Dialogues in Clinical Neuroscience & Mental Health

DOI: 10.26386/obrela.v2i2.115 ISSN 2585-2795

Randomized Controlled Trials: The case of Multiple Sclerosis - Refining the constraints of a treasure, a short outline

Theodoros S. Constantinidis

Abstract

Randomized controlled trials (RCTs) are the most valid methodological tool for establishing causal relationships. Nevertheless, their validity is constrained by various methodological details concerning their design, conduction and implementation. Failure of successful randomization, absence of correction for multiple comparisons, use of noisy scales for measuring a disease parameter, are a few of these constraints. Furthermore, there are constraints inherent in the scientific research methodology, like the use of the p-value as a threshold of statistical significance and in succession of inferential reasoning, as a threshold of truth. In this work, the examples of RCTs illustrating these limitations are drawn from the field of multiple sclerosis (MS). In general, RCTs in MS mostly are well designed, adequately powered, and well conducted. Nevertheless, sometimes there are exceptions, leading to false conclusions and steering clinical practice toward wrong choices.

Keywords: Clinical trials, multiple sclerosis, methodology, disability, p-value, replication.

Special Issue in Demyelinating Diseases

Corresponding author: Theodoros S. Constantinidis MD, PhD, Private Consultant Neurologist, 31 Ermou str, Korinthos 20131,t.constantinidis@gmail.com

DOI: 10.26386/obrela.v2i2.115 ISSN 2585-2795

Introduction

Multiple Sclerosis is a chronic autoimmune disease of the central nervous system (CNS), implicating both inflammatory and neurodegenerative pathogenic mechanisms (1). The main clinical phenotype is the relapsing-remitting type of the disease (RRMS). In the majority of cases it is characterized by acute exacerbations and subsequent various degrees of recovery (2). The typical course of RRMS, after several years of relapses and remissions, is the gradual and continuous (perhaps fluctuating) worsening of disability, marking the transition to the secondary progressive subtype of the disease (SPMS) (3). In a minority of patients, this gradual and continuous worsening of disability begins from the disease onset. This is the defining clinical feature of the primary progressive subtype of MS (PPMS). Both SPMS and PPMS subtypes belong to the spectrum of progressive type of MS (3).

European Medicines Agency (EMA) has already approved 16 drugs for MS (4), which aim to either prevent relapses or treat symptoms (Table 1). In Greece, 14 of these drugs are already being reimbursed from social security services while the other two (Cladribine and Ocrelizumab) will be integrated in the forthcoming months. It should be noted that Ocrelizumab has been approved by EMA for the early stages of PPMS (4). It is the first drug approved with this indication.

TABLE 1

Active Substance	Year of Approval	Administration	Indication
Interferon β 1-b	1995	Injectable SC	RRMS
Interferon β 1-a EM	1997	Injectable SC	RRMS
Interferon β 1-a SC	1998	Injectable SC	RRMS
Mitoxantrone	1998	Injectable IV	RRMS
Glatiramer Acetate	2002	Injectable SC	RRMS
Natalizumab	2006	Injectable IV	RRMS
Fingolimod	2011	Per os	RRMS
Fampridine	2011	Per os	Walking Disability
Cannabidiol / δ-9-tetrahydrocannabinol	2011	Oromucosal Spray	Spasticity
Chronic Pain			
Teriflunomide	2013	Per os	RRMS
Alemtuzumab	2013	Injectable IV	RRMS
Dimethylfumarate	2014	Per os	RRMS
Peginterferon β-1a	2014	Injectable SC	RRMS
Daclizumab	2016 | Withdrawn	Injectable IV	RRMS
Cladribine	2018	Per os	RRMS
Ocrelizumab	2018	Injectable IV	RRMS | PPMS

Randomized Controlled Trials

A prerequisite for drug approval by the authorities is the design and implementation of large scale, multicenter, multinational, double blind, RCTs. The randomization procedure is considered as the most valid and reliable method for balancing all the baseline confounders (effect modifiers), both known (table 2a and 2b) and unknown, between the groups under comparison e.g. novel therapy vs placebo. This is why in the framework of evidence-based medicine, RCTs are graded with the highest level of validity among other types of studies, namely cohort or case control observational studies (5).

TABLE 2a

Effect modifiers usually checked for balancing between groups, after randomization	
NOVEL THERAPY	PLACEBO
Age	Age
Sex (Usually % of females)	Sex (Usually % of females)
Duration from disease onset	Duration from disease onset
No of Relapses the Last 1 or 2 years or ARR*	No of Relapses the Last 1 or 2 years or ARR*
Disability-EDSS** score	EDSS** score
MRI‡- No or volume of T2 hyperintense lesions	MRI‡ - No or volume of T2 hyperintense lesions
MRI‡ – No of Contrast enhancing T1 lesions	MRI‡ – No of Contrast enhancing T1 lesions
Previous use of DMT§	Previous use of DMT§

TABLE 2b

Effect modifiers usually not-checked for balancing between groups, after randomization	
CSF†	CSF†
Cognitive Status	Cognitive Status
Total Brain Volume or Grey or White Matter Volume	Total Brain Volume or Grey or White Matter Volume
Nfl$ in CSF or Blood	Nfl$ in CSF or Blood

*ARR: Annualized Relapse Rate **EDSS: Expanded Disability Status Scale

‡MRI: Magnetic Resonance Imaging §DMT: Disease Modifying Therapy

†CSF: Cerebrospinal Fluid $Neurofilaments (Light Chain)

Does Randomization always succeed?

In the majority of clinical trials randomization works as expected. Nevertheless, sometimes it may result in imbalances, due to mostly unidentified reasons. An example is the pivotal clinical trial of glatiramer acetate 20mg (GA), published on 1995 (6). The randomization was implemented using the SAS statistical package. The baseline EDSS score, after randomization, was 2.8±1.2 (mean± standard deviation) for the GA group and 2.4±1.3 for placebo (6). This difference is imbalanced since a t-test results in t=2.5, p<0.02 (7). The effect of GA on disability progression, as measured by the EDSS change (increase) for 3 or 6 consecutive months, was not significantly different from the effect of placebo. The final conclusion was that GA does not inhibit disability deterioration more significantly than placebo. In addition, for the calculation of the relapse rate, the authors employed a multiple regression statistical technique, using the EDSS change as a covariate in order to adjust for the EDDS imbalance. This adjustment resulted in a significant difference in relapse rate between GA and placebo, with a p-value=0.007. The Food and Drug Administration (FDA) did not accept this statistical adjustment because it was not pre-planned. In its own report only the unadjusted comparisons of relapse rate with a p-value=0.055 are mentioned (8). This pivotal RCT of GA was the only one for several years. As a result, a systematic review of disease modifying therapies (DMTs) in MS, including only double blind RCTs, reported (21) the efficacy of GA as simi-lar to the efficacy of placebo under several outcome measures, without any mention to the EDSS imbalance in the baseline of GA's pivotal RCT (9). For a whole decade after this RCT of GA, there was a growing bibliographic trend of suggesting that GA is as effective as placebo. This trend changed after the publica-tion of two open label, randomized, head-to-head trials, com-paring GA to interferon β-1b (BEYOND) (10) and subcutaneous interferon β-1a (REGARD) (11). According to their conclusion, the efficacy of GA was equivalent to that of the two other inter-ferons-β in all outcome measures.

Multiple Comparisons

All pivotal clinical trials in MS aim to establish a statistically significant outcome of a primary endpoint. This may be a single outcome measure e.g. annualized relapse rate (12) (13). Alternatively, multiple primary endpoints (e.g. multiple dosing schemas) may be investigated. Furthermore, the endpoint may be composite. For example, in the context of time to failure, failure is defined either as the first occurrence of a relapse or the permanent discontinuation of treatment due to any cause

(14). In addition to these primary endpoints several other secondary ones are also included in the trial's list of endpoints, e.g. disability progression, magnetic resonance imaging (MRI) lesion burden, adverse events etc. All these endpoints have to be investigated through a vast number of comparisons. However, increasing the number of comparisons leads to a higher probability of getting falsely lower p-values (at the significance threshold 0.05). In turn, this may lead to a higher number of false positive results (15). In order to tackle the inflation of false positives due to multiple comparisons, several statistical procedures have been developed. The Bonferroni test is the most traditional. Roughly, it divides the p-value by 2 for every successive comparison. Nevertheless, this correction may be dramatic and has been criticized for overcorrection (15). Other more modest corrective procedures are used more frequently, like Hochberg, Benjamini-Hochberg, Hommel, Dunnet etc. The pivotal RCTs in MS include prespecified statistical procedures for controlling false positives due to multiple comparisons, since the regulatory agencies, namely FDA (16) and EMA (17), demand adherence to their guidelines on controlling multiplicity issues. Nevertheless, after the approval of a DMT by the authorities, several post-hoc group comparisons are published in order to assess various aspects of the treatment profile of DMT based on the original data of the pivotal RCT. All these comparisons should be embedded in the succession of comparisons of the initial RCT and should be corrected for multiplicity in order to avoid type I error inflation.

The latest example is the ORATORIO trial of ocrelizumab for PPMS (18). In this trial a prespecified exploratory endpoint (among several others, both primary and secondary) was a 20% confirmed progression in the time of Nine-Hole Peg Test (9-HPT) in all patients with PPMS in order to examine upper extremity function separately. One year after the initial publication of ORATORIO trial in the New England Journal of Medicine, a paper dedicated to the upper extremity function in PPMS patients was published in Multiple Sclerosis (19). This paper is a re-analysis of the ORATORIO upper extremity data. But this time, several more groups of patients were investigated, each with different confirmed progression thresholds in 9-HPT: 25%, 30% and 35%, during three different time periods: 12, 24 and 120 weeks. In addition, confirmed improvement (instead of progression) was used as a grouping factor in order to compare the two groups (improvement versus no improvement) according to two different thresholds of 9-HPT: 15% and 20% (19). All these comparisons were carried out for the total number of patients, and additionally the authors examined two more groups: patients with EDSS ≥ 6 and EDSS < 6. All these post-hoc comparisons were carried out without any multiplicity correction resulting in inflation of type I error. For example, the time to ≥ 25% confirmed progression in 24 weeks was significantly less than the corresponding time for placebo with a p-value 0.027 for both hands and 0.033 for the better hand. These p-values are close to the significance threshold of 0.05 and should be rather insignificant if corrected (at least according to the Bonferroni correction). In spite of the authors of this paper declaring these analyses as exploratory (19), multiplicity testing should be rigorously performed even in exploratory tri-als (15).

Estimation of disability with EDSS score

Permanent, irreversible disability, of any severity, is the greatest concern of all MS patients and the most important outcome measure of RCTs. Relapses cause temporary disability and, if absolutely remitted, cause only a dysfunction for few days or weeks. On the contrary, permanent disability affects decisively all aspects of the patients' daily life, their present and future. All the outcome measures of RCTs, like the relapse rate, the MRI lesion burden or the degree of short-term disability sustained over 3 or 6 months, are surrogate measures of permanent disability.

The EDSS scale is the gold standard tool for the assessment of disability in MS. Besides that, many RCTs also used EDSS for the confirmation of relapse. This is defined by an increase of 0.5 points in EDSS score or 2 points in one functional system or by 1 point in two functional systems of EDSS. To estimate the improvement or the worsening of permanent disability, the threshold for the EDSS score change is 1 point or more, except from the patients with baseline EDSS score 0 and >5.5, for whom a change of 1.5 and 0.5 has to be confirmed, respec-

tively. But the most important parameter for the confirmation of disability as (permanent) progression is the period of time. The vast majority of RCTs use the 3 or/and 6 months period of sustained EDSS score change to define disability progression or improvement. Is this time period sufficient to confirm any EDSS score change as irreversible? A study published in Brain on 2015 (20), investigated the events of EDSS score progression, as defined by the persistence of this score over 3, 6, 12 and 24 months, and calculated the proportion of events sustained over the following five years. The number of patients included in the study was 16.636, extracted from the international MSBASE registry. The proportion of events persisted over 5 years was 70%, 74%, 80% and 89%, for the 3, 6, 12, 24 months of confirmed disability progression (CDP), respectively. That is, the 3 months confirmation for the estimation of permanent disability progression is false in 30% of events, the 6 months estimation in 26% of events, the 12 months estimation in 20% of events and even the 24 months estimation is false in 11% of events. More commonly the restored progression confirmations were recorded in younger patients, those with relapsing-remitting course of MS, small changes in EDSS score and more frequent visits. The prominence of false permanent progression in these subpopulations of patients probably highlights the implication of measurement errors in low EDSS scores. This is in accordance with the finding of Ebers et al (21), who examined the placebo arms of 31 RCTs and concluded that the EDSS score as a surrogate marker of disability progression is totally unreliable in RRMS. Furthermore, a baseline EDSS score < 4, reflects measurement errors, random variations and remitting relapses (21). In contrast, baseline EDSS score > 4, which is usually the case of SPMS, is significantly more reliable and less noisy.

Taking in mind these findings, we may estimate that many disability progression events recorded in RCTs eventually proved to be relapses with delayed remittance. In addition, high potency DMTs (Natalizumab, Alemtuzumab, Ocrelizumab, Cladribine), which were thought to significantly affect the disability status (either higher rate of inhibition of disease worsening or higher rates of disability improvement), may simply act by their potent anti-inflammatory mechanism of action (remitting of relapses) and not against neurodegenerative pathophysiological mechanisms. Besides this disadvantage of EDSS, we have to mention the absence of a cognitive functional system assessment, as well as the absence of fatigue estimation. Both are important sources of disability of MS patients, affecting seriously their activities of daily living and subsequently their quality of life.

No Evidence of Disease Activity (NEDA)

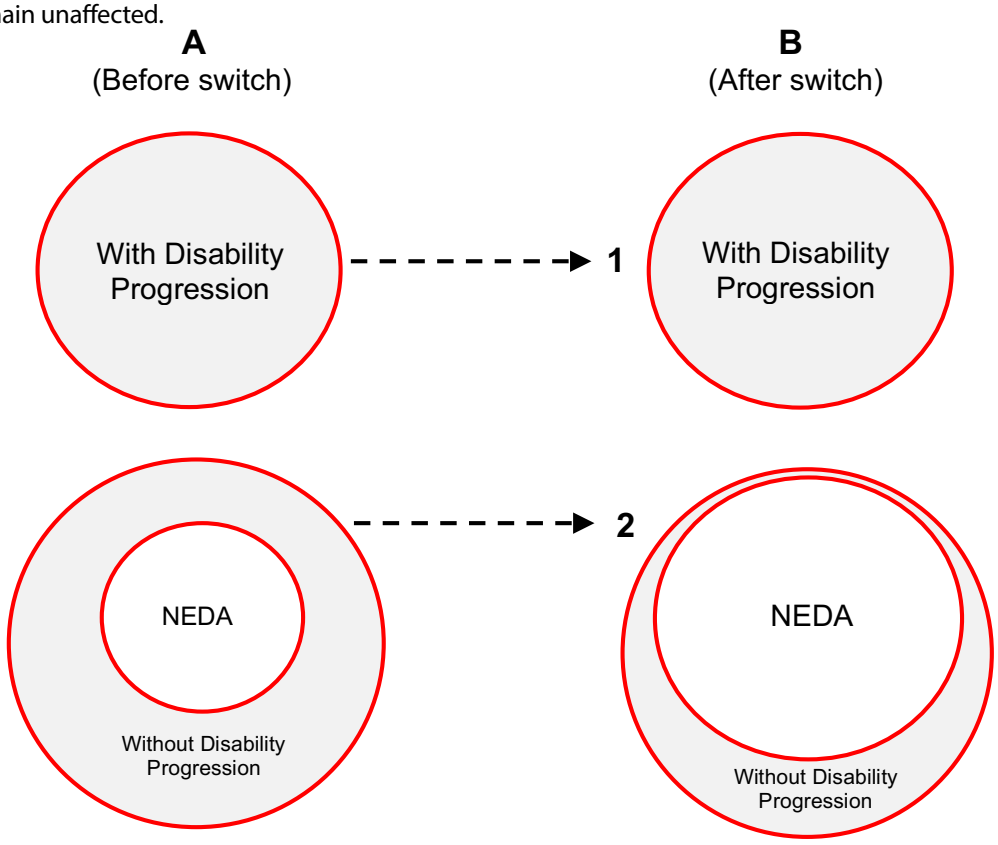
NEDA is a recently proposed treatment target for MS (22) and denotes the absence of any disease activity concerning relapses, disability progression and MRI lesion burden. A NEDA-4 treatment target has also been proposed (23), incorporating the brain volume loss as measured by MRI, as well as NEDA-5, adding to the NEDA-4 the light chain neurofilament levels in cerebrospinal fluid. Among all these proposed treatment targets only NEDA has been widely adopted by the MS community.

In a 7-years longitudinal cohort, 46% of the patients fulfilled NEDA at 1 year, while only 7.9% after 7 years (24). Another similar 10-years cohort found that only 9% of the patients fulfilled NEDA at the 10th year. (25) Thus, it is clear that NEDA could not be maintained in the long-run. It is an acceptable treatment goal but too ambitious at present time, under the current pharmacological armamentarium in MS. The extension of TRANSFORMS trial presents special interest, since it compares two DMTs using NEDA as an outcome measure, and may provide some insights. The TRANFORMS trial compared fingolimod with intramuscular interferon β-1a for one year. At the end of first year, all patients on interferon β-1a switched to fingolimod and followed up over the next 3.5 years (26). At the end of first year, the proportion of patients with NEDA was 44.3% for interferon β-1a and 63.4% for fingolimod. At the end of second year (one year after the switch to fingolimod), the interferon β-1a group showed a statistically significant increase in the proportion of patients with NEDA to 66% (i.e. 21.7% increase). On the contrary, for the fingolimod group, the continuation of the same drug resulted in an insignificant increase, to 69%. Hence, by relying on NEDA as an outcome measure, it seemed that there was a striking difference in the therapeutic

effect between the two DMTs, in favor of fingolimod. However, there is a point demanding special attention: the disability progression. Both criteria of disability confirmation, i.e. persistence over 3 or 6 months, showed statistically insignificant differences between groups throughout the whole period of study. More specifically, after the switch, the proportion of patients with 3 months CDP was 21% for interferon β-1a and 22% for fingolimod. Similarly, with 6 months CPD the proportions were 15% and 17% respectively (26). Therefore, there is a striking discrepancy between NEDA and disability progression. But the latter is included in the former and subsequently NEDA should have been affected by the proportion of patients with disability progression, resulting in a non-significant difference for NEDA as well. In order to clarify this point we propose an alternative interpretation of these findings of the study, as illustrated in figure 1. In short, after exposition to a DMT improving NEDA, the patients with the better course (without disability progression) of their disease continue to improve even more (as estimated by NEDA proportion), while those with the worse course remain unaffected.

Figure 1

Column A represents the patients in the group of interferon β-1a, before switching (at first year of study) and column B after switching (at the second year of study). The patients in column A were divided in two sets: 1. With disability progression, 2. Without disability progression. Group 1, in the first row, remained unchanged (i.e. the proportion of patients with disability progression was equal to fingolimod during first year. After switching, during the second year, the equality was sustained). Group 2, in the second row, includes a subset of patients with NEDA. This subset, during the first year of the study, was already significantly smaller in interferon β-1a group than fingolimod. This subset increased in favour of fingolimod after switching during the second year. That is, the patients with the better course of the disease from the beginning (without disability progression, group A, 2) improved even more.



P-value per se
The meaning of the p-value (the threshold 0.05).

The conceptual definition (without mathematical formulation) of the p-value may be the following. It is the value of probability of the data under examination, or even smaller values, provided the null hypothesis is true (27). This definition may be further divided in two conceptual steps:

1. The observed data may correspond to a probability value (p-value or less) in the tails of probability distribution, given the null hypothesis.

2. There are two hypotheses for testing, before any experimentation: the null hypothesis and the alternative hypothesis. After the calculation of type I (α) and type II (β) error, and using both of them, we define the critical region and accordingly accept/reject the null or alternative hypothesis (28).

The first step was introduced by Ronald A. Fisher and the second by Jersey Neyman and Egon Pearson. Among them, there was a conceptual and methodological-philosophical gap. Nevertheless, the concept of p-value used after them and up to this day is a hybrid of the two conceptual steps. None of the founders intended this interpretation of the p-value (28). According to Steven Goodman, the most pernicious misconception around the p-value is believing that a 0.05 value represents a 5% chance of the null hypothesis to be true (27). This misconception and the definition of the p-value in the beginning of the paragraph differ at the point that the definition considers as given that the null hypothesis is true and does not attribute any probability of being true to either the null or the alternative hypothesis. Nevertheless, the p-value deviated from its original meaning and is used in every day scientific research practice as a threshold of truth.

An extension of the misconception around the p-value is that it denotes the false positives of the null hypothesis. Nevertheless, a method employing Bayes theorem and the likelihood ratio (Bayes factor) has been proposed by Colquhoun (29) in order to calculate more precisely the false positive risk, which is the complementary of positive predictive value. He points out that if you get a p=0.05 from your analyzed data, then the probability of being wrong is at least 30%, and even higher if the study is underpowered (30). Colquhoun constructed a free access web page for the calculation of false positive risk, requiring the user to provide the sample sizes, the level of p-value, the prior probability and the standardized effect size: http://fpr-calc.ucl.ac.uk/.

Replication crisis

During the last twenty years, there has been a growing body of evidence questioning the validity of research findings (31), culminating in a 2005 publication by Ioannidis (32). Ioannidis mentioned that the increased number of false positives due to the use of the p-value threshold is an important factor contributing to the replication crisis. One corrective proposal concerning the p-value was signed by 72 renowned statisticians and epidemiologists (33). Its authors proposed lowering the p-value threshold to 0.005, which is 10 times lower than its current value. With the use of Bayesian statistics, it was shown that the number of false positives decreases down to 5% if the p-value threshold is set to 0.005. In addition, on March 20, 2019, the American Statistician journal dedicated a whole supplement to the subject: "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" (34). On the same day, a plea was published in Nature to "retire statistical significance" (35). The plea was signed by more than 800 statisticians and epidemiologists around the globe. This indicates a significant consensus on the role of the p-value threshold on the depreciation of the validity of research findings.

Significance tests are used in every scientific field and of course in RCTs. Hence a number of comparisons should be false positives, especially those with marginal significance. We would like to mention two large RCTs, for two different DMTs, that did not replicate their own previous results. The FREEDOMS trial of fingolimod showed a significant effect on the disability progression, 30% greater than placebo (HR=70%) (36). Nevertheless, the FREEDOMS II trial failed to reveal any significant effect on the disability progression in comparison to placebo (37). Exactly the same failure of duplicating the dis-

ability progression effect occurred between DEFINE (38) and CONFIRM (39) trials of dimethyfumarate (DMF). The disability progression effect of the twice daily DMF in DEFINE trial was 38% better than placebo, while in CONFIRM trial, the same two comparisons were equally effective. Do these discrepancies represent the noisy EDSS score mentioned above? Or rather the false positive results of the first trial of the couple of trials (FREEDOMS, DEFINE)? Or even the false negative of the second trial of the couples (FREEDOMS II, CONFIRM)? The questions could not be answered. The fact is the absence of replication.

Long-term DMT use and ethical issues

In general, the majority of RCTs in MS, are well designed, adequately powered, well conducted and of long enough duration (usually about 2 years).

The long-term extensions of RCTs in MS, concerning the first line injectables, have shown convincingly, that 15 years after randomization, there was a substantial decrease in the ac-cumulation of disability (EDSS≥4 or ≥6) for the patients taking the injectable DMT (subcutaneous interferon β-1a) consistently (high cumulative dose drug exposure), in comparison to those with intermittent use of DMT (low cumulative dose drug exposure) (40). The patients in the placebo arm were part of the last group of low dose drug exposure. In addition, the stron-gest predictor of the long-term disability accumulation was the EDSS change during the first two years after randomization (40). This change was equivalent to 30% benefit for the DMT group versus placebo (40). Besides that, 21 years after random-ization, a significant number of the patients assigned to sub-cutaneous interferon β-1b showed a considerable reduction in all-cause mortality. In terms of hazard rate, this reduction cor-responded to a 46.8%, in the proportion of deaths among DMT patients compared to placebo.

The majority of RCTs used a placebo arm for comparison to the novel therapy. According to the above-mentioned long-term disability accumulation and survival rates, every RCT designed with a placebo arm condemns the patients in this arm to long-term disability progression and decreased survival. Of course, this is a serious ethical issue.

This article is intended to be only a very short outline of several major methodological and statistical issues, concerning the design, conduct and analysis of RCTs, drawing examples from the field of MS. It aims only to highlight several points of interest in order to facilitate the critical reading from the viewpoint of the clinicians.

Disclosures: The author declares no conflicts of interest relevant to the present article.

References

1. Lassmann H. Pathogenic Mechanisms Associated With Different Clinical Courses of Multiple Sclerosis. Front. Immunol. 2019, 9:3116. Doi:10.3389/fimmu.2018.03116

2. Katz Sand I. Classification, diagnosis, and differential diagnosis of multiple sclerosis. Curr Opin Neurol. 2015, 28:193-205. Doi: 10.1097/WCO.0000000000000206

3. Lublin F. New Multiple Sclerosis Phenotypic Classification. Eur Neurol 2014;72(suppl 1):1-5. Doi:10.1159%2F000367614

4. European Medicines Agency. Ocrevus. Accessed March 25, 2019. https://www.ema.europa.eu/en/medicines/human/EPAR/ocrevus

5. Guyatt G, Rennie D, Meade MO, Cook DJ. Users' Guides to the Medical Literature. Essentials of Evidence Based Clinical Practice. McGraw Hill-JAMA Network. 3rd Edition, pages 31-33, 2015. ISBN: 978-0-07-180872-9.

6. Johnson K.P., Brooks B.R., Cohen J.A., Ford C.C., Goldstein J., Lisak R.P. et al. Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: Results of a phase III multicenter, double-blind, placebo-controlled trial. Neurology 1995;45:1268-1276. Doi: 10.1212/WNL.45.7.1268.

7. Goodin D.S. Disease-modifying therapy in MS: a critical review of the literature. Part I: Analysis of clinical trial errors. J. Neurol. 2004;251:Suppl 5:V/3-V/11. Doi: 10.1007/s00415-004-1503-z.

8. Food and Drug Administration. Copaxone. Accessed March 27, 2019. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2001/020622_S015_COPAXONE_INJECTION_AP.pdf

9. Freedman M.S, Hughes B, Mikol D.D, Bennett R, Cuffel B, Divan V, et al. Efficacy of Disease-Modifying Therapies in Relapsing Remitting Multiple Sclerosis: A Systematic Comparison. Eur Neurol 2008;60:1–11. Doi:10.1159/000127972.

10. Paul O'Connor, Massimo Filippi*, Barry Arnason, Giancarlo Comi, Stuart Cook, Douglas Goodin, Hans-Peter Hartung, Douglas Jeffery, Ludwig Kappos, Francis Boateng, Vitali Filippov, Maria Groth, Volker Knappertz, Christian Kraus, Rupert Sandbrink, Christoph Pohl, Timon Bogumil,for the BEYOND Study Group. 250 µg or 500 µg interferon beta-1b versus 20 mg glatiramer acetate in relapsing-remitting multiple sclerosis: a prospective, randomised, multicentre study. Lancet Neurol 2009; 8:889–97. Doi:10.1016/S1474-4422(09)70226-1.

11. Mikol D.D, Barkhof F, Chang P, Coyle P.K, Jeffery D.R, Schwid S.R, et al, on behalf of the REGARD study group. Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis. Lancet Neurol 2008; 7: 903–14. Doi:10.1016/S1474-4422(08)70200-X.

12. O'Connor P, Wolinsky J, Confavreux C,Comi G, Kappos L, Olsson T, for the TEMSO trial group. Randomized trial of oral teriflunomide for relapsing multiple sclerosis. N Engl J Med 2011;365:1293-303. Doi:10.1056/NEJMoa1014656.

13. Confavreux C, O'Connor P, Comi G, Freedman MS, Miller AE, Olsson TP, et al, for the TOWER Trial Group. Oral teriflunomide for patients with relapsing multiple sclerosis (TOWER): a randomised, double-blind, placebo controlled, phase 3 trial. Lancet Neurol 2014;13:247-256. Doi:10.1016/ S1474-4422(13)70308-9.

14. Vermersch P, Czlonkowska A, Grimaldi LME, Confavreux C, Comi G, Kappos L, et al, for the TENERE trial group. Teriflunomide versus subcutaneous interferon beta-1a in patients with relapsing multiple sclerosis: a randomised, controlled phase 3 trial. Mult Scler 2014;20:6:705-716. Doi: 10.1177/1352458513507821.

15. Dmitrienko A, D'Agostino RB. Multiplicity Considerations in Clinical Trials. N Engl J Med 2018;378:2115-22. Doi: 10.1056/NEJMra1709701.

16. Food and Drug Administration. Multiple endpoints in clinical trials. Guidance for industry. January 2017. Accessed March 30, 2019. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf

17. European Medicines Agency. Guideline on multiplicity issues in clinical trials. 15 Dec 2016. Accessed March 30, 2019. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf

18. Montalban X, Hauser SL, Kappos L, Arnold DL, Bar-Or A, Comi G, et al. Ocrelizumab versus placebo in progressive multiple sclerosis. N Engl J Med 2017;376:3:209-220. Doi:10.1056/NEJMoa1606468.

19. Fox EJ, Markowitz C, Applebee A, Montalban X, Wolinsky JS, Belachew S, et al. Ocrelizumab reduces progression of upper extremity impairment in patients with primary progressive multiple sclerosis: Findings from the phase III randomized ORATORIO trial. Mult Scler. 2018;24(14):1862-1870. doi: 10.1177/1352458518808189.

20. Defining reliable disability outcomes in multiple sclerosis. Brain 2015;138:3287–3298. Doi:10.1093/brain/awv258.

21. Ebers CG, Heigenhauser L, Daumer M, Noseworthy JH. Disability as an outcome in MS clinical trials. Neurology 2008;71:624-631. Doi: 10.1212/01.wnl.0000313034.46883.

22. Banwell B, Giovannoni G, Hawkes C, Lublin F. Editors' welcome and a working definition for a multiple sclerosis cure. Mult. Scler. Relat. Disord. 2013;2:65-67. Doi:10.1016/j.msard.2017.07.011.

23. Kappos L, Radu EW, Freedman MS, Cree BA, Radue EW, Sprenger T, et al. Inclusion of brain volume loss in a revised measure of 'no evidence of disease activity' (NEDA-4) in relapsing-remitting multiple sclerosis. Mult. Scler. 2016;22:10:1297-1305. Doi:10.1177/1352458515616701.

24. Rotstein DL, Healy BC, Malik MT, Chitnis T, Weiner HL. Evaluation of no evidence of disease activity in a 7-year longitudinal multiple sclerosis cohort. JAMA Neurol. 2015;72:152–158. Doi:10.1001/jamaneurol.2014.3537.

25. De Stefano N, Stromillo ML, Giorgio A, Battaglini M, Bartolozzi ML, Amato MP, et al. Long-term assessment of no evidence of disease activity in relapsing-remitting MS. Neurology 2015;85:19:1722-1723. Doi:10.1212/WNL.0000000000002105

26. Cohen JA, Khatri B, Barkhof F, Comi G, Hartung HP, Montalban X, et al. Long-term (up to 4.5 years) treatment with fingolimod in multiple sclerosis: results from the extension of the randomised TRANSFORMS study. J Neurol Neurosurg Psychiatry 2016;87:5:468-475. Doi:10.1136/jnnp-2015-310597.

27. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin. Hematol. 2008;45:135-140. Doi:10.1053/j.seminhematol.2008.04.003.

28. Goodman S. Toward evidence-based medical statistics. 1: the p-value fallacy. Ann. Int. Med. 1999;130:995-1004. Doi:10.7326/0003-4819-130-12-199906150-00008.

29. Colquhoun D. The false positive risk: a proposal concerning what to do with p-values. The American Statistician, 2019;73:sup1:192-201, Doi:10.1080/00031305.2018.1529622.

30. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R. Soc Open Sci, 2014;1:3:140216. Doi:10.1098/rsos.140216

31. Loken E, Andrew Gelman A. Measurement error and the replication crisis. Science 2017;355:6325:584-585. Doi:10.1126/science.aal3618.

32. Ioannidis J. Why Most Published Research Findings Are False. PLoS Med 2005;2:8 e124. Doi:10.1371/journal.pmed.0020124.

33. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nature Human Behavior 2017;2:6-10. Doi: 10.1038/s41562-017-0189-z.

34. Statistical Inference in the 21st Century: A World Beyond p < 0.05. American Statistician 2019;73:1-401. https://www.tandfonline.com/toc/utas20/73/sup1?nav=tocList

35. Amrhein V, Greenland S, McShane and more than 800 signatories. Retire statistical significane. Nature 2019; 567:305-307. Doi:10.1038/d41586-019-00857-9.

36. Kappos L, Radue EW, O'Connor P, Polman C, Hohlfeld R, Calabresi P, et al.

A Placebo-Controlled Trial of Oral Fingolimod in Relapsing Multiple Sclerosis. N Engl J Med 2010; 362:387-401. Doi:10.1056/NEJMoa0909494.

37. Calabresi A, Radue EW, Goodin D, Jeffery D, Rammohan KW, Reder AT, et al. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. Lancet Neurol 2014; 13: 545–56. Doi:10.1016/S1474-4422(14)70049-3.

38. Gold R, Kappos L, Arnold DL, Bar-Or A, Giovannoni G, Selmaj K, et al. for the DEFINE investigators. Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis. N Engl J Med 2012;367:1098-107. Doi:10.1056/NEJMoa1114287.

39. Fox RJ, Miller DH, Phillips TJ, Hutchinson M, Havrdova E, Kita M, et al. for the CONFIRM investigators. Placebo-Controlled Phase 3 Study of Oral BG-12 or Glatiramer in Multiple Sclerosis. J Neurol Neurosurg Psychiatry 2012; 367:1087-1097. Doi:10.1056/NEJMoa1206328.

40. Kappos L, Kuhle J, Multanen J, Kremenchutzky M, Verdun di Cantogno E, Cornelisse P, et al. Factors influencing long-term outcomes in relapsing–remitting multiple sclerosis: PRISMS-15. J Neurol Neurosurg Psychiatry 2015;86:1202-1207. Doi:10.1136/jnnp-2014-310024.

41. Goodin DS, Reder AT, Ebers GC, Cutter G, Kremenchutzky M, Oger J, et al. Survival in MS. A randomized cohort study 21 years after the start of the pivotal IFNβ-1b trial. Neurology 2012;78:1315–1322. Doi:10.1212/WNL.0b013e3182535cf6.