3-1-2022

Wilcoxon-Mann-Whitney odds ratio: A statistical measure for ordinal outcomes such as EDSS

C. W. Howard
University of Manitoba

G. Zou
Western University

S. A. Morrow
Western University

S. Fridman
Western University

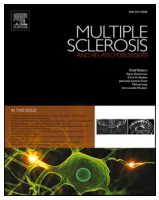
J. M. Racosta
Western University, jracosta@uwo.ca

Multiple Sclerosis and Related Disorders 59 (2022) 103516

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard



Wilcoxon-Mann-Whitney odds ratio: A statistical measure for ordinal outcomes such as EDSS

C.W. Howard [a,*], G. Zou [b,c], S.A. Morrow [d], S. Fridman [d], J.M. Racosta [d,e]

[a] Section of Neurology, Department of Internal Medicine, University of Manitoba, Manitoba, Canada
[b] Dept of Epidemiology and Biostatistics, Western University, London, Canada
[c] Robarts Research Institute, Western University, London, Canada
[d] Department of Clinical Neurological Sciences, Western University, London, Ontario, Canada
[e] The London MS Epidemiology Laboratory. London, Ontario, Canada

ARTICLE INFO

Keywords:
Odds
Multiple sclerosis
Non-parametric
Statistics
Wilcoxon-Mann-Whitney

ABSTRACT

Background: In many clinical situations, ordinal scales afford the primary method of semi-quantifying patient outcomes. In the field of multiple sclerosis, the primary ordinal scale is the Expanded Disability Status Scale. Predominant methods of ordinal scale statistical analysis provide a p-value without effect size or rely heavily on the assumption of proportionality of odds, subjecting them to lack of power and error. The Wilcoxon-Manny-Whitney Odds is a statistical method which provides significant information such as p-value, effect size, number needed to treat, confidence intervals, and is largely assumption-free. However, its utility has not been demonstrated in the field of multiple sclerosis.

Methods: Three clinical studies in the field of multiple sclerosis were selected which utilized ordinal scale outcomes at group or individual levels. Data from these studies was extracted using WebPlotDigitizer, and a custom Wilxocon-Mann-Whitney Odds software was applied to each dataset to re-analyze the main outcomes of the studies.

Results: *Re*-analysis of the manuscript by Muraro et al., 2017 demonstrated that autologous stem cell transplantation for relapsing remitting multiple sclerosis resulted in a 65% chance of improving from any Expanded Disability Status Scale category, although not significant. *Re*-analysis of the manuscript by Songthammawat et al., 2019 demonstrated chance of improvement with intravenous methylprednisolone and concurrent plasma exchange was 185% versus 32% in intravenous methylprednisolone with add-on plasma exchange, although not significant. *Re*-analysis of Kister et al., 2012 demonstrated the chances of mobility or cognition scores generally favored decline at every 5-year increment of study, and although statistically significant, these were smaller effect sizes ranging from an 11% chance of improvement to a 66% chance of decline over a 5-year interval.

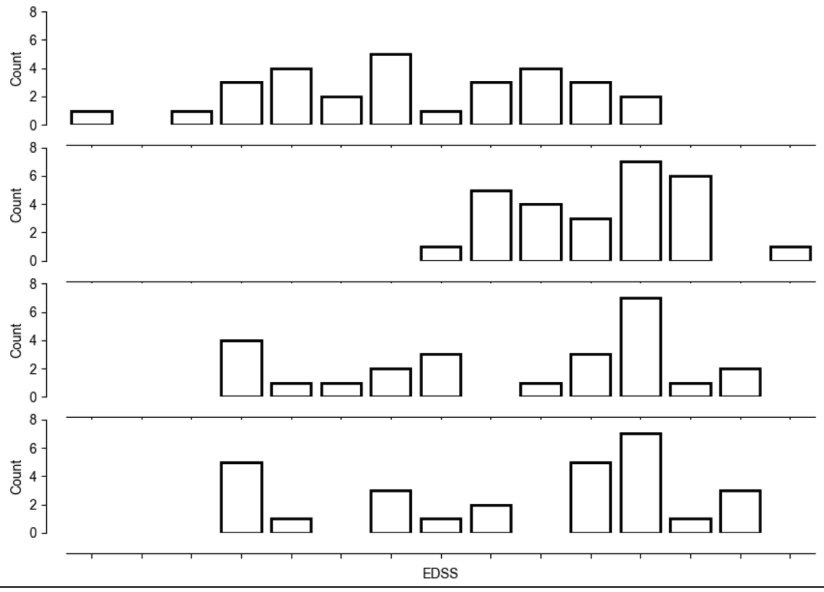
Discussion: The Wilcoxon-Mann-Whitney Odds simplifies ordinal data analysis with its robust largely assumption-free nature. In the place of numerous statistical tests, this single test provides effect size estimate, number needed to treat, p-values, and confidence intervals. Importantly, the Wilcoxon-Mann-Whitney Odds effect size calculation is intuitively applicable to both individual and population-levels. Further, the Wilcoxon-Mann-Whitney Odds allows intuitive description of the progression of large cohorts over time, and we were able to clearly convey the odds of mobility and cognitive decline over 30 years in a large multiple sclerosis cohort. Overall, the Wilcoxon-Mann-Whitney Odds is a powerful and robust statistical test with significant promise within the field of multiple sclerosis.

1. Introduction

The Expanded Disability Status Scale (EDSS) is the most widely used disability scale in clinical and research settings in multiple sclerosis (MS) (Kurtzke, 1983). Due to its ordinal nature, the distance between two consecutive EDSS values is not equivalent (Cumming et al., 2015). Therefore, much information provided by the EDSS is missed or misinterpreted when it is analyzed with statistical methods meant for

* Corresponding author. Health Sciences Center 4th Flr, GC430 820 Sherbrook St. R3A 1R9 Winnipeg, Manitoba, Canada.
E-mail address: howardc1@myumanitoba.ca (C.W. Howard).

https://doi.org/10.1016/j.msard.2022.103516
Received 28 September 2021; Received in revised form 13 December 2021; Accepted 8 January 2022
Available online 10 January 2022



Before Transplant	WMW Odds	IC95	p	NNT
Transplant	5.57	2.7-11.4	<0.01	1.44
1 year FU	1.65	0.9-3.0	0.09	4.09
Last FU	1.49	0.8-2.7	0.17	4.98

Fig. 1. Muraro et al., 2017.

continuous or dichotomic datasets. A common approach in analysis of EDSS scores is calculation of means, which is statistically inappropriate given unequal spacing between EDSS scores. Another approach is artificially dichotomizing EDSS values to facilitate statistical analysis which assumes equivalence of different values within the dichotomized groups, losing much of the inherent information (Cumming et al., 2015; Howard et al., 2013).

These limitations have led to the implementation of other methods for ordinal scale statistical analysis. Among them, the global statistic and responder analysis have been the most widely used; however, they still have significant shortcomings (Saver, 2007). Although the global statistic has the benefit of assessesing treatment effects on multiple outcome measures simultaneously, it fails to translate the measures of effect from the group level to individual level (Saver, 2007). Responder analysis addresses ordinal scales by adjusting outcome thresholds to individual score values at study entry (Saver, 2007). However, this method requires pre-specifying the variables to be adjusted, which is often not feasible in practice.

Two other techniques have been proposed to assess the full range of an ordinal scale. The Cochran–Mantel–Haenszel test is an assumption free based technique for the analysis of ordinal scales. Since the variance is derived under the null hypothesis, this method only provides p-values, without associated effect size estimates (Churilov et al., 2014). The other method, the proportional-odds logistic regression (also known as ordinal regression), does provide an intuitive effect size estimate, but its modeling relies heavily on the assumption of proportionality of odds, often not satisfied in clinical trials (Williams, 2016).

The "Wilcoxon-Mann-Whitney Odds Ratio", derived to relax the proportional odds assumption by Agresti, is an intuitive and statistically appropriate approach for ordinal data analysis (Agresti, 1980). It should be noted that this is technically more consistent with an odds than an odds ratio, and so we note it as the Wilcoxon-Mann-Whitney Odds (WMW-Odds) (Rahlfs and Zimmerman, 2019). The WMW-Odds method involves estimation of the odds that a randomly selected subject from one group will have a better outcome than that of a randomly selected subject from the other group. In situations wherein probability of a good outcome is equivalent between the two subjects, the ties may be split in order to maintain accurate representation. In this study we assessed the feasibility of implementing the WMW-Odds and its application in Multiple Sclerosis research by re-analysing three relevant published studies with publicly available data. In this study we aim at assessing the usability, interpretability and statistical appropriatness of the WMW-Odds for the analysis of ordinal data, with a particular emphasis on EDSS data. Altogether, these goals seek to expand on previous concerns referred to as "missing medians" (Cumming et al., 2015), or inaccurate representation of ordinal data in clinical studies (Cumming et al., 2015).

2. Methods

A Python-based code was developed to apply the WMW-Odds as follows as follows:

$$WMW\ Odds = \frac{[\Pr(Y2 > Y1) + 0.5Pr(Y1 = Y2)]}{[(\Pr(Y2 < Y1) + 0.5\Pr(Y1 = Y2)]}$$

This results in the creation of a generalized odds ratio which splits ties equally across both groups, following a modification of tie handling for the original Agresti's Generalized Odds Ratio (Churilov et al., 2014). To validate the accuracy of the code, the mock data analyzed by Cumming et al. was analyzed, and the WMW-O model's calculated result was compared to WMW-Odds results calculated manually with Excel calculations (Cumming et al., 2015). Further, a second version of the WMW-Odds was developed independently of the first by a second independent author using R software. This model's results were compared to the Python-based WMW-Odds to assess consistency.

The number needed to treat (NNT), defined as the expected number of patients who need to be treated such that one patient would have a better outcome was calculated as:

$$NNT = 1 + \frac{2}{(\text{WMW Odds} - 1]}$$

2.1. Characterisitics and re-analysis of studies

Three MS studies implementing different methodologies were selected. Selection criteria was based on their scientific relevance and public availability of individual or specific group data (Kister et al., 2013; Muraro et al., 2017; Songthammawat et al., 2020). When the data from the studies selected was only available in figures, it was extracted using WebPlotDigitizer (Rohtagi, 2020).

In the first study Muraro et al. evaluated outcomes after autologous hematopoietic stem cell transplantation for treatment of relapsing and progressive forms of multiple sclerosis that failed to respond to standard

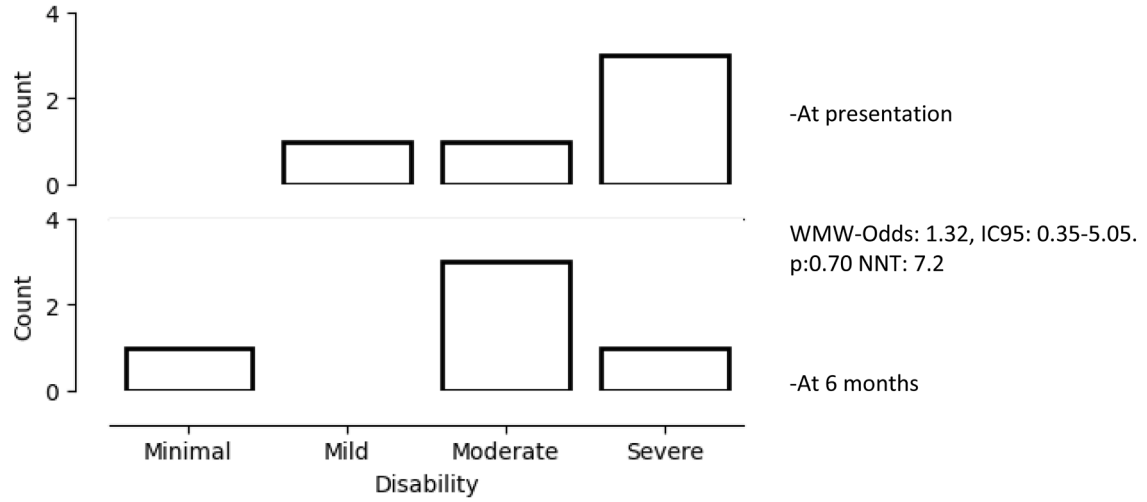


Fig. 2. EDSS severity change in patients using IVMP and add on PLEX.

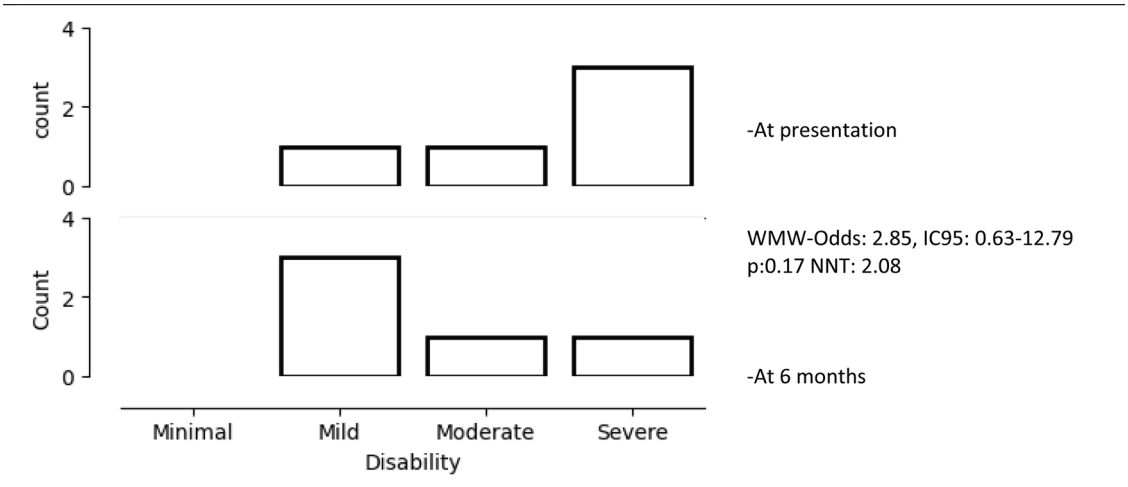
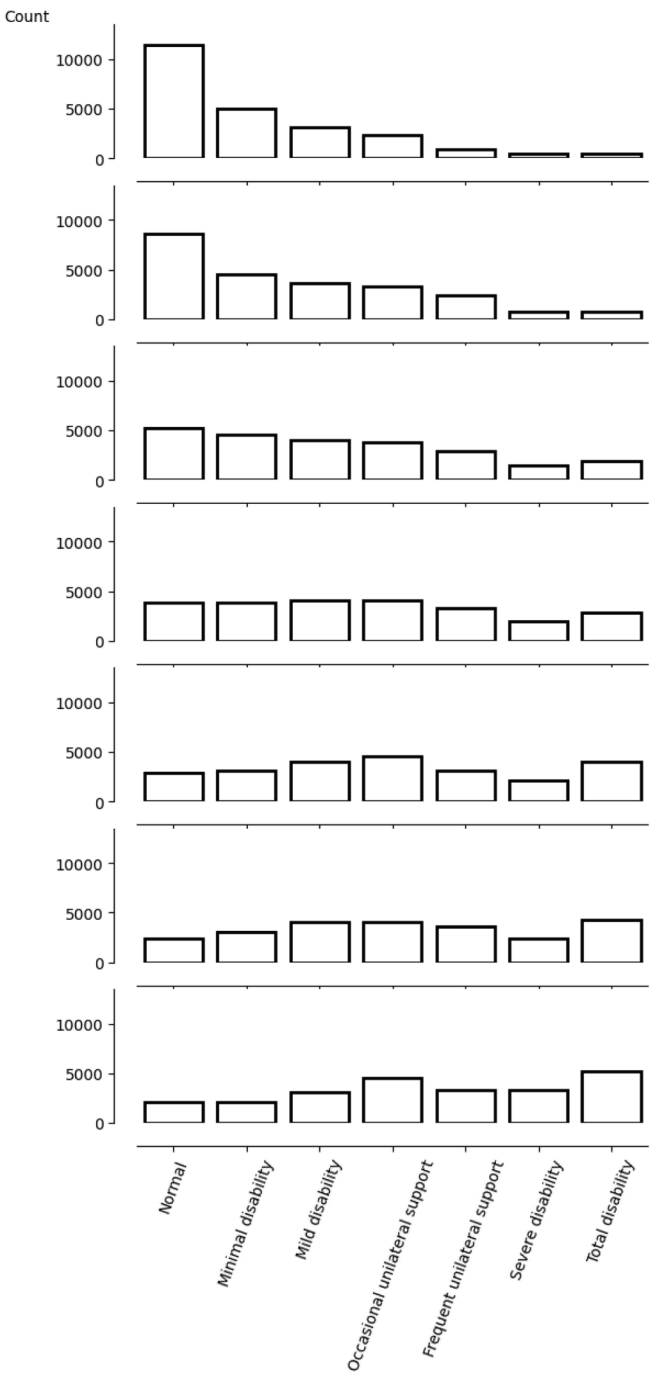


Fig. 3. EDSS severity change in patients using IVMP plus Plex.

therapies (Muraro et al., 2017). Outcomes collected were the EDSS at baseline and at least one follow-up visit/report after transplantation. For each patient having an EDSS assessment 1 year before and 1 year after transplant, the yearly EDSS changes pre- and post-transplant were calculated and compared by a repeated measures analysis of variance (ANOVA) with 2 time points (change pre-transplant vs post-transplant). We re-analyzed the change in EDSS using WMW-Odds in the sub-group of patients with RRMS, including EDSS change from the first follow up to the time of the transplant (1,1 +/- 0.16 years prior transplant), from the time of the transplant to the one year follow up (0.96 +/- 0.19 years after transplant) and from the time of the transplant to last follow up (2.3+/- 0.86 years after). In the second study Songthammawat et al. compared the efficacy between 5 patients with Neuromyelitis Optica Spectrum Disorders (NMOSD) with a severe acute attack who received intravenous methylprednisolone (IVMP) with subsequent add-on plasma exchange (PLEX) versus simultaneous IVMP and PLEX (Songthammawat et al., 2020). We re-analyzed the change in EDSS for the same comparisons using WMW-Odds. In the third study Kister et al. analyzed symptom prevalence in each of the 11 domains included in The North American Research Committee on Multiple Sclerosis (NARCOMS) Registry database for the first 30 years from symptom onset, with more than 35,000 patients recording on symptom from 1996 to June 2011 (Kister et al., 2013). The change over the years was depicted on "symptom prevalence tables". We re-analyzed the changes on two relevant symptoms (mobility and cognition) at 5 years periods during the 30 years reported.

3. Results

Our comparison of the computational codes developed independently by two co-authors (SF and JMR) using different software (R and Python) yielded equal results. In addition, these results were also equal to those generated by manual calculation and using step-wise analytical procedures using Excel software. Our re-analysis of the study by Muraro et al. revealed significant worsening in EDSS from the initial assessment until the time of transplant while the analysis of EDSS after the transplant showed improvement more marked in the first year as compared with the last follow up, despite not being significant in either case (Fig. 1) (Muraro et al., 2017). Our re-analysis of the study by Songthammawat et al. revealed that treatment effect sizes for the groups using IVMP plus PLEX and IVMP add-on PLEX did not reach statistical significance (Figs. 2 and 3) (Songthammawat et al., 2020). The analysis of



WMW-Odds	CI95%	p	NNT
1.66	1.62-1.69	<0.001	4.04
1.49	1.46-1.52	<0.001	5.08
1.29	1.26-1.31	<0.001	7.97
1,22	1.19-1.24	<0.001	10.16
1.08	1.06-1.10	<0.001	26.32
1.23	1.20-1.25	<0.001	9.85

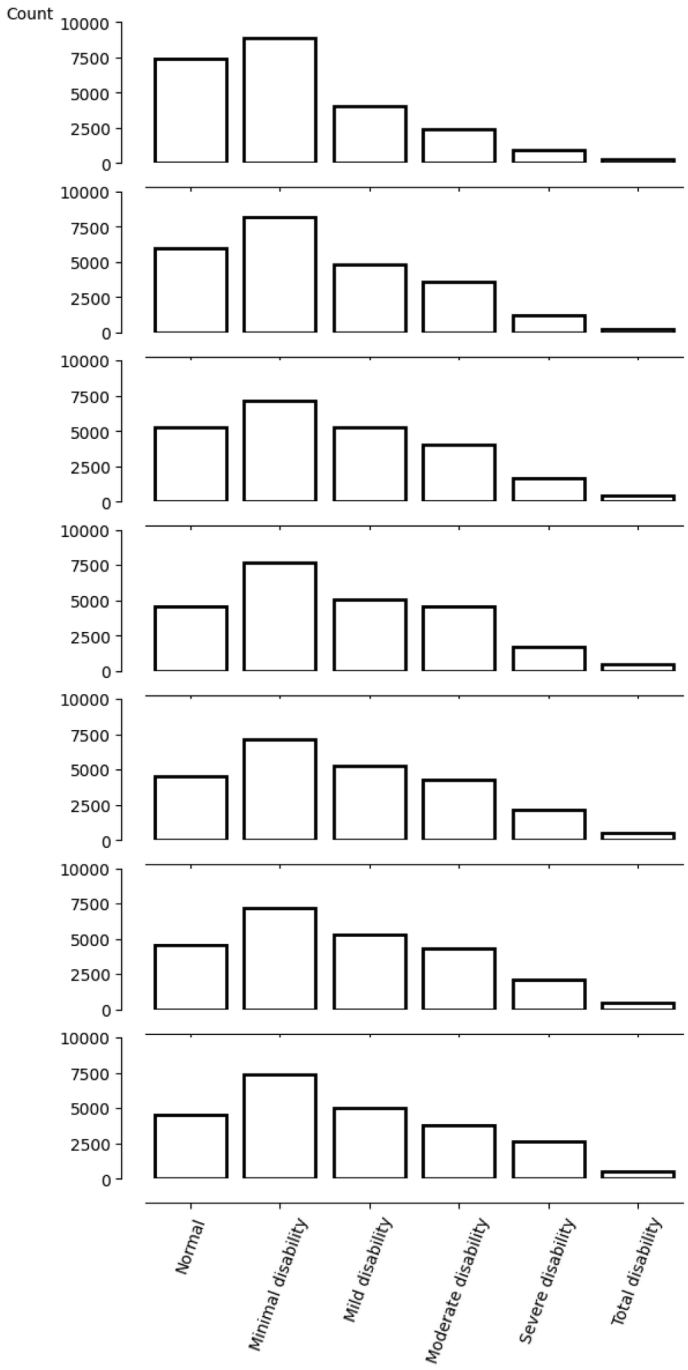
Fig. 4. Kister al 2020 Mobility over 30 year.

the study by Kister et al. quantifies the changes in large scale patient populations over time (Fig. 4 and 5) (Kister et al., 2013).

4. Discussion

Despite the recent advances in the analysis of neurological scales of ordinal nature (The optimising Analysis of Stroke Trials (OAST) Collaboration, 2007), there is yet no general consensus on the most appropriate statistical method to quantify EDSS changes. In this study we analyze the features of WMW-Odds, and its feasibility to quantify EDSS changes.

Up to date, most studies utilize mean and standard deviation values to report EDSS changes (Cumming et al., 2015). However, this approach is inherently inadequate, as distances between points in ordinal scales are unevenly spaced. Among different ordinal scales and patients' populations, this difficulty is particularly notorious when analyzing EDSS changes, as its distribution is rarely normal, and often follow a bimodal pattern (Hohol et al., 1995). Muraro et al. suggested significant mean EDSS increase in people with MS during the 12 months preceding autologous hematopoietic stem cell transplantation (0.94 points, 95%



	WMW-Odds	CI95%	p	NNT
	1.39	1.3-1.41	<0.001	6.18
	1.17	1.15-1.20	<0.001	12.56
	1.06	1.04 -1.08	<0.001	34.97
	1.04	1.02-1.06	<0.001	50.51
	1.13	1.11-1.16	<0.001	15.95
	0.89	0.87-0.90	<0.001	16.42

Fig. 5. Kister al 2020 Cognition over 30 years.

CI= 0.77 to 1.11), and significant decrease following the transplant (0.32, 95%CI= −0.15 to −0.49) (Muraro et al., 2017). Our re-analysis demonstrated that the chances for patients with RRMS to improve from any given EDSS category to a better category at one year after the transplant were 1.65 or 65% over the first year and 1.49 or 49% at the last follow up (Fig. 1). Our analysis yielded statistically significant EDSS worsening before entering the study, but only a tendency for improvement for the 1 year and last follow up ($p = 0.09$ and 0.17). These statistical outcomes are likely limited by the retrospective nature of the study. Similarly, our analysis of the study by Songthammawat et al. indicated similar trends to those suggested by the authors, however, statistical significance was not reached (Fig. 2 and 3) (Songthammawat et al., 2020). Similar to the initial study, the power of the study was largely limited by its small sample size, and limited by its retrospective nature. However, we also believe that the inaccurate impression of statistical significance in the original results was caused by the unfitted statistical approach. This could in turn hamper proper design of future studies (i.e. inaccurate sample size calculation). Nevertheless, it should not be assumed that WMW-Odds is less likely to detect significant treatment effects than parametric analyses. On the contrary, the

experience using other common ordinal scales (i.e. Rankin scale), has demonstrated that parametric and binary outcomes analysis are often less likely to detect significant treatment effects in the context of non-parametric distributions (The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007).

The use of WMW-Odds could be advantageous in the context of large sample sizes. In this circumstance, the analysis of statistical significance alone can be misleading, and the use of appropriate effect size estimation aids in the interpretation of clinical relevance of an effect. Kister et al. assessed the natural history of large MS cohorts, rounding 35,000 patients per cohort, and were able to discern 2 divergent patterns of progression among different symptoms. While in domains such as mobility a steady worsening was observed over the 3 decades of follow up, other domains such as cognition demonstrated little change in the distribution after the initial 15 years of disease. However, the authors only used a visual inspection of their "symptom prevalence tables", without any statistical analysis, only reporting the percentage of change within the groups. Our re-analysis of these two domains (mobility and cognition) implementing WMW-Odds allowed quantifying the change for the patterns observed by Kister et al. (Fig. 4) (Kister et al., 2013). For example, the chance of progressing to a worse category of mobility during the first 5 years of enrollment was 66% (WMW-Odds=1.66), and despite the progression was slower over the following decades, a sizeable worsening was still observed over the last 5 years of the analysis (WMW-Odds=1.23) (Fig. 4). On the contrary, this pattern was not observed in the cognition domain, where worsening was markedly reduced towards last years of follow-up. Regarding the underlying data, a recent study demonstrated 15% of the population in this patient cohort had passed away by 2010, which raises concern for possible frame shifts which may obscure changes in symptoms severity over time by masking changes in patients who passed away (Cutter et al., 2015).

WMW-Odds also enables the estimation of NNT. NNT, defined as the expected number of people who need to be treated for one person to benefit, has an immediate natural relationship to WMW-Odds as previously described (Bath et al., 2011). Specifically, this is a net NNT, which contrasts the odds that a person receiving a given treatment or intervention improves, remains the same, or declines, compared to someone not receiving it. NNT are thought to be generally lower when implementing ordinal approaches, such as WMW-Odds, as compared to binary approaches. This higher sensitivity has been demonstrated by analysing ordinal outcomes from stroke trials, and a similar effect may be expected for trials analyzing EDSS changes. Our calculations of NNT provided complementary insight into the effect sizes for the comparisons (Figs. 1, 4 and 5).

In addition, the use of WMW-Odds allows pooling of ordinal outcomes in meta-analysis using the standard analytical approach for odds ratios (Churilov et al., 2014). The current lack of standard ordinal analysis, might lead to exclusion of MS studies from metanalysis (Cumming et al., 2015). Conversely, as meta-analyses are heavily influenced by assumption of normality, dichotomizing ordinal outcomes is not a feasible alternative. Cumming et al. demonstrated that WMW-Odds can be readily combined in meta-analysis, and stated that it is not be limited to ordinal data, but could also be applied to continuous data (Cumming et al., 2015). Importantly, with simple ordinal tables, a synthetic individual-level dataset can be created and analyzed using the WMW-Odds, allowing analysis of synthetic individual-level data from meta-analysis data.

Overall, the WMW-Odds is a relatively easily computed odds ratio. Being similar to proportional odds models, WMW-Odds does not rely upon the assumption of proportional odds, and therefore its application is not subject to undesirable test-born premises restricting the interpretability of the results(Howard et al., 2013). Nevertheless, this advantage is not in detriment of losing other benefits of proportional odds models, such as the possibility of adjusting for covariates. Despite this particular aspect is not explored in this paper due to the unavailability of individual covariates data in the studies examined, this characteristic has been explicitly demonstrated by Howard et al. (Howard et al., 2013), and we believe it represents a venue for further research.

This study has some limitations. First, extracting individual data from published figures in addition to tables and text provided by the publications might have led to discrepancies. In addition, comparison of effect sizes using the WMW-Odds approach vs. dichotomization of EDSS outcomes was not conducted in this study. Lastly, our study did not collect sufficient data to test the use of WMW-Odds outcomes in metanalysis.

In conclusion, our study demonstrated that WMW-Odds is a convenient and straightforward statistical approach to analyze and report EDSS changes. WMW-Odds provides an assumption-free and easy to compute effect size estimation, that can be implemented in multiple study designs, and can be readily combined in meta-analysis. The use of WMW-Odds might specially benefit progressive MS research, where outcomes are typically based on EDSS. Specifically, the WMW-Odds improves the analysis of EDSS data by respecting the ordinal nature of EDSS data, without compromising its robustness as a statistical tool (Cumming et al., 2015). In the future, it should be tested whether utilizing WMW-Odds improves the detection of significant effects, as it could allow to decrease the sample sizes required in clinical research.

Funding

This work received no funding.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors have no acknowledgements to make.

References

Agresti, A., 1980. Generalized Odds Ratios for Ordinal Data Author. Biometrics 36, 59–67.

Bath, P.M.W., Hogg, C., Tracy, M., Pocock, S., 2011. Calculation of numbers-needed-to-treat in parallel group trials assessing ordinal outcomes: case examples from acute stroke and stroke prevention Optimising the Analysis of Stroke Trials (OAST) Collaboration with the writing committee 6, 472–479. https://doi.org/10.1111/j.1747-4949.2011.00614.x.

Churilov, L., Arnup, S., Johns, H., Leung, T., Roberts, S., Campbell, B.C.V, Davis, S.M., Donnan, G.A., 2014. An improved method for simple, assumption-free ordinal analysis of the modified Rankin Scale using generalized odds ratios. Int. J. Stroke 9, 999–1005. https://doi.org/10.1111/ijs.12364.

Cumming, T.B., Churilov, L., Sena, E.S., 2015. The Missing Medians: exclusion of Ordinal Data from Meta-Analyses. PLoS ONE 10, 1–10. https://doi.org/10.1371/journal.pone.0145580.

Cutter, G.R., Zimmerman, J., Salter, A.R., Knappertz, V., Suarez, G., Waterbor, J., Howard, V.J., Ann Marrie, R., 2015. Causes of death among persons with multiple sclerosis. Mult. Scler. Relat. Disord. 4, 484–490. https://doi.org/10.1016/j.msard.2015.07.008.

Hohol, M., Orav, E., Weiner, H., 1995. Disease Steps in Multiple Sclerosis. Neurology 45, 251–255.

Howard, G., Waller, J.L., Voeks, J.H., Virginia, J., Jauch, E.C., Lees, K.R., Nichols, F.T., Rahlfs, V.W., Hess, D.C., 2013. A simple, assumption-free and clinically interpretable approach for analysis of modified Rankin outcomes. Stroke 43, 1–11. https://doi.org/10.1161/STROKEAHA.111.632935.A.

Kister, I., Bacon, T.E., Chamot, E., Salter, A.R., Cutter, G.R., Kalina, J.T., Otr, L., Herbert, J., 2013. Multiple Sclerosis Symptoms. Int. J. MS Care 15, 146–157. https://doi.org/10.7224/1537-2073.2012-053.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33, 1444–1453.

Muraro, P.A., Pasquini, M., Atkins, H.L., Bowen, J.D., Farge, D., Fassas, A., Freedman, M.S., Georges, G.E., Gualandi, F., Hamerschlak, N., Havrdova, E., Kimiskidis, V.K., Kozak, T., Mancardi, G.L., Massacesi, L., Moraes, D.A., Nash, R.A., Pavletic, S., Ouyang, J., Saiz, A., Simoes, B., Trněný, M., Zhu, L., Badoglio, M., Zhong, X., 2017. Long-term Outcomes After Autologous Hematopoietic Stem Cell Transplantation for Multiple Sclerosis. JAMA Neurol 74, 459–469. https://doi.org/10.1001/jamaneurol.2016.5867.

Rahlfs, V., Zimmerman, H., 2019. Effect size measures and their benchmark values for quantifying benefit or risk of medicinal products. Biomed. J. 61, 973–982.

Rohtagi, A., 2020. WebPlotDigitizer.
Saver, J., 2007. Novel End Point Analytic Techniques and Interpreting Shifts Across the Entire Range of Outcome Scales in Acute Stroke Trials. Stroke 38, 3055–3062.
Songthammawat, T., Srisupa, T., Siritho, S., Kittisares, K., 2020. A pilot study comparing treatments for severe attacks of neuromyelitis optica spectrum disorders: intravenous methylprednisolone (IVMP) with add-on plasma exchange (PLEX) versus simultaneous IVMP and PLEX. Mult. Scler. Relat. Disord. 38, 101506 https://doi.org/10.1016/j.msard.2019.101506.
The Optimising Analysis of Stroke Trials (OAST) Collaboration, 2007. Can We Improve the Statistical Analysis of Stroke Trials? Stroke 38, 1911–1915.
Williams, R., 2016. Understanding and interpreting generalized order logit models. J. Math. Sociol. 40, 7–20.