False positives associated with responder/non-responder analyses based on motor evoked potentials

van de Ruit, Mark; Grey, Michael J.

DOI
10.1016/j.brs.2018.11.015

Publication date
2019

Document Version
Final published version

Published in
Brain Stimulation

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable). Please check the document version above.
Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.
False positives associated with responder/non-responder analyses based on motor evoked potentials

Mark van de Ruit a, Michael J. Grey b, *

a Department of Biomechanical Engineering, Delft University of Technology, Delft, the Netherlands
b Acquired Brain Injury Rehabilitation Alliance, School of Health Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK

ARTICLE INFO

Article history:
Received 15 June 2018
Received in revised form 24 August 2018
Accepted 29 November 2018
Available online 3 December 2018

Keywords:
Variability
MEP
TMS
Plasticity
Corticospinal excitability
Responders

ABSTRACT

Background: A trend in the non-invasive brain stimulation literature is to assess the outcome of an intervention using a responder analysis whereby participants are di- or trichotomised in order that they may be classified as either responders or non-responders.

Objective: Examine the extent of the Type I error in motor evoked potential (MEP) data subjected to responder analyses.

Methods: Seven sets of 30 MEPs were recorded from the first dorsal interosseous muscle in 52 healthy volunteers. Four classification techniques were used to classify the participants as responders or non-responders: (1) the two-step cluster analysis, (2) dichotomised thresholding, (3) relative method and (4) baseline variance method.

Results: Despite the lack of any intervention, a significant number of participants were classified as responders (21–71%).

Conclusion: This study highlights the very large Type I error associated with dichotomising continuous variables such as the TMS MEP.

© 2018 Elsevier Inc. All rights reserved.

Introduction

Similar to many other interventions, the efficacy of non-invasive brain stimulation (NIBS) is limited to a subset of the population and it is important to better understand what proportion of participants might respond. A recent trend in the NIBS literature is to use a responder analysis to classify participants as responders or non-responders following an intervention. This simplifies the statistical analysis, interpretation and presentation of results [1]. In the NIBS literature, this classification is typically performed by di- or trichotomising the motor evoked potential (MEP) produced in response to transcranial magnetic stimulation (TMS) as this is considered a surrogate marker of neuroplasticity [2].

Pellegrini et al., 2018 [3] recently conducted a systematic review of responder analyses in NIBS. They concluded that responder analyses can effectively identify subgroups based on response patterns, and be used to estimate the proportion of participants who might respond to the intervention. However, they also noted a lack of consistency and consensus in the methods by which responders are quantified. Furthermore, they highlighted that many studies in the NIBS literature lack a control group. As a result, the effect of natural variability of the MEP is not accounted for with these analyses. The MEP magnitude has considerable trial-to-trial variability and drift over time, which arise due to controllable and uncontrollable factors of physiological (e.g. cortical rhythms, arousal, etc.) and non-physiological (e.g. TMS coil placement and/or movement) origin [4,5].

Responder analyses methods gained popularity in the early 2000s in the clinical medicine and psychology literature primarily as a means to establish proportions of responders in drug trials and in marketing studies [6–8]. However, these methods were then criticised by methodologists who questioned the validity of dichotomising (or trichotomising) continuous variables. They noted in particular that inferences made from such analyses are susceptible to large Type I error (false positives) that can lead to erroneous conclusions [1,6,9–19]. The aim of the present study was to examine the extent of the Type I error in MEP data that are subjected to different types responder analyses.

* Corresponding author.
E-mail address: m.grey@uea.ac.uk (M.J. Grey).

https://doi.org/10.1016/j.brs.2018.11.015
1935-861X/© 2018 Elsevier Inc. All rights reserved.
Methods

Experimental procedures

Fifty-two healthy participants, without contraindication to TMS and no history of neurological psychiatric disorder, participated in the study (20 ± 2 y, range 18–25, 35 female). Participants visited the laboratory once for ~1 h, during which MEPs were recorded from the first dorsal interosseous (FDI). Participants sat comfortably and were instructed to relax both the hand and arm, and to keep their eyes open for the duration of the experiment. To facilitate this instruction throughout the experiment, interactive feedback of FDI muscle activity was provided on a computer monitor. TMS was delivered through a 90 mm figure-of-8 coil (type: batwing; type no. 15411) using a Magstim Rapid² stimulator (Magstim Ltd, Dyfed, United Kingdom). Coil position and orientation were monitored with frameless stereotaxy (BrainSight 2, Rogue Research Inc, Montreal, Canada). The stimulation intensity required to evoke 1 mV (Stim) peak-to-peak MEPs (MEP_{pp}) was determined by adjusting the intensity until the mean of 30 stimuli produced a 1 mV MEP_{pp} (calibration data set in Fig. 1A). Next, seven sets of 30 stimuli were recorded with a 4 s inter-stimulus interval and 2 min rest between sets. The first set was deemed a baseline to which the remaining 6 data sets would be compared. Fig. 1A summarises the experimental protocol.

Statistical analysis

The MEP_{pp} amplitude was extracted between 20 and 50 ms after stimulation and averaged across all stimuli within a set. The mean MEP_{pp} for each set was then used for statistical analysis and classification either: (1) without any further processing; or (2) after normalisation to the mean MEP_{pp} of the baseline set (B), the ‘grand average (GA) method’. Therefore, each classification method was performed twice on the same data, either the absolute mean MEP_{pp} amplitudes for each set, or the normalised GA data. Before classification, the continuous data was analysed using a repeated measures analysis of variance (RM-ANOVA) across sets for the mean absolute MEP_{pp} values. Subsequently, the participants were classified using the four common methods found in the NIBS literature. Following classification, a mixed RM-ANOVA was performed on the absolute MEP_{pp} data with the within-factor ‘set’ and between-subjects factor ‘group’ (i.e. the result of the classification method). In addition, a one-way RM-ANOVA was performed for each group individually on the absolute MEP_{pp} data to classify groups of participants as either:

- (+) responders: significant increase in MEP_{pp} across set
- (−) responders: significant decrease in MEP_{pp} across set
- (0) responders or non-responders: no significant change in MEP_{pp} across set

If Mauchly’s Test of Sphericity indicated that the assumption of sphericity had been violated, a Greenhouse-Geisser correction (GG) was performed. All statistical tests were performed using SPSS, with significance accepted at p < 0.05.

Responder analysis methods

1) Two-step cluster analysis: This SPSS method uses a two-step clustering approach that allows automatic detection of the optimal number of clusters. In the first step all cases are scanned and pre-clustered based on a predefined distance criterion (e.g. squared Euclidian distance or log-likelihood) that specifies either the difference or similarity between cases. In the second step, the algorithm uses agglomerative hierarchical clustering to merge the sub clusters resulting from the first step into a smaller number of clusters. In the present study we allowed the algorithm to automatically determine the number of clusters rather than specifying two or three clusters. This is a commonly used method in NIBS literature [20–26].

2) Dichotomised thresholding: This method separates data into two groups based on a predefined threshold. For GA data, participants were categorised using the mean GA of sets (in our case sets T1-T6). Participants were then classified as negative responders for mean GA < 1 and positive responders for mean GA > 1. This analysis was also performed on absolute MEP_{pp} data. With absolute MEP_{pp} data this method can be applied either on a group level or individually. For the group level analysis, the mean MEP_{pp} amplitude across all participants was chosen as the threshold (1.35 mV in this study). For the individual analysis, the threshold is set to the mean MEP_{pp} of the baseline set for each participant individually. Next, each participant is classified as a positive responder if the mean MEP_{pp} across T1-T6 is greater than the threshold and a negative responder if the mean MEP_{pp} across T1-T6 is less than the threshold. Dichotomised thresholding is a common method of subgrouping normalised MEP data [22,24–33].

3) Relative method: This method is used to classify participants into three groups based on a predefined percent change from baseline threshold. This method has been used in several studies to trichotomise participants using a threshold of 10% [23,34], 15% [35], 20% [20] or 50% [36]. In the present study we used a conservative approach by choosing 20% change from baseline as the threshold. For the GA data, participants are classified as negative responders for mean GA across sets T1-T6 < 0.8, positive responders for mean GA > 1.2 and non-responders between 0.8 and 1.2. Likewise for the absolute MEP_{pp} data the threshold was 1.35 ± 0.27 mV as for the collected data the group mean of the baseline set B was 1.35 mV. This procedure was also performed on an individual level, in which case the threshold was individually determined based on the mean MEP_{pp} amplitude of set B.

4) Baseline variance method: In this method participants are trichotomised based on the variance of the baseline measure. For the GA data, the standard error (SE) of the GA of the baseline set was 0.14 across all participants. Therefore, a participant was classified as a (−) or (+) responder if the mean GA across sets T1-T6 was smaller or greater than 1.27 (95% confidence limit (CL) 1.00 ± 0.27) and a non-responder otherwise. Similarly, for MEP_{pp} data the SE of the baseline set was 0.17 across all participants (95% CL 1.35 ± 0.36 mV) and therefore a participant was a (+) responder when above this upper limit, a (−) when below the lower limit or a non-responder otherwise. The same analysis was also performed on the level of each individual, i.e. the CL of the baseline set was determined individually to assign the participant to the correct group. This method has been used in several studies [28,33,37–41].

Results

A one-way RM-ANOVA applied across all seven data sets (B-T6) before dichotomisation revealed neither a significant difference in mean MEP_{pp} amplitude across these data sets (F(4,76,242.75) = 1.27, GG = 0.28) nor in GA (F(4,74,241.73) = 1.31, GG = 0.26; Fig. 1B). The results for the subgrouping methods are presented in Table 1 and for the group level analysis visualized in Fig. 1C. The SPSS two-step cluster analysis determined two clusters to best separate the data. For the MEP_{pp} data 11 participants (−21%) were classified as responders, showing a significant increase in MEP_{pp} amplitude across T1-T6. For the GA data, 17 participants (−32%) were classified as responders, showing a significant increase in GA amplitude across T1-T6.
Fig. 1. Responder/non-responder analysis across TMS MEP testing sets. (A) Seven sets of 30 MEPs were acquired at a stimulation intensity selected to producing a mean 1 mV peak-to-peak MEP amplitude (mean SI1mV: 56 ± 10% of maximum stimulator output). The first set was considered the baseline to which the remaining six sets would be compared. (B) MEPpp amplitude across all participants and all sets. No effect of set on MEPpp amplitude observed for these data. (C) MEPpp amplitude is shown across each of the seven data sets, with the participants di- or tricotomised using a two-step cluster analysis, dichotomised thresholding, relative threshold method or baseline variance method on a group level. In this way participants are classified as either (+) responders (light grey lines), showing an increase in MEPpp amplitude compared to baseline, (0)- or non-responders (grey lines), no change in MEPpp amplitude across set, or (−) responders (black lines), a decrease in absolute MEPpp across set. The left column presents results when the classification was based on absolute MEPpp data, the right column when based on GA data. All data are presented as Mean ± S.D. The number of participants for each group can be found in Table 1.
(p < 0.01) across time, and 41 participants (79%) were classified as non-responders (p = 0.96). The same groups were identified using the GA data but with 19 responders (p < 0.01) and 33 non-responders (p = 0.22). The MEPpp and GA across time for each group is illustrated in Fig. 1C.

Using the dichotomised thresholding method on MEPpp data and a group level, 33 participants (63%) were classified as (+) responders (p < 0.01) and 19 participants (37%) as non-responders (p = 0.88). For the GA data, 28 participants (54%) were classified as (+) responders (GA > 1, p < 0.01) and 24 participants (46%) were classified as (−) responders (GA < 1, p = 0.01) (Fig. 1D).

The relative and baseline variance methods produced similar proportions of responders when performed irrespective of the group or individual level analysis. Generally, more participants were classified as non-responders for the GA data (40–58%) than the MEPpp data (29–52%). Moreover, the baseline variance method resulted in more non-responders (46–58%) than the relative method (29–40%).

Discussion

The present study followed a typical intervention design where TMS MEP data are collected at baseline and then again at predesigned times following the intervention. However, in the present study the participants were not exposed to an intervention. Therefore, subject to normal MEP variability, the ‘post-intervention’ data sets would not be expected to be different from baseline. As expected, parametric statistics performed on this continuous data set revealed no significant difference with time. However, when the data were subjected to different responder analyses, between 21 and 71% of the participants were classified as responders, thus revealing a large number of false positives.

The responder analysis has been used throughout the clinical medicine and psychology literature because it simplifies the analysis and interpretation of experimental results. Proponents of the responder analysis highlight its usefulness in clinical decision making [7]. However, for more than two decades methodologists have argued that the dichotomisation of continuous variables is not valid for hypothesis testing [1,9–14,16–18]. The dichotomisation of continuous variables results in significant loss of information (~35–50% depending on the distribution of the data), reduced power of the statistical tests, high probability of Type I error, biased parameter estimates and erroneously small variances (for detailed discussion see [113,16]).

The specific objective of the present study was to investigate the Type I error associated with responder analyses when MEP data are used to classify participants. In general, we observed substantial Type I errors with all of the responder analyses methods. Our results suggest that at best, 20% of the participants who have been classified as responders will have been classified erroneously. It may be valid to use a responder analysis to compare an intervention with a control group, but the specific response rates may be overestimated.

Conflicts of interest

We have no conflicts of interest to declare.

Ethical approval

The study was carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) and informed consent was obtained from all participants recruited to the study. Ethical approval for the study was granted from the University of Birmingham’s Science, Technology, Engineering and Mathematics ethics committee (ERN_13-0701).

Funding

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

We would like to thank Mr. Chris W. Wright for his assistance with the data collection, and Dr. Allan Clark for valuable discussions with respect to the responder analysis and statistical processing.

References