Measuring and Monitoring Cognition in the Postoperative Period

Lucy H. Piggin[a]

Stanton P. Newman[a*]

[a] City, University of London, United Kingdom (UK)

Corresponding Author

Professor Stanton P. Newman

City University of London,

Northampton Square,

London,

United Kingdom,

EC1V OHB

stanton.newman.1@city.ac.uk

Abstract

It is common for patients of all ages to experience some degree of cognitive disturbance following surgery. In most cases, impairment appears mild and is restricted to the acute post-operative period, resolving steadily and speedily. In a small number of cases, however, deficits may be more pronounced and/or endure for longer periods, significantly delaying recovery and increasing the risk of serious clinical complications. The ability to accurately measure postoperative cognition, and track recovery of function, is an important clinical task. This review explores practical and methodological issues that may confound this process, examining how best to obtain reliable and meaningful measures of cognition before and after surgery. It considers neuropsychological test selection, administration, analysis and interpretation and offers evidence-based practice points for clinicians and researchers.

Keywords: postoperative cognition, recovery, measurement, cognitive testing

Patients of all ages may experience detrimental changes in cognitive function following surgery – changes often referred to as *postoperative cognitive deficits* (POCD). In the absence of any formal clinical definition, this term has been used to describe an array of presentations: from transient cognitive disturbances in the acute postoperative period, to more persistent deficits in the months and years that follow. Despite a long history of empirical study, there is still much about POCD that appears unclear and ill-defined [1]. There remains ambiguity surrounding aetiology, risk factors, incidence rates, timelines, and even the profile of impairments that may be experienced [2,3]. Many of the disparities between studies that have emerged in the literature may be attributable to differences in surgery (e.g. the type of surgery or anaesthesia) or to patient factors (e.g. age or number of comorbidities) [4,5]. Besides these empirical factors, many conceptual and methodological differences between studies have also generated differences in findings. The most important of these being differences in how POCD is defined and measured [3,6,7].

There is a growing consensus that whilst many patients experience some degree of mild cognitive change immediately after surgery, most will see impairments resolve steadily and speedily, leaving only a small number who experience delayed or incomplete recovery over the longer term [8]. It is, however, understood that both early and enduring impairments may have serious clinical implications, with robust evidence linking POCD to prolonged hospital stays [9], increased disability [10], poorer quality of life (QoL) [11], and higher rates of mortality [5]. The ability to accurately measure postoperative cognition and to track recovery of function is, understandably, recognised as an important clinical task [12]. Here we explore practical and methodological issues that may complicate this process, examining approaches to the measurement of cognition before and after surgery.

Definitions

The term POCD applies to newly-acquired, *objectively* discernible cognitive impairments known to have emerged after a surgical procedure: a classification requiring preoperative and postoperative neuropsychological assessment. Formal testing should always be preceded by efforts to identify clinically distinct conditions, including transient confusional states (e.g. postoperative *delirium*). Delirium, as defined by the *Diagnostic and Statistical Manual for Mental Disorders* (DSM-5) [13], is identified as an acute episode of reduced or fluctuating awareness and attention, typically accompanied by diffuse deficits in cognitive function [7]: a state far more likely to occur in the immediate postoperative period, remitting relatively quickly. Any disturbance of cognition associated with delirium is likely to show rapid recovery and is not typically considered to reflect POCD [14]. There remains an interesting issue of the relationship between delirium and POCD; some have suggested an increase in POCD with those experiencing delirium [15,16], whilst others have asserted that no such relationship exits [17].

In the absence of agreed clinical criteria, attempts have recently been made to align POCD with pre-existing definitions of cognitive dysfunction. It has, for example, been suggested that that POCD should be categorised as a form of *mild neurocognitive disorder* (MCD): i.e. a new deficit in one or more neurocognitive domains (e.g. memory, attention) that does not interfere with capacity for independence in everyday activities [13,18,19]. In this model, the temporary cognitive disturbances commonly experienced in the days after would be labelled *delayed neurocognitive recovery*. This would transition to MCD after 30 days, in which time it would be expected that most patients would have seen near-complete recovery of function [18]. Cognitive recovery, in this context, is defined as a return to baseline function or better, the rates of which may vary according to factors such as age and comorbidity [20,21]. There is continued debate as to which definitions of POCD should be accepted and how they might be applied [22].

Cognitive Assessments

Establishing the presence of POCD, and accurately monitoring its trajectory over time, is reliant on appropriate and timely assessment before and after surgery. It is common for patients, especially older patients, to self-report cognitive changes in the days immediately after surgery [23], triggering debate as to whether *subjective* measures should be included in assessments. Informal inquiries about postoperative changes in cognition are certainly to be encouraged and documented; however, caution is urged in the interpretation of patients' own appraisals of cognition, as subjective reports often demonstrate extremely poor correlation with objective performance on standardised tests of cognition [24,25,26]. Historically, subjective evaluations of cognitive status have been far more robustly associated with *psychological* wellbeing after surgery [26,27] (Khatri et al., 1999; Johnson et al., 2002),. Individuals who demonstrate elevated levels of anxiety and depression are significantly more likely to notice and report cognitive change [28,23]. Subjective reports are also known to be highly susceptible to influence from other psychological factors (e.g. expectation and self-esteem) [29], further undermining reliability of this form of assessment.

It is important to recognise that without an accurate preoperative baseline, it is impossible to discern changes in cognition attributable to surgery or to monitor recovery of function. We will therefore focus on measurement of cognition using objective neuropsychological tests, exploring methods for ensuring that assessments are valid and reliable estimates of cognition. These are issues first addressed in an official Statement of Consensus within the field of cardiac surgery [30], which offered recommendations for test selection, procedures, timing of assessments, and interpretation of results – the underlying principles of which are expanded upon here.

Test Selection

Neuropsychological tests are only clinically useful where they are standardised, reliable, and validated measures of cognition. As such, considerable importance is placed on the *selection* of appropriate tests, with particular focus on cognitive specificity and psychometric properties.

Cognitive Domains

Formal tests of cognition can offer scores indicative of general intellectual abilities or allow finer analysis of performance within discrete cognitive domains. As POCD typically presents as relatively mild impairment, often affecting a number of specific abilities, it is advised that assessments allow for detailed analysis [31]. Vast numbers of neuropsychological tests purport to gauge performance within specific domains: namely, memory and learning, attention/concentration, perception, verbal abilities, psychomotor skills, and executive functions [32]. Although POCD may be defined as persistent impairment in one or more cognitive domains, many researchers have been reticent to specify which domains are at greatest risk [31]. The original Consensus Statement made specific recommendation about the minimum domains to test in POCD research, which included verbal memory, attention, and psychomotor ability [30]. There remains significant variability in the domains that are assessed and the tests used to assess them [3,33,34]. Most studies have included measures of memory and learning, attention/concentration, and executive function, whilst verbal abilities, psychomotor skills, and perception appear far less frequently. Tests should be combined to cover all domains recommended in the Consensus Statement.

Number of Tests

Ambitions to assess multiple cognitive domains must be balanced with practical considerations. Given the time-constraints imposed by imminent surgery and potential physical limitations that might also be present, care must be taken to ensure that tests are not overly burdensome. Assessments that are too long or demanding increase the risk that performance will be adversely affected by factors such as pain or fatigue and/or that patients will disengage or reduce effort. Many commonly used batteries may indeed prove too cumbersome and resource intensive for use in routine clinical practice. It is also possible, and may prove useful in some contexts, to use brief screening batteries that include cognitive assessments, especially where these have been designed specifically for ease of administration (e.g. the Postoperative Quality of Recovery Scale [Postop-QRS]) [20]. However, these do only offer a screen of cognitive performance and will need to be followed up with more details assessment of cognition.

It is also important to be aware of the risk of inflating type I errors by over testing. The probability of concluding that a patient has POCD significantly increases as the number of tests increase, with greater numbers of test parameters generating variability in performance that can lead to the attribution of impairment by chance alone: Lewis et al. (2006) found that POCD incidence rose from 13.3% on two tasks to 49.4% on seven tasks when uncorrected analysis was applied to the same data [35]. There have been various attempts made to reduce the risk of type I error by grouping and analysing tests according to cognitive domains [11] and also by applying statistical corrections that take into account the number of tests used [36]; however, caution is urged here, as these techniques often assume that measures are independent. Although there is a tendency to describe cognitive functions as separate entities (e.g. memory, attention), they are in reality highly interrelated abilities; performance on tests purporting to measure one discrete function will almost certainly draw on multiple cognitive resources. Tests administered to measure pre- and postoperative cognition are often highly correlated and thus require cautious interpretation [37].

Test Characteristics

All tests should be fit for purpose in respect to reliability and validity, demonstrating sufficient *sensitivity* (i.e. high probability that even small cognitive changes will be correctly identified in impaired individuals) and *specificity* (i.e. high probability that normal function will be correctly identified in non-impaired individuals; [38, 39], both being important in the measurement of cognitive recovery. As tests must demonstrate sufficient sensitivity to detect relatively small, discrete impairments or improvements, all should be appraised in respect to *floor-effects,* which emerge in tests so difficult that they cause the majority of participants to produce low scores, reducing the likelihood of impairment being detected, and *ceiling-effects,* seen in tests that are so simple that most patients will achieve high scores and be able to reproduce these scores even if some degree of impairment occurs [40]. Both floor- and ceiling-effects reduce variability and thus discriminability: scores become artefacts of the test rather than a reflection of patient ability.

Unlike many tests used in routine clinical practice, those used to measure postoperative function will also be subject to repeated administration and must be sufficiently sensitive to detect even small change. It is important that tests demonstrate appropriate test-retest reliability, especially over the typically shorter intervals between testing used to monitor postoperative recovery [20]. Care should also be taken in respect to language and culture [41]: most neuropsychological tests have been developed and standardised in English-speaking populations and may not be appropriate for all patient populations. Any translations made should be validated in healthy control groups.

Huge numbers of tests have been used to assess POCD, with variable levels of reliability and validity (Berger et al., 2015). The Consensus Statement recommends a core battery: the Rey Auditory Verbal Learning Test [42] Trail-Making Tests (Part A and B) [43], and the Grooved Pegboard [44]. Other commonly used tests include Digit Span and Digit Symbol Substitution [45] and The Stroop Test [46] – additions made by the International Study of Postoperative Cognitive Dysfunction (ISPOCD) [4]. These remain robust selections.

Practice Effects

Changes in scores cannot be assumed to reflect true change in cognitive function. Even on seemingly reliable tests, participants often demonstrate improvement with repeated administration, attributable to prior exposure to testing materials or what is termed *practice effects*. Practice effects typically derive from learning that is both *declarative* (e.g. remembering specific items on a test) and *procedural* (e.g. remembering how to do a test). Improvement between tests is to some extent dependent on the number of times the test has been administered and intervals between administrations [47, 48], with learning observed for intervals as long as two years [49]. Practice effects are a particular concern when assessing POCD, where test-retest intervals are typically much shorter than in most clinical contexts.

As practice effects are not typically uniform across measures [50], it is advisable to select tests known to be less susceptible to learning. It is also desirable to select tests with parallel forms (i.e. alternate versions of test material), which are known to significantly reduce learning effects. It should, however, be noted that these equivalent forms address familiarity with *content* only and learning about test *procedures* may still apply, potentially encouraging “test wise” participants to adopt a change in strategy over repeated administrations [51]. It is for this reasons that additional statistical methods for addressing practice effects are often needed, typically incorporating control group data.

Control Groups

Well-matched controls are generally considered the most efficient means of estimating practice effects. These groups are used to gain a measure of average improvement between assessments (i.e. an estimate of practice effects for a given test). Studies that include non-surgical patient controls offer eloquent demonstrations of practice effects in POCD research, highlighting the potential for distortion in interpretation when learning is not taken into account [52]. Control groups may be made up of healthy volunteers, patients with other physical health conditions, or patients with the same condition who have not had surgery, either as a result of clinical decision-making or through personal choice [3]. It is important, however, to recognise that the occurrence of learning will reduce the likelihood and frequency of identifying individuals with POCD.

It has been suggested that change scores from *healthy* controls may not provide entirely appropriate comparison when clinical cases are being assessed: Heaton et al. (2001) suggest potential differences in change scores associated with different physical health conditions, with these unique change scores offering useful clinical information [53]. Although many studies do opt to use patient rather than healthy controls, care must be taken to ensure that patient groups are matched closely. It is possible that any number of extraneous clinical variables may confound results and even surgical and nonsurgical patients with the same clinical condition may differ in respect to the physical and/or psychological factors shaping decisions about proceeding to surgery. Most surgical patients also present with multiple co-morbidities, further complicating the matching process.

There is a reasonable expectation that healthy controls will experience stable cognitive function during the testing period, offering good insight into different types of performance variability not attributable to either physical conditions or events such as surgery. This provides an opportunity for tests and testing procedures to be appraised in relation to other factors (e.g. sensitivity, specificity, the impact of different retest intervals; [20]. In sophisticated designs, multiple control groups (i.e. non-surgical patients and healthy volunteers), may be assessed alongside surgical patients. In all cases, controls should be appropriately matched for gender, age, ethnicity, education, and measures of physical and psychological health.

Patient Variability

Test performance may vary according to a number of additional patient variables, including age, education, intelligence quotient (IQ), disease status, number of previous surgeries, and physical co-morbidities [3,54,55]. It is advisable to consider these variables during assessment and in analysis. Formal assessments of education and/or IQ should be made by short-form tests of IQ or measures that enable estimates of IQ to be made (e.g. the National Adult Reading Test [NART]) [56].

Test Administration

In the case of POCD, the ‘usual rules’ of neuropsychological testing apply and the aim of assessment is to enable patients to achieve optimal performance. It is recommended that assessments are conducted by a suitably qualified and trained individual, administering tests in an appropriate setting and in a standardised manner [30]. Testers should pay close attention to the development of rapport, both to minimise test anxiety and as a means of keeping patients engaged with the assessment/research process – an important concern in longitudinal studies where attrition rates can be high. Ideally, all tests should be performed by the same individual in the same location and at the same time of day. Testing locations should be private, quiet, and free from external distraction [57].

Psychological and Physical Wellbeing

As performance on neuropsychological tests can be adversely influenced by mood states, concomitant assessments of anxiety and depression should be undertaken [30]. Anxiety may be heightened by the imminence of surgery *and* the test situation [58,59]. Mood states can influence test performance directly, impeding cognition, and indirectly (e.g. by reducing motivation to perform). Given the medical context, it is recommended that assessments of physiological and pharmacological factors are also undertaken. This should include measures of pain, fatigue, insomnia, and analgesic medication that may impact on test performance [5,60,61]. Particular care may be needed to address sensory impairments or other physical restraints associated with surgery (e.g. cannulas and infusions in hands that may impede movement and dexterity). Formal tests of functional abilities may also be useful [39,18].

Timing of Assessments

Preoperative Assessments

Without an accurate baseline, postoperative deterioration may go undetected. First assessments should be timed to ensure optimal performance. Assessments performed too close to surgery may be particularly susceptible to the influence of anxiety, whilst tests scheduled alongside pre-surgery checks may be influenced by stress and/or fatigue [39,58]. Although surgical interventions may be scheduled at short-notice, it is recommended, where possible, that baseline assessments occur 1-2 weeks before surgery and, if co-ordinated to coincide with other pre-surgery checks, are performed prior to other tests and procedures [58]. It is recognised that pre-existing illness or injury – in most cases the precipitant of the surgical intervention – may affect performance.

Postoperative Assessments

Timing of postoperative assessments emerges as one of the most significant factors shaping POCD incidence rates following cardiac surgery [62]. Research follow-up periods have ranged from one day to five years post-surgery, capturing changes defined as acute (<1 week), intermediate (<3 months), and long-term (≥1-2 years) [3, 63]. In the context of recovery, assessments typically occur at much shorter intervals such as hours and days after surgery [21]. A focus on cognitive aspects of recovery may be more readily prioritised once the patient is physiologically stable and being evaluated for home-readiness [64].

As POCD typically demonstrates gradual improvement over time, it is recommended that postoperative tests are performed within the first week: a time when the risk of POCD is greatest [3,65]. Some degree of impairment appears to affect almost all patients during this acute postoperative period, resolving faster in younger patients [66]. It provides an opportunity to look at individual dfferences in cognition a a response to surgery. As most surgical patients are discharged within one week [67], it is recommended that tests are re-administered prior to discharge; this is likely to be a point at which there is a reasonable expectation that most patients are beginning to resume many usual routines and activities. Tests at this time should consider the potential effects of post-surgical pain, sleep disturbance, fatigue, nausea, and the effects of opioid and sedative medications, which may not be entirely resolved by discharge and may increase the risk of 'false positives' [68].

Leaving assessment until after this acute period minimises these potential confounds, however, it also increases the risk that cognitive deficits will be missed, thereby depriving clinicians of valuable information about recovery [69]. Research suggests that some patients, typically those who are older and/or have greater impairment at baseline, become increasingly likely to decline follow-up assessments as time from surgery increases [70]. This should be an added incentive to retest early. Although there has been considerable variability in the timing of reassessments, most studies conduct follow-ups within 10 days of surgery [3]. During this period, Monk et al. (2008) found evidence of POCD in 36.6% of patients aged 18-39 years, 30.4% aged 40-59 years, and 41.4% aged ≥60 years [5].

The number and timing of any additional assessments has also differed between studies. Many researchers have elected to conduct only one follow-up, offering little information about the duration of cognitive change and/or rates of recovery. A period of three months is commonly applied, which has been considered an interval that coincides with more complete physical recovery. Approximately 10% of patients demonstrate persistent impairment at this time [4,5]. Longer-term follow-ups in studies are much rarer; however, a number of studies have reassessed at one to two years, some as many as five years, after cardiac and noncardiac surgery [11,71,72,73]. Two collective reviews of these studies have found little evidence of lasting change resulting from surgery [74,75]. Indeed, the concept of persistent POCD remains somewhat controversial. Clinically, the need for longer-term follow-up is likely to be determined by the results of earlier assessments and whether a patient is deemed to have sufficiently recovered function.

Analysis and Interpretation

The analysis and interpretation of assessment data, and the criteria by which POCD is qualitatively and quantitatively described, are issues of huge practical and conceptual importance. Although any analysis always involves comparison of preoperative and postoperative performance, with discrepancy between scores being of central concern, there has been limited agreement regarding the appropriate statistical approach to use to discern whether any observed changes are clinically significant [76]. Indeed, this remains a controversial subject and a number of different methods appear across the literature.

Dichotomisation and 'Caseness'

Many researchers have opted to characterise POCD by dichotomising test scores (i.e. assigning them to categories that indicate whether POCD is present or absent). This type of categorisation, and the use of established 'cut-offs' to define caseness, is familiar to medical practice, where scores on continuous measures are often transformed in this way (e.g. measures of blood pressure that define a patient as 'hypertensive') [77]. One of the most commonly used methods of defining POCD caseness involves evidencing a percentage decline from baseline performance on a specified number of cognitive tests, often defined as a decline of around 20% on ≥2 tests. Categorisations have also been expressed as a number of standard deviations (SDs) decline from either an individual baseline or a reference population mean. There is continued debate as to whether SDs of 1, 1.5, or 2 are appropriate and on how many tests this should apply [34]. Both procedures recognise the expectation of some degree of error within the assessment process, acknowledging that a proportion of change is likely to reflect normal variability in performance. All attempt to establish a magnitude of change that *exceeds* estimates of normal variability (i.e. significant and case-worthy change).

Categorising in this way can be useful in the clinical assessment of individual patients, where there are often established norms but no appropriate control group. It is also an approach that has encouraged greater consistency within studies, systematising the identification of cases to allow incidence to be more easily recorded and thus expanding insight into levels of clinical need. There are, however, significant limitations to categorisation. It has, for example, proven difficult to directly compare POCD studies where different methods of categorisation are applied; chiefly because different methods and criteria can produce quite different incidence rates [3]. These differences have been bought to our attention most starkly in studies applying multiple criteria to the same data. Mahanna et al. (1996) observed that percentage thresholds generated consistently higher estimates of POCD than did SD thresholds at discharge (66% vs 35%), six weeks post-surgery (34% vs 13.8%), and again six months later (19.4% vs 4.4%) [78]. The ISPOCD study went further, applying different percentage criteria to *normative* population data, finding vast differences in the incidence of cognitive dysfunction: from 0% when applying a criterion of 20% decline on two tests, to 40.3% when seeking 25% decline on a single test [4].

Dichotomising continuous variables is also associated with information loss that reduces statistical power [39,77,79,]. This is a significant consequence given the relatively small samples recruited to many POCD studies [3]. The risk of identifying false positives increases dramatically under these methods, especially where test scores in large batteries are analysed individually; in these cases, the probability of participants demonstrating impairment on at least one measure in a battery rises exponentially with each test used [35].

There is also significant masking of variability in this process. Dichotomisation provides definitive numbers of patients demonstrating caseness but offers little insight into the *range* of performance within categories. In some instances, useful qualitative descriptors of cognitive status can be provided to aid interpretation, signifying impairment that is mild, moderate, or severe; however, such nuance is rarely reported and is often no less arbitrary than grosser dichotomies. In respect to impairment and recovery, it is important to consider subtler variations in performance (i.e. change within as well as across categories). During categorisation, scores close to the boundaries become particularly problematic: two patients producing scores that are very close in absolute terms may be classified quite differently if these scores fall on either side of an imposed boundary. Small changes in scores close to designated cut-offs may significantly alter rates of POCD within samples, even where such changes are not statistically significant, whilst substantial decline or improvement further away from the boundary may go undetected if the threshold is not crossed [39].

The arbitrary nature of these thresholds must be acknowledged. There is no conceptually derived rationale for where cut-offs should be placed and how they should be interpreted. To decipher the meaning of change scores, it is necessary to keep in mind the purpose of assessment. One must consider whether it is most important to detect the occurrence of change *per se* or whether one is seeking only to establish change of a designated magnitude. This can be a difficult judgement to make given the lack of consensus on how to differentiate between statistically significant and clinically meaningful change.

Continuous Change

As an alternative to categorisation, scores may be retained as continuous data. This allows analysis of degrees of cognitive decline or improvement. Traditionally, the primary aim of categorisation has been the identification of caseness within the individual or group; continuous data analysis more readily facilitates subtler comparison between groups. This may include standard control groups or groups formed of patients receiving different types of anaesthetic and/or surgical procedures [80,81]. This provides a useful shift in focus away from caseness, and the need to meet arbitrary thresholds of impairment, towards the concepts of relative impairment and rates of recovery.

Comparative assessment of performance using mean scores is arguably the simplest and most intuitive method to examine and compare performance in this way: this approach calculates discrepancy scores based on differences in pre- and postoperative performance (i.e. absolute change over time). This approach does, however, also have limitations. The most significant of these is the considerable variability lost when data are combined. When analysed collectively, scores from patients demonstrating improvement due to learning effects *and* decline due to cognitive impairment may produce stable arithmetic means [82]. The identification of particular risk factors for POCD (e.g. age, previous surgeries) indicate that subgroups are likely to exist even before wider groups are considered; the ability to detect variable levels of decline, improvement, and stability within and between groups is imperative.

The interpretation of simple change scores is further complicated by testing confounds (e.g. practice effects and measurement error) and by established statistical artefacts (e.g. regression to the mean; RTM) [83]. RTM is a well-observed statistical tendency for scores to regress towards the population mean at re-administration; this means that lower scores at baseline generally increase towards the mean, whilst higher scores decrease. These changes occur even in the absence of any true change in function. As Duff (2010) notes, RTM is often most evident in cognitively stable patients, where high scores at first administration drift down on repeat testing [84]. This change may easily be misinterpreted as cognitive decline attributable to surgery. To address the problem of RTM, it is possible to compute change scores adjusted for baseline variance (i.e. taking into account relative starting points for each case). These residualised change scores partial out the proportion of follow-up data that is linearly predictable from baseline, usually achieved through multiple regression modelling [85]. This allows calculation of change scores that control for variable levels of preoperative neuropsychological performance [86]. In subsequent analysis, these residualised change scores do not offer measures of change *per se* but instead provide indications of whether a post-test score is larger or smaller than the value predicted by the individual's baseline assessment.

It can also be difficult to discern from simple discrepancy scores whether change is clinically significant. This problem has traditionally been addressed by calculating a reliable change index (RCI); a method first applied to psychometric data to determine whether meaningful change had occurred following psychotherapy [87]. This technique produces a standardised z-score that communicates the number of SDs a score falls above or below a population mean. Transforming scores in this way allows comparison to normal distributions, offering an indication of whether change is statistically significant within a particular population. As it is well-established that the interpretation of change scores is further complicated by testing confounds, there are also modified versions of the RCI that take practice effects into account – forming the RCI_{PE} [88]. This was the preferred method in the ISPOCD studies, where non-surgical controls were tested to provide estimates of practice effects [4]. When compared to percentage and SD changes, the RCI_{PE} method has been shown to have superior sensitivity and specificity [37].

Although modified RCIs have the obvious advantage of considering practice effects, it is noted that they assume uniformity of such effects across the sample. They also fail to account for differences in the baseline performance, leaving them susceptible to RTM effects. Further use of regression-based analysis, enabling consideration of a wide range of individual differences, has thus been presented as a more sophisticated approach to these problems [89, 90]. Multiple-regression methods are used to build statistical models that can predict postoperative performance based on preoperative scores *plus* any other relevant variables known to influence change to different degrees (e.g. age, education, retest intervals, RTM, practice effects) [50]. Although large samples are needed to formulate accurate estimates of change using these methods, they do enable the development of models that account for a large number of the myriad confounding variables identified in POCD research, allowing closer inspection of many potential causes of variability within the data [76].

In recent attempts to align POCD with DSM-V criteria for neurocognitive disorders, it has been suggested that caseness should be based on corrected z-scores, with a decline of 1 to 2 SDs below either normative or control data characterised as mild disorder and >2SDs as major disorder [18], leaving open the question of whether modified RCIs or regression-based change scores should be used. There is a growing movement within the wider neuropsychological literature towards the use of regression-based methods [91]. Certainly, utilising regression-based approaches in research is enabling development and validation of interesting new measures that can be applied reliably in clinical practice, facilitating the meaningful interpretation of individual scores based on normative data collected from wider populations [92]. The use of regression modelling may also help to redefine narratives surrounding cognitive impairment and recovery after surgery. These approaches encourage consideration of finer degrees of change in individuals and groups, moving beyond the rather narrow focus on absolute deficits and the notion of a POCD ‘diagnosis’. It is certainly hoped these advances will encourage researchers and clinicians to consider postoperative cognitive function in more fluid terms, reflecting on the potential for cognitive *change* rather than deficits alone.

Discussion

This review has addressed the measurement of postoperative cognition and the monitoring of recovery in function over time. It is recognised that research and clinical practice has been impeded by a lack of clarity surrounding clinical definitions of POCD. It continues to be hindered by the use of inappropriate tests, variability in the timing of assessments, the absence of appropriate control groups, and lack of agreement on statistical and clinical definitions. Although there is often marked consternation at these continued inconsistencies, there are a number of important points of agreement that have emerged and should form the basis of good clinical and research practice:

Practice Points

- Cognitive disturbances after surgery are common and typically resolve within days/weeks; however, for some patients, impairments may persist.
- Cognitive recovery is measured as a return to pre-surgery baseline or better.
- Diagnosis of POCD *always* requires objective neuropsychological assessment that includes a pre-surgery baseline and post-surgery reassessment.
- Tests used to assess POCD must be reliable and valid measures capable of capturing subtle change in relatively short intervals.
- Post-surgery reassessments should be undertaken <1 week from surgery, with at least one follow-up assessments within three months.

Research Agenda

- Further work is needed to expand our understanding of POCD in different populations: this should include different patient groups and surgical procedures *and* allow for finer analysis of clinical and demographic factors.
- More robust methodological approaches are needed: studies should be longitudinal and include larger samples with matched-controls; cognitive tests should be valid and reliable, analysed using appropriate statistical methods.
- Sharper focus should be afforded to the concept of cognitive recovery, its trajectory, and the factors that facilitate or impede the process. This may mark a move away from diagnostic ‘all or nothing’ conceptualisations of POCD.

Conflict of Interest Statement

Conflict of interest: none.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

References

[1] An N, Yu WF. Difficulties in Understanding Postoperative Cognitive Dysfunction. J Anesth Perioper Med. 2017; 4(2): 87-94.

[2] Rasmussen LS, O'Brien JT, Silverstein JH, et al. Is peri-operative cortisol secretion related to post-operative cognitive dysfunction? Acta Anaesthesiol Scand. 2005 Oct;49(9):1225-31.

[3] Newman S, Stygall J, Hirani S, et al. Postoperative Cognitive Dysfunction after Noncardiac Surgery: A Systematic Review. Anesthesiology. 2007; 106(3): 572-90

[4] Moller JT, Cluitmans P, Rasmussen LS, et al. Long-term postoperative cognitive dysfunction in the elderly: ISPOCD1 study. Lancet. 1998; 351(9106): 857-61.

[5] Monk TG, Weldon BC, Garvan CW, et al. Predictors of cognitive dysfunction after major noncardiac surgery. Anesthesiology. 2008;108(1):18-30.

[6] Rasmussen LS. Postoperative cognitive dysfunction: incidence and prevention. Best Pract Res Clin Anaesthesiol. 2006; 20(2): 315-30.

[7] Deiner S, Silverstein JH. Postoperative delirium and cognitive dysfunction. BJA. 2009; 103: i41-6.

[8] Berger M, Nadler JW, Browndyke J, et al. Postoperative cognitive dysfunction: minding the gaps in our knowledge of a common postoperative complication in the elderly. Anesthesiology Clinics. 2015; 33(3): 517-50.

[9] Silbert BS, Scott DA, Evered LA, et al. A comparison of the effect of high-and low-dose fentanyl on the incidence of postoperative cognitive dysfunction after coronary artery bypass surgery in the elderly. Anesthesiology. 2006; 104(6): 1137-45.

[10] Steinmetz J, Christensen KB, Lund T, et al. Long-term consequences of postoperative cognitive dysfunction. Anesthesiology. 2009; 110(3): 548-55.

[11] Newman MF, Grocott HP, Mathew JP, et al. Report of the substudy assessing the impact of neurocognitive function on quality of life 5 years after cardiac surgery. Stroke. 2001; 32(12): 2874-81.

[12] Feldman LS, Lee L, Fiore J. What outcomes are important in the assessment of Enhanced Recovery After Surgery (ERAS) pathways? Can J Anaesth. 2015; 62(2): 120-30.

[13] American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[14] Fines DP, Severn AM. Anaesthesia and cognitive disturbance in the elderly. BJA - CEPD Reviews. 2006; 6(1): 37-40.

[15] Rudolph J, Marcantonio ER, Culley DJ, et al. Delirium is associated with early postoperative cognitive dysfunction. Anaesthesia. 2008; 63(9): 941-7.

[16] Sauër AC, Veldhuijzen DS, Ottens TH, et al. Association between delirium and cognitive change after cardiac surgery. BJA. 2017; 119(2): 308-15.

[17] Franck M, Nerlich K, Neuner B, et al. No convincing association between post-operative delirium and post-operative cognitive dysfunction: a secondary analysis. Acta Anaesthesiol Scand. 2016; 60(10): 1404-14.

[18] Evered LA, Silbert B, Knopman D. Recommended nomenclature for perioperative cognitive disorders. Br J Anaesth. 2018.

[19] Hughes CG, Brown IV CH. Postoperative Cognitive Issues. In Ryan Barnett S & Neves SE (eds) Perioperative Care of the Elderly Patient. pp 59-73. Cambridge: Cambridge University Press, 2018.

[20] Royse CF, Newman S, Chung F, et al. Development and Feasibility of a Scale to Assess Postoperative Recovery: The Post-operative Quality Recovery Scale. Anesthesiology. 2010; 113(4): 892-905.

[21] Bowyer A, Jakobsson J, Ljungqvist O, et al. A review of the scope and measurement of postoperative quality of recovery. Anaesthesia. 2014; 69(11): 1266-78.

[22] Hogue CW, Grafman J. Aligning nomenclature for cognitive changes associated with anaesthesia and surgery with broader diagnostic classifications of non-surgical populations: a needed first step. BJA. 2018.

[23] McKhann GM, Selnes OA, Grega MA, et al. Subjective memory symptoms in surgical and nonsurgical coronary artery patients: 6-year follow-up. Ann Thorac Surg. 2009; 87(1): 27-34.

[24] Newman S. The incidence and nature of neuropsychological morbidity following cardiac surgery. Perfusion. 1989; 4(2): 93-100.

[25] Vingerhoets G, De Soete G, Jannes C. Subjective complaints versus neuropsychological test performance after cardiopulmonary bypass. J Psychosom Res. 1995; 39(7): 843-53.

[26] Khatri P, Babyak M, Clancy C, et al. Perception of cognitive function in older adults following coronary artery bypass surgery. Health Psychol. 1999;18(3):301.

[27] Johnson T, Monk T, Rasmussen LS, et al. Postoperative cognitive dysfunction in middle-aged patients. Anesthesiology. 2002;96(6):1351-7.

[28] Newman S, Klinger L, Venn G, et al. The persistence of neuropsychological deficits twelve months after coronary artery bypass surgery. In Wilner AE & Rodewald G (eds) Impact of Cardiac Surgery on the Quality of Life: Neurological and Psychological Aspects. pp. 173-179. Boston, MA: Springer, 1990.

[29] Rabbitt P, Abson V. ‘Lost and Found’: Some logical and methodological limitations of self-report questionnaires as tools to study cognitive ageing. Br J Psychol. 1990;81(1):1-6.

[30] Murkin JM, Newman SP, Stump DA, et al. Statement of consensus on assessment of neurobehavioral outcomes after cardiac surgery. Ann Thorac Surg. 1995;59(5):1289-95.

[31] Hovens IB, Schoemaker RG, van der Zee EA, et al. Thinking through postoperative cognitive dysfunction: how to bridge the gap between clinical and pre-clinical perspectives. Brain Behav Immun. 2012;26(7):1169-79.

[32] Lezak MD, Howieson DB, Loring DW, et al. Neuropsychological Assessment. Oxford University Press, USA; 2004.

[33] Dijkstra JB, Jolles J. Postoperative cognitive dysfunction versus complaints: a discrepancy in long-term findings. Neuropsychol Rev. 2002;12(1):1-4.

[34] Rudolph JL, Schreiber KA, Culley DJ, et al. Measurement of post-operative cognitive dysfunction after cardiac surgery: a systematic review. Acta Anaesthesiol Scand. 2010;54(6):663-77.

[35] Lewis MS, Maruff P, Silbert BS, et al. Detection of postoperative cognitive decline after coronary artery bypass graft surgery is affected by the number of neuropsychological tests in the assessment battery. Ann Thorac Surg. 2006;81(6):2097-104.

[36] Ingraham LJ, Aiken CB. An empirical approach to determining criteria for abnormality in test batteries with multiple measures. Neuropsychology. 1996;10(1):120.

[37] Lewis MS, Maruff P, Silbert BS, et al. The sensitivity and specificity of three common statistical rules for the classification of post-operative cognitive dysfunction following coronary artery bypass graft surgery. Acta Anaesthesiol Scand. 2006;50(1):50-7.

[38] Funder KS, Steinmetz J, Rasmussen LS. Methodological issues of postoperative cognitive dysfunction research. Semin Cardiothorac Vasc Anesth. 2010; 14(2):119-122.

[39] Ghoneim MM, Block RI. Clinical, methodological and theoretical issues in the assessment of cognition after anaesthesia and surgery: a review. Eur J Anaesthesiol. 2012;29(9):409-22.

[40] Goldstein LH, McNeil JE. Clinical neuropsychology: A practical guide to assessment and management for clinicians. John Wiley & Sons; 2012.

[41] Valentin LS, Pietrobon R, Junior A, et al. Definition and application of neuropsychological test battery to evaluate postoperative cognitive dysfunction. Einstein (São Paulo). 2015;13(1):20-6.

[42] Schmidt M. Rey Auditory Verbal Learning Test: A Handbook. Los Angeles, CA: Western Psychological Services; 1996.

[43] Reitan RM, Wolfson D. The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation. Reitan Neuropsychology; 1985.

[44] Spreen O, Strauss E. A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. Oxford University Press; 1998.

[45] Wechsler D, Coalson DL, Raiford SE. WAIS-III: Wechsler Adult Intelligence Scale. San Antonio, TX: Psychological Corporation; 1997.

[46] Trenerry MR, Crosson B, DeBoe J, et al. Stroop Neuropsychological Screening Test. Odessa, FL: Psychological Assessment Resources. 1989.

[47] Catron DW. Immediate test-retest changes in WAIS scores among college males. Psychol Rep. 1978;43(1):279-90.

[48] Catron DW, Thompson CC. Test-retest gains in WAIS scores after four retest intervals. J Clin Psychol. 1979;35(2):352-7.

[49] Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology. 2010;24(5):563.

[50] Duff K. Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Arch Clin Neuropsychol. 2012;27(3):248-61.

[51] Irvine CD, Gardner FV, Davies AH, et al. Cognitive testing in patients undergoing carotid endarterectomy. Eur J Vasc Endovasc Surg. 1998;15(3):195-204.

[52] Marley CJ, Sinnott A, Hall JE, et al. Failure to account for practice effects leads to clinical misinterpretation of cognitive outcome following carotid endarterectomy. Physiol Rep. 2017;5(11).

[53] Heaton RK, Temkin N, Dikmen S, et al. Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. Arch Clin Neuropsychol. 2001;16(1):75-91.

[54] Elkins JS, Longstreth WT, Manolio TA, et al. Education and the cognitive decline associated with MRI-defined brain infarct. Neurology. 2006;67(3):435-40.

[55] Feinkohl I, Winterer G, Spies CD, et al. Cognitive reserve and the risk of postoperative cognitive dysfunction: a systematic review and meta-analysis. Dtsch Arztebl Int. 2017;114(7):110.

[56] Nelson HE, Willison J. National adult reading test (NART). Windsor: NFER-Nelson; 1991.

[57] Murkin JM, McKhann G. Defining dysfunction: group means versus incidence analysis—a statement of consensus. Ann Thorac Surg. 1997;64(3):904-5.

[58] Rasmussen LS, Larsen K, Houx P, et al. The assessment of postoperative cognitive function. Acta Anaesthesiol Scand. 2001;45(3):275-89.

[59] Perks A, Chakravarti S, Manninen P. Preoperative anxiety in neurosurgical patients. J Neurosurg Anesthesiol. 2009;21(2):127-30.

[60] Cleeland CS, Nakamura Y, Howland EW, et al. Effects of oral morphine on cold pressor tolerance time and neuropsychological performance. Neuropsychopharmacology. 1996;15(3):252.

[61] Heyer, EJ, Sharma, R, Winfree, CJ et al. Severe pain confounds neuropsychological test performance. J Clin Exp Neuropsychol. 2000;22(5):633-9.

[62] Newman SP, Stygall J. Neuropsychological outcome following cardiac surgery. In Newman, SP and Harrison, MJG, (eds**)** The Brain and Cardiac Surgery: Causes of Neurological Complications and their Prevention. pp. 1-351. London: Harwood Academic Publishers.

[63] Tsai TL, Sands LP, Leung JM. An update on postoperative cognitive dysfunction. Adv Anesth. 2010;28(1):269-84.

[64] Bowyer AJ, Royse CF. Postoperative recovery and outcomes–what are we measuring and for whom? Anaesthesia. 2016;71(S1):72-7.

[65] Rasmussen L, Stygall J, Newman, S. Cognitive dysfunction and other long-term complications of surgery and anesthesia. In Miller RD, Eriksson, LI, Fleisher LA (eds) Miller's Anesthesia. pp 2805-2841. Churchill Livingston.

[66] Krenk L, Rasmussen LS, Kehlet H. New insights into the pathophysiology of postoperative cognitive dysfunction. Acta Anaesthesiol Scand. 2010;54(8):951-6.

[67] Fleischmann KE, Goldman L, Young B, et al. Association between cardiac and noncardiac complications in patients undergoing noncardiac surgery: outcomes and effects on length of stay. Am J Med. 2003;115(7):515-20.

[68] Sauër AM, Kalkman C, van Dijk D. Postoperative cognitive decline. J Anesth. 2009;23(2):256-9.

[69] Rohan D, Buggy DJ, Crowley S, et al. Increased incidence of postoperative cognitive dysfunction 24 hr after minor surgery in the elderly. Can J Anaesth. 2005;52(2):137-42.

[70] Rudolph J, Marcantonio ER, Culley DJ, et al. Delirium is associated with early postoperative cognitive dysfunction. Anaesthesia. 2008;63(9):941-7.

[71] Selnes OA, Royall RM, Grega MA, et al. Cognitive changes 5 years after coronary artery bypass grafting: is there evidence of late decline? Arch Neurol. 2001;58(4):598-604.

[72] Müllges W, Babin–Ebell J, Reents W, et al. Cognitive performance after coronary artery bypass grafting: a follow-up study. Neurology. 2002;59(5):741-3.

[73] Stygall J, Newman SP, Fitzgerald G, et al. Cognitive change 5 years after coronary artery bypass surgery. Health Psychol. 2003;22(6):579.

[74] Avidan MS, Evers AS. Review of clinical evidence for persistent cognitive decline or incident dementia attributable to surgery or general anesthesia. J Alzheimers Dis. 2011;24(2):201-16.

[75] Silbert B, Evered L, Scott DA. Cognitive decline in the elderly: is anaesthesia implicated? Best Pract Res Clin Anaesthesiol. 2011;25(3):379-93.

[76] Collie A, Darby DG, Falleti MG, et al. Determining the extent of cognitive change after coronary surgery: a review of statistical procedures. Ann Thorac Surg. 2002;73(6):2005-11.

[77] Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.

[78] Mahanna EP, Blumenthal JA, White WD, et al. Defining neuropsychological dysfunction after coronary artery bypass grafting. Ann Thorac Surg. 1996;61(5):1342-7.

[79] Dawson NV, Weiss R. Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Making. 2012;32(2):225-226.

[80] Kong R, Aveling W, Harrison MJ, et al. A clinical trial of clomethiazole as a neuroprotectant in coronary artery bypass grafting. Perfusion. 2001;16(3):251.

[81] Arrowsmith JE, Harrison MJ, Newman SP, et al. Neuroprotection of the brain during cardiopulmonary bypass: a randomized trial of remacemide during coronary artery bypass in 171 patients. Stroke. 1998;29(11):2357-62.

[82] Newman SP. Analysis and interpretation of neuropsychologic tests in cardiac surgery. Ann Thorac Surg. 1995;59(5):1351-5.

[83] Browne SM, Halligan PW, Wade DT, et al. Cognitive performance after cardiac operation: implications of regression toward the mean. J Thorac Cardiovasc Surg. 1999;117(3):481-5.

[84] Duff K, Beglinger LJ, Moser DJ, e al. Predicting cognitive change in older adults: the relative contribution of practice effects. Arch Clin Neuropsychol. 2010;25(2):81-8.

[85] Castro-Schilo L, Grimm KJ. Using residualized change versus difference scores for longitudinal research. J Soc Pers Relat. 2018;35(1):32-58.

[86] Griva K, Newman SP, Harrison MJ, et al. Acute neuropsychological changes in hemodialysis and peritoneal dialysis patients. Health Psychol. 2003 Nov;22(6):570.

[87] Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991;59(1):12.

[88] Chelune GJ, Naugle RI, Lüders H, et al. Individual change after epilepsy surgery: Practice effects and base-rate information. Neuropsychology. 1993;7(1):41.

[89] McSweeny AJ, Naugle RI, Chelune GJ, et al. “T scores for change”: An illustration of a regression approach to depicting change in clinical neuropsychology. Clin Neuropsychol. 1993;7(3):300-12.

[90] Temkin NR, Heaton RK, Grant I, et al. Detecting significant change in neuropsychological test performance: A comparison of four models. J Int Neuropsychol Soc. 1999 May;5(4):357-69.

[91] Cysique LA, Franklin Jr D, Abramson I, et al. Normative data and validation of a regression-based summary score for assessing meaningful neuropsychological change. J Clin Exp Neuropsychol. 2011;33(5):505-22.

[92] Royse CF, Chung F, Newman S, et al. Predictors of patient satisfaction with anaesthesia and surgery care: a cohort study using the Postoperative Quality of Recovery Scale. Eur J Anaesthesiol. 2013;30(3):106-10.