Detection of Postoperative Cognitive Decline After Coronary Artery Bypass Graft Surgery is Affected by the Number of Neuropsychological Tests in the Assessment Battery

Matthew S. Lewis, BAppSc Hons, Paul Maruff, PhD, Brendan S. Silbert, MBBS, FANZCA, Lis A. Evered, BS, and David A. Scott, MBBS, FANZCA, PhD

Background. The assessment of postoperative cognitive dysfunction after coronary artery bypass graft surgery is made with the repeated administration of cognitive tests. This classification is vulnerable to error, and it has been suggested that increasing the number of tests in a battery while maintaining constant inclusion criteria for postoperative cognitive dysfunction increases the rate of false positive classification of deterioration. The current study tested this by applying a constant rule for cognitive dysfunction using combinations of two to seven cognitive tests.

Methods. Two hundred and four coronary artery bypass graft patients (surgical) and 90 healthy nonsurgical controls aged 55 years or older completed a battery of cognitive tests at baseline (preoperative) and 1 week later (postoperative). In both groups, postoperative cognitive dysfunction was classified using all unique combinations of two to seven cognitive tests when performance deteriorated on two or more tests by at least the value of the baseline standard deviation.

Results. The average incidence of cognitive dysfunction progressively increased in both groups as the number of cognitive tests increased from two (surgical: 13.3%; control: 3.1%) to seven tests (surgical: 49.4%; control: 41.1%).

Conclusions. Increasing the number of tests used to classify postoperative cognitive dysfunction appears to increase the sensitivity to change in the coronary artery bypass graft group. However, accompanying false positive classifications suggest that this improved sensitivity reflected increased error. Future rules for postoperative cognitive dysfunction need to account for this error and include a control group.


A fter coronary artery bypass graft (CABG) surgery some patients demonstrate subtle cognitive decline relative to their preoperative performance. As the cognitive systems affected most in postoperative cognitive dysfunction (POCD) have not been identified, consensus statements recommend that the neuropsychological evaluation of patients undergoing CABG assess a wide range of cognitive functions before and after surgery. However, this recommendation must be balanced against clinical constraints such as the medical management of the patient and the limited time available for assessment. Consequently there has been considerable variation in the composition of neuropsychological batteries used to classify POCD in CABG with some including as few as one test and others as many as 27 tests. Although batteries containing larger numbers of neuropsychological tests are considered to have greater sensitivity to POCD, one unwanted consequence of increasing the number of tests in a battery is that the probability of falsely classifying cognitive change (ie, type I error) also increases. Thus, studies of CABG that use large neuropsychological test batteries are likely to classify some individuals with POCD when they were normal. Some researchers have acknowledged that false positive error increases with the number of tests used in test batteries and have sought to minimize this risk by decreasing the number of cognitive endpoints through the use of theoretically or statistically derived composite or factor scores. However, no study using composite scores to classify POCD has demonstrated that the modification to the number of outcome values reduced the associated false positive error rate to acceptable levels (<5%). A less understood, but more reliable method to reduce type I error associated with the classification of POCD is to alter the POCD criteria to account for the level of performance change classified as abnormal on individual tests, the number of tests or outcome measures in the battery, and whether the hypothesis is one- or two-tailed.

Accepted for publication Jan 10, 2006.

Address correspondence to Dr Lewis, Department of Anaesthesia, St. Vincent’s Hospital, Victoria Parade, Melbourne 3065, Australia; e-mail: matt.lewis2@gmail.com.

© 2006 by The Society of Thoracic Surgeons

Published by Elsevier Inc
doi:10.1016/j.athoracsur.2006.01.044
Aiken 7 provided a theoretical illustration of this using simulations based on binomial probability equations. These simulations demonstrate that while type I error rate increases with the number of neuropsychological tasks in a battery, the statistical rules used to classify abnormality can be modified to maintain an acceptable level of type I error (ie, <5%) as the number of tests increase. For example, a classification of abnormality that requires performance to be one standard deviation (SD) below a control mean on two tests on a battery of only two tasks has a false positive rate of less than 5%. The false positive error rate associated with this same classification increases to 70% if the number of tests in the battery is increased to 15. However, this very high probability of false positive classification can be reduced to less than 5% by changing the criteria for abnormality to require performance to be two SD below the control mean on two tests.

Although relevant, the applicability of the Ingraham and Aiken simulations to the classification of POCD is unclear. First, the simulations assume that each neuropsychological measure is independent 7, a condition unlikely to exist in most neuropsychological test batteries. Second, the simulations reflect decisions about abnormality when individual performance is compared with a normative range while the classification of POCD is usually based on decisions about the change in performance on multiple measures between baseline and post-surgical assessments.

In the current study, we used seven cognitive tests taken from the AustraliaN Trial Investigating Post Operative Deficit, Early extubation and Survival (ANTIPODES), a prospective randomized controlled trial investigating the cognitive impact of early extubation after CABG, and cognitive data from a separate nonsurgical, healthy control group. Postoperative cognitive dysfunction was classified if performance on two or more tests declined by more than the value of the baseline SD. This rule is commonly used in POCD research 10–12 and allows direct comparison to the published simulations of type I error rates which also used the same rule 7. The rate of POCD was then computed for all possible unique combinations of neuropsychological tests when the number of the tests in the battery was systematically varied from two to seven in both the CABG and a healthy control group.

### Patients and Methods

#### CABG Patients

Two hundred and four patients aged 55 years or older scheduled to undergo first time elective CABG surgery were drawn from two sites involved ANTIPODES. Patients were ineligible if they had poor ventricular function, (ejection fraction < 30%) associated major systemic illness, preexisting neurologic disease, or anticipated difficulty with neuropsychological assessment (eg, difficulty with eyesight or hearing, poor English comprehension, hemiparesis). Demographic details are shown in Table 1.

Conventional surgical techniques were used. Anesthesia consisted of temazepam premedication, midazolam, fentanyl (either 10 μg/kg or 50 μg/kg), propofol, and rocuronium. No inhalational anesthetics were administered. Arrest of the heart was managed using a combination of antegrade and retrograde tepid blood cardioplegia without active cooling. Proximal anastomoses were performed under aortic cross-clamping. Cardiopulmonary bypass was undertaken with a membrane oxygenator and roller or centrifugal pump with continuous flow of 2.0 to 2.4 L · min⁻¹ · m⁻², moderate hypothermia (32°C to 34°C), and mean systemic pressures maintained at 60 to 80 mm Hg. Further surgical details can be viewed in Table 2.

#### Controls

Ninety healthy control participants completed the neuropsychological test battery on two consecutive occasions, at baseline and 1 week later (mean, 7 days; range, 5 to 15). Control subjects were recruited as part of our ongoing studies in CABG. All were recruited from the general public or from among spouses or friends of patients undergoing CABG. Before entry in the study each control was assessed by a neurologist and a neuropsychologist to ensure that they were cognitively normal and neurologically healthy. Each control participant also underwent a general medical examination to ensure that they were free from cardiovascular disease and other systemic illnesses. Control participants were excluded if they had a history of respiratory, circulatory or endocrine

### Table 1. Demographic Details of the Coronary Artery Bypass Graft Patients and Control Participants

<table>
<thead>
<tr>
<th></th>
<th>Control Group</th>
<th>CABG Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>68.8 (7.0)</td>
<td>67.8 (7.9)</td>
</tr>
<tr>
<td>Sex (male/female)</td>
<td>65/25</td>
<td>152/52</td>
</tr>
<tr>
<td>NART estimated IQ</td>
<td>108.9 (11.0)</td>
<td>106.4 (18.4)</td>
</tr>
</tbody>
</table>

* Data are reported as mean (standard deviation).

CABG = coronary artery bypass graft; NART = National Adult Reading Test.

### Table 2. Surgical Data for Coronary Artery Bypass Graft Patients (n = 204)

<table>
<thead>
<tr>
<th></th>
<th>Median (IQR)</th>
<th>Range (Min:Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number grafts</td>
<td>3 (1)</td>
<td>1:6</td>
</tr>
<tr>
<td>Cross-clamp time</td>
<td>78 (27.3)</td>
<td>32:132</td>
</tr>
<tr>
<td>CBP time</td>
<td>98 (34)</td>
<td>43:180</td>
</tr>
<tr>
<td>OPCABG (n)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Stroke/mortality</td>
<td>1 CVA, day 7 postoperative (no residual symptoms); 2 deaths, 1 on postoperative day 2, the other within 3 months postoperatively</td>
<td></td>
</tr>
</tbody>
</table>

CBP = cardiopulmonary bypass; CVA = cerebrovascular accident; IQR = interquartile range; min = minimum; max = maximum; OPCABG = off-pump coronary artery bypass graft.
disease, personal or family history of psychiatric illness, head injury or substance abuse, Mini Mental Status Examination score less than 27, or if English was not the primary language (see Table 1).

Procedure

After Institutional Ethics approval (March, 2001) and informed consent, participants completed a neuropsychological assessment battery at two time points: baseline (preoperative assessment) and again 1 week later (postoperative day 6—typically the day of discharge in the CABG group). At baseline, all participants completed the National Adult Reading Test (NART) to provide an indication of each patient’s IQ.

At both times, participants completed a series of neuropsychological assessments yielding seven outcome variables: the CERAD (Consortium to Establish a Registry for Alzheimer’s Disease) word learning task (WLT [total number of words recalled on immediate trials, maximum score 30]), the Trail Making Task Part A (TM TA [time (seconds) required to complete]), the Trail Making Task Part B (TMTB [time (seconds) required to complete]), the Digit Symbol Substitution Task (DSST [number of symbols correctly transcribed in 90 seconds]), the Controlled Oral Word Association task (COWAT [total number of words generated for the three 60-second letter presentations]), the Grooved Peg Board Task, dominant hand (GPD [number of seconds to completion]), and the Grooved Peg Board Task, nondominant hand (GPND [number of seconds to completion]). The tasks were administered according to protocol by trained investigators and are comprehensively described elsewhere.

Table 3. Correlations Between Neuropsychological Tests for the Coronary Artery Bypass Graft Group and Control Group

<table>
<thead>
<tr>
<th></th>
<th>Baseline Neuropsychological Data</th>
<th>Day 7 Neuropsychological Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSST TMTA TMTB COWAT GPD GPND</td>
<td>DSST TMTA TMTB COWAT GPD GPND</td>
</tr>
<tr>
<td>Coronary Artery Bypass Graft Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLT</td>
<td>0.377* -0.228* -0.293* 0.400* -0.174* -0.201*</td>
<td>0.554* -0.482* -0.479* 0.466* -0.439* -0.449*</td>
</tr>
<tr>
<td>DSST</td>
<td>-0.573* -0.630* 0.366* -0.438* -0.340*</td>
<td>-0.672* -0.677* 0.469* -0.541* -0.541*</td>
</tr>
<tr>
<td>TMTA</td>
<td>0.628* -0.323* 0.353* 0.367*</td>
<td>0.812* -0.460* 0.653* 0.558*</td>
</tr>
<tr>
<td>TMTB</td>
<td>-0.335* 0.349* 0.315*</td>
<td>-0.431* 0.616* 0.484*</td>
</tr>
<tr>
<td>COWAT</td>
<td>-0.148b -0.124</td>
<td>-0.317* -0.339*</td>
</tr>
<tr>
<td>GPD</td>
<td>0.850*</td>
<td>0.726*</td>
</tr>
<tr>
<td>Control Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLT</td>
<td>0.104 0.040 -0.141 -0.089 0.173</td>
<td>0.206 0.030 -0.044 -0.079 -0.131 -0.076 0.121</td>
</tr>
<tr>
<td>DSST</td>
<td>-0.246b -0.663* 0.184 0.014 0.014</td>
<td>0.064 -0.292* -0.197 0.131 0.162 -0.020</td>
</tr>
<tr>
<td>TMTA</td>
<td>0.296* 0.027 -0.120 -0.079</td>
<td>0.335* 0.078 0.039 0.443*</td>
</tr>
<tr>
<td>TMTB</td>
<td>-0.207 -0.192 -0.155</td>
<td>-0.059 -0.040 -0.095</td>
</tr>
<tr>
<td>COWAT</td>
<td>-0.111 -0.247b</td>
<td>0.028 0.135</td>
</tr>
<tr>
<td>GPD</td>
<td>0.475*</td>
<td>0.229b</td>
</tr>
</tbody>
</table>

* Correlation is significant at the 0.01 level (two-tailed).  
  b Correlation is significant at the 0.05 level (two-tailed).


Data Analysis

For all analyses, the direction of data was corrected so that positive changes indicated improvement, whereas negative changes indicated deterioration.

INTERCORRELATION BETWEEN TESTS. To determine the way that individual neuropsychological tasks were intercorrelated, Pearson’s r analysis was conducted on the neuropsychological data at both baseline and day 7 assessments. These are shown in Table 3.

GENERATION OF INCIDENCE CURVES. Data analysis comprised two stages. The first stage involved classifying the rate of abnormal cognitive decline for each neuropsychological test in the CABG and control groups. For each task, abnormal cognitive decline was classified if performance had declined by 1 SD or more of the baseline performance for that group (see Table 4). Postoperative cognitive dysfunction was classified if the patient showed decline of 2 or more tasks. The incidence of abnormal cognitive decline detected on individual tests is shown in Table 4. The term “abnormal cognitive decline” has been used throughout this manuscript to signify a deterioration in performance on a given task as indicated by the statistical criteria outlined above.

As the values in Table 4 indicated that the individual tests had variable sensitivity to cognitive change, the second stage involved computing the incidence of POCD classified for each unique combination of two to seven neuropsychological tests administered in the...
CABG and control groups. The total number of unique possible neuropsychological test combinations expressed algebraically as \((n^2) / [(n-y)^2y^2]\), where \(n\) = total number of tests in the battery, and \(y\) = number of tests removed. With all seven tests, there was one possible combination; for six of seven tests, seven unique combinations; for five of seven tests, 21 unique combinations; for four of seven tests, 35 unique combinations; and for three of seven tests, 35 possible combinations. Each unique test combination yielded an incidence rate of POCD (see Fig 1), which was averaged for each combination of tests of a specific number. The mean and SD of the CABG and control group incidence rates are shown in Table 5 and have been graphed together in Figure 2.

### Table 4. Standard Deviations and Incidence of Decline Detected for the Coronary Artery Bypass Graft and Control Group on Individual Neuropsychological Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Baseline Standard Deviation</th>
<th>Dysfunction Incidence Rates on Individual Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>CABG</td>
</tr>
<tr>
<td>WLT</td>
<td>3.8</td>
<td>4.0</td>
</tr>
<tr>
<td>DSST</td>
<td>10.3</td>
<td>10.0</td>
</tr>
<tr>
<td>TMTA</td>
<td>12.9</td>
<td>23.8a</td>
</tr>
<tr>
<td>TMTB</td>
<td>21.9</td>
<td>61.5a</td>
</tr>
<tr>
<td>COWAT</td>
<td>9.5</td>
<td>12.1b</td>
</tr>
<tr>
<td>GPD</td>
<td>13.0</td>
<td>34.5a</td>
</tr>
<tr>
<td>GPND</td>
<td>16.5</td>
<td>40.0a</td>
</tr>
</tbody>
</table>

*Levene’s test \(p < 0.001\).  
*Levene’s test \(p < 0.01\). Mann-Whitney analysis, control versus CABG control SD  
\(^c \ p < 0.05\) and  
\(^d \ p < 0.001.\)

CABG = coronary artery bypass graft group;  
CABG–CABG SD = CABG group with CABG group baseline SD;  
CABG–Cont SD = CABG group with control baseline SD;  
COWAT = Controlled Oral Word Association Task;  
DSST = Digit Symbol Substitution Task;  
GPD = Grooved Pegboard Dominant Hand Condition;  
GPND = Grooved Pegboard Non Dominant Hand Condition;  
TMTA = Trail Making Task Part A;  
TMTB = Trail Making Task Part B;  
WLT = Consortium to Establish a Registry for Alzheimer’s Disease (CERAD) Word Learning Task.

---

**Fig 1.** Incidence of decline (percentages) detected using every unique combination of neuropsychological tests for each number of tests. (CABG = coronary artery bypass graft; SD = standard deviation.)
Results

Intercorrelations

All neuropsychological tests were correlated moderately and significantly with one another at both the baseline and day 7 assessments in CABG patients. This pattern of intercorrelations was not observed in the control group at either assessment (Table 3).

Calculation of Cognitive Dysfunction on Individual Tests

The CABG group baseline SD was larger than the control group baseline SD for all neuropsychological tests excepting the DSST. Levene’s test indicated that the differences in baseline SD between the two groups were significant for all tasks with the exception of the WLT and DSST. For the TMTB, GPD, and GPND, the CABG SD was more than two times that of the control group (see Table 4). As the SDs for the neuropsychological tasks differed between the CABG and control groups, a second calculation of dysfunction was conducted using the control group’s SD as the threshold for abnormal cognitive decline for CABG patients (CABG-cont), allowing the change to be expressed in the same terms for both groups. The incidence rates for cognitive dysfunction in the CABG group using the CABG group baseline SD are included, but statistical analysis focused on the data calculated using the control baseline SD as the common impairment threshold.

Table 4 shows the incidence of abnormal cognitive decline detected by each of the seven neuropsychological tests for the control group, the CABG group using the CABG baseline SD (CABG) and the control group’s baseline SD (CABG-cont). The CABG-cont criteria classified a higher incidence of abnormal cognitive decline on all tasks than the control group. Mann-Whitney U tests showed that the CABG-cont group rates of abnormal cognitive decline were significantly greater than for controls for the TMTA, TMTB, and GPND tasks (see Table 4). These tests showed the greatest sensitivity and specificity to POCD in the CABG-cont group as the false positive classification in the control group was low, and the rates of abnormal cognitive decline in the CABG-cont group were high.

Effect of Neuropsychological Test Battery Size and Detected POCD

The rates of POCD computed for each combination of tests for batteries of two to seven tests in the control and CABG groups are shown in Table 5. As the number of tests in a battery was systematically increased there was an increase in the detection of POCD. In the CABG group, the incidence of POCD detected with two tasks was 13.3% and increased up to 49.4% using all seven tasks. In the control group, the classification of POCD with two tasks was 3.1% and increased to 41.1% with all seven tasks included. For batteries of two to six tasks, the incidence of POCD significantly differed between the CABG-cont and control groups. With seven tasks, Mann-Whitney U analysis indicated that this difference was not significant. Figure 2 graphs the POCD incidence curves against the number of tests for CABG and control groups, and shows that while both demonstrate an increasing classification of POCD with the number of tasks, the curves have different forms.
Comment

The current study demonstrated that classification of POCD in the CABG population increased as the number of neuropsychological tests used to make the classification increased. However, classification of POCD also increased in healthy controls as the size of the neuropsychological test battery increased. As none of the healthy controls had undergone surgery or demonstrated any new cognitive impairment over the study period, the classification of abnormal cognitive change in this group provided an estimate of the type I error rate associated with the classification rule and number of cognitive tests. For batteries of more than two tests, the type I error rate exceeded accepted levels of error (eg, 5%; \( p < 0.05 \), one tailed). When the classification of POCD in the two groups is considered together, the current data suggest strongly that many POCD cases detected in the CABG group with neuropsychological batteries of more than two tests were in fact type I error. Therefore, while increasing the number of neuropsychological tests in a battery does provide the appearance of greater sensitivity to POCD after CABG, it also results in a greater number of false positive classifications if the statistical rule used to classify POCD is held constant.

In the current study, abnormal cognitive dysfunction classified in the CABG and control groups was based on a rule that required performance on the neuropsychological battery to have declined by at least one SD (drawn from the baseline assessment) on two or more tests. This rule has been used commonly to classify POCD after CABG (eg, Mullges and coworkers 10, Silbert and coworkers 11, and Zamvar and colleagues 12) and allowed comparison to the simulations of Ingraham and Aiken [7]. Interestingly, the SDs in performance at the baseline were not equivalent between controls and CABG patients, indicating that the baseline performance of CABG patients was more variable than that of controls (see Table 4). Given this difference, we used the control group baseline SD as the threshold for change on each neuropsychological measure in both groups, as this allowed the direct comparison of the rates of abnormal cognitive change between the control and CABG groups. Using this threshold, the estimated incidence of POCD ranged from 13% (two tests) to 49% (seven tests) in the CABG sample, and for the controls ranged from 3% (two tests) to 41% (seven tests). While the rate of abnormality was consistently higher in the CABG group than in controls, the progressive increase in the classifications of abnormality as the number of tests increased was qualitatively similar in both groups. The differences between the CABG and control groups calculated from the values presented in Table 5 suggest that the true incidence of POCD ranged from 8% to 17.5%.

The increased probability of classifying POCD with larger numbers of tests in a neuropsychological battery is consistent with theoretical models of the relationship between type I error and the number of tests on which classifications of abnormality are based [7]. Several studies have considered the possibility that using larger numbers of cognitive outcome variables raises the risk of making false positive classifications 6, 8, 9, 16. However, the relationship between number of tests, rate of POCD, and type I error rate are not commonly considered and the actual false positive rate associated with the classification are difficult to estimate. For example, using the 1 SD on two or more task decline criteria and a battery of eight neuropsychological tests, Treasure and colleagues 17 classified POCD in 73% of the sample 8 days postoperatively. Our false positive data associated with the application of the same rule to seven tests suggests that at least 41% of the Treasure and coworkers sample may have been classified incorrectly as having POCD.

One strategy to reduce the likelihood of false positive classifications using large numbers of neuropsychological tests has been to group like tasks under cognitive domains and generate composite scores for each domain [8, 9, 18–20]. Although this will undoubtedly reduce error, the current results suggest that if the classification of POCD was based on a 1 SD change on two or more of the factors, then the type I error rates may remain unacceptably high. To illustrate, Newman and colleagues [19] reduced 11 neuropsychological tests to four cognitive domains and classified POCD if patients deteriorated by more than 1 SD on one or more of these cognitive domain scores. While the impact of the type I error rate will be substantially reduced using four outcome measures rather than 11, our data indicate that the false positive classification rate associated with this strategy with the more conservative 1 SD decline on two or more tests rule may remain unacceptably high at 15.4% (see Table 5), whereas the data of Ingraham and Aiken suggest the error rate could be over 40% [7]. However, the equivalence of applying classification rules for POCD to factor or composite scores relative to individual tests needs to be better understood. In the meantime, we believe that applying the POCD rule to cognitive performance data from a healthy control group is the best method to determine the false positive rates.

Minimizing the false classification of POCD is central to gaining a full understanding of the etiology of POCD, as these studies rely on the accurate differentiation of true POCD from normal cognitive performance. If error exists, either as the false classification of POCD when it has not occurred, or the missed diagnosis of true POCD, the ability to identify reliable biomarkers, clinicopathologic correlations, or even outcome variables is reduced substantially. Ideally, both types of error will be constrained and patients will be classified accurately. If analysis was conducted using the full battery and the current POCD rule, the incidence of POCD would largely consist of false positive classifications. It could be expected that the true incidence of POCD using the 1 SD decline criteria is 8.3% in the current study, and that the rest represents error. The reliability and predictive validity of these comparisons decrease with type I error as the groups are not representing true cases of either normal cognition or POCD.

The effect of type I error on statistical decisions (ie, significance testing) is better understood in group com-
parisons than for the study of change in individuals. For example, in null hypothesis significance testing of group change, researchers will define the level of tolerance they have for type I error and express this as some value of alpha. Setting alpha to the standard 0.05 allows a 5% chance that the observed difference may be due to error and thus a 95% probability that the difference will be correct. Where multiple measures are used and therefore more than one statistical decision being made, the alpha value for each decision is reduced accounting for type I error. An example of this is Bonferroni correction [21], which restricts the effects of type I error by making the acceptance criteria of significance more restrictive as the number of statistical comparisons increase. These same principles apply to statistical decisions about the presence of true cognitive change in individuals. In the case of individual patients, most decisions about abnormal cognitive change recognize that some error is represented in the change data and set a change threshold, such as 1 SD change on two or more tasks, rather than just examining relative performance. However, more needs to be done to minimize type I error.

Because standard protections against type I error in group change studies are not readily applicable to individual change designs, it may be more important to address the impact of type I error in individual change studies as they are less powered and are more prone to error than group designs [22]. These concerns can be addressed utilizing statistical change criteria that accurately reflect all potential sources of error, and that have a demonstrated ability to maintain type I error at acceptable levels so that true change can be accurately classified. Protections can be gained by increasing the threshold for abnormal cognitive decline or classifying POCD on the basis of abnormal cognitive decline of more than the two tasks required to meet the criteria in the current study [7].

The impact of type I error on individual change designs affect other fields where cognitive dysfunction may affect patient subgroups or specific risk groups. For example, in studies of therapeutic agents for psychiatric illness it is known that not all patients successfully “respond” to the medication. In attention deficit hyperactivity disorder, treatment response can be classified on performance change on a neuropsychological test battery given before and after the medication [23]. Similarly, in persons who suffer a head injury resulting from participation in contact sports, cognitive test performance can be compared before and after the injury and guide the management of these athletes, aiding decisions about when they should return to play [24]. In both cases, the detection of cognitive change is based on the application of a statistical rule to individual change on a neuropsychological test battery. We have found that reliable decisions about the presence or absence of true cognitive change in these settings require that the statistical rule and the number of tests be taken into account, and that these can be refined through the application of these same rules to some matched control group where there truly has been no cognitive change.

Using a group of healthy controls to determine the false positive rate associated with the classification rule is similar to the methods used by Rasmussen and colleagues [16]. It was demonstrated that the control group could be used to calibrate the POCD rule so that type I error was constrained to acceptable levels. This group recognized the increased probability of type I error in large neuropsychological test batteries and believed that it was not possible to distinguish true POCD from error without assessing a comparable nonsurgical sample. The criteria were set so POCD would be identified in the worst performing 2.5% of a normative sample. This POCD rule is not directly applicable to all studies, as it may be specific to their neuropsychological assessment battery. However, the methods employed by Rasmussen and colleagues provide a good basis for the development of statistical rules for POCD in future studies.

We recognize that the 1 SD decline rule has been criticized by a number of sources [16, 25, 26], but it was used here because it provided a ready comparison to the theoretical model of Ingraham and Aiken [7] and has been used previously to assess POCD [10–12]. However, the principles that underlie the increased incidence of POCD as the neuropsychological test battery size increased will affect any POCD rule and reiterate the need for researchers to develop and calibrate their POCD rule relative to their own neuropsychological assessment batteries.

Our findings have two important implications for the interpretation and conduct of studies that classify neuropsychological change in individuals. The first is that results of different studies cannot be directly compared even when the statistical criteria for change are the same. Results need to be more cautiously examined relative to the size and composition of the neuropsychological assessment battery, and it needs to be recognized that the application of a consistent statistical POCD rule to test batteries of different sizes will result in different incidences of POCD. Secondly, it is recommended that future studies calibrate their statistical criteria for individual change against a control group to ensure that type I error is maintained at an acceptable level before applying it against an experimental sample.

Funding was provided by the National Health and Medical Research Council, Australia (Project Grant 140510).

References