Using Mahalanobis distance to evaluate recovery in acute stroke

Hannah Tehan

Kate Witteveen

G. Anne Tolan

School of Psychology, Australian Catholic University, Banyo, Queensland, Australia

Gerald Tehan

Graeme J. Senior

School of Psychology and Counselling, University of Southern Queensland, Springfield, Queensland, Australia

Word Count: 2646

Corresponding Author:
Gerry Tehan
School of Psychology and Counselling
University of Southern Queensland, Springfield, Australia
e-mail: tehan@usq.edu.au

Abstract

Objective

In the weeks immediately following a stroke, impairments across multiple cognitive domains are pervasive yet there is little literature that explores cognitive recovery during this period. This paper evaluates the use of Mahalanobis distance as a means of statistically evaluating cognitive change at the individual level.

Method

A small battery of standardised neuropsychological tests was administered on five or six occasions across a two week period to the participants recovering from a stroke and a non-stroke control group. Mahalanobis distance was used to evaluate the change profile of those who were recovering from a stroke relative to the non-stroke control.

Results

The outcomes of three patients show that Mahalanobis distance could statistically differentiate recovery, no change, and deterioration from normal repetition effects.

Discussion

In the acute phase of stroke using Mahalanobis distance it is possible to distinguish between recovery, normal learning, and gerneralised learning deficits thereby identifying likely candidates for further cognitive assessment and rehabilitation.

Large scale studies assessing cognitive abilities in people who have recently suffered a stroke, have shown that upwards of 80% of patients have impairment in at least one area of cognition, with 65% of patients show mutiple impairments across diverse cognitive domains (Jaillard, Naegele, Trabucco-Miguel, LeBas, & Hommel, 2009; Nys, van Zandvoort, de Kort, Jansen, de Haan, & Kappelle, 2007). While early cognitive impairment is pervasive, there are few published studies that attempt to track multi-domian cognitive recovery over the first three months post stroke. Tracking the recovery process involves multiple test sessions, which produces its own set of problems, and logically, it makes sense to distinguish between those domains that have been impaired and those that have not. Recovery should be limited to those domains that have been affected by the stroke. In this study we exlore the use of Mahalanobis distance (MD) as a means of statistically determining the extent of recovery in acute stroke at the individual level.

On most standardised neuropsychological tests performance improves on a second administration (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010). Such repetition (practice) effects have traditionally been seen to introduce unwanted noise into measurement. However, as a number of authors have recently suggested, the presence, absence, or differential strength of repetition effects have the potential to provide clinically useful information (Darby et al., 2002; Duff, Callister, Dennett, & Tometich, 2012). Discriminating recovery from practice is currently hampered by the lack of normative information regarding changes across multiple test administrations over a brief period. Ideally, the recovery issue could be resolved if normative repetition effects were available and there was a method of comparing the individual's pattern of performance over multiple test sessions to that of a normative sample. We argue that Mahalanobis distance is one method that has the potential to solve the current problems of discriminating between patterns of normal and abnormal behaviour change.

In other scientific disciplines, Mahalanobis distance has been widely used as a means of generalised pattern analysis to establish clusters of data points or classify data points into different groups. As such, it has emerged as a central application in protocols for computerised face recognition and more general aspects of computerised image analysis. A secondary application of Mahalabobis distance involves the indentification of outlier patterns. For example, using cluster analysis, traffic flow conditions on a multi-lane highway can be grouped into a set of normal patterns that vary by time of day, weather conditions, etc. Mahalonobis distance-based algorithms have been used to identify abnormal, outlier flow patterns caused by accidents (Warren, Smith, & Cybenko, 2011). In psychology, the used of Mahalanobis distance has typically been used in this secondary role of identifying multivariate outliers, that is identifying patterns of data that do not belong to identified clusters or classified groups. Conceptually, Mahalanobis distance is the equivalent of a multi-dimensional z-score, and serves a similar function in being able to classify individual scores as either members or outliers of a parent population. In the context of stroke, patients may be outliers in terms of overall levels of performance, repetition effects, or a combination of absolute levels and repetition effects. In what follows we explore the utility of Mahalanobis distance for assessing patterns of behaviour change associated with repeated testing on a small battery of standardised neuropsychological tests, using the data from three participants recovering from a stroke. We make the distinction between intact and impaired cognitive domains on the assumption that change patterns might differ in each case.

Method

Measures

The full test battery consisted of seven cognitive tasks. The five standardised neuropsychological tests were the Stroop Colour and Word Test (Golden, 1978) as measures of attentional functioning, general cognitive efficiency or resistance to interference; the

WAIS-IV digit span sub-tests (Wechsler, 2008) as a measure of attention and working memory; the Verbal Associative Fluency Test (Spreen & Benton, 1969) and the Animal Naming Test (Goodglass & Kaplan, 1972) as executive measures of cognitive organization, initiation, and maintenance of effort; and the Rey Tangled Lines Test (RTLT) (Rey, 1958) as a measure of visual tracking under interference. Full descriptions of the battery can be found at https://osf.io/2qpxs/

Procedure

The battery was administered on five or six sessions, usually within a two-week period, to patients in the rehabilitation ward of a metropolitan hospital in Brisbane, Australia, who had all suffered a stroke in the preceding twelve weeks. The length of each session was limited to approximately 40 minutes to control for fatigue, illness and problems in concentration (Nys et al., 2005). The battery was also administered to a control group on six occasions across a two week period.

Given that at the individual level, not all cognitive domains are impaired it makes intuitive sense to compare change profiles on tasks that are deemed to be impaired on baseline testing with change profiles on preserved tasks. Standard repetition effects might be expected on tasks that have been preserved, but recovery may or may not emerge on impaired tasks. It is also possible that a stroke might produce a generalised deficit in which case the profile depicting little or no change could emerge on both preserved and impaired tests. Because composites of tests are more reliable than individual tests, in what follows we have formed two composite scores for each patient, based upon performance on the first test session. While composite scores are usually constructed on the basis common processes, we have formed composites by converting the raw data of each test to scaled scores (Mean =10, SD = 3) based on the mean and standard deviations of the control group. Then for each individual the scaled scores were averaged across the tasks that were preserved to form the

persevered composite, and similarly, the scaled scores on the tasks that were impaired were averaged to form the impaired composite. The scaled scores for impaired and preserved composites were calculated for each session, and the composite scores across sessions formed the basis for the Mahalanobis distance analyses.

Participants

Non-stroke Participants: The control group consisted of 26 volunteers (10 male and 16 female) whose ages ranged from 55 to 87 (M = 65.77, SD = 8.13), a range that covers 87% of strokes in Australia. A subset of this sample served as age-matched controls, but as the outcomes did not change as a function of full or part sample, the full sample is used as a reference point. All participants were in a current state of good health, all lived independently and none had a diagnosis that was associated with impaired cognition.

Patients with Stroke: **KS,** a 55-year-old male, was first tested 53 days after a right middle cerebral artery ischaemic stroke that left him densely hemiplegic on the left side of his body, causing left-side facial droop, and significant mobility impairment. On first testing KS was deemed to be impaired on all three trials of the Stroop task and Semantic fluency measures, but not on digit span, phonemic fluency, or RTLT.

MF, an 82-year-old female, was first tested 3 days after a right frontal lobe subacute infarct in her cingulate cortex, resulting in reduced power in her lower limbs, and cognitive problems with naming and abstract reasoning. On initial testing MF was impaired on the three Stroop measures, and on semantic fluency tasks. She was not impaired on digit span, phonemic fluency, or RTLT.

GL, an 82-year-old male, was first tested 5 days after suffering a subdural haematoma causing damage to both hemispheres of his brain resulting in upper limb weakness, slurring of his speech and difficulty moving, but no cognitive issues. On initial testing GL was

impaired on two of the three Stroop measures, on the RTLT, and semantic fluency. He was not impaired on the Colour-Word Stroop measure, digit span, or phonemic fluency.

Measures of Change

Two Mahalanobis distance measures were taken. The first, MD, assessed overall performance. The second, and more important, evaluated the change profile slopes (MD-R[ecovery]), independent of overall levels of performance. To create each measure, the control sample performance across the six sessions was used to generate an inverse covariance matrix which constitutes the denominator in the MD computations. The numerator differed for the MD and the MD-R. The numerator for the MD was computed by subtracting the individual patient's scores over each of the testing sessions from the corresponding mean for the control sample. A difficulty with this, however, is that individuals who have substantially lower than normal scores would register as a multivariate outlier based on the magnitude of the difference alone and not because of the pattern of change scores over time. For this reason, the numerator for the MD-R was computed and evaluated in which the individual's difference scores were adjusted to be identical to that of the control sample on the first test session. By adding this difference to all subsequent scores, the individual's scores are anchored to the control group's first score. In phase shifting the pattern of scores over trials to that of the control group (see Figure 1), the influence of magnitude of score differences is minimised and the focus of the multivariate outlier analysis is focused solely on the pattern of change over time. The probability of the obtained distance (d_m) can be evaluated statistically because d_m^2 is chi-square distributed. Consequently, d_m^2 will be reported in all analyses.

Results

Performance of the control group is summarised in Table 1. Repetition effects in the control group were analysed by means of one-way repeated measures ANOVAs conducted

on each measure and the outcomes are also presented in Table 1. For all measures performance improved above baseline performance, such that a linear function accounted for over 90% of variance across sessions in each of the measures.

Composite Scores In order to evaluate change across test sessions, the scores for each task were converted to scaled scores that have a mean of 10, a standard deviation of 3 and a cut-off for impairment a score of 4 (corresponding to a percentile ranking of 2 or a z-score of -2.00). The scaled scores were then averaged across tasks. With the participants recovering from a stroke, separate averages were calculated for initially impaired performance and initially intact measures (see Method section), and performance was compared to the control sample on the same set of tests. The bottom line in Figure 1 presents the composite scores (MD) for initially impaired and unimpaired measures for each of the stoke patients. Recovery (MD-R) using the same composites is also presented in each panel, where initial values of patient and control group have been equated.

In terms of the MD measure KS was deemed to be an outlier on the initially unimpaired tasks, d_m^2= 9.64, p = .045, and the impaired tasks, d_m^2= 50.39, p <.001. The MD-R measure showed normal repetition effects on the unimpaired tasks in that improvement across sessions is similar to that observed in the control group, d_m^2= 5.21, p = .267. On the initially impaired measures, the change profile is steeper than for the control group, d_m^2 = 10.07, p = .039. KS shows accelerated learning indicative of recovery.

MF was also considered to be a multivariate outlier in terms of the MD for both unimpaired, d_m^2= 14.79, p = .005, and impaired composites, d_m^2= 15.11, p = .004. However for the MD-R measure, MF showed normal repetition effects on the unimpaired tasks in that improvement across sessions is similar to that observed in the control group, d_m^2= 6.18, p = .19. However, on the impaired tasks, MF shows an outlier profile, d_m^2= 9.85, p = .043, as represented by a deterioration in performance on the fourth and fifth sessions.

GL is a multivariate outlier on the MD measure for both initially unimpaired, d_m^2= 28.52, p <.001, and impaired composites, d_m^2= 58.02, p <.001. For the MD-R GL is statistical outlier for the unimpaired measures, d_m^2 = 28.52, p <.001, but not for the impaired measures, d_m^2= 7.32, p = .19. For both composites, there is a deterioration in Session 2 with no change in outcomes on the remaining four sessions. In short, there is no evidence for either normal repetition effects, nor for recovery.

Discussion

While the three patients presented here were chosen because they reflect different change profiles, they share the general characteristics of large-scale studies, in that each patient shows preserved function in some tasks and multiple impairments across other tasks (Jaillard et al., 2009; Nys et al, 2007). Moreover, the precise profile of impaired and preserved tests differed from person to person.

The primary aim of this paper was to determine the effectiveness of using Mahalanobis distance as a way of evaluating recovery in acute stroke where composite scores were utilised to increase the reliability of the measures. When MD was calculated, the combined influence of absolute levels of performance and behavioural change across the test sessions resulted in large deviations from the control group for all three patients in both impaired and unimpaired composites. The adoption of the MD-R measure was aimed at eliminating the initial difference in absolute levels of performance between control group and patient so that behaviour change became the metric that was evaluated. Here we observed three distinct patterns of performance. For KS and MF, normal repetition effects were observed on the composite consisting of tests that were initially unimpaired. In the case of KS, there was accelerated improvement on the impaired tests compared to the control group, which is consistent with the recovery process. For MF, performance on the impaired tasks showed initial improvement but was followed by two sessions in which performance

deteriorated substantially. For GL, there was no evidence of improvement for either impaired or unimpaired domains. His performance is indicative of a generalised learning difficulty. GL's performance also points to a possible limitation to the utility of this measure. While it is clear that GL does not improve on the impaired scores, the MD-R was not significantly different to that of the control group. Thus, it is still the case that absence of change in mean levels of performance in the MD-R measure is still the primary piece of evidence for evaluating the recovery process. In the case of GL there was no behaviour change but for MF and KS scores did improve over baseline performance.

There are clear limits to the current research. The tasks on test battery were selected for the brevity of administration time and consequently were predominantly speeded response tasks. Other cognitive domains like visual and verbal memory were not evaluated. Moreover, it is not certain that the relatively clear patterns in the current study would emerge if a different set of cognitive tasks were employed. Normal repetition effects were not equivalent across tasks, with statistical significance from baseline only emerging on Session 5 on the RTLT. Future work in this areas should explore those tasks that do show substantial repetition effects. Finally, while three participants are sufficient to demonstrate that Mahalanobis distance can be used to identify distinct recovery profiles, larger numbers of participants need to be examined before the clinical utility of the measure can be firmly established.

In conclusion, the study is best thought of as a proof-of-concept demonstration that in a hospital setting the repeated administration of a battery of brief standardised tests can produce stable data; cognitive domains that have been impaired by a stroke and those that are unimpaired can be identified; recovery process in impaired domains (or lack thereof) can be identified at the individual patient level thereby affirming the importance of practice effects as an additional marker of cognitive impairment (Darby et al., 2002; Duff et al., 2012). The

study confirms that Mahalanobis distance is a useful method for evaluating normal and abnormal behaviour change in acute stroke. Early detection of recovery is possible, and it is possible to identify, at an early stage, the likely candidates for further cognitive assessment and rehabilitation.

References

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience, 11* doi:10.1186/1471-2202-11-118

Darby, D., Maruff, P., Collie, A., & McStephen, M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology, 59,* 1042-1046.

Duff, K., Callister, C., Dennet, K., & Tometich, D. (2012). Practice effects: A unique cognitive variable. *The Clinical Neurologist, 26,* 1117-1127. doi: 10.1080/13854046.2012.722685

Golden, C. (1978). *Stroop color and word test*. Illinois: Stoelting Company

Goodglass, H. & Kaplan, E. (1972). *Assessment of Aphasia and Related Disorders.* Philadelphia: Lea and Febiger.

Jaillard, A., Naegele, B., Trabucco-Miguel, S., LeBas, J.F., & Hommel, M. (2009). Hidden dysfunctioning in subacute stroke. *Stroke, 40,* 2473-2479. doi: 10.1161/STROKEAHA.108.541144

Nys, G. M. S., Van Zandvoort, M. J. E., de Kort, P. L. M., Jansen, B. P. W, Van Der Worp, Kappelle, L. J., & de Haan, E. H. (2005). Domain-specific cognitive recovery after first-ever stroke: A follow-up study of 111 cases. *Journal of the International Neuropsychological Society, 11,* 795-806. doi: 10.1017/S1355617705050952

Nys, G. M. S., Van Zandvoort, M. J. E., de Kort, P. L. M., Jansen, B. P. W., de Haan, E. H., & Kappelle, L. J. (2007). Cognitive disorders in acute stroke: Prevalence and clinical determinants. *Cerebrovascular Diseases, 23,* 408-416. doi: 10.1159/000101464

Rey, A. (1958). L'examen clinique en psychologie. Paris, France: Press Universitaires de France.

Spreen O., Benton D. F. (1969). Neurosensory Center of Comprehensive Examination for Aphasia: Manual of directions. Victoria, BC: Neuropsychology Laboratory, University of Victoria.

Warren, R., Smith, R. E., & Cybenko, A. K. (2011). Use of Mahalanobis Distance for detecting outliers and outlier clusters in markedly non-normal data: a vehicular traffic example. Interim Technical Report, United States Air Force. http://www.dtic.mil/dtic/tr/fulltext/u2/a545834.pdf

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition.* San Antonio, TX: Pearson.

Table 1.

Mean performance of control group on test battery as a function of test session.

		Session							
		1	2	3	4	5	6	F	η_p^2
Phonemic Fluency	Mean	49.58	53.54	56.62	58.42	62.65	64.23	20.91	.45
	SD	15.61	16.94	14.70	16.04	17.82	18.08		
Semantic Fluency	Mean	74.58	79.85	83.92	86.81	87.23	91.19	24.82	.49
	SD	15.02	17.11	15.91	17.93	19.65	18.86		
Stroop - Word	Mean	102.65	106.69	108.81	110.27	110.19	112.15	9.37	.22
	SD	11.65	12.27	11.48	12.44	12.87	12.68		
Stroop - Colour	Mean	80.19	83.73	84.12	87.04	86.88	88.65	11.51	.31
	SD	11.07	14.26	13.67	14.39	13.23	16.23		
Stroop – Colour Wo	Mean	49.15	52.27	54.50	56.88	57.85	61.27	26.70	.52
	SD	13.45	13.27	12.72	15.46	14.41	15.04		
Digit Forward	Mean	11.50	11.50	11.19	12.08	12.27	12.54	14.01	.15
	SD	2.08	2.23	2.12	2.33	1.85	1.98		
Digit Backward	Mean	8.81	9.08	10.12	9.96	10.12	10.65	23.32	.47
	SD	2.25	2.06	2.36	2.42	2.25	2.17		
Rey Tangled Lines	Mean	8.40	8.47	8.23	7.97	7.88	7.90	2.62	.09
	SD	1.78	1.67	1.69	1.89	1.61	1.66		

Note: degrees of freedom for all ANOVAs (5,130); p =<.001 for all tests except Rey Tangled Lines where p = .02

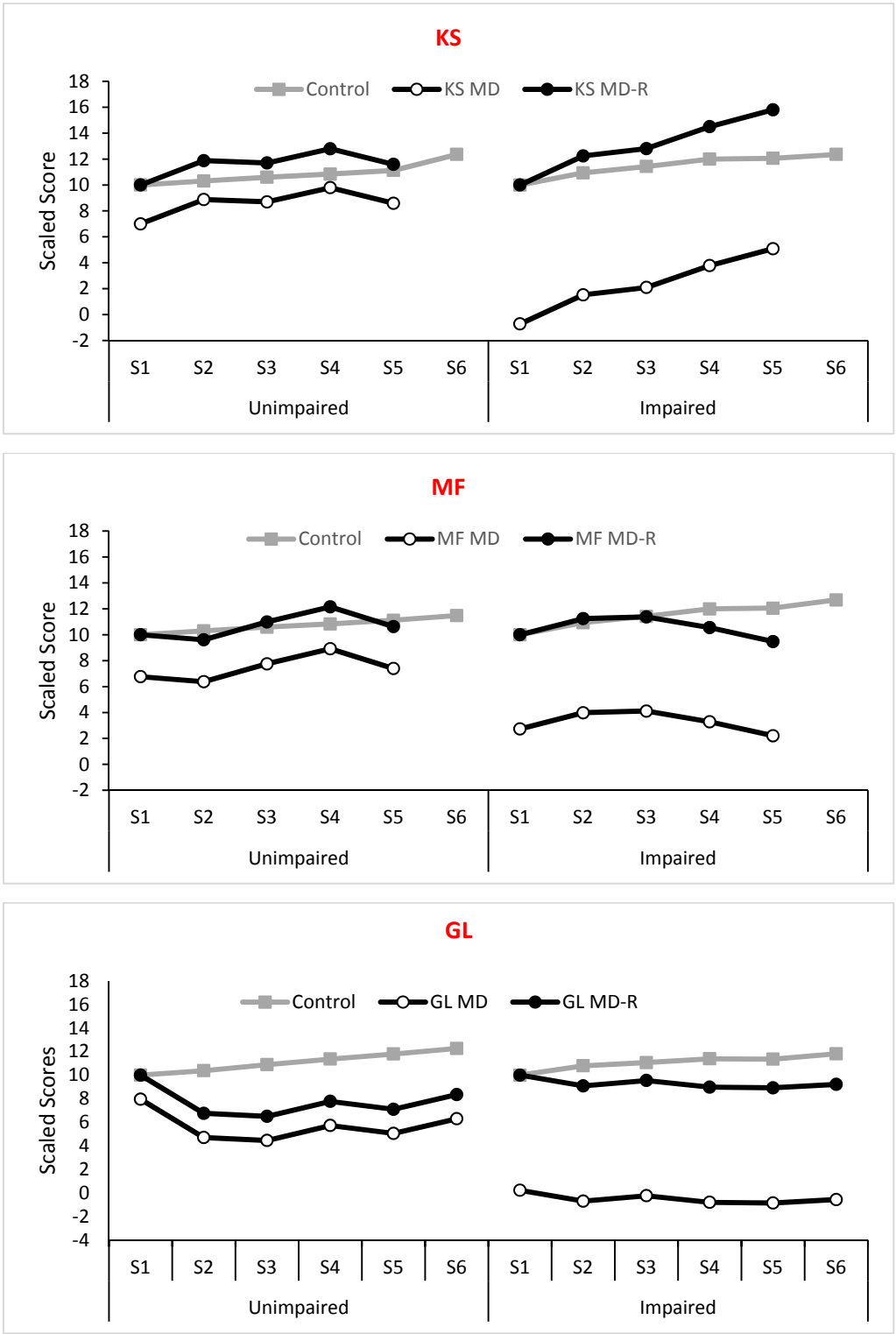

Figure 1. Changes in composite scores for initially impaired and unimpaired performance over test sessions for KS, MF, and GL.