_{max}.

Classification of treatment related cognitive change in individual subjects

Cognitive change scores were used to classify individuals as showing treatment (placebo or alprazolam) related cognitive impairment in two ways. First, previous simulations indicated that for nine outcome measures, a change in performance of 1.96 or greater on two or more tests would retain the Type I error for classification of impairment at less than 5% under a two-tailed hypothesis (Ingraham and Aiken 1996; although these simulations assume tests are independent). Because the cognitive change scores had been standardized using the WSD, cut-scores for abnormality could be derived directly from the known properties of normal distributions. Therefore, individuals who showed cognitive change scores of ≤−1.96 or ≥1.96 on two or more outcome measures were classified as showing treatment-related cognitive change ($p<0.05$, two tailed).

The second method defined treatment -related cognitive change by identifying whether subtle impairment in cognitive function occurred across all measures (Rasmussen et al. 2001). A composite cognitive change score was computed by summing the cognitive change scores across the nine performance measures for each individual. Because the composite cognitive change score required that cognitive change scores be summed, the characteristics of the data distribution changed. Therefore, it was necessary to restandardize these scores to use the properties of a normal distribution to guide decisions about abnormality. A group mean and standard deviation of composite cognitive change scores was then computed for the placebo group and the alprazolam group. The composite cognitive change scores for all subjects was then standardized using the mean and SD of the composite cognitive change score in the placebo group. A subject in the alprazolam group with a composite cognitive change score of greater than the placebo group mean ≤−1.96 or ≥1.96 standard deviations ($p<0.05$, two tailed) was classified as showing treatment-related cognitive impairment. Classifications of treatment-related cognitive change in the placebo group were deemed to be false positive classifications; the false positive rate observed for the placebo group was used to check the accuracy of the theoretically derived false positive rates (e.g., Ingraham and Aiken 1996). The number of individuals who met the criterion for treatment-related cognitive change in each group was compared using Fisher's exact test.

Table 3 Number of subjects classified as showing cognitive change for each performance measure under the placebo and alprazolam conditions

Performance measure	Placebo (n=12)	Alprazolam (n=14)	P
Detect (speed)	1	6	0.06
Monitor (speed)	3	1	0.3
Identify (speed)	0	4	0.1
Match (speed)	0	6	0.01
Remember one back (speed)	0	4	0.06
Remember one back (acc)	1	4	0.3
Maze learn (efficiency)	0	4	0.1
Sort (acc)	1	4	0.3
Learn (acc)	2	8	0.03

P derived from Fisher's exact probability (two tailed)

Results

Stability of cognitive performance over time

Group means and their relative WSD for the three baselines in the 26 subjects are shown in Table 2. The ANOVAs indicated that no practice effects occurred on any of the measures (F values given in Table 2). The estimates of stability (WSD) and the coefficient of variation for each measure are also shown in Table 2. Multiplication of the coefficient of variation by 100 allows the stability of each performance measure to be expressed as a percentage. The speed measures performance varied less than 5% over assessments. The stability of accuracy measures was a little lower with variance generally around 10%.

Classification of treatment-related cognitive change in individual subjects

Table 3 shows the number cognitive change scores ≤1.96 for each performance measure in the alprazolam and placebo conditions. For no measure did performance improve after either placebo or alprazolam. Compared to the placebo condition, the number of individuals in the alprazolam group showing reliable change on a cognitive measure was significantly greater only for the CogState Match and Learn performance scores. Table 4 shows the number of individuals classified as showing treatment-related cognitive change cognitively impaired (two or more cognitive change scores ≤−1.96 or a standardized composite cognitive change score of ≤−1.96). When treatment-

Table 4 Number (%) of individuals who meet the criteria for impairment in the placebo and alprazolam conditions

	Criterion 1: two or more measures ≤−1.96	Criterion 2: composite ≤−1.96	Criterion 3: either
Placebo (*n*=12)	0	1 (8.3)	1 (8.3)
Alprazolam (*n*=14)	10 (71.4)	13 (92.9)	13 (92.9)

related cognitive change was defined as cognitive change scores of ≤−1.96 on two or more measures, 71% of individuals in the alprazolam condition (hits) and no subjects in the placebo condition (false positive) met the criteria for impairment. When treatment-related cognitive change was defined as a composite cognitive change score of ≤1.96, the hit rate increased to 93% although this was accompanied by an increase in the false positive rate to 8% (i.e., one subject in the placebo condition was classified as having impairment). When either criterion was sufficient, these rates remained the same with 93% of subjects in the alprazolam condition and 8% of subjects in the placebo group classified as showing treatment related cognitive change. Finally, the between group comparison of the composite cognitive change scores was significant [$t(28)=5.1$; $p<0.01$] yielding an effect size (Cohen's d) of 2.0 for the magnitude of the difference.

Discussion

The aim of this study was to determine whether a novel but theoretically derived statistical and methodological approach to the classification of cognitive change in individuals could identify cognitive change in individuals who had received an acute 1 mg dose of alprazolam. The dose and compound were selected because their sedative actions are well known and because previous studies had shown that relatively low doses reliably induce cognitive deterioration in healthy adults. Thus, we expected that performance on the cognitive tasks would become worse after treatment. By deriving change scores for each cognitive performance measure and a composite score of all measures, statistically reliable cognitive impairment was identified for 93% of individuals who had received alprazolam. The same method and analysis identified only 8% of individuals who had received only placebo. Therefore, statistical rules that required change on a subset of change scores derived from the nine performance measures or from a composite score computed from the sum of change scores from all nine measures showed excellent sensitivity to cognitive change while maintaining false classification rates at acceptable levels. Thus, the current method and analysis showed an excellent sensitivity and specificity for alprazolam related cognitive deterioration in individuals. These findings, although preliminary, suggest that statistically reliable decisions about the effects of benzodiazepine drugs on cognition can be made for individuals.

The reliable identification of cognitive impairment in individuals required both the statistical methods described in this study and the methodological soundness of the study design. First, the cognitive tasks used in the study were chosen because they have been designed for rapid administration without giving rise to practice effects (Collie et al. 2003a,b; Falleti et al. 2003, Mollica et al. 2005). Therefore, the test battery could be given repeatedly at relatively short retest intervals (days) to allow estimation of performance stability before randomization. The multiple baselines also allowed a stable estimate of baseline performance to be estimated, against which the effect of alprazolam could be compared. The performance measures derived from the tasks used in this study were chosen to measure cognitive change because they have been shown to be distributed normally (after transformation), not limited by floor or ceiling effects, not subject to range restriction, and contain no skew (Mollica et al. 2004). These good metric properties and the repeatability of tasks meant that each performance measure chosen showed good stability over time according to the method of Bland and Altman (1996). In fact, each estimate of variation in performance over time within individuals (i.e., WSD) was systematically smaller than the corresponding estimates of variability in performance between individuals at any one time (i.e., the group standard deviation on any one assessment, see Table 2). In the past, some studies seeking to identify individuals for whom significant posttreatment cognitive change has occurred have used the pretreatment group standard deviation as an estimate of population stability for their statistical decision rules (e.g., CABG studies, see Collie et al. 2003a,b; for a review). This method may be the only available for estimating of normal variability over time when there is a single pretreatment assessment. However, had we have used the group baseline standard deviation to estimate the background variability in cognitive test performance, then the sensitivity of the statistical analysis and method would have been reduced considerably. This observation emphasizes the importance of multiple baseline or pretreatment assessments.

A group analysis of this same data indicated that performance on all of the measures used in this study was significantly worse than baseline after the single acute dose of 1 mg of alprazolam, whereas there were no significant changes in performance over the same assessments for group who received placebo (Snyder et al. 2005). For the alprazolam group, the magnitude of deterioration in performance was largest for the measures of attention (effect sizes of approximately 0.8). Deterioration in memory and executive functions was not as severe (effect sizes of approximately 0.5). Interestingly, unlike this analysis, and indeed other group studies of the effects of alprazolam (Snyder et al. 2005; Verster et al. 2002; Vermeeren et al. 1995), the analysis of individual performance identified no performance measures that were uniquely sensitive to the effects of alprazolam on cognitive function. For example,

compared to the placebo condition, individuals treated with alprazolam showed greater rates of abnormality on each of the cognitive measures used. However, this difference was significant only for measures of speed on the detection and matching tasks and for accuracy on the learning task. Even for the measures with a significantly greater sensitivity to alprazolam-related cognitive change than placebo, the rate of abnormality was never greater than 53% (eight subjects). Despite this nonspecificity, 71% of individuals met the criterion for cognitive change that required that abnormal change to occur on two or more of the performance measures. This indicates that alprazolam-related cognitive impairment occurred on different performance measures in different individuals. This nonspecificity for individuals is also reflected in the finding that the composite change score criterion was more sensitive to the effects of alprazolam than the two abnormal score criterion. Thus, it appears that alprazolam did cause performance on each of the tasks to deteriorate, although the magnitude of this deterioration varied across performance measures in different individuals. In the majority of cases, the magnitude of deterioration was not statistically significant (i.e.,<−1.96). However, when summed across performance measures, these subtle but systematic changes combined to yield statistically significant deterioration in the composite cognitive change scores for all but two individuals who had received alprazolam and only one individual who had received placebo. One likely reason for the qualitatively different patterns of abnormality across the performance measures between the group and individual analyses is that the superior statistical power of hypotheses tested for groups allows a more detailed analysis of the effect of the drug on different tests. At the individual level, there are fewer observations of performance and therefore greater potential for error to obscure results. In this case, the composite score acts to maximize the statistical power by concentrating all of the observations conducted on the same individual into a single measure. Interestingly, when the composite score for each individual was treated as an outcome measure and compared in a group analysis, the magnitude of the difference detected (d=2) was much larger than the difference between groups for any of the individual measures (see Snyder et al. 2005). This suggests that the use of the composite score may also bring greater statistical power in the analyses of cognitive change in groups.

Despite the sensitivity of the cognitive tasks and the analytic method, two individuals in the alprazolam group did not show the expected deterioration in cognitive function (Fig. 1). This could reflect either error in the performance measures or the cognitive tasks from which they were derived, error in statistical criteria used, noncompliance with the pre- or posttreatment testing requirements, or that these individuals' cognitive function truly did not change in response to the drug. We believe the latter explanation to be correct. In both of these individuals, the composite score was positive (Fig. 1), suggesting that cognitive performance improved with alprazolam. Inspection of their raw data indicated no unusually slow or variable responses or unusually slow performance that would suggest lack of understanding, low motivation, or noncompliance. Furthermore, neither of these subjects showed significant cognitive deterioration on any one of the nine performance measures. Therefore, it appears that these two subjects are true nonresponders, at least after 1 mg alprazolam. Clinical studies in patients with panic disorder suggest that nonresponse rates to alprazolam can be as high as 20% (Verster and Volkerts 2004). Of course, this is decided on clinical variables rather than cognitive outcomes as used in this study. However, the method developed in this study could be applied to any data, including that from standardized clinical rating scales. Such an application may provide a more rigorous basis to assist clinical decisions of nonresponse or response. Because most psychopharmacological studies of alprazolam have compared groups of subjects, there are no estimates of the rate of nonresponders. It is possible that increases in dose would lead the cognitive nonresponders identified in this study to show some cognitive deterioration. It is important to note that the aim of the current study was to validate the method for detecting and classifying cognitive change in individuals. Therefore, further study is

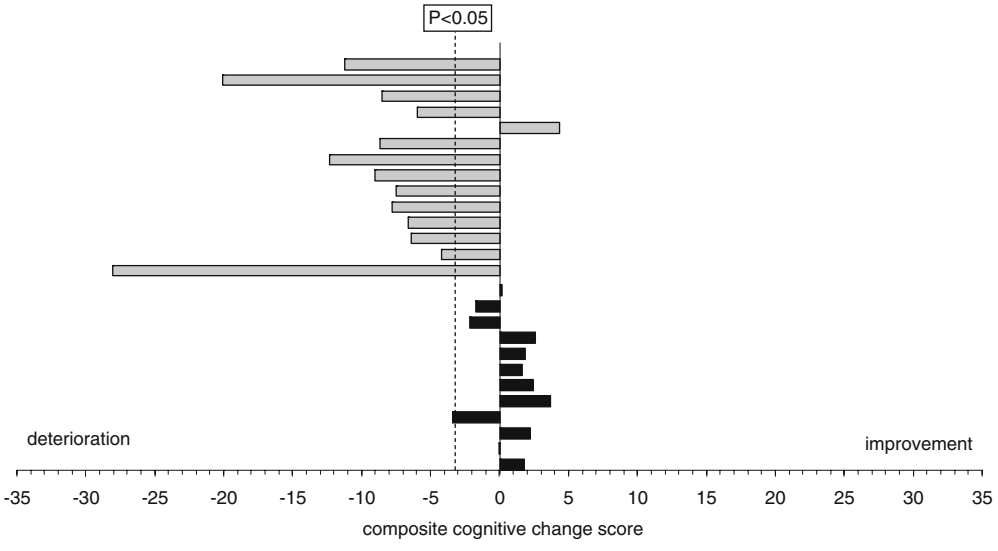

Fig. 1 Distribution of raw composite cognitive change scores in the placebo (*black bars*) and alprazolam (*grey bars*) conditions. *Dotted line* indicating p<0.05 (two tailed) is set at the mean composite cognitive change score in the placebo group mean minus 1.96 SD units, also derived from the placebo group

required to determine whether the nonresponders identified in this study would show cognitive deterioration at higher doses of alprazolam or whether a cognitive endpoint might also be useful in determining treatment response to alprazolam in individuals with anxiety disorder.

The methodology applied in this study also classified one individual who received placebo as showing cognitive decline. This positive classification (8%) was similar with the error rate of the classification rules used (e.g., $p<0.05$). Had the score required for abnormality been lowered further to avoid this false positive classification (e.g., to $p<0.01$), then the sensitivity of the method to alprazolam-related cognitive decline would have also decreased (e.g., at $p<0.01$ the hit rate for the composite score was 11/14 or 78.5%). Therefore, it appears that the criteria chosen for abnormality provided an appropriate balance between sensitivity and specificity to alprazolam-related cognitive decline in healthy adults.

In a previous study we applied these same cognitive tests and statistical methods to measure treatment response in a group of children with attention deficit disorder who had been treated with 7.5 mg of D-amphetamine (Mollica et al. 2004). It is worth considering this study for two reasons. First, the aim of the study of Attention Deficit Hyperactivity Disorder was to detect improvement in cognitive function in the children after treatment (as opposed to the cognitive deterioration expected in this study). The results of these two studies suggest strongly that the CogState test battery, and the statistical method developed in this study can be used to identify both positive and negative treatment response. Second, in the study of treatment response to D-amphetamine, we applied the statistical method developed in this study to selected clinical rating scales and found a high degree of agreement between classifications of cognitive response and symptomatic response. These data illustrate that the method described has good clinical validity and can be applied to other types of outcome measure. There were a number of limitations operating in the current study. First, this was a reanalysis of previous data in which the results were known. Hence, it is necessary to now apply the method detailed in this study in a de novo study of another compound known to alter cognitive function. Second, the current study used a parallel group design rather than a complete crossover. Although this would be unlikely to have affected the results obtained in this study, a crossover design would require the subjects to complete twice as many reassessments (i.e., in both treatment conditions) and to provide two baselines. This should improve further the estimates of stability of the pretreatment performance. Third, the current method should be applied to a study with a larger pool of subjects, ideally those who also vary on characteristics that are known to influence treatment response (e.g., weight, gender, genotype, etc.). The method for classifying cognitive change developed in this study should be sensitive and specific in these newer studies, but outcome measures such as the composite score should covary in a predictable manner with these factors in the treatment group. Nevertheless, despite these limitations, the results of the current study provide a firm basis for refinement through further challenge.

In conclusion, a novel methodological and statistical development to classify treatment-related cognitive decline in individuals did adequately identify those who showed true cognitive change, while keeping false positive classification rates at acceptably low levels. When the cognitive performance of individuals alone was examined, measures that contain good metric properties provide reliable estimates of performance. With measures that contain metric properties appropriate for measuring change, certain statistical methods can be applied to determine whether cognitive impairment has occurred at the individual level. As the results demonstrate, more individuals were classified as impaired following 1 mg alprazolam compared to placebo, and this was most evident in the composite cognitive change scores that were derived across all the performance measures, rather than the single scores derived for each measure. This suggests that for future studies, the calculation of a composite cognitive change score across appropriate measures is one method that could be applied.

References

Anastasi A, Urbina S (1997) Psychological Testing, 7th edn. Prentice Hall, Upper Saddle River, NJ

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. Br Med J 310:170

Bland JM, Altman DG (1996) Measurement error. Br Med J 313:744

Calabrese JR, Vieta E, El-Mallakh R, Findling RL, Youngstrom EA, Elhaj O, Gajwani P, Pies R (2004) Mood state at study entry as predictor of the polarity of relapse in bipolar disorder. Biol Psychiatry 56:957–963

Collie A, Darby DG, Falleti MG, Silbert BS, Maruff P (2002) Determining the extent of cognitive change after coronary surgery: a review of statistical procedures. Ann Thorac Surg 73:2005–2011

Collie A, Maruff P, Darby DG, McStephen M (2003a) The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc 9:419–428

Collie A, Maruff P, McStephen M, Darby DG (2003b) Psychometric issues associated with computerised neuropsychological assessment of concussed athletes. Br J Sports Med 37:556–559

Darby D, Walsh K (2005) Walsh's Neuropsychology: a clinical approach. 5th edn. Elsevier Churchill Livingstone, London

de Visser SJ, van der Post JP, de Waal PP, Cornet F, Cohen AF, van Gerven JM (2003) Biomarkers for the effects of benzodiazepines in healthy volunteers. Br J Clin Pharmacol 55:39–50

D'Souza DL, Levasseur LM, Nezamis J, Robbins DK, Simms L, Koch (2001) Effect of alosetron on the pharmacokinetics of alprazolam. J Clin Pharmacol 41:452–4554

Falleti MG, Maruff P, Collie A, Darby DG, McStephen M (2003) Qualitative similarities in cognitive impairment associated with 24 h of sustained wakefulness and a blood alcohol concentration of 0.05%. J Sleep Res 12:265–274

Guyatt GH, Walter SD, Norman GR (1997) Measuring change over time; assessing the usefulness of evaluative instruments. J Chronic Dis 40:171–178

Hinkle DE, Wiersma W, Jurs SG (1996) Applied statistics for the behavioural sciences. Houghton Mifflin, Boston

Ingraham LJ, Aiken CB (1996) An empirical approach to determining criteria for abnormality in test batteries with multiple measures. Neuropsychology 10:120–124

Jacobson NS, Truax P (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 51:12–19

Jenkinson C (1995) Evaluating the efficacy of medical treatment; possibilities and limitations. Soc Sci Med 74:68–80

Lesch KP, Gutknecht L (2004) Focus on The 5-HT1A receptor: emerging role of a gene regulatory variant in psychopathology and pharmacogenetics. Int J Neuropsychopharmacol 7:381–385

Macher J-P, Corcq M-A (2004) Treatment goals: response and nonresponse. Dialogues Clin Neurosci 6:83–91

Mollica C, Maruff P, Vance A (2004) Statistical method for detection of cognitive improvement in individual children with ADHD-CT. Hum Psychopharmacol 19:445–456

Mollica CM, Maruff P, Collie A, Vance A (2005) Repeated assessment of cognition in children and the measurement of performance change. Neuropsychol Dev Cogn C Child Neuropsychol 11:303–310

Rasmussen LS, Larsen K, Houx P, Skovgaard LT, Hanning CD, Moller JT, the ISPOCD group (2001) The assessment of postoperative cognitive function. Acta Anaesthesiol Scand 45:275–289

Redelmeier DA, Tversky A (1990) Discrepancy between medical decisions for individual patients and for groups. N Engl J Med 322:1162–1164

Schroder MD, Snyder PJ, Sielski I, Mayes L (2004) Impaired performance of children exposed in utero to cocaine on a novel test of visuospatial working memory. Brain Cogn 55:409–412

Snyder PJ, Werth J, Giordani B, Caveny A, Feltner D, Maruff P (2005) A method for determining the magnitude of change across different cognitive functions in clinical trials: The effects of acute administration of two different doses alprazolam. Hum Psychopharmacol 20:263–273

Stewart SA (2005) The effects of benzodiazepines on cognition. J Clin Psychiatry 66(Supp 2):9–13

Vermeeren A, Jackson JL, Muntjewerff ND, Quint PJ, Harrison EM, O'Hanlon JF (1995) Comparison of acute alprazolam (0.25, 0.50 and 1.0 mg) effects versus those of lorazepam 2 mg and placebo on memory in healthy volunteers using laboratory and telephone tests. Psychopharmacology (Berl) 118:1–9

Verster JC, Volkerts ER (2004) Clinical pharmacology, clinical efficacy, and behavioral toxicity of alprazolam: a review of the literature. CNS Drug Rev 10:45–76

Verster JC, Volkerts ER, Verbaten MN (2002) Effects of alprazolam on driving ability, memory functioning and psychomotor performance: a randomized, placebo-controlled study. Neuropsychopharmacology 27:260–269

Williams JL, Naylor CD (1992) How should health status be assessed? Cautionary notes on procrustean frameworks. J Clin Epidemiol 45:1347–1351

Wong SL, Locke C, Staser J, Granneman GR (1998) Lack of multiple dosing effect of sertindole on the pharmacokinetics of alprazolam in healthy volunteers. Psychopharmacology 135:236–241