BJA

British Journal of Anaesthesia, 126 (6): 1119–1127 (2021)

doi: 10.1016/j.bja.2021.01.035
Advance Access Publication Date: 2 April 2021
Review Article

NEUROSCIENCE AND NEUROANAESTHESIA

Methodology of measuring postoperative cognitive dysfunction: a systematic review

Friedrich Borchers[1], Claudia D. Spies[1], Insa Feinkohl[2], Wolf-Rüdiger Brockhaus[1], Antje Kraft[1], Petra Kozma[1], Marinus Fislage[1], Simone Kühn[3,4], Catinca Ionescu[1], Saya Speidel[1], Daniel Hadzidiakos[1], Dieuwke S. Veldhuijzen[5,6], Fatima Yürek[1], Lisbeth A. Evered[7,8] and Thomas H. Ottens[9,*]

[1]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany, [2]Max-Delbrück Center for Molecular Medicine, Berlin, Germany, [3]Universitätsklinik Hamburg–Eppendorf, Hamburg, Germany, [4]Max Planck Institute for Human Development, Berlin, Germany, [5]Leiden University, Leiden, the Netherlands, [6]Leiden Institute for Brain and Cognition, Leiden, the Netherlands, [7]St. Vincent's Hospital Melbourne, Fitzroy, Victoria, Australia, [8]Department of Anesthesiology, Weill Cornell Medicine, New York, NY, USA and [9]Haga Teaching Hospital, Department of Intensive Care Medicine, The Hague, the Netherlands

*Corresponding author. E-mail: t.ottens@hagaziekenhuis.nl

Abstract

Background: Postoperative cognitive dysfunction (POCD) is an adverse outcome that impacts patients' quality of life. Its diagnosis relies on formal cognitive testing performed before and after surgery. The substantial heterogeneity in methodology limits comparability and meta-analysis of studies. This systematic review critically appraises the methodology of studies on POCD published since the 1995 Consensus Statement and aims to provide guidance to future authors by providing recommendations that may improve comparability between future studies.
Methods: This systematic review of literature published between 1995 and 2019 included studies that used baseline cognitive testing and a structured cognitive test battery, and had a minimal follow-up of 1 month. For cohorts with multiple publications, data from the primary publication were supplemented with available data from later follow-up studies.
Results: A total of 274 unique studies were included in the analysis. In the included studies, 259 different cognitive tests were used. Studies varied considerably in timing of assessment, follow-up duration, definition of POCD, and use of control groups. Of the 274 included studies, 70 reported POCD as a dichotomous outcome at 1 to <3 months, with a pooled incidence of 2998/10 335 patients (29.0%).
Conclusions: We found an overwhelming heterogeneity in methodology used to study POCD since the publication of the 1995 Consensus Statement. Future authors could improve study quality and comparability through optimal timing of assessment, the use of commonly used cognitive tests including the Consensus Statement 'core battery', application of appropriate cut-offs and diagnostic rules, and detailed reporting of the methods used.
PROSPERO registry number: CRD42016039293.

Keywords: diagnostic criteria; methodology; neurocognitive disorders; neuropsychological testing; perioperative cognition; postoperative cognitive dysfunction

Received: 1 October 2020; **Accepted:** 26 January 2021

Editor's key points

- The authors compared the methodology of studies on postoperative cognitive decline (POCD) to the criteria published in the 1995 Statement of Consensus on Assessment of Neurobehavioral Outcomes After Cardiac Surgery.
- From more than 8000 studies published after the Consensus Statement, only 274 used baseline cognitive testing and followed patients for at least 1 month. The authors identified more than 250 different cognitive tests and a large variety of diagnostic rules.
- The authors conclude that poor compliance with the Consensus Statement has resulted in a body of literature that is difficult to interpret. The authors provide suggestions on study design to improve the comparability of future studies.

In a Statement of Consensus (CS) from 1995, Murkin and colleagues[1] took effort to improve the quality of postoperative cognitive dysfunction (POCD) studies. Despite widespread acceptance and frequent citation of the CS, the methodology of papers in this field has remained highly heterogeneous, most importantly because of differences in the composition of cognitive test batteries, timing of follow-up testing, cut-offs, and diagnostic rules used to adjudicate cognitive outcome. This heterogeneity creates substantial challenges in the interpretation of the current body of evidence.[2,3]

The CS recommends that studies on POCD have at least one cognitive follow-up assessment when cognitive function has 'stabilised', not earlier than 1 month postoperatively. Regarding the composition of the cognitive test battery, the CS recommends the use of a 'core battery' of tests that include the Rey Auditory Verbal Learning Test, Trailmaking Test, and Grooved Pegboard Test. These tests broadly cover the domains of verbal memory, divided attention, and motor skills. It further recommends methods that minimise practice effects and effects of natural variability in test performance over time. This not only requires cognitive tests that are robust to practice effects, but also includes conservative mathematical definitions of POCD. Such methods, referred to as the 'Reliable Change Index', require the inclusion of a non-surgical control group.[1,4]

Several studies have pointed out that the composition of the cognitive test battery, timing of follow-up testing, and mathematical definitions strongly influence the POCD incidence.[5,6] The choice to use more conservative mathematical methods, such as the reliable change index (RCI), also impacts study designs because they require data from non-surgical controls to correct for learning effects and natural variability in cognitive test performance. The choice of RCI variant also significantly affects the outcome.[7] Patient selection, especially exclusion of patients with preoperative cognitive disorders, is also likely to affect findings. Comparability between studies is difficult because both continuous and dichotomous models are currently used to describe and compare cognitive outcomes between groups, making meta-analysis particularly difficult.[8]

However, authors currently have little guidance when designing studies on perioperative cognition. This review aims to provide guidance to future authors. We describe how the methodological aspects specific to the field of POCD research have been applied since the CS. We critically appraise those methods and provide recommendations that may improve comparability between studies. We will discuss screening for preoperative cognitive functioning, timing of cognitive follow-up testing, the composition of cognitive test batteries, the use of control groups, and the various mathematic definitions of POCD.

Methods

Design

The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement recommendations for conducting a systematic review.[9] The study protocol and the full search strategy were registered within the PROSPERO International prospective register of systematic reviews as CRD42016039293. We originally aimed to also study the influence of the different methods on the POCD outcome estimate. However, because of the large number of studies that matched the inclusion criteria and the overwhelming variability of study methods, we decided to focus on critically appraising the most commonly used methods. Because we did not perform weighted meta-analyses and give detailed descriptions of methodology, we omitted other forms of formal bias assessment.

Eligible studies

Searches were conducted in July 2016 and repeated in July 2019 in three databases (Pubmed, EMBASE, and Cochrane) for articles published after release of the CS.[1] We structured the search presented in Supplementary Table S1 for domain (postoperative period) and the outcome (POCD). We also searched for references in reviews, meta-analyses, and the authors' personal files.

Study selection

Search results, abstracts, and titles were screened in duplicate by a team of four authors (FB, THO, WRB, and IF). At this stage, discussions about inclusion of studies were resolved by a neuropsychologist (AK). Articles were included if they were based on original research data from prospective studies on adult patients undergoing surgery (except intracranial neurosurgery), published from May 1995 onwards in English, German, Dutch, or French. Articles were excluded if there was no cognitive assessment before surgery, the study focused on cognitive improvement after surgery, or if follow-up duration was less than 1 month. This cut-off was chosen to avoid inclusion of studies that focus on early postoperative cognitive disorders such as delirium and *delayed neurocognitive recovery*.[10] We also excluded articles that did not quantify cognitive change, used self-reporting, informant reporting, or screening instruments (e.g. Mini Mental State Examination [MMSE]) exclusively to assess cognitive decline. Furthermore, we excluded studies with a primary focus on postoperative delirium, quality of life, or any other construct that has been explicitly designed for another research context (e.g. quality of life questionnaires).

Articles selected for full-text reading were assessed for inclusion by six teams of two assessors. Disputes about inclusions were resolved by the main authors (FB and THO).

Data extraction

Data extraction was performed independently by each author using a predefined, structured data extraction sheet, which is provided in Supplementary Table S2. In short, we extracted data on methodological aspects relevant to determine adherence to the CS. We also extracted data on the sample size, timing of testing, use and handling of control groups, definitions, and cut-offs for POCD and the POCD incidence estimate.

If data from the same cohort appeared in multiple papers with different research questions or follow-up durations, but with the same cognitive test methods, we treated the series of papers as one study and extracted the POCD incidence estimates from all available follow-up time points.

In the tables and analyses presented in this paper, we presented the highest incidence of POCD in cases in which the authors reported different severity grades (e.g. mild, moderate, or severe POCD). We reported the incidence calculated with the most restrictive algorithm if alternative results were presented.[5,11]

Results

Study identification

From 8829 unique articles, 8461 references were excluded based on reviewing title and abstract. We decided to exclude studies with a clear focus on cognitive improvement (mainly within the field of bariatric, transplant, or cataract/ophthalmic surgery). In our search, we identified 173 reviews and meta-analyses, of which 79 were relevant to the research question. From the references of these reviews and meta-analyses and one further narrative review, we included an additional 30 references for full-text review. A total of 398 articles entered full-text review.

Of the 398 selected articles, eight were not accessible for full-text reading and could not be retrieved through the corresponding author. Nineteen articles were not based on original research data from prospective study designs. Of the remaining 371 articles, 41 fulfilled one or more of the exclusion criteria.

Fifty papers reported on the same cohort as a previously published paper. Only the index papers of these cohorts were included for data extraction on methodology. We included the data of these cohorts for reporting POCD incidence on long-term cognitive follow-up. Six studies were excluded for miscellaneous reasons. Therefore, a total of 274 studies were included for systematic review (see Fig. 1). A full reference list is provided in the Supplementary information.

Study characteristics

We identified 169/274 (61.7%) observational and 105/274 (38.3%) interventional studies. Table 1 presents the study characteristics. Sample sizes varied widely over the studies (11–1218 patients for observational studies and 10–1277 for interventional studies). Cardiac surgery patients represented the most frequent type of cohort in both observational and interventional studies.

Interventions were either related to surgery, specific cardiac surgery, and cardiopulmonary bypass techniques, anaesthesia techniques, or medication. Supplementary Table S3 lists the drugs investigated in the interventional studies. Of the 28 RCTs, only a few investigated the same drugs or drugs with similar types of action (lidocaine, magnesium, cyclo-oxygenase-2 [COX-2] inhibitors, and dexamethasone).

Screening for cognitive impairment at baseline assessment

In 104/274 studies (38.0%), patients were screened for pre-existing cognitive impairment using the MMSE or equivalent screening tools. Screening results were used to exclude patients with scores indicating some degree of pre-existing impairment in 88/274 (32.1%) cases.

Timing of testing and association with POCD incidence estimates

Of the included studies, 85/274 (31.0%) scheduled a single cognitive follow-up. In most cases testing was performed between 1 and less than 3 months after surgery (48/85, 56.5%), followed by 31/85 (36.5%) with cognitive testing from 3 months up to 1 yr after surgery. A single follow-up rarely occurred later than 1 yr postoperatively (5/85, 5.9%).

Most included studies (189/274; 69.0%) scheduled multiple follow-ups. From these 189 studies, 87 (46.03%) started testing within 7 days or before/at discharge from hospital. Furthermore, 152/189 (80.4%) performed two follow-ups, 30/189 (15.9%) three, and only 7/189 (3.7%) scheduled more than three follow-ups. The median number of follow-up visits was 2 (inter-quartile range [IQR], 2–2; min 2, max 8). The distribution of follow-up times is presented in Figure 2. Most studies (144/189, 76.2%) with multiple follow-ups failed to report which of their follow-up tests was used to classify patients as having POCD or not. The effect of the timing for testing on the POCD outcome estimate is presented in Figure 3.

In Supplementary Table S4, large cohorts that generated several publications are listed along with their maximum follow-up duration and POCD incidence (if available).

Neuropsychological testing

The complete 'core set' – or very similar tests – as recommended in the CS were included in the test battery in 78/274 (28.5%) of the included studies. Test batteries were composed of a median of six tests (IQR, 4–7). Some authors used sets of standalone neuropsychological tests whereas others used pre-specified batteries.

We identified at least 259 unique cognitive tests across the included studies. The number of tests was difficult to determine because some authors omitted the exact description of their test battery. Of the 259 unique tests, 140 were used by one study only.

Table 2 lists the 15 most commonly used neuropsychological tests. All cognitive tests reported are listed in Supplementary Table S5.

The CS advises the use of tests that are robust to practice effects and ideally have parallel versions for follow-up testing.[1] Still, broad screening tools and dementia screening tests were often part of the test battery. The MMSE was used as part of the cognitive battery in 42 studies (15%). Of note, studies that relied solely on MMSE or similar screening tools and had no additional cognitive tests to identify cognitive change over time were excluded from this review.

A majority of studies (212/274, 77.4%) used conventional 'paper-and-pencil' methods to administer cognitive tests. Overall, 37/274 studies (13.5%) used computerised cognitive

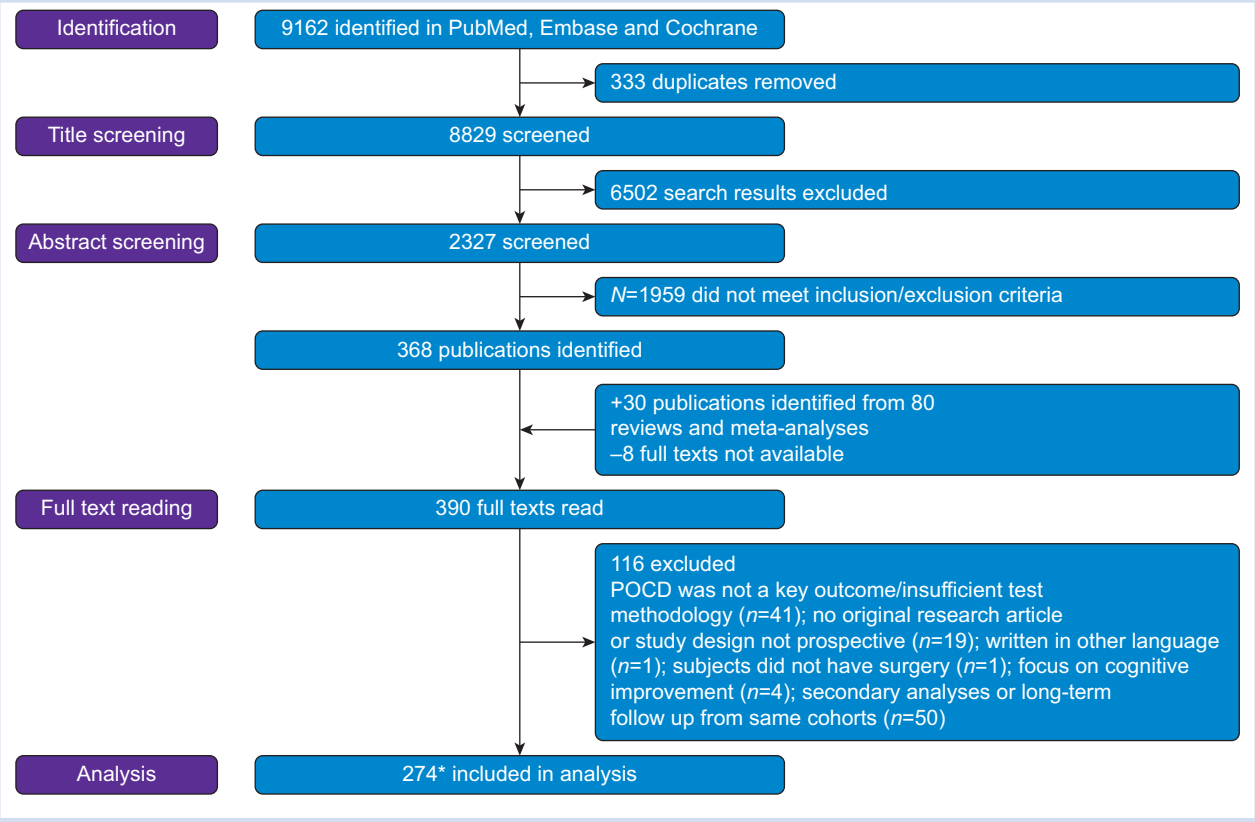


Fig 1. Inclusion chart. *Fifty studies were based on data from the same cohort as a previously published index study. We included these studies for sub-analyses on long-term postoperative cognitive dysfunction (POCD) incidence but not for complete data extraction.

tests or a mix of both. Meanwhile, 25/274 studies (9.1%) did not report type of testing.

Definition of POCD and non-surgical control groups

Only 52 (28.6%) from the 182 studies with a dichotomous POCD endpoint included a control group. Matching of controls and patients was reported by 27 of 52 studies (51.9%). Demographic details about the control group were reported in 38/52 (73.1%) cases. There was considerable variation in the ratio between patients and controls in the studies. Most (34/52, 65.4%) included fewer controls than patients, with ratios ranging from 1 control per 2–14 patients (median, 3; IQR, 3–5). Nine of 52 (17.3%) studies included as many non-surgical controls as surgical patients (ratio 1:1), and 9/52 (17.3%) included more than one non-surgical control per surgical patient (median, 4; IQR, 2.5–7; range, 2–12).

Studies reporting POCD as a dichotomous outcome

POCD was reported as dichotomous outcome in 182/274 (66.4%) studies. The remaining cases compared cognitive test results between groups as a continuous outcome rather than adjudicating a POCD diagnosis, or did not make any comparisons. In the studies with a dichotomous POCD outcome, the methods to define relevant cognitive decline varied broadly.

The RCI uses data from a non-surgical control group to adjust for natural variability and learning effects. This method subtracts the expected learning effect and natural variability (both derived from the non-surgical controls) from a patient's test performance change from baseline to follow-up. If a patient's change in test performance from baseline to follow-up exceeds a predefined cut-off, the POCD diagnosis is adjudicated.[7]

Studies did not consistently report details on their analysis method, and the exact number of studies using an RCI-based analysis could not be determined. The majority of studies (110/182, 60.4%) relied on simple analysis methods that compared change to the patient or study population baseline, and did not account for natural variability and learning effects. Deviation from population norms was used to define relevant change in 17/182 (9.3%) studies.

After defining if the cognitive change in an individual test parameter is relevant, studies require a diagnostic rule to adjudicate the POCD diagnosis. Most studies applied simple diagnostic rules based on a particular number of tests with relevant change. Studies less commonly used composite scores calculated from all the cognitive tests in the battery (10/182, 5.5%) or combinations of the two rules (26/182, 14.3%). Even more complex analysis methods involved component analysis and adjudication of the POCD diagnosis based on decline in one or more predefined cognitive domains. Often, the exact number of cognitive test parameters to which the diagnostic rule was applied was not reported. The diagnostic rule itself was not reported in 5/182 (2.7%) cases. Some authors who described cut-off values in cognitive test parameters for

Table 1 Study characteristics. All values presented as median (inter-quartile range) and *n* (%). *Lung surgery (VATS), Egawa 2016.[12] CABG, coronary artery bypass grafting; N/a, not applicable; VATS, video-assisted thoracoscopic surgery.

Characteristic	All studies (*n*=274)	Observational studies (*n*=169)	Interventional studies (*n*=105)
Sample size of surgical cohort	101 (range, 51–194)	82 (range, 40–168)	134 (range, 65–234)
Type of cohort			
Mix of different non-cardiac surgery patients	28 (10.2)	19 (11.2)	9 (8.6)
Cardiac surgery (CABG, valve repair, valve replacement, surgery on the thoracic aorta)	173 (63.1)	95 (56.2)	78 (74.3)
Vascular surgery (including carotid, surgery for peripheral arterial occlusive disease)	35 (12.8)	30 (17.8)	5 (4.8)
Peripheral surgery (breast, orthopaedic, eye, ear–nose–throat, spine, peripheral nerves, plastic surgery, dermatology)	30 (10.9)	20 (11.8)	10 (9.5)
Abdominal surgery (gastrointestinal, liver, gall bladder, kidney, adrenal, stomach, oesophagus, pancreas, spleen)	7 (2.6)	5 (3.0)	2 (1.9)
Other*	1 (0.4)	0 (0.0)	1 (0.9)
Types of interventions	N/a	N/a	
Anaesthesia related			25 (23.8)
Surgical technique related			20 (19.0)
Cardiac bypass related			24 (22.9)
Randomised drug trials			28 (26.7)
Other			2 (1.9)

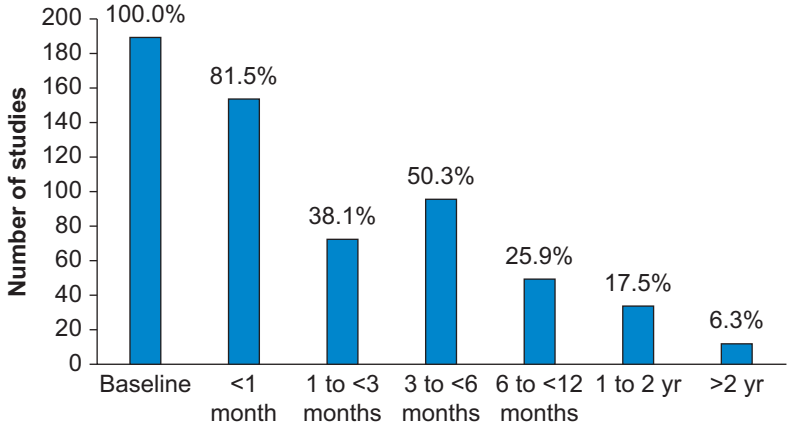


Fig 2. Distribution of testing time points in studies with more than one follow-up.

dichotomisation did not report POCD incidence on an individual patient level.[14–18]

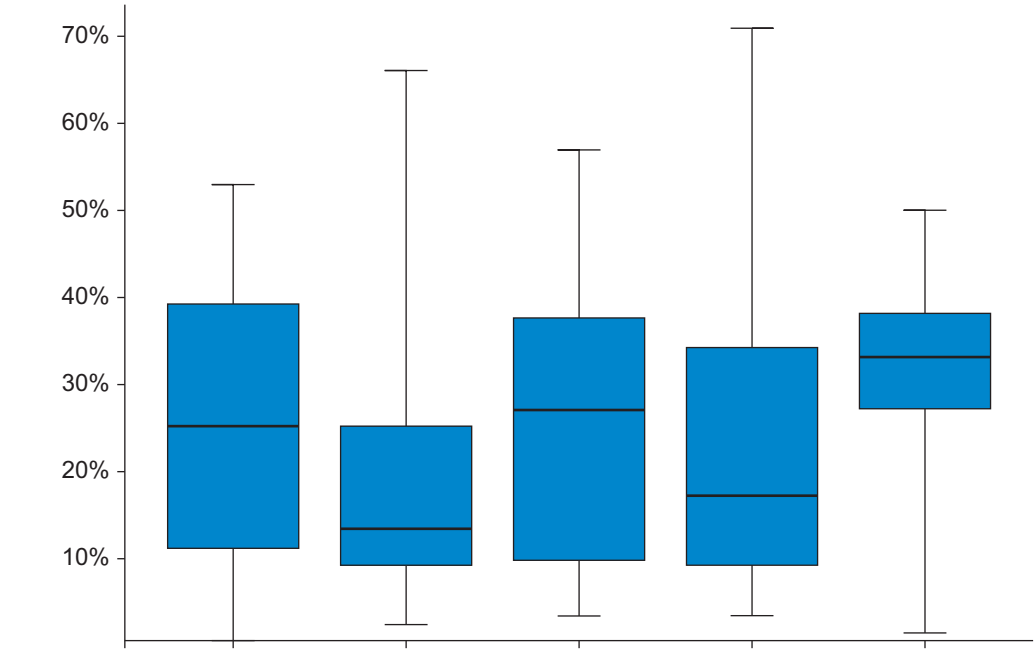
Table 3 shows the different cut-off values and diagnostic rules applied, from more liberal to more restrictive.

Studies reporting POCD as a continuous outcome.

Ninety-two of the 274 studies (33.6%) exclusively reported a continuous cognitive outcome. Most of these calculated Z-score changes, simple group differences, performed cluster analyses to define cognitive domain changes or applied analysis of variance (anova), multivariate analysis of covariance (mancova) or mancova techniques to either investigate predictors of change or to adjust for age, sex, and educational level.

Discussion

As expected, the methodology used in the 169 observational and 105 interventional studies included in this



	1 to <3 months	3 to <6 months	6 to <12 months	1 to <2 yr	2 yr or later
N (%) studies reporting POCD at this follow-up time	70 (45.5)	63 (40.9)	24 (15.6)	17 (11.0)	9 (5.8)
Pooled *N* of participants	10 335	12 011	4094	4528	1577
Pooled *N* (%) patients classified as POCD	2998 (29.0)	1695 (14.1)	1107 (27.0)	779 (17.2)	386 (24.5)

Fig 3. Pooled analysis of studies reporting postoperative cognitive dysfunction (POCD) as a dichotomous outcome and with available data at least one follow-up time after 1 month (n=154). Box plot represents median reported POCD incidence, 25th and 75th percentile, whiskers represent 1.5*inter-quartile range below 25th and above 75th percentile.

systematic review was highly heterogenous. Even though only studies that met certain quality criteria (baseline testing, structured neuropsychological testing, minimum 1 month follow-up) were included, adherence to the CS guidance was poor. Only 28.5% of studies included the recommended 'core' cognitive tests in their cognitive battery. The composition of test batteries differed largely: from 259 unique cognitive tests, 140 were used only once in the included studies. This heterogeneity hinders pooled analyses and other forms of data syntheses across studies. Despite the CS recommendation to include non-surgical control groups, these were used in just 52 of the 182 studies that calculated a dichotomous incidence. Comparability of outcome estimates between studies is further compromised by the diversity of definitions of relevant cognitive change and different diagnostic rules in studies reporting dichotomous outcomes.

Timing of testing

Our findings and those of previous studies show that the incidence of POCD detected is dependent on follow-up time.[5,6] In accordance with the 2018 Nomenclature Recommendations, POCD should be named postoperative neurocognitive disorder (p-NCD) in future studies. Cognitive change diagnosed between 1 and 12 months postoperatively should be termed p-NCD exclusively. In order to study p-NCD, at least one postoperative follow-up between 1 and 12 months is thus required. Our findings correlate well with previous reports on the natural course of cognitive dysfunction over time. Long-term follow-up studies have shown that cognitive function is most affected between 3 and 6 months postoperatively but often recovers later on.[5] Authors designing new studies can expect the highest incidence between 1 and 3 months (29%) and the lowest incidence between 3 and 6 months (14.1%) postoperatively.

Table 2 Most frequently used neuropsychological tests. For test manuals, refer to Lezak and colleagues.[13] WAIS, Wechsler Adult Intelligence Scale; WMS, Wechsler Memory Scale.

Test	Type of test/ main domain	Frequency of use in included studies (n=274)
Trailmaking test	Executive function test	163 (59%)
Digit span subtest of the WAIS	Memory test	108 (39%)
Digit Symbol Substitution Test of the WAIS	Broad cognitive function test	90 (33%)
Rey Auditory Verbal Learning Test	Verbal memory test	89 (32%)
Grooved Pegboard Test	Motor skill test	70 (26%)
Stroop Color–Word Interference Test	Executive function test	60 (22%)
Controlled Oral Word Association Test	Fluency test	43 (16%)
Mini Mental State Exam	Dementia screening battery	42 (15%)
Rey Complex Figure Test	Non-verbal memory test	28 (10%)
Finger Tapping Test	Motor skill test	21 (8%)
Category type verbal fluency tests	Fluency test	19 (7%)
Symbol–Digit Modalities Test	Broad cognitive function test	19 (7%)
Boston Naming Test	Language test	18 (7%)
Modified Visual Reproduction test of the WMS	Intelligence test	18 (7%)
Visual Verbal Learning Test	Verbal memory test	17 (6%)

Table 3 Cut-off values and diagnostic rules to define POCD in studies reporting a dichotomous outcome. All values presented as n (%). RCI, reliable change index; sd, standard deviation; POCD, postoperative cognitive dysfunction.

Cut-offs	
20% change from patients own baseline	28 (15.4)
<1 sd change from baseline	5 (2.7)
1–2 sd change from baseline OR corresponding RCI	77 (42.3)
≥2 sd change from baseline OR corresponding RCI	43 (23.6)
Reported multiple cut-offs	29 (15.9)
Cut-off not reported	3 (1.6)
POCD diagnostic rules	
Relevant decline in at least 1 test/test parameter	39 (21.4)
Relevant decline in at least 2 tests/test parameters	53 (29.1)
Relevant decline in at least 3 tests/test parameters	6 (3.3)
Relevant decline in at least 20% of test parameters	17 (9.3)
Significant change in composite score	10 (5.5)
Combination of change in individual tests OR composite score	26 (14.3)
Component analysis, domain specific rules	19 (10.4)
Combination of analysis strategies	5 (2.7)
Diagnostic rule not reported	5 (2.7)

Sometimes authors reported more than one follow-up time point for cognitive testing but only reported one dichotomous outcome. This, however, was not clearly attributed to one of the follow-up time points.[19] We recommend that authors using multiple postoperative follow-ups clearly report which follow-up was used to calculate the p-NCD incidence and which follow-up was used as the primary outcome. The number of follow-ups will depend on the research question and available resources. The Successful Aging after Elective Surgery (SAGES) study is an example of the feasibility of studies that combine reliable methods with a large number of follow-up visits.[20]

The exclusion of patients based on results of a preoperative cognitive screening was fairly common in the studies (32%) but may select a population that is at a lower risk of further cognitive decline.[21] We recommend that authors clearly report if patients were excluded from study participation based on preoperative tests. Patients' preoperative cognitive functioning and the proportion of patients impaired at baseline should also be reported.

Composition of cognitive test batteries

The 1995 Consensus Statement recommends authors include the Trailmaking Test (A and B), the Grooved Pegboard Test, and the Ray Auditory Verbal Learning Test. The cognitive test batteries in studies included in this review were composed of a median 6 (IQR 4–7) tests. Most studies used a diagnostic rule that adjudicates the POCD diagnosis based on relevant change in at least one or two of the tests in the battery. Authors should be aware that increasing the number of tests in the battery will increase sensitivity at the expense of specificity.[5,6]

To improve comparability between studies on perioperative cognition, we recommend authors choose commonly used cognitive tests, such as those presented in Table 2, in addition to the recommended core battery, if necessary, to address their specific research question. We recommend authors clearly report which test parameters were derived from each test and used in the diagnostic model.

Cut-offs and diagnostic rules

Reporting of cut-offs and diagnostic rules was often incomplete in the studies, which hinders comparability.

We suggest authors make results on all cognitive test parameters included in their diagnostic methods accessible in order to facilitate advanced meta-analyses.

The CS recommends the use of methods that take learning effects and natural variability in cognitive test performance over time into account. In studies included in this review, this

practice was still uncommon: 60.2% calculated a dichotomous POCD incidence with use of a simple analysis method only.

In order to comply with the CS and to improve comparability between studies, we recommend that authors use an RCI-based method to define relevant change in cognitive test performance over time. In line with the 2018 Nomenclature Recommendations, multiple severity levels should be discerned. We recommend authors specify mild decline (1–2 standard deviation [sd]) and major decline (>2 sd) for each test result. In addition, advancement of p-NCD to a clinical diagnosis according to DSM-5 criteria may help to distinguish if the change measured interferes with independence in everyday activities.

Neither the 1995 Consensus Statement nor the 2018 Nomenclature recommendations clearly state what the preferred diagnostic rule should be. We recommend future authors align their study design with the most commonly used diagnostic rule from our large sample of studies: p-NCD is diagnosed when relevant decline is present in at least two test parameters observed between 1 and 12 months postoperatively. Other studies have shown this practice also prevents overestimation of the p-NCD incidence.[5]

Control groups

The use of the RCI method requires inclusion of a control group.[5] The exemplary methodology of handling control group data to calculate a dichotomous incidence using RCI and a diagnostic rule is found in the methods of the ISPOCD-1 study.[22]

In our sample, only 28.6% of studies reporting dichotomous outcome included non-surgical controls. The control group composition, recruitment, and matching procedures were often not precisely described. The cognitive test performance in the control group is highly relevant and likely to be influenced by demographic factors and alignment with the surgical patient cohort. We recommend authors include control groups that are comparable with the surgical group, in order to reliably predict learning effects and natural variability in cognitive test performance over time.

Computerised testing

A substantial proportion of studies already used computerised testing. The computerised batteries commonly use their own specific cognitive tests, sometimes based on traditional paper-and-pencil tests.[23,24]

Computerised testing offers advantages in terms of data processing, avoids missing data, and reduces assessor effects by automating and standardising test instructions. In the future, it could be used to test patients more frequently and in the comfort of their own environment, providing a more accurate assessment of actual day-to-day functioning (ecological momentary assessment). However, this approach has disadvantages, such as reduced control over the testing environment. As it is more easily accessible, computerised testing has the potential to introduce cognitive testing into routine preoperative assessment. This would facilitate identification of patients at risk, so they may receive preventive measures. In follow-up, potential future treatment may be possible should patients develop cognitive decline.

In our opinion, the ideal future computerised cognitive test battery should have

1. An interface suitable for subjects of all ages and abilities
2. Appropriate normative data for the population under investigation (e.g. cardiac surgery cohorts require different normative data compared with orthopaedic surgery cohorts, considering confounding risk factors for cognitive decline)
3. Produce data that can be analysed with sound mathematical methods such as RCI

Conclusions

We conclude that, despite the advent of a Consensus Statement in 1995, the immense heterogeneity in methods used to study cognitive decline after surgery has remained. Many authors still insufficiently describe and adapt their methods to the specific research question in terms of patient selection, cognitive test batteries, follow-up timing, definitions of relevant change over time, and reporting of outcome. Non-surgical control groups are frequently not included in dichotomisation algorithms. The body of literature that has resulted thus is difficult to interpret and hinders meta-analysis. Meta-analysis though is urgently needed to reach the goal of identifying modifiable risk factors or evaluating potential preventive or therapeutic interventions for perioperative cognition. The inconsistent use of nomenclature poses further challenges. The large number of studies excluded from this review because of methods that do not allow for reliable conclusions on clinically relevant cognitive change after anaesthesia and surgery should sensitise the perioperative research community to the importance of sensible allocation of resources and ethics involved in subjecting patients to cognitive outcome studies.

The low compliance with the CS standards that we found in the published literature after its publication could indicate that there is either a requirement for more precise guidelines or that the ‘old’ consensus leaves room for very broad interpretation. We suggest working towards an international, interdisciplinary task force that could define standards for perioperative cognition research, including diagnostic criteria and updated standards for composition of cognitive test batteries. The implementation of high-quality research standards, and an interdisciplinary strategic approach should be able to advance this field to the next level: developing treatment options for affected patients who are at risk of a compromised quality of life after surgery.

Authors’ contributions

Idea and study concept: FB, AK, TO
Study design: FB, TO
Data extraction: FB, IF, WB, PK, MF, SK, CI, SS, DH, DV, FY, TO
Data management: FB, TO
Project management: FB
Preparation and submission of the manuscript: FB, TO
Critical revision of manuscript: CS, IF, WB, PK, MF, SK, CI, SS, DH, DV, FY, LE
Project supervision: CS, TO
Resolution of disagreements during data extraction: AK
Specialist consultation on neuropsychology: AK, DV
Expert consultation on perioperative cognition: LE

Acknowledgements

The authors thank A.J. Slooter, Simone van Montfort and Ilse Kant (University Medical Center Utrecht, The Netherlands) for their assistance in the screening process.

Declarations of interest

The authors declare that they have no conflicts of interest.

Funding

Institutional sources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bja.2021.01.035.

References

1. Murkin JM, Newman SP, Stump DA, Blumenthal JA. Statement of consensus on assessment of neurobehavioral outcomes after cardiac surgery. *Ann Thorac Surg* 1995; **59**: 1289–95
2. Hovaguimian F, Tschopp C, Beck-Schimmer B, Puhan M. Intraoperative ketamine administration to prevent delirium or postoperative cognitive dysfunction: a systematic review and meta-analysis. *Acta Anaesthesiol Scand* 2018; **62**: 1182–93
3. Miller D, Lewis SR, Pritchard MW, et al. Intravenous versus inhalational maintenance of anaesthesia for postoperative cognitive outcomes in elderly people undergoing non-cardiac surgery. *Cochrane Database Syst Rev* 2018; **8**: CD012317
4. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991; **59**: 12–9
5. Keizer AM, Hijman R, Kalkman CJ, Kahn RS, Dijk D van, Octopus Study Group. The incidence of cognitive decline after (not) undergoing coronary artery bypass grafting: the impact of a controlled definition. *Acta Anaesthesiol Scand* 2005; **49**: 1232–5
6. Nadelson MR, Sanders RD, Avidan MS. Perioperative cognitive trajectory in adults. *Br J Anaesth* 2014; **112**: 440–51
7. Lewis MS, Maruff P, Silbert BS, Evered LA, Scott DA. The influence of different error estimates in the detection of post-operative cognitive dysfunction using reliable change indices with correction for practice effects. *Arch Clin Neuropsychol* 2006; **21**: 421–7
8. Selnes OA, Grega MA, Bailey MM, et al. Neurocognitive outcomes 3 years after coronary artery bypass graft surgery: a controlled study. *Ann Thorac Surg* 2007; **84**: 1885–96
9. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009; **62**: 1006–12
10. Evered L, Silbert B, Knopman DS, et al. Recommendations for the nomenclature of cognitive change associated with anaesthesia and surgery. *Br J Anaesth* 2018; **121**: 1005–12
11. Ballard C, Jones E, Gauge N, et al. Optimised anaesthesia to reduce post operative cognitive decline (POCD) in older patients undergoing elective surgery, a randomised controlled trial. *PLoS One* 2012; **7**, e37410
12. Egawa J, Inoue S, Nishiwada T, et al. Effects of anesthetics on early postoperative cognitive outcome and intraoperative cerebral oxygen balance in patients undergoing lung surgery: a randomized clinical trial. Effets des agents anesthésiques sur l'issue cognitive postopératoire précoce et l'équilibre peropératoire d'oxygène cérébral chez les patients subissant une chirurgie pulmonaire: une étude clinique randomisée. *Can J Anaesth* 2016; **63**: 1161–9. https://doi.org/10.1007/s12630-016-0700-4
13. Lezak MD, Howieson DB, Bigler ED, Tranel D. *Neuropsychological assessment*. 5th edn. New York: Oxford University Press; 2012
14. Dogan S, Aybek T, Risteski PS, et al. Minimally invasive port access versus conventional mitral valve surgery: prospective randomized study. *Ann Thorac Surg* 2005; **79**: 492–8
15. Ernest CS, Worcester MUC, Tatoulis J, et al. Neurocognitive outcomes in off-pump versus on-pump bypass surgery: a randomized controlled trial. *Ann Thorac Surg* 2006; **81**: 2105–14
16. Farag E, Chelune GJ, Schubert A, Mascha EJ. Is depth of anaesthesia, as assessed by the Bispectral Index, related to postoperative cognitive dysfunction and recovery? *Anesth Analg* 2006; **103**: 633–40
17. Forrest CM, Mackay GM, Oxford L, et al. Kynurenine metabolism predicts cognitive function in patients following cardiac bypass and thoracic surgery. *J Neurochem* 2011; **119**: 136–52
18. Hudetz JA, Patterson KM, Pagel PS. Comparison of preexisting cognitive impairment, amnesic mild cognitive impairment, and multiple domain mild cognitive impairment in men scheduled for coronary artery surgery. *Eur J Anaesthesiol* 2012; **29**: 320–5
19. Ganguly G, Dixit V, Patrikar S, Venkatraman R, Gorthi SP, Tiwari N. Carbon dioxide insufflation and neurocognitive outcome of open heart surgery. *Asian Cardiovasc Thorac Ann* 2015; **23**: 774–80
20. Inouye SK, Marcantonio ER, Kosar CM, et al. The short-term and long-term relationship between delirium and cognitive trajectory in older surgical patients. *Alzheimers Dement* 2016; **12**. 766–65
21. Evered L, Scott DA, Silbert B. Cognitive decline associated with anaesthesia and surgery in the elderly: does this contribute to dementia prevalence? *Curr Opin Psychiatry* 2017; **30**: 220–6
22. Moller JT, Cluitmans P, Rasmussen LS, et al. Long-term postoperative cognitive dysfunction in the elderly ISPOCD1 study. *Lancet* 1998; **351**: 857–61
23. Ancelin ML, De Roquefeuil G, Scali J, et al. Long term postoperative cognitive decline in the elderly: the effects of anesthesia type, apolipoprotein E genotype, and clinical antecedents. *J Alzheimers Dis* 2010; **22**: S105–13
24. Askar FZ, Cetin HY, Kumral E, et al. Apolipoprotein E ε4 allele and neurobehavioral status after on-pump coronary artery bypass grafting. *J Card Surg* 2005; **20**: 501–5

Handling editor: Jonathan Hardman