Johns Hopkins University, Dept. of Biostatistics Working Papers

6-28-2004

Quantitative Methods for Tracking Cognitive Change 3 Years After Coronary Artery Bypass Surgery

Sarah Barry
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, sbarry@jhsph.edu

Scott L. Zeger
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu

Ola A. Selnes
Department of Neurology, The Johns Hopkins School of Medicine, oselnes@jhmi.edu

Maura A. Grega
Division of Cardiac Surgery, The Johns Hopkins School of Medicine, mgrega@csurg.jhmi.jhu.edu

Louis M. Borowicz, Jr.
The Zanvyl Krieger Mind Brain Institute, The Johns Hopkins University, lborowi1@jhmi.edu

See next page for additional authors

Authors
Sarah Barry; Scott L. Zeger; Ola A. Selnes; Maura A. Grega; Louis M. Borowicz, Jr.; and Guy M. McKhann

NS 035610-07

Quantitative Methods for Tracking Cognitive Change 3 Years after Coronary Artery Bypass Surgery

Sarah J.E. Barry, MS[1], Scott L. Zeger, PhD[1], Ola A. Selnes, PhD[2], Maura A. Grega, MSN[3], Louis M. Borowicz, Jr., MS[4], Guy M. McKhann, MD[2,4,5]

From Department of Biostatistics[1], The Johns Hopkins Bloomberg School of Public Health; Departments of Neurology[2], Surgery, Division of Cardiac Surgery[3], and Neuroscience[5], The Johns Hopkins University School of Medicine; and the Zanvyl Krieger Mind Brain Institute[4], The Johns Hopkins University

Word count: 2850

Corresponding Author: Sarah J.E. Barry, M.S.
Department of Biostatistics
Johns Hopkins School of Public Health
615 N. Wolfe Street
Baltimore
MD 21205
Phone: 410-614-1892
Fax: 410-955-0958
Email: sbarry@jhsph.edu

Abstract

Background: The analysis and interpretation of change in cognitive function test scores after Coronary Artery Bypass Grafting (CABG) present considerable statistical challenges. Application of hierarchical linear statistical models can estimate the effects of a surgical intervention on the time course of multiple biomarkers.

Methods: We use an "analyze then summarize" approach whereby we estimate the intervention effects separately for each cognitive test and then pool them, taking appropriate account of their statistical correlations. The model accounts for dropouts at follow-up, the chance of which may be related to past cognitive score, by implicitly imputing the missing data from individuals' past scores and group patterns.

We apply this approach to a study of the effects of CABG on the time course of cognitive function as measured by 16 separate neuropsychological test scores, clustered into 8 cognitive domains. The study includes measurements on 140 CABG patients and 92 nonsurgical controls at baseline, and 3, 12, and 36 months. Including a nonsurgical control group allows comparison of changes in cognition over time between the surgery group and patients with similar risk factors, controlling for potential effects of aging and vascular disease.

Results: CABG patients have longitudinal changes from baseline in cognitive function similar to those observed for nonsurgical controls. Any small differences tend to favor greater improvement in CABG patients than in the nonsurgical controls.

Conclusions: The methods used have application to a wide range of intervention studies in which multiple biomarkers are followed over time to quantify health effects. Software to implement the methods in commonly used statistical packages is available from the authors at http://www.biostat.jhsph.edu/research/software.shtml .

Introduction

The establishment of an association between the surgical procedure (CABG), and either short- or long-term cognitive change has been hampered by the use of studies involving only the CABG population, without comparison with suitable control groups. Short-term studies have compared the outcomes of CABG with those of other surgical procedures, such as orthopedic operations (1), but these controls do not have the high incidence of underlying risk factors for vascular disease that occurs in the CABG population. The lack of appropriate controls is particularly problematic for interpretation of studies concerning long-term cognitive performance after CABG, where possible decline could be related to the surgical procedure, age-related change, underlying cerebrovascular disease (2), or a combination of these factors.

We are involved in an ongoing study that allows comparison of patients receiving CABG with a group of individuals that have established coronary artery disease, but do not have surgery; these nonsurgical controls (NSC) have an incidence of risk factors for vascular disease similar to that of the CABG group. In the accompanying paper by Selnes et al. we compare the longitudinal performance of these two groups at baseline and 3, 12, and 36 months post surgery or enrollment. A first question is whether the pattern of cognitive change in the CABG group differs from that observed in the NSC group. A second question is whether any differences are likely caused by the surgery.

Determining whether there are different short-term or long-term declines in cognitive function for CABG as compared to a control group is challenging for at least three reasons. First, cognitive function is a multidimensional construct, which is assayed by numerous cognitive tests designed to assess different aspects of cognition such as memory, language, or executive function. Second, measurements of cognitive function over time will be affected by several factors, including practice effects, age-related change, error of measurement, and any intervention effect. Finally, the demonstrated medical efficacy of CABG makes a randomized treatment trial comparing surgical and nonsurgical treatments difficult.

In this paper, we discuss a hierarchical statistical model that can be used to quantify differences in change in cognitive function over time between the CABG and control groups. We use the statistical model to estimate the average cognitive function over time for the surgery and control groups, after adjusting for known differences in potential confounding variables, specifically age, gender, education, and the presence of symptoms of depression.

After a summary of the main ideas for hierarchical linear models, we estimate both short-term and long-term effects of CABG on cognitive function, while controlling for differences between the two populations at baseline and differences due to age, gender, education, and level of depressive symptoms. Finally, we combine the estimates of the surgery effects across many measures into domain-specific estimates of group differences. This method relies upon prior knowledge about the domains of cognitive function measured by each test.

The statistical approaches we have used for evaluation of prospective, longitudinal data comparing patients after CABG with a nonsurgical control group have general applicability to other clinical studies in which the goal is the evaluation of the impact of an intervention.

Methods

Study Design

This is an observational study of 140 patients undergoing CABG and 92 nonsurgical cardiac controls. Surgical patients (CABG) and nonsurgical controls were recruited from September 1997 through March 1999 at the Johns Hopkins Cardiac Unit. The NSC group was identified by Johns Hopkins cardiologists as potential patients who were diagnosed with coronary artery disease by cardiac catheterization.

Study participants were administered a battery of standardized neuropsychological tests at baseline, and 3, 12 and 36 months. Appendix 1 presents the 16 cognitive measures that are organized into 8 domains of cognitive function based principally on face validity. Patients were also administered The Center for Epidemiological Studies Depression scale (CES-D) at baseline and follow-up (3), in order to adjust cognitive test scores for possible effects of depressed mood. See the accompanying paper by Selnes et al. (2004) for a detailed description of the patient population and study design.

Hierarchical Linear Statistical Model (4)

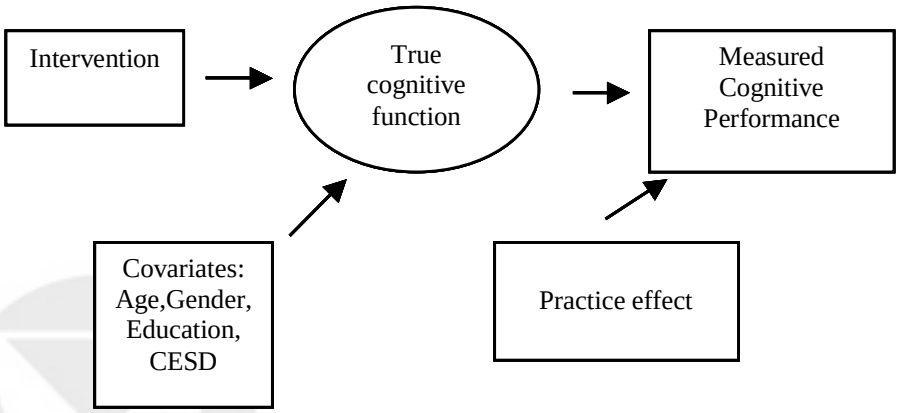
This section describes a now standard statistical model designed to capture the key components of the change in cognitive function over time for the individuals in our study and for their population, and to compare the typical change for persons who do and do not receive an intervention such as CABG. As an example we focus on a single cognitive domain, Verbal Memory; below we present a method for pooling results across tests to obtain the domain values.

The model is specified by the following assumptions:

- Each person has a unique level and time trend of cognitive function.

- Over periods of time, such as a few years, true cognitive function changes gradually and can be approximated by a smooth function of time, such as a low order polynomial (5).
- The intervention may affect people in the short term by immediately increasing or decreasing their function; and over the longer term by changing their pre-intervention trend. The short-term and long-term effects of intervention may vary across individuals.
- The level of cognitive function is influenced, possibly in a nonlinear way, by other factors such as age, gender, education, and level of depressive symptoms.
- Measurements of cognitive function are subject to a practice effect whereby a study participant's scores on quantitative tests could improve with repetition, particularly from the first to second testing, absent a change in actual cognitive function level.

Below is presented a schematic of this model. The goal is to estimate the effects of an intervention from a dataset comprising repeated observations on cognitive tests over time for persons who have received the intervention and other similar persons who have not.



Schematic representation of the statistical model for estimating the effects of an intervention (e.g., CABG) on a single measure of cognitive function. Ovals represent unobserved variables. See Appendix 2 for details.

The model illustrated above is made precise by the equations in Appendix 2. This model is implemented for a single cognitive function measure by using random effects software available in most standard statistical software packages.

The proposed model has two degrees of freedom to quantify a possible effect of CABG: the rise from 0 to 3 months (short-term, or learning effect); and a difference in the slope from 3 to 36 months (long-term effect). We use a Wald test (6) of the null hypothesis of no CABG effect by testing whether both are equal to 0.0.

The model has the ability to reduce bias caused by differential missing data between the groups when dropout is related to past cognitive score, resulting in individuals seen at all follow-up points having a different distribution of scores to the entire group. The model uses information from previous time-points and group patterns to internally impute missing data at later follow-up points and make more precise estimates of the true group means.

Estimating Natural Heterogeneity

In addition to estimating the mean curves for each intervention group, the model is used to estimate the variance among the true levels and trends in a cognitive test score among persons within groups (5). We allow this degree of variation to differ between the two intervention groups. Evidence for this natural variation derives from the correlation among repeated observations on each individual.

Pooling Intervention Effects Across Cognitive Measures

The hierarchical model in Appendix 2 is estimated separately for each of the 16 measures of cognitive function. This produces, for each measure, a short- and long-term intervention effect estimate and a 2x2 covariance matrix that quantifies their statistical error. We estimate the mean

short- or long-term “effect” for a domain as the mean of the effects for the tests in that domain. Since the multiple test scores for an individual are correlated with one another, estimates of the CABG effects for the different measures are also correlated. To correctly estimate the standard errors of these domain-specific or overall effects, we must take this correlation into account. We use bootstrapping (7), re-sampling individuals, to estimate the joint covariance matrix among the 16 pairs of intervention effect estimates and to obtain valid standard errors for the domain and overall intervention effect estimates. We draw with replacement a random sample of 140 CABG and 92 NSC subjects, refit all 16 models to get test-specific short- and long-term effect estimates, average these to obtain domain and overall effect estimates, and then repeat this process 1000 times. The variance among the 1000 bootstrapped replicates of the domain and overall effect estimates gives a valid estimate of statistical uncertainty, used to calculate the confidence intervals of the effect estimates, as it takes appropriate account of the correlation among multiple cognitive test scores for the same individual.

Results

The analyses were performed using the R software package (8).

Figure 1 is a spaghetti plot of the standardized and covariate-adjusted data for the Verbal Memory domain stratified by intervention group. The cognitive test scores were standardized such that the NSC group had a mean score of zero and standard deviation of one at baseline, and were adjusted for age, gender, education level and depressive symptoms. The Verbal Memory domain is made up of the total score, delayed recall (trial 8), retention score and corrected recognition from the Rey Auditory Verbal Learning Test (9). The group mean scores at each time are also shown. A learning effect is evidenced by the increase in mean score in both groups from baseline to 3 months. There is little change from 3 to 36 months in the mean response for either group after the initial rise.

Figure 2 illustrates the steps taken in fitting the hierarchical model detailed in Appendix 2 to the Verbal Memory data. Panel A (upper left) shows the mean Verbal Memory response for each treatment group and time, adjusted for covariates. Note that at baseline, the CABG group was nearly 0.4 of a standard deviation below the NSC group, even after correcting for age, gender, education, and depressive symptoms. At 3 months, both groups increase substantially as would be expected from a practice effect. There is some narrowing of the group difference. At 12 and 36 months, the means change comparatively little from the 3-month level in either group and although both decrease slightly, neither go below their baseline scores.

Panel B is obtained from Panel A by subtracting from all times the baseline values separately for each group to obtain *change scores* by time and group. Note that each group has the value 0.0 at baseline by definition. This step removes any differences between the two groups that are constant over time.

Panel C is a plot of the difference between the intervention (CABG) and control group (NSC) curves in Panel B at each time. These differences are the essential evidence relevant to assessing

the intervention (CABG) effect, as they show the disparities between the cognitive scores of the group that has had surgery and a group with similar risk factors that has not. Without this comparison, we cannot attribute any change in scores for the CABG group to the surgery. They still require some adjustment, however, because they are averages of observed data and do not take account of the fact that the number of dropouts may differ between the two groups.

Panel D presents improved estimates of the difference at each time between the CABG and NSC group curves shown in C. The results in D are obtained from the hierarchical model in Appendix 2. The small differences in the curves in Panels C and D reflect changes from taking appropriate account of missing data. In the case of this domain, the CABG patients remaining at 36 months are those with poorer cognitive scores at baseline than the individuals that dropped out, with the opposite effect in the NSC group, a phenomenon likely to bias the mean group score at 36 months towards lower values. The hierarchical model implicitly imputes the missing data by using the information for the individuals at previous times and the patterns for their group. The means in Panel C ignore the missing data and are biased unless the chance of dropping out is unrelated to past cognitive score, which is an unlikely situation.

Panel D shows evidence of a difference in population mean Verbal Memory value between the intervention and control groups, with the CABG group having a greater improvement from baseline than the NSC group, both in the short-term at 3 months and in the long-term at 36 months. The test of the null hypothesis that both the short- and long-term effects are zero has a p-value of 0.01, indicating that this hypothesis can be rejected in favor of the CABG group.

Figure 3 shows the estimated difference in the cognitive function time course between the intervention (CABG) and NSC groups (Panel D in Figure 2 for Verbal Memory) for all 8 of the cognitive domain measures. The accompanying paper by Selnes et al. (2004) presents the corresponding figure for the 16 cognitive subtests. The p-values on the plots in Figure 3 again result from tests of zero difference in the trends over time between the groups and make it clear

that in this dataset, there is little or no evidence consistent with a detrimental effect of CABG on cognitive function as measured by these 8 scores.

Discussion

We have used a hierarchical linear statistical model to quantify the evidence relevant to assessing whether CABG causes short- or long-term decline in cognitive functioning. Our approach was to estimate this model separately for each of the cognitive measures and to pool the results across measures into cognitive domain effects. We take appropriate account of the correlation among repeated measures for an individual when setting confidence intervals for the average domain effects.

The findings in this study emphasize the importance of having a control group for at least two reasons. First, there is strong evidence in both groups of an improvement in the mean score from baseline to 3 months indicative of a learning effect. In 15 out of the 16 cognitive test measures, the mean score increased over this initial period in the CABG group. If only the CABG data were available, what appears to be a learning effect might be mistaken as a benefit of the intervention. Second, there is a mix of positive and negative trends over the 36 months in the CABG group for the different cognitive tests and domains. But these trends were statistically different from the trends observed in the NSC group for only two domain measures and in both cases showed a positive effect of CABG at 36 months. Hence a trend in the CABG group data should not be mistaken for evidence of a treatment effect without comparison with controls.

The model described here estimates the average difference between the CABG and control groups in the **change** in cognitive function from baseline. We use the model to adjust for baseline differences between the groups in test scores and for differences over time that are attributable to demographics and depression symptoms, the latter as measured by the CESD.

However, no model can adjust for unmeasured differences between the two groups that are more likely to arise in observational studies where subjects choose their treatment in consultation with their physician rather than having it assigned by a known, random mechanism. Hence, we

must be cautious in our interpretation of the evidence, asking what other factors might account for the differences or lack thereof between the two groups.

The hierarchical model allows one to take appropriate account of dropouts, a common phenomenon in longitudinal studies such as this one. The model includes terms that acknowledge the correlation among repeated observations for each individual. Having done so, it can internally impute missing values by predictions based upon the earlier responses and other covariates (5). Failure to use a model that accounts for within-person correlation can lead to biased estimates of treatment effects except when the dropout process is independent of the past responses, which is unlikely.

We have presented an approach to the difficult problem of how to estimate the effect of CABG on the performance of 16 cognitive measures by first analyzing each of them separately with an hierarchical model and then pooling the effect estimates to obtain domain effects. We refer to this as the “analyze then summarize” approach. An alternative is to “summarize then analyze” the data by using factor analysis (10) or some other method to create summary scores from the 16 test results and then to use hierarchical models with the summary measures. We prefer our approach because it produces a separate treatment effect for every measure so that unanticipated patterns can be discovered. It avoids the difficult problem of how to choose the best summary scores. Typically, summarization is based upon the correlation among test results at one time and does not take appropriate account of the longitudinal information.

The methods used here have wide application to a variety of longitudinal studies comparing intervention groups or groups defined in other ways when the outcome is multivariate. To facilitate the application of these methods, software to implement the analyses presented here has been posted to our webpage (http://www.biostat.jhsph.edu/research/software.shtml).

Acknowledgments:

This study was supported by grant 35610 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD; by the Charles A. Dana Foundation, New York, N.Y. and the Johns Hopkins Medical Institution GCRC grant RR 00052.

Special thanks are extended to Maryanne Bailey, Catherine Cristinzio, Sarah Moeller, and Sharon Owens, who assisted in the neuropsychological assessments, and to our study participants who volunteered their time and energy to make this study possible. We thank Pamela Talalay, Ph.D. for her help during the preparation of this manuscript. We would also like to thank the cardiologists, cardiac surgeons, and anesthesiologists at our institution as well as Johns Hopkins Bayview Medical Center who helped with this study.

References

1. Williams-Russo P, Sharrock, NE, Mattis S, Szatrowski TP, Charlson ME. Cognitive effects after epidural vs general anesthesia in older adults. JAMA 1995; 274:44-50.
2. Knopman DS, Boland Ll, Mosley T, Howard G, Liao D, Szklo M, et al. Cardiovascular risk factors and cognitive decline in middle-aged adults. Neurology 2001; 56: 42-48.
3. Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Measurement 1977; 1:385-401.
4. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982; 38:963-74.
5. Diggle PJ, Heagerty P, Liang KY, Zeger SL. The analysis of longitudinal data 2nd ed. Oxford, England: Oxford University Press 2002.
6. Agresti A. Categorical data analysis 2nd ed. Wiley-Interscience 2002.
7. Efron B, Tibshirani R. Improvements on cross-validation: The Bootstrap method. J Am Stat Assoc 1997; 92:548-60.
8. Ihaka R, Gentleman R. A language for data analysis and graphics. J Computational Graphical Stats 1996; 5:299-314.
9. Rey A. L'examen clinique en psychologie. Paris: Presses Universitaires de France 1964.
10. DeVellis RF. Scale Development: Theory and Applications. Newbury Park, California: SAGE Publications, Inc. 1991.

Fig 1. Spaghetti plots of the standardized and adjusted (for age, gender, education and depressive symptoms) Verbal Memory scores across time, stratified by treatment group.

Fig 2. Illustration of the steps taken to fit the hierarchical model to the Verbal Memory data; Upper left Panel A – average curves for each intervention group; Upper right Panel B – average change from baseline for each intervention group; Lower left Panel C – difference in mean change from baseline between the CABG and NSC groups; Lower right Panel D – model estimates of group difference in change from baseline that take appropriate account of missing data.

Fig 3. Model estimates of the difference in cognitive function change between the CABG and NSC groups over time for all 8 domains, where a positive difference in the solid line indicates a greater improvement from baseline in the CABG group than in the NSC group. Note that where there is only one test in a domain, the estimate for that particular test is used instead of bootstrapping.

Appendix 1: Details on cognitive function tests.

Test names, abbreviations and descriptions, along with the domains that are made up of them

Test Name	Abbreviation	Description	Domain
Rey Auditory Verbal Learning Test		A word-list learning task assessing verbal learning, retention and recognition memory	Verbal Memory
Total score	RVLTTOT		
Delayed recall (Trial 8)	TR8		
Retention score	RETEN		
Recognition (corrected)	RECCORR		
Rey Complex Figure		A measure of ability to recall a complex visual design previously copied	Visual Memory
Retention score (%)	RCFRET		
Delayed recall	RCFDR		
Rey Complex Figure Copy	RCFC	A measure of visuospatial abilities requiring subject to copy a complex visual design	Visuo-construction
Block Design	BLOCKS		
Boston Naming Test (short form)	BNT	A measure of visual confrontation naming requiring subject to name a series of 30 line drawings	Language
Grooved Pegboard		Test of motor speed measuring how quickly subject is able to place 25 keyed pegs in an array of 5 x 5 holes with randomly positioned slots	Motor Speed
Dominant hand	GPDOM		
Non-dominant hand	GPNDOM		
Trail Making Test – Part A	TRAILA	Timed task that requires subject to connect numbered circles in sequence as quickly as possible	Psychomotor Speed
Written alphabet	WA	Timed measure of psychomotor speed in which subject is asked to write letters of alphabet as quickly as possible	
Rey Auditory Verbal Learning Test		A word-list learning task assessing verbal learning, retention and recognition memory	Attention
Trial 1	TR1		
Mini-Mental Status Examination		Attention score from Mini-Mental State Exam	
Attention score	MMATT		
Trail Making Test – Part B	TRAILB	Timed test of psychomotor speed that requires participant to connect numbered and lettered circles alternately in sequential (numerical and alphabetical) order	Executive Function

Appendix 2: Hierarchical statistical model for a single biomarker

This appendix specifies the hierarchical linear statistical model to quantify the difference in average time-curve for intervention and control groups and to estimate an individual's time curve acknowledging the natural heterogeneity among persons in the level and trend of the cognitive functioning. To be precise, we make the following definitions:

η_{it} – true cognitive function level for person i at time t

μ_{it} – true level of cognitive function absent the intervention

X_i = 1 if intervention received and 0 if not

δ_{it} – intervention effect for person i at time t

Y_{it} – measurement of cognitive function

Z_{it} – covariate information for including: age, gender, education level, CESD depression index

$Post_t$ = 1 for post-intervention visits and 0 at baseline

β_p – change in score from the first to second measurement due to a learning effect

We assume that the true cognitive function level for person i at time t is the sum of their value μ_{it} that would be expected absent the intervention plus an intervention effect δ_{it}:

$$\eta_{it} = \mu_{it} + X_i \delta_{it} . \qquad (1)$$

We then assume that the baseline curve can be approximated by a linear function with intercept and trend that are specific to each person (i):

$$\mu_{it} = \beta_{0i} + \beta_{1i} t . \qquad (2)$$

We let the level at baseline β_{0i} be influenced in possibly non-linear ways by covariates such as age, gender, education level, and depression symptoms (CESD) and to have a person-specific

deviation b_{0i} from the population average reflecting other unmeasured influences in cognitive function.

$$\beta_{0i} = \beta_{00} + b_{0i} + S(age_i) + \ldots + S(CESD_i) \qquad (3a)$$

Similarly, the trends are allowed to vary among persons with each person having his or her own deviation b_{1i}:

$$\beta_{1i} = \beta_{10} + b_{1i}. \qquad (3b)$$

We assume that the treatment effect δ_{it} is also linear with subject-specific intercept δ_{0i} and trend δ_{1i}:

$$\delta_{it} = \delta_{0i} + \delta_{1i}t \qquad (4)$$

We assume that collectively, the intercepts and trends for the cognitive function of time absent intervention (b_{0i},b_{1i}) and for the intervention effect $(\delta_{0i},\delta_{1i})$ can be thought of as samples from a Gaussian distribution with variances G_b and G_d.

Finally, we acknowledge that the measured cognitive test score includes a learning effect assumed to be roughly constant after the baseline measure plus statistical noise that is assumed to be independent from one time to the next, that is:

$$Y_{it} = \eta_{it} + Post_i \beta_p + \varepsilon_{it}. \qquad (5)$$

The model represented by equations (1) - (5) can be readily simplified to deal with the case of a single group. This reduced model is defined by equations (2), (3) and (5) and includes random effects (b_{0i},b_{1i}) only.

Figure 1

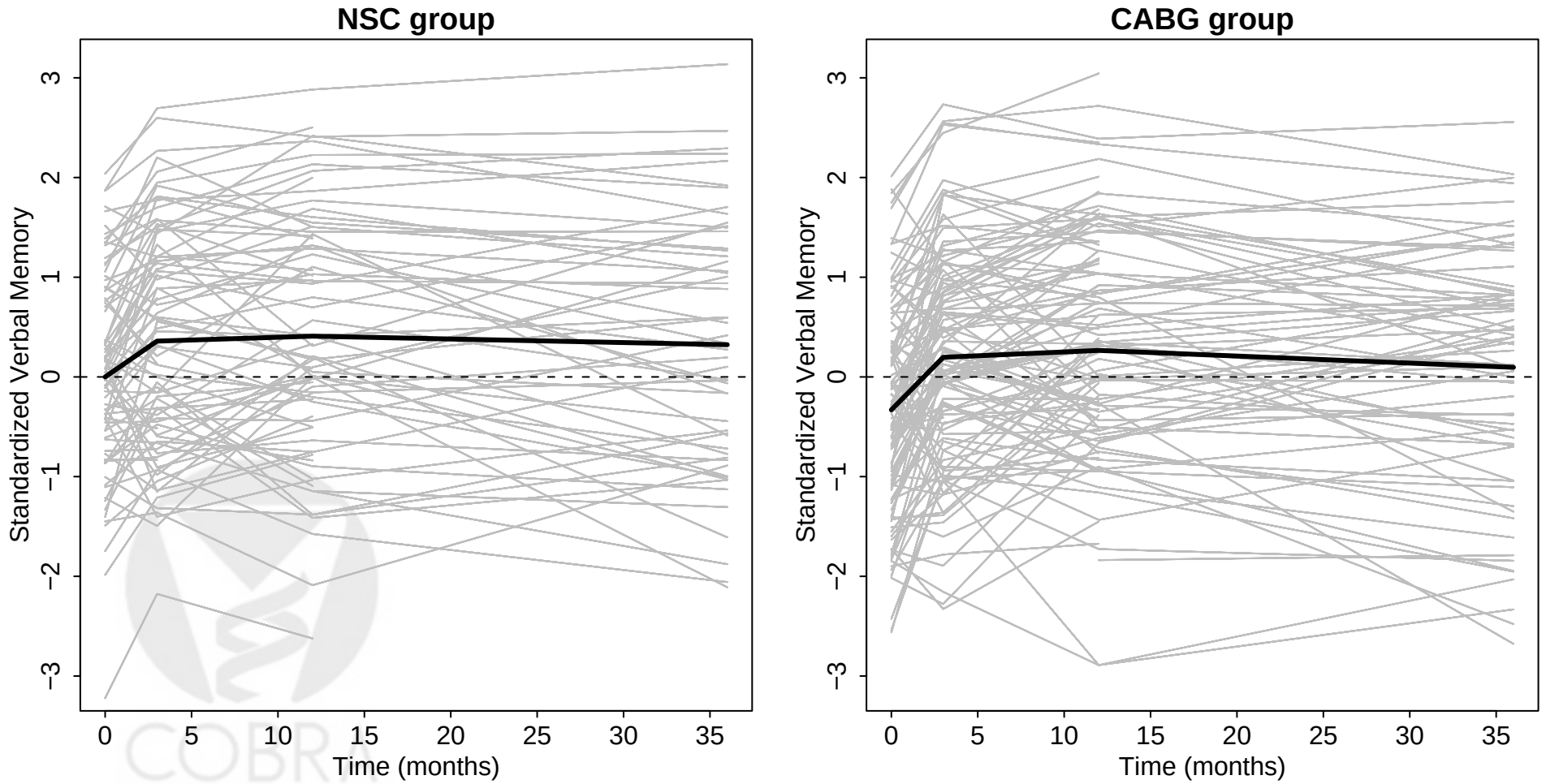


Figure 2

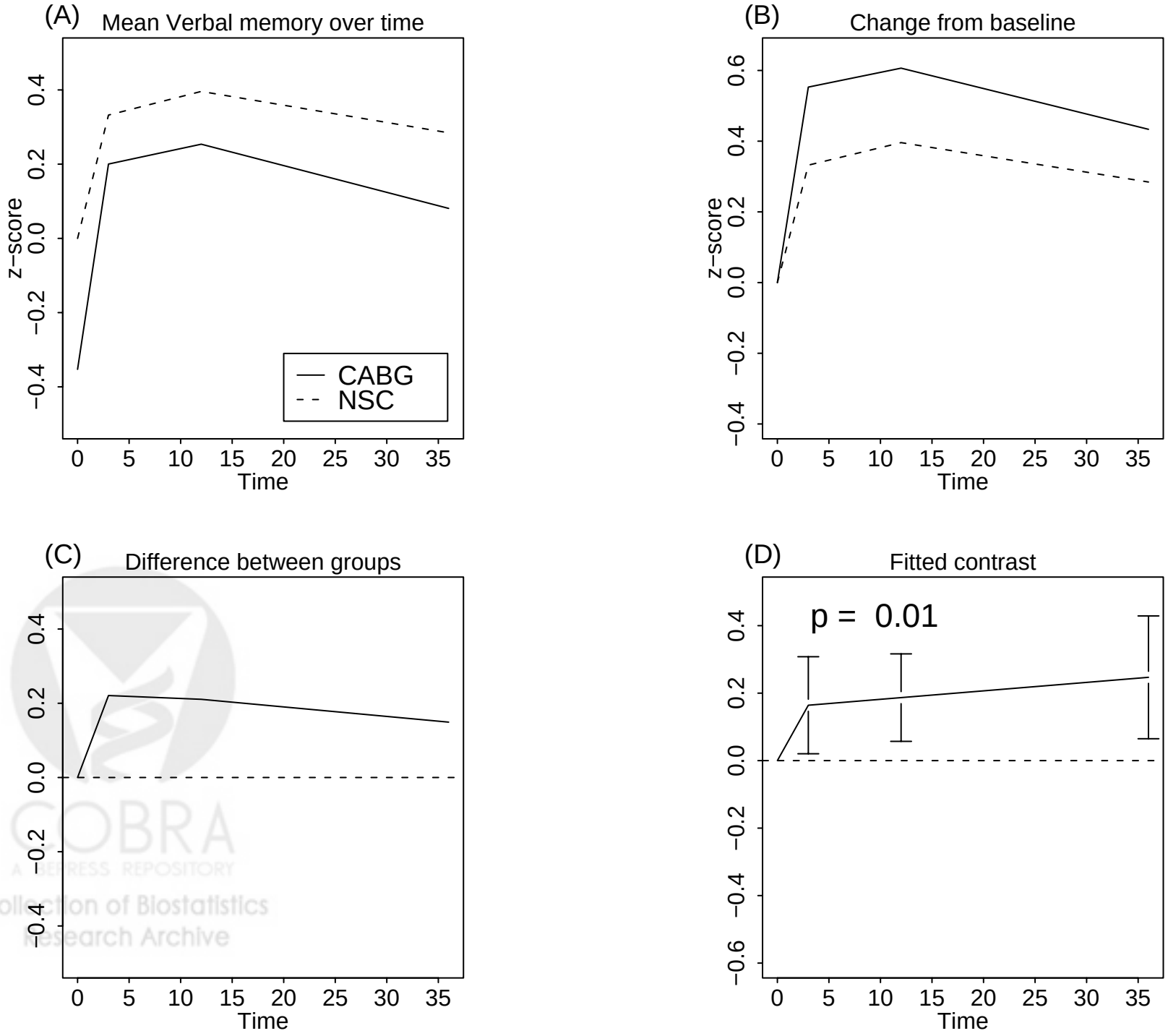


Figure 3