



Published in final edited form as:

Psychol Assess. 2022 August ; 34(8): 731–741. doi:10.1037/pas0001138.

Logistic versus linear regression-based Reliable Change Index: a simulation study with implications for clinical studies with different sample sizes

Rafael De Andrade Moral¹, Unai Díaz-Orueta², Javier Oltra-Cucarella³,
Alzheimer's Disease Neuroimaging Initiative*

¹Department of Mathematics and Statistics, Maynooth University, Ireland

²Department of Psychology, Maynooth University, Ireland

³Department of Health Psychology, Miguel Hernandez University

Abstract

The linear regression-based Reliable Change Index (RCI) is widely used to identify memory impairments through longitudinal assessment. However, the minimum sample size required for estimates to be reliable has never been specified. Using data from 920 participants from the Alzheimer's Disease Neuroimaging Initiative data as true parameters, we run 12000 simulations for samples of size 10 to 1000 and analyzed the percentage of times the estimates are significant, their coverage rate, and the accuracy of the models including both the True Positive Rate and the True Negative Rate. We compared the linear RCI with a logistic RCI for discrete, bounded scores. We found that the logistic RCI is more accurate than the linear RCI overall, with the linear RCI approximating the logistic RCI for samples of size 200 or greater. We provide an R package to compute the Logistic RCI, which can be downloaded from the Comprehensive R Archive Network (CRAN) at <https://cran.r-project.org/web/packages/LogisticRCI/>, and the code to reproduce all results in this paper at <https://github.com/rafamoral/LogisticRCIpaper/>.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding autor: Dr. Javier Oltra-Cucarella, Department of Health Psychology, Miguel Hernandez University, Avda de la Universidad s/n, Edificio Altamira, 03202 Elche (Alicante, Spain), joltra@umh.es.

Author contribution

Conceptualization: RM, UD-O, JO-C

Data curation: RM, UD-O, JO-C

Formal analyses: RM, UD-O, JO-C

Methodology: RM

Software: RM

Visualization: RM, UD-O, JO-C

Writing – original draft: RM, UD-O, JO-C

Writing – review & editing: RM, UD-O, JO-C

Declarations of interests: none

Data used for the simulations can be accessed at www.adni.info.org, and cannot be shared by researchers as per the agreement with the ADNI project. This work was not preregistered, but a preprint version of this manuscript can be found at <https://psyarxiv.com/gq7az/>

Keywords

assessment; dementia; mild cognitive impairment; reliable change index; statistical models

Standard verbal memory tests are essential in the neuropsychological assessment of memory functioning, and are necessary to capture memory impairments in people with mild cognitive impairment (MCI) due to Alzheimer's disease (AD) who are at a greater risk of developing AD (Albert et al., 2011; Winblad et al., 2004). In the standard assessment of memory functioning, individuals are administered a verbal memory task and raw scores are compared against normative data obtained from a reference group (Strauss et al., 2006). However, performance on verbal memory tests can be interpreted using statistical techniques derived for serial assessment. These techniques, encompassed under the term Reliable Change Index (Duff, 2012), were developed to identify change in longitudinal assessment that exceeds normal variability (Hinton-Bayre, 2011; McSweeney et al., 1993). Among several RCI methods that are available for the interpretation of significant change (Calamia et al., 2012; Duff, 2012), one of the most statistically developed technique is the standard regression-based RCI (RCI_{RB}). With the RCI_{RB} , a comparison group is used to predict scores on the second assessment using scores on the first assessment (McSweeney et al., 1993). In order to interpret whether reliable change has occurred, observed scores at the second assessment are subtracted from the expected scores based on the regression equation, and the discrepancy is standardized using the standard error of the regression equation (SEE).

The RCI_{RB} has been applied to analyze memory impairments in different samples, including high school athletes (Brett et al., 2016), patients with epilepsy (Busch et al., 2015), traumatic brain injury (Metcalf et al., 2019), migraine (Roebuckspencer et al., 2007), cancer (Ouimet et al., 2009), HIV (Cysique et al., 2011), cardiac surgery (Sweet et al., 2008), schizophrenia (Roseberry & Kristian Hill, 2014) or psychosis (Sánchez-Torres et al., 2018), dementia (O'Connell et al., 2019) or MCI (Campos-Magdaleno et al., 2017; Duff et al., 2017), older adults with total joint replacement (Scott et al., 2017), as well as healthy individuals of different ages (Bouman et al., 2015; Crockford et al., 2018; Elbin et al., 2019; Frerichs & Tuokko, 2006; Gavett et al., 2015; Gonçalves et al., 2016; Raymond et al., 2006b; Salinsky et al., 2001; Schatz & Ferris, 2013; Temkin et al., 1999; Van Der Elst et al., 2008), and non-demented older adults (Duff, 2014; Sánchez-Benavides et al., 2016).

As with any other statistical technique, assumptions about linear regression analyses must be met for the regression equation to be accurate. Assumptions of regression analysis are that the residuals (differences between obtained and predicted scores on the dependent variable -DV) are normally distributed around the predicted DV scores, are independent for each value of the predictor (Tabachnick & Fidell, 2013), and that the variance of the residuals is the same for all values of the predictors, especially for small sample sizes (Williams et al., 2013). Following the Gauss-Markov theorem, even when residuals are not normally distributed, ordinary least squared parametric estimates are the best linear unbiased estimates (Williams et al., 2013). The violation of the assumption about normality of residuals affects significance tests and confidence interval of regression coefficients, even if they are still

unbiased (Williams et al., 2013). This means that regression coefficients close to the real parameter might go undetected (increasing the false negative rate) if confidence intervals are too large (i.e., include the 0), or that regression coefficients that deviate from the real parameter might reach statistical significance (increasing the false positive rate), especially in small samples (Williams et al., 2013).

Heterogeneity in sample sizes is quite large in studies using the RCI_{RB} . Although most of the studies reported sample sizes equal or lower than 200 (Bouman et al., 2015; Busch et al., 2015; Campos-Magdaleno et al., 2017; Crockford et al., 2018; Cysique et al., 2011; Duff, 2014; Duff et al., 2010; Elbin et al., 2019; Frerichs & Tuokko, 2006; Gonçalves et al., 2016; Hermann et al., 1996; Kashyap et al., 2014; Martin et al., 2002, 2006; Meekes et al., 2013, 2014; Raymond et al., 2006a, 2006b; Salinsky et al., 2001; Sánchez-Benavides et al., 2016; Sánchez-Torres et al., 2018; Scott et al., 2017; Sweet et al., 2008; Temkin et al., 1999; Womble et al., 2016), some reported samples larger than 500 (Brett et al., 2016; Gavett et al., 2015; Tombaugh, 2005; Van Der Elst et al., 2008) or lower than 30 (Metcalf et al., 2019; Nakhutina et al., 2010; Ouimet et al., 2009; Roebuckspencer et al., 2007; Roseberry & Kristian Hill, 2014; Schatz & Ferris, 2013; Sherman et al., 2003, 2003). However, residuals have barely been tested or plotted, and thus the probability of using unreliable estimates is unknown.

Previous studies have analyzed, through Montecarlo simulations, the Type I error rate (i.e., wrongly concluding that the patient has a deficit) associated with several reliable change methods for different sample sizes. Crawford and Garthwaite (2012) showed that the Type I error rate when using z-scores obtained from means and standard deviations doubled the nominal rate for small samples, and approximated the expected 5% (one-tailed) for samples of size 50 and above. When reliable change was calculated with the regression-based RCI, Crawford and Garthwaite (2006) showed that the Type I error rate varied with different test-retest correlation and different sample sizes, and showed that using the standard error for a new case maintained the error rate close to 5%. However, neither the Type II error rate nor the influence of other covariates in the regression model were analyzed.

Additionally, one of the methodological topics that to our knowledge has never been analyzed is the use of statistical methods according to the nature of the data. The RCI_{RB} is calculated using a linear regression model, which is intended to be used with continuous data. However, if the RCI_{RB} is to be used to identify reliable memory decline, then scores obtained on memory tests will be used. The scores obtained on memory tests are discrete rather than continuous (e.g., it is not possible to recall 1.5 items), and are bounded between lower (typically 0) and upper (maximum number of items) possible values. For example, the Rey's Auditory Verbal Learning Test (Rey, 1964) includes 15 words, and thus performance is bounded between 0 and 15. The linear regression model assumes that the response is continuous and unbounded, therefore not being the most suitable approach for this type of analysis. Binomial generalized linear models (McCullagh & Nelder, 1989), however, do accommodate the discrete and bounded nature of a response variable, and therefore represent a more suitable alternative to linear regression.

The aim of the present work was to analyze the sample size needed to increase the number of true positives and to reduce the number of false negatives to a minimum, in order to identify correctly individuals with objective longitudinal memory decline using the RCI_{RB} . As a further step, we provide the `LogisticRCI R` package with an alternative method to model cognitive scores when analyzing reliable change with discrete, bounded scores from memory tests.

Methods

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The first ADNI period (ADNI1) was updated in the ADNIGO and ADNI2 grant periods. Information about magnetic resonance imaging, positron emission tomography, other biological markers and clinical and neuropsychological assessment are available for more than 1,000 normal controls, individuals with MCI, and individuals with mild dementia (Petersen et al., 2010) (www.adni-info.org). The ethical committee at each participating site approved the project. All ADNI participants provided written consent. All participants received physical and neurological examinations, screening laboratory tests, and provided blood samples for DNA and APOE testing. We used data from 920 participants across 1840 visits. Four hundred and fifty participants were labeled as cognitively normal and 470 as having MCI at baseline. Cognitively normal participants had no memory complaints, Clinical Dementia Rating (Hughes et al., 1982) scale = 0, Mini-Mental State Examination (Folstein et al., 1975) scores ≥ 24 , normal education-corrected Logical Memory subtest scores, and no significant impairments in activities of daily living. Participants with MCI met Petersen (1999) criteria: they had memory complaints, Clinical Dementia Rating (Hughes et al., 1982) scale = 0.5, Mental State Examination (Folstein et al., 1975) scores ≥ 24 , education-corrected Logical Memory subtest scores equal or lower than 1.5 standard deviations below the mean of a normative sample, no significant impairments in activities of daily living, and did not meet criteria for dementia.

Differences between groups on demographics and cognitive variables were analyzed with χ^2 and independent t-tests. Cohen's d was calculated as a measure of effect size for continuous variables, with values of .20, .50 and .80 indicating a small, medium and large effect size respectively (Cohen, 1992). Test-retest reliability was calculated with the Pearson correlation coefficient, with values of .10, .30 and .50 indicating small, medium and large associations respectively (Cohen, 1992).

Linear Regression-Based Reliable Change Index (Linear RCI)

Let Y_i be the random variable representing the score obtained by individual i , $i = 1, \dots, n$. We begin by assuming the distribution of Y_i is normal with mean μ_i and variance σ^2 , with $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where \mathbf{x}_i^T is the i -th row of the design matrix and $\boldsymbol{\beta}$ is the vector of regression coefficients. This is a standard multiple linear regression model, with $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ and $\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - p}$ the well-known least squares estimators for $\boldsymbol{\beta}$ and σ^2 , respectively,

where p is the dimension of the β vector. Crawford and Garthwaite (2007) refer to $\hat{\sigma}$ as the SEE. The Linear Regression-Based Reliable Change Index (Linear RCI) is given by

$$\text{Linear RCI} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}$$

where $\hat{\mu}_i = \mathbf{x}_i^T \hat{\beta}$ is the predicted mean score for each individual (Crawford & Garthwaite, 2007). Assuming the model is well fitted, the residuals are meant to follow a normal distribution. Since the Linear RCI is a standardised version of the raw residuals, for a well fitted model it is assumed to follow a standard normal distribution. Therefore, values that fall in the lower tail of the distribution are assumed to represent reliable decline. In this work, we used the 5th percentile (one-tailed) of the standard normal distribution (i.e., z-score -1.64) as a threshold to detect reliable decline (Crawford & Garthwaite, 2007). The effects of the terms in the model were assessed via F tests.

Logistic Regression-Based Reliable Change Index (Logistic RCI)

The response variable in this study, the Auditory Verbal learning Test Delayed Recall score (AVLT-DR) (Rey, 1964), is a discrete score that is bounded between zero and 15 that reflects the number of items correctly recalled following a 20 minute delay. Therefore, its nature is of a discrete proportion, and a sensible modelling approach would involve binomial generalized linear models (GLMs) and extensions. Here we propose a new RCI based on logistic regression: a binomial GLM with a logit link. Let Y_i be the random variable representing the number of items correctly recalled by individual i . We may assume that the distribution of Y_i is Binomial(m_i, π_i), where $m_i = 15$ is the denominator of the distribution and π_i is the probability of an item being recalled for individual i . We model π_i as a function of different predictors, in the logit (or log-odds) scale, i.e.

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \mathbf{x}_i^T \beta$$

where \mathbf{x}_i^T is the i -th row of the design matrix and β is the vector of regression coefficients. Typically, the design matrix includes an intercept and the effects of baseline score, and may also include other covariates such as age, gender and education level. By fitting a logistic regression model, we are able to estimate the regression coefficients and the linear predictor $\hat{\eta}_i = \mathbf{x}_i^T \hat{\beta}$. Consequently, the probability of a question being correct would be represented as $\hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}}$.

The Logistic Regression-Based Reliable Change Index (Logistic RCI) is based on the Anscombe residuals for generalized linear models, combined with the correction proposed by Cox and Snell (1968) to stabilise the asymptotic variance of their distribution. We may write it as

$$\text{Logistic RCI} = \frac{\delta\left(\frac{y_i}{m_i}\right) - \delta\left(\hat{\pi}_i - \frac{1 - 2\hat{\pi}_i}{6m_i}\right)}{\{\hat{\pi}_i(1 - \hat{\pi}_i)\}^{-\frac{1}{6}}}$$

, where $\delta(x) = \int_0^x \{t(1-t)\}^{-\frac{1}{3}} dt$ is the incomplete beta function. Asymptotically, the Logistic RCI has a normal distribution with mean zero and variance that depends on $\hat{\pi}_i$. We may standardise it to obtain z-scores by dividing the logistic RCI by its observed standard deviation, therefore yielding an asymptotic $N(0, 1)$ distribution. Again, we used the 5th percentile of the standard normal distribution as a threshold to detect reliable decline (i.e. z-score -1.64).

Since both linear and logistic regression are generalized linear models, the assumption of linearity at the linear predictor scale applies to both methods. Because for linear regression we typically use the canonical identity link, this translates directly into linearity between the independent and dependent variables. For logistic regression, we now assume linearity in the logit scale. We fit the binomial GLMs allowing $\text{Var}(Y_i) = \phi m_i \pi_i (1 - \pi_i)$, and estimated ϕ via a quasi-likelihood approach using the Pearson residuals, thus allowing for a more flexible variance function that accommodates extra variability, should that be present in the data. Finally, we assessed the significance of the model effects using F tests, since the dispersion parameter ϕ had been estimated.

Simulation Study

We began by fitting a linear and a logistic regression model to the ADNI dataset including baseline score, age, education and gender in the linear predictor, and treated the estimated parameters as the true population parameters. Then, we simulated a population of 1 million individuals based on these estimated parameters and predefined distributions for baseline score, age, education and gender that are similar to the ones observed in the ADNI dataset. After that, we drew 1,000 samples of sizes 10, 20, 30, 40, 50, 100, 150, 200, 250, 500, 750 and 1,000 from the simulated dataset (a total of 12,000 simulated samples), re-fitted the linear and logistic regression models to each sample, and calculated the linear and the logistic RCI. We then identified individuals with z-score -1.64 as showing reliable decline. Individuals were labelled as true positives if they showed a discrepancy between observed and predicted scores equal or lower than -1.64 both in the simulated dataset and for each sample of different sizes. Individuals showing reliable decline in the simulated dataset but not for the smaller sample sizes were labelled as false negatives.

To assess model overall goodness-of-fit, we produced half-normal plots with a simulated envelope for the studentised residuals for the linear RCI and deviance residuals for the logistic RCI. This was obtained by plotting the ordered absolute values of the residuals versus the expected order statistics of a half-normal distribution (Moral et al., 2017). By simulating data from the fitted models, refitting the models and obtaining the ordered absolute values of the residuals, we could obtain an envelope by computing the 97.5th and 2.5th percentile for each order statistics. The envelope was such that for a well-fitting model,

we would expect most points to lie within it. We assessed the normality of the linear RCI and the logistic RCI by producing the half-normal plot with a simulated envelope and counting the number of points that fell outside of the envelope (the more points that fall outside, the bigger the departure from normality).

Lastly, we assessed (i) the significance of the effects in the linear predictor, (ii) the normality of the regression-based RCI (based on the half-normal plot with a simulation envelope) (Moral et al., 2017), (iii) the percentage of points outside the envelope of the half-normal plot of the residuals (as a measure of overall goodness-of-fit), (iv) the percentage coverage of the 90%, 95% and 99% confidence intervals for each true parameter value, and (v) the accuracy (true positive and true negative rates) for detecting reliable decline at different thresholds.

For the logistic RCI, the baseline scores were simulated from a Beta(1.59,1.36) distribution. This distribution was obtained by fitting a beta model to 15 minus the baseline scores observed in the ADNI dataset, and scaling them to be between 0 and 1. The age variable was simulated from a N(73.41,46.78) distribution, which was obtained by fitting a normal model to the observed ages in the ADNI dataset. The gender variable was simulated from a Bernoulli(0.5) distribution, so that approximately half of the individuals were male and half female. Finally, the education level variable was simulated from a discrete uniform distribution, ranging from 4 to 20, the range of the education variable in the ADNI dataset. Although the education variable in the ADNI dataset is skewed towards higher levels, we opted to represent all education levels equally in the large simulated dataset. All simulations and visualisations were produced using R (R Core Team, 2021), and all associated code is made available at <https://github.com/rafamoral/LogisticRCIpaper>. Data used for the simulations can be accessed at www.adni-info.org. This work was not preregistered, but a preprint version of this manuscript can be found at <https://psyarxiv.com/gq7az>.

Results

Descriptive statistics of the sample used for the simulation can be found in Table 1. Compared to the MCI group, the NC group had a higher percentage of males (58.7% vs. 48.9%, $p = .003$), was slightly older ($p < .001$), had a higher level of education ($p = .018$), and had higher MMSE ($p < .001$), baseline AVLT-DR ($p < .001$) and follow-up AVLT-DR ($p < .001$) scores. Effect sizes were negligible for education, small for age, and medium to large for MMSE and AVLT scores. According to the guidelines reported by Strauss, Sherman and Spreen (2006), the test-retest coefficient for the AVLT-DR scores was adequate ($r = .71$, 95% CI = .68, .74).

Simulation study

The estimated parameters treated as the true population parameters for simulations are shown in Table 2. Looking at the percentage of times the F test was significant for each effect (see Figure 1), linear and logistic models yielded very similar results. For the baseline score, even with a sample size as small as 20, we already observed significance (associated p-value less than .05) for 100% of the samples. The effect of baseline score, which was the slope of the curve, was the largest in magnitude (see Table 2), and therefore even with a

small sample size it was not difficult to obtain a significant estimate. Education, however, was of a smaller magnitude and significance was attained for 100% of samples of size 150 or larger. When looking at the effects of age and gender, it seems that it was very difficult for the method to obtain significant estimates, even with a sample as large as 1000, especially for the effect of gender.

When studying the coverage of the 90%, 95% and 99% confidence intervals (Figure S1), again we observed very similar results between the linear and logistic models. The coverage for the intercept, baseline and age effects was very close to the nominal coverage rate for samples as small as 30. For the gender effect, coverage was systematically above the nominal rate. This is because the gender effect is typically associated with a large standard error, and therefore the confidence intervals are inflated. The education level effect, on the other hand, presented coverage systematically below the nominal rate, although very close to it. This is because not only did education have a small numerical effect, but also the continuous education covariate had to be discretised, which makes it more difficult to estimate its effect and the uncertainty around the estimate. Consequently, the confidence intervals were slightly narrower than what they should have been to provide the nominal coverage rate. The distribution of the linear and logistic regression-based RCI was considered to be normal for most simulated datasets, at very similar rates, based on the half-normal plot with a simulated envelope (Figure S2). Model goodness-of-fit, however, was systematically better for the logistic model when compared to the linear model, although as discussed above, inferential power seemed to be very similar for both modelling approaches. As can be seen, there was a higher percentage of points outside of the envelope (>20%) for sample sizes above 25. For instance, we observed 47% of points outside of the envelope for the linear regression and 18% for the logistic regression at the largest sample size of 1,000. This means that the distribution of the residuals indicated that the models did not fit the data well, according to the half-normal plot with a simulated envelope. This was expected because as the sample size increases, the envelope bands become narrower when close to zero and depending on the simulated sample, many points will be outside the envelope bands in that region, increasing the overall average. However, most samples presented a satisfactory fit for the logistic regression model (a median of 9% for the logistic compared to 49% for the linear at a sample of size 1,000, with much lower values for smaller sample sizes). This indicates that the logistic regression is a suitable alternative to analyze this type of data.

When attempting to identify reliable change, the true positive rates (TPR) for the logistic RCI were systematically greater than the TPR for the linear RCI for smaller sample sizes (200 or less, see Fig. 2, top-left panel). The true negative rates (TNR), however, were very similar for both approaches, and close to 100%, although the TNR for the linear RCI was slightly lower. It became clear that the overall accuracy (bottom panels of Fig. 2) was dictated by the TPR in this case, and the logistic RCI presented better performance overall regardless of the threshold chosen. We would like to highlight that the normality of the RCI is based on an approximation, both for the linear and logistic RCI. As the sample size increases, this approximation becomes more evident and the departure from normality is clearer. However, this is not a problem in terms of true/false positive rates for detecting

reliable decline, especially if different thresholds are used depending on the objectives of the study.

In order to analyze whether the results might be biased because of including both cognitively normal and participants with MCI in the sample used to obtain the estimates for the simulations, we re-ran the analyses including diagnosis as a covariate both for the linear and logistic regression models and found that: (i) diagnosis was a significant covariate for both models; (ii) when the linear predictor includes diagnosis, fewer individuals with MCI presented $RCI < -1.64$ when compared to the model where diagnosis was not included as a covariate; and (iii) when diagnosis was not included as a covariate the mean RCI for individuals with MCI was smaller than the mean RCI for individuals in the control group. This was all expected, since when we include diagnosis as a covariate, we are effectively fitting lines with different intercepts (one for control, another for MCI). Since the RCI depends on the residuals, because we correct for diagnosis the RCI for both groups will be smaller in magnitude (closer to zero), and therefore fewer individuals will be identified as showing reliable decline.

Discussion

The present work aimed to analyze the sample size needed to obtain reliable estimates when assessing memory decline with the standard linear RCI. Additionally, we analyzed through simulation whether the identification of reliable decline was as accurate (true positives and true negatives) when using a linear model as when using a logistic model. We used the AVLT delayed recall scores, because delayed recall scores are believed to reflect the memory consolidation processes that occur within the hippocampus to a larger extent than immediate recall (Belleville et al., 2017), has good to excellent psychometric properties to discriminate patients with AD and healthy controls (Cerami et al., 2017), and has good predictive power to identify cognitive decline in healthy older adults (Wearn et al., 2020), and to identify individuals at a higher risk of progression from MCI to AD (Cerami et al., 2017; Fleisher et al., 2007). The simulation showed that both models give similar results for samples sizes of 200 or greater, with the logistic RCI presenting better performance with smaller sample sizes.

The main implications raise when evaluating the accuracy in terms of sensitivity (detection of true positives) and specificity (detection of true negatives). For samples smaller than 200, which are quite frequent in the literature of clinical research in older adults and dementia, the TPR for the logistic RCI are systematically greater than the TPR for the linear RCI, whereas the TNR for the linear RCI are only slightly lower. In other words, this implies that traditionally estimated linear RCI has higher rates of both false positives (individuals incorrectly identified as having reliable decline) and false negatives (individuals incorrectly identified as not having reliable decline). For the subject of false positives, Klekociuk et al. (2014) stated that the rate of recovery observed in MCI indicates that existing MCI diagnostic criteria comprise an unacceptably high rate of false positive diagnoses and lack adequate sensitivity and specificity. Edmonds et al. (2015) reported similar concerns as their results showed that a significant proportion of individuals in the ADNI/MCI sample are cognitively normal if more detailed testing is taken into account, and that a subset of

individuals from the cluster-derived normal group could be at risk of developing MCI (thus uncovering the relevance of potential false negative cases).

It has been highlighted in the literature that the potential impact of false negatives has remained largely ignored (Vadillo et al., 2016), and the over-interpretation of null results is even more dangerous than the prevalence of false positives in some areas of research, since null results (1) are inherently ambiguous, (2) they are silent about the amount of support for the null hypothesis, and (3) they are surprisingly easy to obtain by mere statistical artefacts (e.g., using a small sample or a noisy measure can suffice to produce a false negative). As Edmonds et al. (2016) state, the impact of “missed” cases of MCI because of being wrongly discarded as healthy has a direct impact in clinical practice, but also in research studies and clinical trials targeting prodromal AD.

Knowing the impact that high FNR have on clinical practice, the linear RCI seems to be underpowered for small sample sizes. Using a logistic RCI that decreased the FNR would be extremely helpful in the sense that less individuals would miss opportunities for intervention (in the form of cognitive rehabilitation) and they would subsequently engage in potentially beneficial treatments from early stages, with clinicians being more confident in the type of recommendations provided to them and their families. In addition, early interventions targeting risk factors by encouraging both physical and cognitive activities would reach a higher percentage of population. Where necessary, compensatory strategies in the form of external aids would be applied at earlier stages, as well as referrals to other professionals for further assessment (Edmonds et al., 2016). This positive scenario of higher accuracy in the detection of reliable decline can be achieved if traditional reliance on linear models for RCI estimation is overcome. Our results suggest that for typical sample sizes used in many clinical studies (an average $n < 200$ individuals), a logistic model can more accurately both identify actual clinical cases and discard healthy individuals.

Additionally, our results suggest an alternative way of improving the accuracy of the RCI. Researchers and clinicians may be less conservative when setting up the threshold to identify reliable change. This yields improved TPR at the expense of a smaller TNR. As shown in the middle panels in figure 2, when changing the threshold from $z = -1.64$ to a more liberal one of $z = -1.28$ the TPR is above 95% for the logistic RCI even for samples as small as 10, rising to almost 100% for samples of size 100 and larger, when compared to a TPR of around 85% when using the more conservative threshold. This improvement is obtained at the expense of lowering the TNR and overall accuracy from close to 100% to around 98%. With regards to the linear RCI, lowering the threshold to a more liberal $z = -1.28$ rises the TPR to values higher than 90% for samples of size 50 while maintaining the TNR above 95%. For small sample sizes, the linear model seems to be unreliable for either threshold.

However, caution is needed when modifying the threshold used to identify reliable decline. Lowering the threshold to -1.28 will allow classifying more observations as showing reliable decline, which will increase the number of true positives, but also the number of false positives. If we select even more liberal criteria, say $z = -1.04$ (right-hand panels on Fig. 2) we see that the improvement in TPR is not that different from when using $z = -1.28$, however the TNR and overall accuracy now fall to around 95%. If it is more

important to identify those individuals who present reliable decline, at the expense of obtaining a few more false positives, then we recommend relaxing the lower bound to a value greater than -1.64 . We are not, however, advocating for a hard threshold of -1.28 . We are simply pointing out that if we assume that in the population the 5th lower percentile are representative of reliable decline, when analysing smaller samples it could be a good idea to look at a higher percentile of the samples to identify reliable change (e.g., the 10th percentile). This would increase the TPR at a small expense of lowering the TNR only a little. This is in line with the common use of normative data to interpret performance on neuropsychological tests (e.g., in MCI research), where the 7th percentile ($z = -1.5$) is used to identify low scores. However, it is important that the cut-off point be defined prior to testing, rather than examining multiple thresholds after the test results are known.

The implications of our results is that, based on the higher FNR associated with the linear RCI, several studies with small sample sizes seem unreliable to identify reliable decline. There is no way to know whether the estimates reported in previous studies are false, as it is not possible to gather the real parameters in the population, but the results from our simulation suggesting that the probability of having left unidentified a large proportion of individuals with cognitive impairment is high raises concerns about their conclusions. Replication studies with sample of sizes larger than 200 are needed.

This study comprises limitations as it involves a modest approach to detection of changes based on discrete scores of an episodic memory test. Our results are only applicable to the Rey AVLT. It is noteworthy to mention that the AVLT has a high resolution because it subdivides the 0–1 interval in 16 different values. This is helpful in terms of detecting reliable decline, and also in obtaining a reasonably good approximation using the linear regression model. Other studies may use shorter list-learning tasks (e.g. CERAD, HVLT), which could make detecting reliable decline more difficult, especially when approximating the behavior of the discrete response variable with a linear regression model.

This approach uses only one test to detect reliable decline. Klekociuk et al. (2014) highlighted the importance of using comprehensive test batteries to enhance sensitivity and specificity in MCI classification by including both memory and non-memory assessments. In addition, Blanco-Campal et al. (2019) suggested that it would be interesting to identify cases with a raw score below or above the standard cut-point but whose qualitative performance may point in the opposite direction (e.g., score above the cut-point with indications of decline of clinical relevance). In any case, our study has shown that the application of a logistic RCI can increase the accuracy of identifying reliable decline and improve true positive and true negative rates, and has provided a pathway to identify the relevance of incorporating moderate to high sample sizes in future clinical studies. It is reasonable to assume that this same model will show more accurate results when data from multiple sources, both memory and non-memory tests, are taken into consideration.

Lastly, our results are based on the analysis of test-retest in two assessment points. However, neuropsychological assessment is typically performed multiple times in order to assess longitudinal change. Our results prompt future research in which we explore how the logistic RCI might outperform its linear counterpart in studies that assess individuals at multiple

time points. These would require the accommodation of serially correlated measures, e.g. through a random-effects approach as a binomial generalized linear mixed model. The RCI formulae would have to be adapted to incorporate the extra correlation parameters. Additional research is also needed to determine whether logistic models using additional scores from multiple tests and diverse clinical samples can improve the estimation of an accurate RCI even further.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. ADNI had no involvement in study design; in the analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

References

- Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, & Phelps CH (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 270–279. 10.1016/j.jalz.2011.03.008
- Belleville S, Fouquet C, Hudon C, Zomahoun HTV, Croteau J, & disease-Quebec, C. for the E. I. of A. (2017). Neuropsychological Measures that Predict Progression from Mild Cognitive Impairment to Alzheimer's type dementia in Older Adults: A Systematic Review and Meta-Analysis. *Neuropsychology Review*, 27(4), 328–353. 10.1007/s11065-017-9361-5 [PubMed: 29019061]
- Blanco-Campal A, Diaz-Orueta U, Navarro-Prados AB, Burke T, Libon DJ, & Lamar M (2019). Features and psychometric properties of the Montreal Cognitive Assessment: Review and proposal of a process-based approach version (MoCA-PA). *Applied Neuropsychology: Adult*, 1–15. 10.1080/23279095.2019.1681996
- Bouman Z, Hendriks MPH, Aldenkamp AP, & Kessels RPC (2015). Temporal Stability of the Dutch Version of the Wechsler Memory Scale—Fourth Edition (WMS-IV-NL). *The Clinical Neuropsychologist*, 29(sup1), 30–46. 10.1080/13854046.2015.1137354 [PubMed: 26911865]
- Brett BL, Smyk N, Solomon G, Baughman BC, & Schatz P (2016). Long-term Stability and Reliability of Baseline Cognitive Assessments in High School Athletes Using ImPACT at 1-, 2-, and 3-year Test–Retest Intervals. *Archives of Clinical Neuropsychology*, acn;acw055v1. 10.1093/arclin/acw055
- Busch RM, Lineweaver TT, Ferguson L, & Haut JS (2015). Reliable change indices and standardized regression-based change score norms for evaluating neuropsychological change in children with epilepsy. *Epilepsy & Behavior*, 47, 45–54. 10.1016/j.yebeh.2015.04.052 [PubMed: 26043163]

- Calamia M, Markon K, & Tranel D (2012). Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. *The Clinical Neuropsychologist*, 26(4), 543–570. 10.1080/13854046.2012.680913 [PubMed: 22540222]
- Campos-Magdalenó M, Facal D, Lojo-Seoane C, Pereiro AX, & Juncos-Rabadán O (2017). Longitudinal Assessment of Verbal Learning and Memory in Amnesic Mild Cognitive Impairment: Practice Effects and Meaningful Changes. *Frontiers in Psychology*, 8. 10.3389/fpsyg.2017.01231
- Cerami C, Dubois B, Boccardi M, Monsch AU, Demonet JF, & Cappa SF (2017). Clinical validity of delayed recall tests as a gateway biomarker for Alzheimer's disease in the context of a structured 5-phase development framework. *Neurobiology of Aging*, 52, 153–166. 10.1016/j.neurobiolaging.2016.03.034 [PubMed: 28317646]
- Cohen J (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. 10.1037/0033-2909.112.1.155 [PubMed: 19565683]
- Cox DR, & Snell EJ (1968). A General Definition of Residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2), 248–275.
- Crawford JR, & Garthwaite PH (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, 20(3), 259–271. 10.1037/0894-4105.20.3.259 [PubMed: 16719619]
- Crawford JR, & Garthwaite PH (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, 21(5), 611–620. 10.1037/0894-4105.21.5.611 [PubMed: 17784809]
- Crawford JR, & Garthwaite PH (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex*, 48(8), 1009–1016. 10.1016/j.cortex.2011.06.021 [PubMed: 21843884]
- Crockford C, Newton J, Lonergan K, Madden C, Mays I, O'Sullivan M, Costello E, Pinto-Grau M, Vajda A, Heverin M, Pender N, Al-Chalabi A, Hardiman O, & Abrahams S (2018). Measuring reliable change in cognition using the Edinburgh Cognitive and Behavioural ALS Screen (ECAS). *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 19(1–2), 65–73. 10.1080/21678421.2017.1407794 [PubMed: 29214872]
- Cysique LA, Franklin D, Abramson I, Ellis RJ, Letendre S, Collier A, Clifford D, Gelman B, McArthur J, Morgello S, Simpson D, McCutchan JA, Grant I, Heaton RK, the CHARTER group, & the HNRC group. (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology*, 33(5), 505–522. 10.1080/13803395.2010.535504 [PubMed: 21391011]
- Duff K (2012). Evidence-Based Indicators of Neuropsychological Change in the Individual Patient: Relevant Concepts and Methods. *Archives of Clinical Neuropsychology*, 27(3), 248–261. 10.1093/arclin/acr120 [PubMed: 22382384]
- Duff K (2014). One-Week Practice Effects in Older Adults: Tools for Assessing Cognitive Change. *The Clinical Neuropsychologist*, 28(5), 714–725. 10.1080/13854046.2014.920923 [PubMed: 24882553]
- Duff K, Atkinson TJ, Suhrie KR, Dalley BCA, Schaefer SY, & Hammers DB (2017). Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. *Journal of Clinical and Experimental Neuropsychology*, 39(4), 396–407. 10.1080/13803395.2016.1230596 [PubMed: 27646966]
- Duff K, Beglinger LJ, Moser DJ, & Paulsen JS (2010). Predicting cognitive change within domains. *The Clinical Neuropsychologist*, 24(5), 779–792. 10.1080/13854041003627795 [PubMed: 20358479]
- Edmonds EC, Delano-Wood L, Clark LR, Jak AJ, Nation DA, McDonald CR, Libon DJ, Au R, Galasko D, Salmon DP, & Bondi MW (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia*, 11(4), 415–424. 10.1016/j.jalz.2014.03.005
- Edmonds EC, Delano-Wood L, Jak AJ, Galasko DR, Salmon DP, & Bondi MW (2016). "Missed" Mild Cognitive Impairment: High False-Negative Error Rate Based on Conventional Diagnostic Criteria. *Journal of Alzheimer's Disease*, 52(2), 685–691. 10.3233/JAD-150986

- Elbin RJ, Fazio-Sumrok V, Anderson MN, D'Amico NR, Said A, Grosse L, Schatz P, Lipinski D, & Womble M (2019). Evaluating the suitability of the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) computerized neurocognitive battery for short-term, serial assessment of neurocognitive functioning. *Journal of Clinical Neuroscience*, 62, 138–141. 10.1016/j.jocn.2018.11.041 [PubMed: 30472342]
- Fleisher AS, Sowell BB, Taylor C, Gamst AC, Petersen RC, Thal LJ, & for the Alzheimer's Disease Cooperative Study. (2007). Clinical predictors of progression to Alzheimer disease in amnesic mild cognitive impairment. *Neurology*, 68(19), 1588–1595. 10.1212/01.wnl.0000258542.58725.4c [PubMed: 17287448]
- Folstein MF, Folstein SE, & McHugh PR (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatry Research*, 12(3), 189–198. 10.1016/0022-3956(75)90026-6
- Frerichs RJ, & Tuokko HA (2006). Reliable change scores and their relation to perceived change in memory: Implications for the diagnosis of mild cognitive impairment. *Arch. Clin. Neuropsychol*, 21(1), 109–115. 10.1016/j.acn.2005.08.003 [PubMed: 16216465]
- Gavett BE, Ashendorf L, & Gurnani AS (2015). Reliable Change on Neuropsychological Tests in the Uniform Data Set. *Journal of the International Neuropsychological Society*, 21(7), 558–567. 10.1017/S1355617715000582 [PubMed: 26234918]
- Gonçalves MM, Pinho MS, & Simões MR (2016). Test-retest reliability analysis of the Cambridge Neuropsychological Automated Tests for the assessment of dementia in older people living in retirement homes. *Applied Neuropsychology: Adult*, 23(4), 251–263. 10.1080/23279095.2015.1053889 [PubMed: 26574661]
- Hermann BP, Seidenberg M, Schoenfeld J, Peterson J, Leveroni C, & Wyler AR (1996). Empirical Techniques for Determining the Reliability, Magnitude, and Pattern of Neuropsychological Change After Epilepsy Surgery. *Epilepsia*, 37(10), 942–950. 10.1111/j.1528-1157.1996.tb00531.x [PubMed: 8822692]
- Hinton-Bayre AD (2011). Specificity of Reliable Change Models and Review of the Within-subjects Standard Deviation as an Error Term. *Archives of Clinical Neuropsychology*, 26(1), 67–75. 10.1093/arclin/acq087 [PubMed: 21147862]
- Hughes CP, Berg L, Danziger WL, Coben LA, & Martin RL (1982). A new clinical scale for the staging of dementia. *British Journal of Psychiatry*, 140(6), 566–572. 10.1192/bjp.140.6.566
- Kashyap M, Belleville S, Mulsant BH, Hilmer SN, Paquette A, Tu LM, & Tannenbaum C (2014). Methodological Challenges in Determining Longitudinal Associations Between Anticholinergic Drug Use and Incident Cognitive Decline. *Journal of the American Geriatrics Society*, 62(2), 336–341. 10.1111/jgs.12632 [PubMed: 24417438]
- Klekociuk SZ, Summers JJ, Vickers JC, & Summers MJ (2014). Reducing false positive diagnoses in mild cognitive impairment: The importance of comprehensive neuropsychological assessment. *European Journal of Neurology*, 21(10), 1330–e83. 10.1111/ene.12488 [PubMed: 24943259]
- Martin R, Griffith HR, Sawrie S, Knowlton R, & Faught E (2006). Determining empirically based self-reported cognitive change: Development of reliable change indices and standardized regression-based change norms for the multiple abilities self-report questionnaire in an epilepsy sample. *Epilepsy & Behavior*, 8(1), 239–245. 10.1016/j.yebeh.2005.10.004 [PubMed: 16368275]
- Martin R, Sawrie S, Gilliam F, Mackey M, Faught E, Knowlton R, & Kuzniecky R (2002). Determining Reliable Cognitive Change after Epilepsy Surgery: Development of Reliable Change Indices and Standardized Regression-based Change Norms for the WMS-III and WAIS-III. *Epilepsia*, 43(12), 1551–1558. 10.1046/j.1528-1157.2002.23602.x [PubMed: 12460258]
- McCullagh P, & Nelder JA (1989). *Generalized Linear Models*. Springer US. 10.1007/978-1-4899-3242-6
- McSweeney AJ, Naugle RI, Chelune GJ, & Lüders H (1993). "TScores for Change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, 7(3), 300–312. 10.1080/13854049308401901
- Meekes J, Braams OB, Braun KPJ, Jennekens-Schinkel A, van Rijen PC, Alpherts WCJ, Hendriks MPH, & van Nieuwenhuizen O (2014). Visual memory after epilepsy surgery in children: A standardized regression-based analysis of group and individual outcomes. *Epilepsy & Behavior*, 36, 57–67. 10.1016/j.yebeh.2014.04.016 [PubMed: 24857810]

- Meekes J, Braams O, Braun KPJ, Jennekens-Schinkel A, & van Nieuwenhuizen O (2013). Verbal memory after epilepsy surgery in childhood. *Epilepsy Research*, 107(1–2), 146–155. 10.1016/j.epilepsyres.2013.08.017 [PubMed: 24042124]
- Metcalfe K, Sabaz M, Daher M, & Simpson G (2019). Measuring reliable change in traumatic brain injury (TBI): The pitfalls of using readily available formulae. *Applied Neuropsychology: Adult*, 1–10. 10.1080/23279095.2018.1559166
- Moral RA, Hinde J, & Demétrio CGB (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*, 81(1), 1–23. 10.18637/jss.v081.i10
- Nakhutina L, Pramataris P, Morrison C, Devinsky O, & Barr WB (2010). Reliable Change Indices and Regression-Based Measures for the Rey-Osterreith Complex Figure Test in Partial Epilepsy Patients. *The Clinical Neuropsychologist*, 24(1), 38–44. 10.1080/13854040902960091 [PubMed: 19606396]
- O’Connell ME, Gould B, Ursenbach J, Enright J, & Morgan DG (2019). Reliable change and minimum clinically important difference (MCID) of the Repeatable Battery for the Assessment of Neuropsychology Status (RBANS) in a heterogeneous dementia sample: Support for reliable change methods but not the MCID. *Applied Neuropsychology: Adult*, 26(3), 268–274. 10.1080/23279095.2017.1413575 [PubMed: 29319337]
- Quimet LA, Stewart A, Collins B, Schindler D, & Bielajew C (2009). Measuring neuropsychological change following breast cancer treatment: An analysis of statistical models. *Journal of Clinical and Experimental Neuropsychology*, 31(1), 73–89. 10.1080/13803390801992725 [PubMed: 18608651]
- Petersen R, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, & Weiner MW (2010). Alzheimer’s Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, 74(3), 201–209. 10.1212/WNL.0b013e3181cb3e25 [PubMed: 20042704]
- Petersen R, Smith G, Waring S, Ivnik R, Tangalos T, & Kokmen E (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, 56(3), 303–308. 10.1001/archneur.56.3.303 [PubMed: 10190820]
- Raymond PD, Hinton-Bayre AD, Radel M, Ray MJ, & Marsh NA (2006a). Assessment of statistical change criteria used to define significant change in neuropsychological test performance following cardiac surgery. *European Journal of Cardio-Thoracic Surgery*, 29(1), 82–88. 10.1016/j.ejcts.2005.10.016 [PubMed: 16337395]
- Raymond PD, Hinton-Bayre AD, Radel M, Ray MJ, & Marsh NA (2006b). Test–Retest Norms and Reliable Change Indices for the MicroCog Battery in a Healthy Community Population Over 50 Years of Age. *The Clinical Neuropsychologist*, 20(2), 261–270. 10.1080/13854040590947416 [PubMed: 16690546]
- Rey A (1964). *L’Examen Clinique en Psychologie*. Presses Universitaires de France.
- Roebuckspencer T, Sun W, Cernich A, Farmer K, & Bleiberg J (2007). Assessing change with the Automated Neuropsychological Assessment Metrics (ANAM): Issues and challenges. *Archives of Clinical Neuropsychology*, 22, 79–87. 10.1016/j.acn.2006.10.011
- Roseberry JE, & Kristian Hill S (2014). Limited practice effects and evaluation of expectation for change: MATRICS Consensus Cognitive Battery. *Schizophrenia Research*, 159(1), 188–192. 10.1016/j.schres.2014.08.004 [PubMed: 25182539]
- Salinsky MC, Storzbach D, Dodrill CB, & Binder LM (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16-week period. *Journal of the International Neuropsychological Society: JINS*, 7(5), 597–605. [PubMed: 11459111]
- Sánchez-Benavides G, Peña-Casanova J, Casals-Coll M, Gramunt N, Manero RM, Puig-Pijoan A, Aguilar M, Robles A, Antúnez C, Frank-García A, Fernández-Martínez M, Blesa R, NEURONORMA Study Team, Oliva R, Gómez-Ansón B, Monte G, Alayrach E, Sainz A, Caprile C, ... Ribas R (2016). One-Year Reference Norms of Cognitive Change in Spanish Old Adults: Data from the NEURONORMA Sample. *Archives of Clinical Neuropsychology*, 31(4), 378–388. 10.1093/arclin/acw018 [PubMed: 27193368]
- Sánchez-Torres AM, Moreno-Izco L, Lorente-Omeñaca R, Cabrera B, Lobo A, González-Pinto AM, Merchán-Naranjo J, Corripio I, Vieta E, de la Serna E, Butjosa A, Contreras F, Sarró S, Mezquida G, Ribeiro M, Bernardo M, & Cuesta MJ (2018). Individual trajectories of cognitive performance

- in first episode psychosis: A 2-year follow-up study. *European Archives of Psychiatry and Clinical Neuroscience*, 268(7), 699–711. 10.1007/s00406-017-0857-z [PubMed: 29164332]
- Schatz P, & Ferris CS (2013). One-Month Test-Retest Reliability of the ImPACT Test Battery. *Archives of Clinical Neuropsychology*, 28(5), 499–504. 10.1093/arclin/act034 [PubMed: 23702678]
- Scott JE, Mathias JL, Kneebone AC, & Krishnan J (2017). Postoperative cognitive dysfunction and its relationship to cognitive reserve in elderly total joint replacement patients. *Journal of Clinical and Experimental Neuropsychology*, 39(5), 459–472. 10.1080/13803395.2016.1233940 [PubMed: 27676314]
- Sherman EMS, Slick DJ, Connolly MB, Steinbok P, Martin R, Strauss E, Chelune GJ, & Farrell K (2003). Reexamining the effects of epilepsy surgery on IQ in children: Use of regression-based change scores. *Journal of the International Neuropsychological Society*, 9(6), 879–886. 10.1017/S1355617703960085 [PubMed: 14632247]
- Strauss EH, Sherman EHS, & Spreen O (2006). *A Compendium of Neuropsychological Tests. Administration, Norms and Comments*. Oxford University Press.
- Sweet JJ, Finnin E, Wolfe PL, Beaumont JL, Hahn E, Marymont J, Sanborn T, & Rosengart TK (2008). Absence of Cognitive Decline One Year After Coronary Bypass Surgery: Comparison to Nonsurgical and Healthy Controls. *The Annals of Thoracic Surgery*, 85(5), 1571–1578. 10.1016/j.athoracsur.2008.01.090 [PubMed: 18442540]
- Tabachnick BG, & Fidell LS (2013). *Using Multivariate Statistics*. (Sixth Ed.). Pearson Education Inc.
- Temkin NR, Heaton RK, Grant I, & Dikmen SS (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society: JINS*, 5(4), 357–369. [PubMed: 10349298]
- Tombaugh T (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Archives of Clinical Neuropsychology*, 20(4), 485–503. 10.1016/j.acn.2004.11.004 [PubMed: 15896562]
- Vadillo MA, Konstantinidis E, & Shanks DR (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, 23(1), 87–102. 10.3758/s13423-015-0892-6 [PubMed: 26122896]
- Van Der Elst W, Van Boxtel MPJ, Van Breukelen GJP, & Jolles J (2008). Detecting the significance of changes in performance on the Stroop Color-Word Test, Rey's Verbal Learning Test, and the Letter Digit Substitution Test: The regression-based change approach. *Journal of the International Neuropsychological Society*, 14(1), 71–80. 10.1017/S1355617708080028 [PubMed: 18078533]
- Wearn AR, Saunders-Jennings E, Nurdal V, Hadley E, Knight MJ, Newson M, Kauppinen RA, & Coulthard EJ (2020). Accelerated long-term forgetting in healthy older adults predicts cognitive decline over 1 year. *Alzheimer's Research & Therapy*, 12(1), 119. 10.1186/s13195-020-00693-4
- Williams MN, Gómez Grajales CA, & Kurkiewicz D (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation*, 18(11), 1–14.
- Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund L-O, Nordberg A, Bäckman L, Albert M, Almkvist O, Arai H, Basun H, Blennow K, de Leon M, DeCarli C, Erkinjuntti T, Giacobini E, Graff C, Hardy J, ... Petersen RC (2004). Mild cognitive impairment—beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, 256(3), 240–246. 10.1111/j.1365-2796.2004.01380.x [PubMed: 15324367]
- Womble MN, Reynolds E, Schatz P, Shah KM, & Kontos AP (2016). Test-Retest Reliability of Computerized Neurocognitive Testing in Youth Ice Hockey Players. *Archives of Clinical Neuropsychology*, 31(4), 305–312. 10.1093/arclin/acw011 [PubMed: 27084734]

Public Significance Statements

This simulation shows the accuracy of the linear Regression-based Reliable Change Index (RCI) to identify longitudinal decline, and provides an R package to calculate reliable change with both the linear regression and an alternative logistic RCI for clinicians and researchers. Our simulations showed that the linear and the logistic models approximate with samples of size 200 or above, with the logistic model outperforming the linear model for smaller samples.

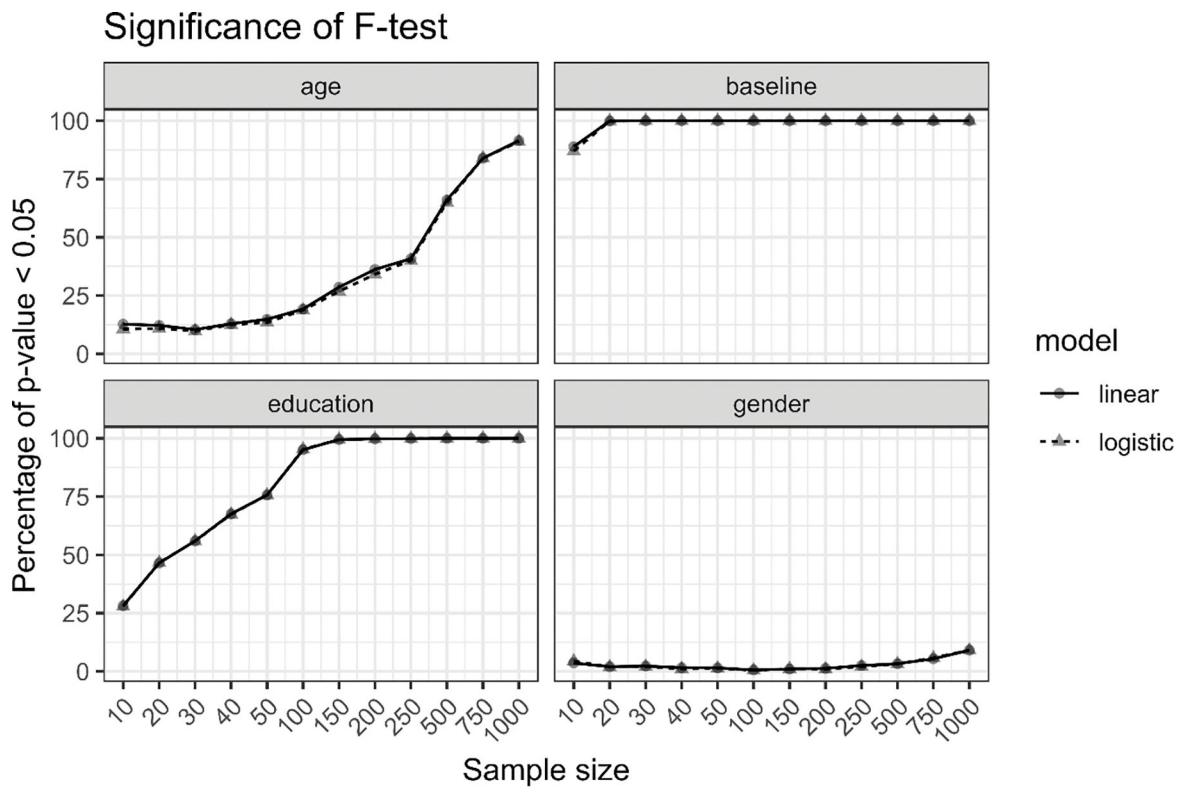


Figure 1:

Percentage of times the p-value associated to the F-test was less than 0.05, i.e. indicated a significant effect, for age, the baseline score, education and gender, for the linear and logistic regression models fitted to the simulated datasets. We drew 1,000 samples of sizes 10, 20, 30, 40, 50, 100, 150, 200, 250, 500, 750 and 1,000 at random from a simulated population comprising of one million individuals. True parameter values are indicated in Table 1.

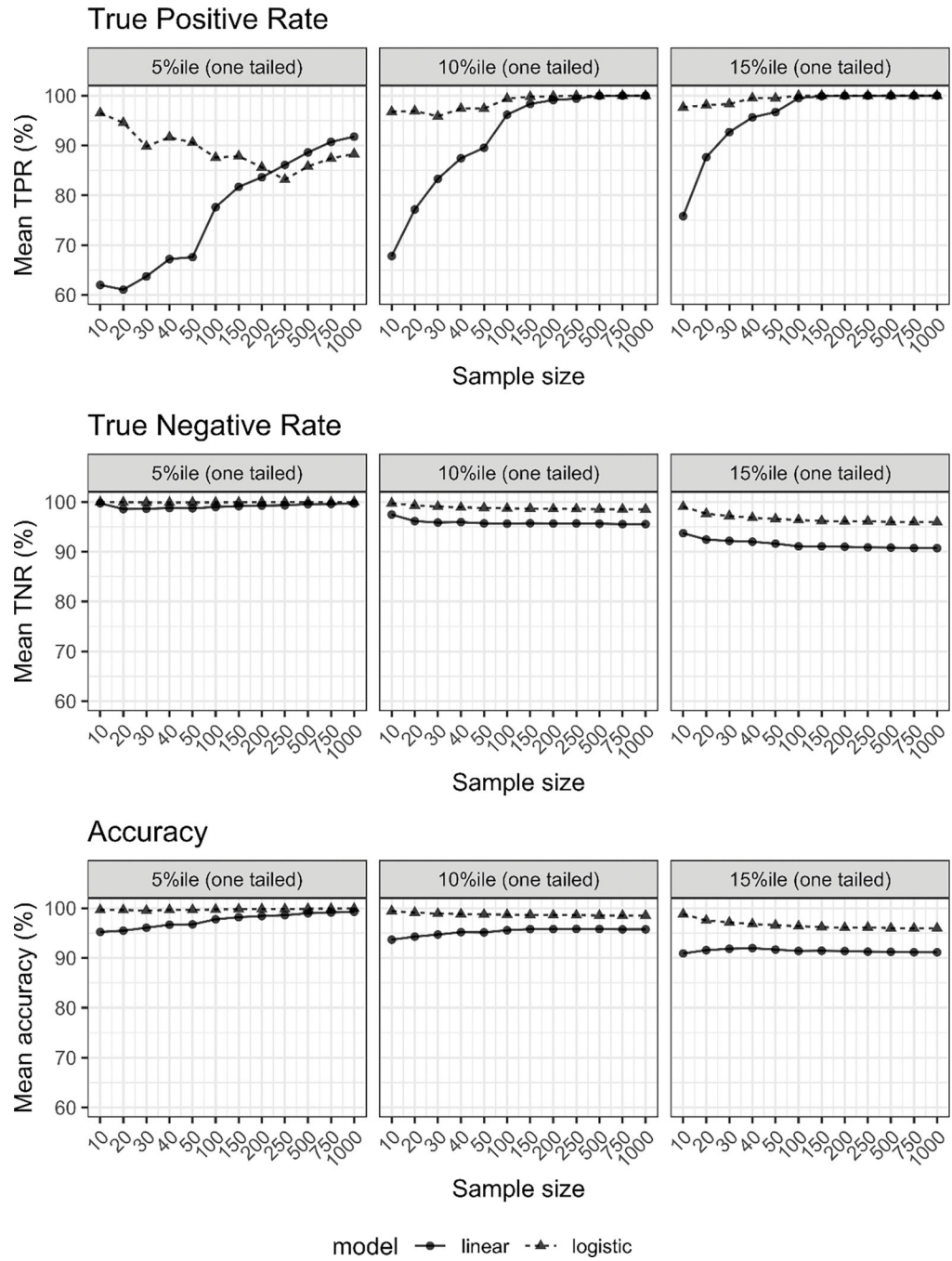


Figure 2: True positive rates, true negative rates and overall accuracy for identifying reliable decline for the linear and logistic reliable change index, using different threshold values: -1.64 corresponding to the 5% percentile of the standard normal distribution, -1.28 corresponding to the 10% percentile and -1.04 corresponding to the 15% percentile. Reliable decline is identified when the z-score associated to the reliable change index is less than or equal to the threshold. These results are averaged across 1,000 simulated datasets of sizes 10, 20, 30, 40,

50, 100, 150, 200, 250, 500, 750 and 1,000 drawn at random from a simulated population comprising of one million individuals.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Descriptive statistics

	Group	95% Confidence Interval			SD	Min	Max	Cohen's d	95% Confidence Interval	
		Mean	Lower	Upper					Lower	Upper
Age	NC	74.26	73.72	74.80	5.86	56.2	90.1	0.24	0.11	0.37
	MCI	72.60	71.92	73.29	7.59	55.0	91.4			
Education	NC	16.43	16.18	16.68	2.73	6	20	0.16	0.03	0.28
	MCI	16.00	15.75	16.25	2.78	4	20			
MMSE	NC	29.07	28.97	29.17	1.12	24	30	0.74	0.59	0.87
	MCI	28.01	27.86	28.17	1.68	24	30			
Baseline										
AVLT-DR	NC	8.01	7.68	8.34	3.57	1	15	0.59	0.46	0.73
	MCI	5.86	5.53	6.19	3.64	1	15			
Follow-up										
AVLT-DR	NC	7.15	6.82	7.48	3.53	1	15	0.58	0.40	0.71
	MCI	5.10	4.78	5.43	3.57	1	15			

NC: normal control group. MCI: mild cognitive impairment. SD: standard deviation

Table 2:

Parameter values estimated from the ADNI dataset from a linear regression model and a logistic regression model

Parameter	Linear regression estimate	Logistic regression estimate
Intercept	2.3102	-.5782
Baseline score	0.6726	0.2010
Age	-0.0285	-0.0087
Education level	0.0686	0.0224
Gender – female	0.2837	0.0894

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript