UNIVERSITY OF THESSALY
SCHOOL OF HEALTH SCIENCES
DEPARTMENT OF MEDICINE
LABORATORY OF BIOMATHEMATICS

MASTER PROGRAM IN

RESEARCH METHODOLOGY IN BIOMEDICINE, BIOSTATISTICS AND CLINICAL BIOINFORMATICS

MASTER THESIS

ASSESS THE REPORTING QUALITY OF STUDIES INVESTIGATING THE DIAGNOSTIC ACCURACY OF NEUROFILAMENT LIGHT CHAIN SERUM LEVELS IN THE DIAGNOSIS OF MULTIPLE SCLEROSIS PUBLISHED FROM 2000 TO 2019 USING THE STARD STATEMENT

ARETI ZORMPA

EVALUATION COMMITTEE:

STEFANIDIS IOANNIS, PROFESSOR, SUPERVISOR

DOXANI CHRYSOULA, RESEARCH FELLOW

ZINTZARAS ELIAS, PROFESSOR

LARISSA, 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΪΑΤΡΙΚΗ ΕΡΕΥΝΑΣ, ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΞΙΟΛΟΓΗΣΤΕ ΤΗΝ ΠΟΙΟΤΗΤΑ ΑΝΑΦΟΡΑΣ ΤΩΝ ΜΕΛΕΤΩΝ ΔΙΑΓΝΩΣΤΙΚΗΣ ΑΚΡΙΒΕΙΑΣ ΠΟΥ ΔΙΕΡΕΥΝΟΥΝ ΤΑ ΕΠΙΠΕΔΑ ΤΗΣ ΕΛΑΦΡΑΣ ΑΛΥΣΟΥ ΤΩΝ ΝΕΥΡΟΪΝΙΔΙΩΝ ΣΤΟΝ ΟΡΟ ΓΙΑ ΤΗ ΔΙΑΓΝΩΣΗ ΤΗΣ ΠΟΛΛΑΠΛΗΣ ΣΚΛΗΡΥΝΣΗΣ ΚΑΙ ΔΗΜΟΣΙΕΥΘΗΚΑΝ ΑΠΟ ΤΟ 2000 ΕΩΣ ΤΟ 2019, ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗ ΔΗΛΩΣΗ STARD.

ΤΗΣ ΑΡΕΤΗΣ Ε. ΖΟΡΜΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ:
ΣΤΕΦΑΝΙΔΗΣ ΙΩΑΝΝΗΣ, ΚΑΘΗΓΗΤΗΣ, ΕΠΙΒΛΕΠΩΝ
ΔΟΞΑΝΗ ΧΡΥΣΟΥΛΑ, ΕΠΙΣΤΗΜΟΝΙΚΟΣ ΣΥΝΕΡΓΑΤΗΣ
ΖΙΝΤΖΑΡΑΣ ΗΛΙΑΣ, ΚΑΘΗΓΗΤΗΣ

ΛΑΡΙΣΑ, 2019

CONTENTS

ΠΕΡΙΛΗΨΗ

Εισαγωγή: Η διάγνωση της πολλαπλής σκλήρυνσης (ΠΣ) βασίζεται στα κριτήρια Mc Donald. Ωστόσο, η εισαγωγή βιοδεικτών, όπως η ελαφρά άλυσος των νευροϊνιδίων (NfL), μπορούν να αναβαθμίσουν τα κριτήρια.

Στόχοι: Σκοπός της μελέτης είναι να αξιολογήσει την ποιότητα αναφοράς των μελετών που διερευνούν την διαγνωστική αξία του NfL στον ορό σε ασθενείς με ΠΣ χρησιμοποιώντας τις STARD κατευθυντήριες οδηγίες.

Μέθοδοι: Χρησιμοποιήθηκε η βάση δεδομένων PubMed. Τα άρθρα και οι περιλήψεις τους αξιολογήθηκαν αν ανταποκρίνονται στις STARD οδηγίες. Διερευνήθηκαν η συνολική ποιότητα αναφοράς και οι διαφορές μεταξύ των μελετών υψηλής και χαμηλής ποιότητας. Επίσης, διερευνήθηκε η συσχέτιση της ποιότητας αναφοράς με τον impact factor, το έτος δημοσίευσης και της υποστήριξης των οδηγιών από τα περιοδικά.

Αποτελέσματα: Εκτιμήθηκαν 24 μελέτες. Η συνολική ποιότητα αναφοράς ήταν μέτρια για τα άρθρα και τις περιλήψεις και τα ερευνητικά ερωτήματα απαντήθηκαν με εξαιρετικά μεγάλη ετερογένεια (0-100%). Η ποιότητα αναφοράς μεταξύ των ομάδων υψηλής ή χαμηλής ποιότητας ήταν στατιστικά σημαντικά διαφορετική (p <0,05), αλλά δεν σχετίζονταν με τον impact factor (p = 0,090), το έτος δημοσίευσης (p = 0,236) και την υποστήριξη των οδηγιών από τα περιοδικά (p = 0,360).

Συμπεράσματα: Υπάρχει ανάγκη βελτίωσης στην αναφορά των μελετών διαγνωστικής ακρίβειας ώστε να διευκολυνθεί η ιατρική έρευνα.

ABSTRACT

Background: The diagnosis of multiple sclerosis (MS) is based on Mc Donald criteria. Nevertheless, the introduction of biomarkers could upgrade these criteria. Neurofilament light protein (NfL), a degenerative biomarker, have diagnostic value in MS.

Objective: The aim of this study was to evaluate the reporting quality of diagnostic accuracy studies investigating NfL in serum in patients with MS using the STARD statement.

Methods: The research was conducted in PubMed Database. The studies and their abstracts were evaluated for their adherence to STARD statement. The overall reporting quality and the differences between high and low quality studies were explored. Also, the effect of adherence to impact factor, publication year and STARD endorsement were investigated.

Results: 24 studies were evaluated. The overall quality of reporting was moderate for articles and abstracts, with a large variability in adherence across investigating items (0 - 100%). The quality of reporting in high versus low quality articles/ abstracts was statistically significant different (p<0.05), but didn't relate to impact factor (p=0.090), publication year (p=0.236) or to STARD endorsement (p=0.360).

Conclusions: The completeness of reporting in diagnostic accuracy studies still has a long way to go in order to facilitate medical research.

INTRODUCTION

Multiple Sclerosis (MS) is a chronic autoimmune disease of the Central Nervous System (CNS) with a variety of neurological symptoms that affect young and middle-aged people. It constitutes an important morbidity factor because it results in chronicity but mostly in disability. MS is considered as a "disease with many faces". Four main types are recognized, Clinicaly isolated syndrome (CIS), Relapsing Remmiting MS (RRMS), Primarly Progressive MS (PPMS) and Secondary Progressive MS (SPMS), which differ in their stages or progression. Nowadays, updated Mc Donald criteria consider a reliable method for the diagnosis of the disease[1]. Nevertheless, the need for further research to refine the criteria includes the introduction of body fluid markers.

Neurofilaments are cytoskeletal proteins of neurons that are significantly plentiful in axons. Their role lies to provide structural support and maintenance of size, shape, and caliber of the axons [2]. They constitute of three parts that differ in molecular size: a light chain, an intermediate chain, and a heavy chain. After axonal damage in the CNS, neurofilament proteins discharge into cerebrospinal fluid (CSF) and offer a sign of axonal damage and neuronal death[3]. The scientific interest is above neurofilament research and neurofilament levels are under investigation as markers of disease activity and progression in a variety of different neurological conditions, like MS. The last years several studies confirm that the concentration of Neurofilament light (NfL) is increased in Cerebrospinal Fluid (CSF) in patients with MS [4–6] and that serum neurofilament light (sNfL) chain levels closely reflect the concentration of CSF NfL in MS patients[7–10]. The fact that lumbar puncture is a relatively invasive procedure limits the value of CSF NfL in routine clinical practice and makes sNfL a more appealing approach. Findings that further support the significance of sNfL levels as a biomarker of tissue damage in MS are the following: sNfL levels appear elevated in MS patients compared to healthy controls or in patients who experienced recent relapses, sNfL levels are positively associated with magnetic resonance imaging (MRI) or disability scores (EDSS) and are lower when disease-modifying therapies (DMTs) last longer[7,9–11].

When searching for studies concerning the diagnostic accuracy of sNfL levels in MS on databases such as PubMed, the reader comes across with abundant articles. In order to evaluate these studies “Standards for Reporting of Diagnostic Accuracy ” (STARD) statement was formed and published originally in 2003 and updated in 2015 [12]. The objective of the STARD initiative is to enhance the completeness and transparency of reporting of the studies regarding diagnostic accuracy, to help readers to assess the potential for bias within the study (internal validity) and to judge its generalisability (external validity). It consists of a 30 items checklist that covers all the article’s sessions (abstract, introduction, methods, results, discussion and other information). Specially, for evaluating the reporting quality for abstracts “STARD for abstracts”, an 11 items checklist, was proposed.

The aim of this study was to evaluate the reporting quality of studies investigating the diagnostic accuracy of neurofilament light chain serum levels in the diagnosis of multiple sclerosis published from 2000 to 2019 using the STARD statement.

METHODS

Data Sources, Search Strategies and Studies Selection

PubMed was searched for clinical studies, published from 2000 to July 31, 2019. The search used the following strategy: from advanced search we typed (((neurofilament light OR NFL)) AND (serum OR blood)) AND (multiple sclerosis OR MS) and we filtered the results by putting “English” in Languages.

We read the abstracts and /or full articles to recognize the eligible studies. Inclusion criteria was: measurement of NFL levels, in serum, in patients with MS. Exclusion criteria were: reviews, irrelevant to the topic articles, measurement of NFL levels only in CSF, articles that evaluate NfL antibodies, studies on animals, meta-analysis and scientific commentaries.

Data Extraction and Reporting Assessment Tool

As assessment tool for quality of reporting, we used the updated STARD 2015 checklist, which includes a 30-item questionnaire (http://www.equator-network.org/reporting-guidelines/stard/). The evaluation of the reporting quality of abstracts of diagnostic accuracy studies was based on the STARD for abstracts, an 11-item questionnaire (http://www.equator-network.org/reporting-guidelines/stard-abstracts/). In order to clarify whether an item is accurately reported in the articles or abstracts, we took into account the guidance provided by the STRARD Explanation and Elaboration document [13,14]. All items were investigated in terms of whether they were reported, not whether they were actually carried out during the study. Items were scored as ‘‘yes’’ if they were reported in enough detail to allow the reader to judge that the definition had been met. Alternative responses (‘‘no’’) and unclear responses to each question were coded as negative responses.

Additional data

In order to find out which journals endorse STARD statement, we checked the section "guidelines for authors" in each journal. Also, the journal’s impact factor the year that the articles were published was recorded. Moreover, data as the origin of study’s population, studies setting and which assays were used, also recorded.

Data analysis

Studies that included more than one independent cohort were regarded as different studies. The overall percentages of reported STARD statement items in both questionnaires were explored. Also, the quality scores were estimated using the following strategy. All items in each STARD checklist were considered equally important and the quality score was calculated by summing the score of the reported items. The items were scored as 1 when ‘‘yes’’ was the answer and 0 when ‘‘no’’ or ‘‘unclear’’ were the answers. The second item of the questionnaire: "*Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)*" was excluded, because for abstracts we applied the questionnaire "STARD for Abstracts". Studies were classified as high quality of reporting when quality score was > 17 and as lower quality when quality score was ≤ 17. The choice of quality score = 17 as cut-off was decided because it was the median of the overall quality scores of studies. Abstracts were classified as higher or lower quality of reporting by putting 5 as a cut-off point, with the same thinking. Then, the quality of reporting in high quality articles/abstracts versus lower quality articles/abstracts was compared using chi square test. Furthermore, the proportion of articles that was published in high-ranked or lower rank journals was estimated. To do this, we divided the studies into two teams depending if the impact factor (IF) was lower than (<) or equal to /greater than (≥) 6. The choice of IF = 6 as cut-off was made because the top 5% of journals have impact factors approximately ≥ 6[15]. A univariant general linear model was applied to examine the relationship between total score and impact factor, and also examined the effect of publication year in this relationship, considering publication year as a bivalent variable (2015-2017, 2018-2019). The choice of these two categories was made because 2018 was the median for publication year. When we examined the relationship between total score and STARD endorsement, we considered both variables as bivalent variables (lower or higher quality articles and yes or no, respectively) and conducted a chi square test. Microsoft Excel 2007 and SPSS software version 25 were used to analyze the data and p values below 0.05 were considered significant.

RESULTS

Eligible studies

The literature review identified 91 articles that met the search criteria in PubMed. Afterwards, these articles were retrieved and screened for eligibility and 22 articles remained. Two articles that included two independent cohorts each were regarded as different studies, reaching the final number of eligible studies to 24 (Table 1). Figure 1 presents a flow diagram of retrieved articles and articles excluded with specification of reasons.

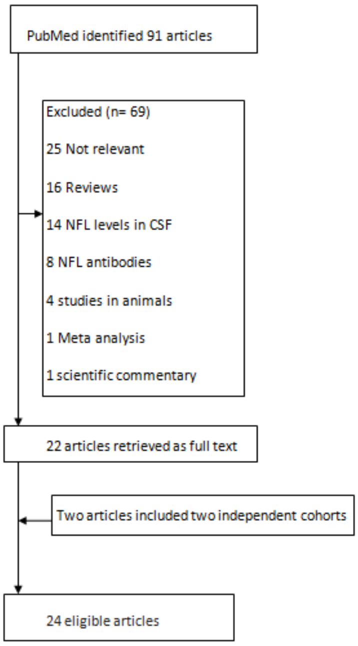


Figure 1: Flow diagram of citations through the retrieval and screening process

Study characteristics

The characteristics of studies included in the analysis are shown in Table 1

Table 1: The characteristics of studies included in the analysis.

Study	Year	Journal	Impact Factor	STARD endorsement	Population	Setting	Assay for index test
Disanto et al[16]	2015	J Neurol Neurosurg Psychiatry	6.431	yes*	17 different countries	Multicenter	ELC immunoassay
Kuhle et al[17]	2016	Mult Scler	4.840	yes	Switzerland	Neurology of Lausanne University Hospital	ELC immunoassay
Bergman et al[18]	2016	Neurology	7.592	yes	Sweden	Not mentioned	simoa
Al-Temaimi et al[19]	2017	Exp Mol Pathol	2.566	no	Kuwaiti	Dasman Diabetes Institute's MS clinic	ELISA
Novakova et al[10]	2017	Neurology	7.609	yes	sweden	4 Swedish University hospitals	simoa
Disanto et al[9]	2017	Ann Neurol	10.244	yes*	Switzerland	Neurocenter of Southern Switzerland	simoa
Disanto et al[9]	2017	Ann Neurol	10.244	yes*	Switzerland	Neurologic Clinic and Policlinic, University Hospital Basel	simoa
Kuhle et al[11]	2017	Neurology	7.609	yes	Not mentioned	Not mentioned	ELC immunoassay
Varhaug et al[20]	2017	Neurology: Neuroimmunology & Neuroinflammation	7.353	yes	Not mentioned	multicenter	simoa
Piehl et al[21]	2018	Multiple Sclerosis Journal	5.649	yes	sweden	Department of Neurology at Karolinska University Hospital	simoa
Piehl et al[21]	2018	Multiple Sclerosis Journal	5.649	yes	sweden	Not mentioned	simoa
Håkansson et al[22]	2018	J Neuroinflammation	5.193	yes	sweden	Department of Neurology, University Hospital of Linköping, Sweden	simoa
Barro et al[23]	2018	Brain	11.814	yes	Switzerland	Neurologic Clinic and Policlinic, University Hospital Basel	simoa
Chitnis et al[24]	2018	Annals of Clinical and Translational Neurology	4.649	yes	Massachusetts	Brigham and Women's Hospital, Boston, Massachusetts	simoa
Browne et al[25]	2019	Journal of Clinical Lipidology	3.581	no	New York	MS Center of the State University of New York at Buffalo	simoa
Sehr et al[26]	2019	Journal of Molecular Medicine	3.340	no	Germany	MS centre Dresden, Germany	simoa

Siller et al[27]	2017	Multiple Sclerosis Journal	5.649	yes	Not mentioned	Not mentioned	simoa
Akgun et al[28]	2019	Neurology: Neuroimmunology & Neuroinflammation	7.353	yes	Germany	Department of Neurology University Hospital, Dresden, Germany;	simoa
Cuello et al[29]	2019	Eur J Neurol.	4.621	no	Spain	Hospital General Universitario, Madrid, Spain	simoa
Abdelhak et al[30]	2019	Frontiers in Neurology	3.508	yes*	Germany	4 University Hospitals in Germany	simoa
Hyun et al[31]	2019	Multiple Sclerosis Journal	5.649	yes	Korea	National Cancer Centre in Korea	simoa
Dalla Costa et al[32]	2019	Neurology	8.689	yes	Italy	Department of Neurology, San Raffaele Hospital, Milan, Italy	ELC immunoassay
Ferraro et al[33]	2019	Acta Neurol Scand	3.126	yes	Not mentioned	Not mentioned	simoa
Kuhle et al[34]	2019	Neurology	8.689	yes	Not mentioned	Not mentioned	simoa

yes*: journals that supported other reporting guidelines, such as for clinical trials (CONSORT) and systematic reviews (PRISMA)

The eligible articles were published during the period 2015–2019. Consequently, all the eligible articles were published after the introduction of STARD statement (i.e. 2003). Eighteen out of twenty four studies (75%) were published in journals that endorse STARD statement or 71.4% of the included journals endorse STARD statement (10 out of 14 journals). Most of the participants derived from European countries, (58.4 %) and afterwards from United States of America (8.3%) and Asia (8.3%). In six articles the nationality of the population isn’t mentioned (25%). Most of the articles refer to studies conducted in university hospitals (9 articles, 37.5%). Second in place comes MS Centers (12%), although in 6 articles (25%) there is no information regarding studies’ setting. In 20 out of 24 studies, the measurement of sNFL conducted with a single-molecule array (Simoa) (83.3%), in 4 out of 24 with electrochemiluminescence (ELC) immunoassay (16.6%) and in 1 with ELISA (4.2%). The lower limit quantification of the index test was 4 times higher in ELC immunoassay compared with simoa technique. Eleven articles (45.8%) were published in high quality articles (STARD score > 17) and 13 articles (54.2%) in lower quality articles (STARD score ≤ 17) (Table 2). Also, 12 abstracts (50.0%) were published in high quality abstracts (STARD score > 5) and 12 abstracts (50.0%) in lower quality abstracts (STARD score ≤ 5) (Table 3). Moreover, 11 articles (45.8%) were published in high-ranked journals (impact factor [IF] ≥ 6) and 13 articles (54.2%) in journals with lower rank (IF < 6).

Adherence of Articles to STARD Statement

The adherence of the 24 studies to STARD statement, in total, in lower and in higher quality articles along with the p-value derived from the comparison between higher and lower quality articles is shown in Table 2.

Table 2: Proportion of reporting the items of STARD statement for the three groups (all studies, lower quality and in higher quality articles)

Section & Topic	No	Item	Overall % of reporting item n = 24	% of reporting item in lower quality articles (score ≤18) n = 13	% of reporting item in higher quality articles (score > 18) n = 11	P-value
TITLE OR ABSTRACT	**1**	**Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)**	12.5	0.0	27.3	**0.044**
ABSTRACT	**2**	**Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)**	50.8 (see STARD for abstracts)	41.7 (see STARD for abstracts)	59.8 (see STARD for abstracts)	**0.003** (see STARD for abstracts)
INTRODUCTION	**3**	**Scientific and clinical background, including the intended use and clinical role of the index test**	95.8	92.3	100.0	0.347
	4	**Study objectives and hypotheses**	100.0	100.0	100.0	1.000
METHODS ***Study design***	**5**	**Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)**	50.0	46.2	54.5	0.681
	6	**Eligibility criteria**	87.5	100.0	72.7	**0.044**
Participants	**7**	**On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)**	54.2	53.8	54.5	0.974
	8	**Where and when potentially eligible participants were identified (setting, location and dates)**	75.0	61.5	90.9	0.097
	9	**Whether participants formed a consecutive, random or convenience series**	62.5	53.8	72.7	0.341
Test methods	**10a**	**Index test, in sufficient detail to allow replication**	100.0	100.0	100.0	1.000
	10b	**Reference standard, in sufficient detail to allow replication**	75.0	76.9	72.7	0.813
	11	**Rationale for choosing the reference standard (if alternatives exist)**	54.2	53.8	54.5	0.974
	12a	**Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory**	33.3	15.4	54.5	**0.042**
	12b	**Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory**	29.2	7.7	54.5	**0.011**
	13a	**Whether clinical information and reference standard results were available**	25.0	7.7	45.5	**0.033**

		to the performers/readers of the index test				
	13b	**Whether clinical information and index test results were available to the assessors of the reference standard**	12.5	7.7	18.2	0.438
Analysis	**14**	**Methods for estimating or comparing measures of diagnostic accuracy**	100.0	100.0	100.0	1.000
	15	**How indeterminate index test or reference standard results were handled**	12.5	15,4	9.1	0.642
	16	**How missing data on the index test and reference standard were handled**	25.0	23.1	27.3	0.813
	17	**Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory**	79.2	69.2	90.9	0.192
	18	**Intended sample size and how it was determined**	0.0	0.0	0.0	1.000
RESULTS ***Participants***	**19**	**Flow of participants, using a diagram**	12.5	15.4	9.1	0.642
	20	**Baseline demographic and clinical characteristics of participants**	100.0	100.0	100.0	1.000
	21a	**Distribution of severity of disease in those with the target condition**	83.3	69.2	100.0	**0.043**
	21b	**Distribution ofalternative diagnoses in those without the target condition**	50.0	38.5	63.6	0.219
	22	**Time interval and any clinical interventions between index test and reference standard**	62.5	46.2	81.8	0.072
Test results	**23**	**Cross tabulation of the index test results (or their distribution) by the results of the reference standard**	20.8	0.0	45.5	**0.006**
	24	**Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)**	41.7	15.4	72.7	**0.004**
	25	**Any adverse events from performing the index test or the reference standard**	0.0	0.0	0.0	1.000
DISCUSSION	**26**	**Study limitations, including sources of potential bias, statistical uncertainty, and generalisability**	70.8	61.5	81.8	0.276
	27	**Implications for practice, including the intended use and clinical role of the index test**	100.0	100.0	100.0	1.000
OTHER INFORMATION	**28**	**Registration number and name of registry**	20.8	23.1	18.2	0.768
	29	**Where the full study protocol can be accessed**	4.2	7.7	0.0	0.347
	30	**Sources of funding and other support; role of funders**	95.8	92.3	100.0	0.347
Total adherence to STARD checklist			49.7	47.1	59.8	**<0.001**

Overall adherence is 49.7% for the 24 studies, 47.1% for the lower quality articles and 59.8% for the higher quality articles, rating across 33 items. A large variability in reporting STARD items is detected, ranging from 0 to 100% in all three groups.

In the group which include all the studies, nine items are adequately reported in more than 80% of the studies (items 3, 4, 6, 10a, 14, 20, 21a, 27, 30) and five of them were reported in every study (items 4, 10a, 14, 20, 27). Six items are reported in 60-80% of the studies (item 8, 9, 10b, 11, 22, 26), five items in 40-60% of the studies (items 5,7, 11, 21b, 24) and six items in 20-40% of the studies (items 12a, 12b, 13a, 16, 23, 28). Seven items are reported in less than 20% of the studies (items 1, 13b, 18, 19, 25, 29) and two of them aren't reported at all in any of the included studies (items 18, 25).

In the group of the lower quality articles, eight items are adequately reported in more than 80% of the articles (items 3, 4, 6,10a, 14, 20, 27, 30) and six of them are reported in every article (items 4, 10a, 14, 20, 27). Five items are reported in 60-80% of the studies (items 8, 10b, 17, 21a, 26), five items in 40-60% of the articles (items 5, 7, 9, 11, 22) and three items in 20-40% of the articles (items 16, 21b, 28). Twelve items are reported in less than 20% of the lower quality articles (items 1, 12a, 12b, 13a, 13b, 15, 18, 19, 23, 24, 25, 29) and one third of them aren't reported at all in any of the included articles (items 1, 18, 23, 25).

In the group of the higher quality articles, twelve items are adequately reported in more than 80% of the articles (items 3, 4, 8, 10a, 14, 17, 20, 21a,22, 26, 27, 30) and eight of them are reported in every article (items 3, 4, 10a, 14, 20, 21a, 27, 30). Five items are reported in 60-80% of the articles (items 6, 9, 10b, 21b, 24), seven items in 40-60% of the articles (items 5, 7, 11, 12a, 12b, 13a, 23) and two items in 20-40% of the articles (items 1, 16). Seven items are reported in less than 20% of the articles (items 13b, 15, 18, 19, 25, 28, 29) and five of them aren't reported at all in any of the included articles (items 18, 19, 29). The bar chart below shows how many items present 0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100% adherence to STARD statement, for all three groups (Figure 2).

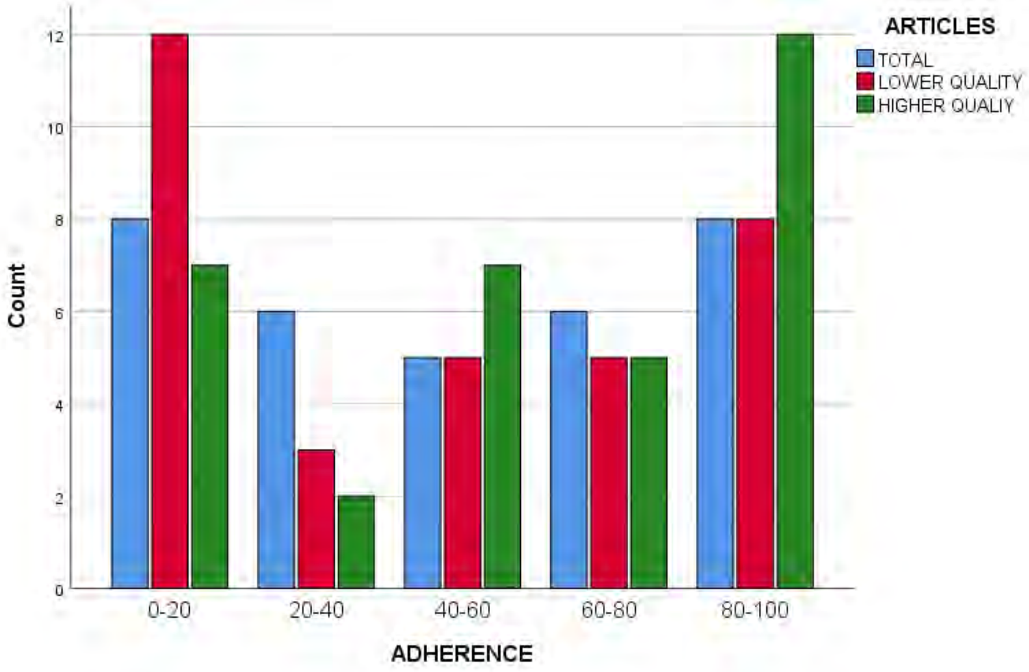


Figure 2: Number of items that present 0-20%, 20-40%, 40-60%, 60-80%, 80-100% adherence to STARD statement for the three groups

When we compared the two groups (lower and higher quality articles) for every item and for the total adherence to STARD statement it came up that for seven items (items 1, 6, 12a, 12b, 21a, 23, 24) along with the total adherence to STARD statement the p value was <0.05, meaning that there is statistically significant difference between the two groups for these items. In all these items, except item 6, high quality articles showed better performance.

Adherence of Abstracts to STARD Statement

The adherence of the 24 studies to STARD statement for abstracts, in total, in lower and in higher quality abstracts along with the p-value derived from the comparison between higher and lower quality articles, is shown in Table 3.

Table 3: Proportion of reporting the items of STARD statement for abstracts, for the three groups (all studies, lower quality and in higher quality articles)

Section & Topic	No	Item	Overall % of reporting item n = 24	% of reporting item in lower quality articles (score ≤ 5) n=12	% of reporting item in higher quality articles ((score > 5) n=12	p-value
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	12.5	0.0	25.0	**0.052**
Background and Objectives	2	Study objectives	100.0	100.0	100.0	1.000
Methods	3	Data collection: whether this was a prospective or retrospective study	12.5	0.0	25.0	0.064
	4	Eligibility criteria for participants and settings where the data were collected	12.5	0.0	25.0	0.064
	5	Whether participants formed a consecutive, random, or convenience series	45.8	25.0	66.7	**0.040**
	6	Description of the index test and reference standard	70.8	66.7	75.0	0.653
Results	7	Number of participants with and without the target condition included in the analysis	95.8	91.7	100.0	0.307
	8	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	12.5	0.0	25.0	0.064
Discussion	9	General interpretation of the results	95.8	91.7	100.0	0.307
	10	Implications for practice, including the intended use of the index test	87.5	75.0	100.0	0.064
Registration	11	Registration number and name of registry	12.5	8.3	16.7	0.537

Total adherence to STARD checklist	50.8	41.7	59.8	**0.003**

Overall adherence was 50.80% for the 24 abstracts, 41.7% for the lower quality abstracts, 59.8% for the higher quality abstracts, rating across 11 items. A large variability in reporting STARD items was detected, ranging from 0 to 100% in all three groups.

In the group which include all the abstracts, four items were adequately reported in more than 80% of the abstracts (items 2, 7, 9, 10) and one of them were reported in every abstract (item 2). One item was reported in 60-80% of the abstracts (item 6) and one item in 40-60% of the abstracts (item 5). Five items were reported in less than 20% of the abstracts (items 1, 3, 4, 8, 11).

In the group of the lower quality abstracts, four items were adequately reported in more than 80% of the articles (items 2, 7, 9, 10) and one of them is reported in every abstract (item 2). One item was reported in 60-80% of the abstracts (item 6) and one item in 20-40% of the abstracts (item 5). Five items were reported in less than 20% of the abstracts (items 1, 3, 4, 8, 11) and four of them weren’t reported at all in any of the included abstracts (items 1, 3, 4, 8).

In the group of the higher quality abstracts, four items were adequately reported in more than 80% of the abstracts (items 2, 7, 9, 10). It’s notable that all of them were reported in every abstract. Two items were reported in 60-80% of the abstracts (items 5 and 6), and four items in 20-40% of the abstracts (items 1, 3, 4, 8). One item was reported in less than 20% of the abstracts (item 11). The bar chart below shows how many items present 0%-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100% adherence to STARD for abstracts, for all three groups (Figure 3).

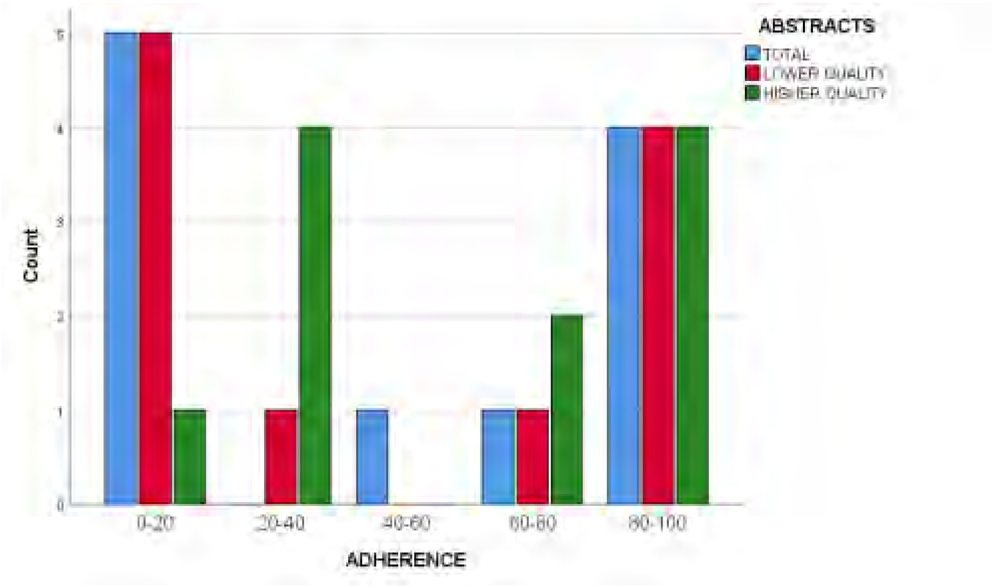
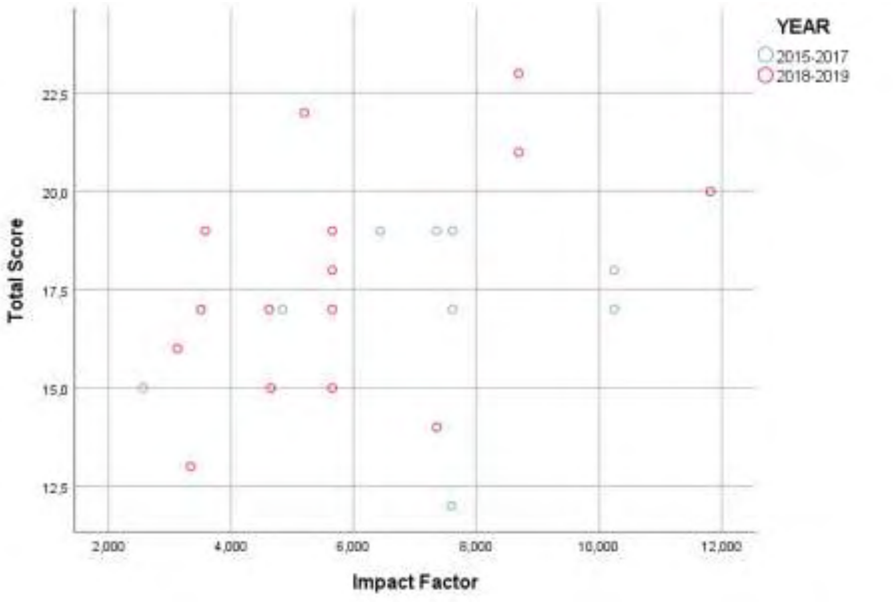


Figure 3: Number of items that present 0-20%, 20-40%, 40-60%, 60-80%, 80-100% adherence to STARD for abstracts for the three groups

When we compared the two groups (lower and higher quality abstracts) for every item and for the total adherence to "STARD for abstracts" it came up that for one item (item 5) along with the total adherence to "STARD for abstracts" the p value was <0.05, meaning that there is statistically significant difference between the two groups for these items. It’s notable that for item 1 we have marginal statistically significant difference between the two groups (p=0.052). In all these items high quality articles showed better performance.

Effect of total score on impact factor

We examined if publication year has an effect on the relationship between impact factor and total score.



There is a slightly indication of relationship between impact factor and total score, but the effect of publication year is not obvious.

We examined the relationship between impact factor and total score after adjusting for publication year. Impact factor is marginal not significantly related to total score (p=0.052). Also, the publication year effect is not significant (p=0,236 >0.05).

Tests of Between-Subjects Effects

Dependent Variable: Total Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	30,380[a]	2	15,190	2,353	,120	,183
Intercept	555,277	1	555,277	86,008	,000	,804
IM	27,355	1	27,355	4,237	**,052**	,168
YEAR	9,608	1	9,608	1,488	**,236**	,066
Error	135,579	21	6,456			
Total	7481,000	24				
Corrected Total	165,958	23				

a. R Squared = ,183 (Adjusted R Squared = ,105)

Thus, we omitted the effect of publication year from the model and the analysis is repeated.

When we examined the relationship between impact factor and total score the p-value for impact factor was p=0.090 (>0.05). Thus, total score is not related to impact factor.

Tests of Between-Subjects Effects

Dependent Variable: Total Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	20,772[a]	1	20,772	3,148	,090
Intercept	684,598	1	684,598	103,737	,000
IM	20,772	1	20,772	3,148	**,090**
Error	145,186	22	6,599		
Total	7481,000	24			
Corrected Total	165,958	23			

a. R Squared = ,125 (Adjusted R Squared = ,085)

Impact of total score on STARD endorsement

When we examined the relationship between total score and STARD endorsement the p-value for total score was p=0.360 (>0.05). Thus, total score is not related to STARD endorsement.

STARD_ENDORSEMENT * ARTICLES Crosstabulation

Count

		ARTICLES		
		LOWER QUALITY ARTICLES	HIGHER QUALITY ARTICLES	Total
STARD_ENDORSEMENT	no	3	1	4
	yes	10	10	20
Total		13	11	24

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,839[a]	1	**,360**		
Continuity Correction[b]	,134	1	,714		
Likelihood Ratio	,880	1	,348		
Fisher's Exact Test				,596	,363
Linear-by-Linear Association	,804	1	,370		
N of Valid Cases	24				

a. 2 cells (50,0%) have expected count less than 5. The minimum expected count is 1,83.

b. Computed only for a 2x2 table

CONCLUSIONS:

The present study investigated the reporting quality of studies regarding the diagnostic accuracy of sNfL levels in MS according to the STARD statement. The studies divided in high and low quality and the differences among them were explored. Moreover, we assessed the quality of reporting of the abstracts of the eligible studies.

On the whole, the quality of reporting was moderate and extremely variable across items. In particular the overall adherence to STARD statement was 49.7% indicating that STARD statement wasn't followed properly in the presentation of the studies. Across the 33 items that were examined in our study (from a 30 item questionnaire, item 2 excluded and items 10, 12, 13 and 21 were divided into 10a, 10b, 12a, 12b, 13a, 13b, 21a and 21b) some items showed very high adherence to STARD statement and some others very low.

Among the items with the poorest reporting are: *identification as a study of diagnostic accuracy using at least one measure of accuracy (item 1), whether clinical information and index test results were available to the assessors of the reference standard (item 13b), flow of participants, using a diagram (item 19), where the full study protocol can be accessed (item 29)*. STARD statement recommends to authors to use minimum one measure of accuracy in title or abstract, in order to facilitate the retrieval of their article. To use flow diagram of participants or to report the source from where the full protocol can be assessed could facilitate reader's comprehension of study design. Also, if the reader is aware of whether or not the results of the index test are known to the evaluator of the reference standard might help him decide if there's potential bias.

Two items aren't reported at all in any of the included studies: *intended sample size and how it was determined (item 18) and any adverse events from performing the index test or the reference standard (item 25)*. By not performing calculations to determine the sample size of the study results in lack of precision. Many of the included studies were small (<100 participants), and the possibility to be imprecise, with wide CIs around them, was a huge disadvantage[13]. Regarding the adverse events it comes with no surprise that weren't reported at all for the index test since phlebotomy is a non invasive procedure with extremely rare adverse events.

Low reporting levels that makes difficult for the reader to assess the validity of a study are also found in the following items: *definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory (item 12a), definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory (item 12b), whether clinical information and reference standard results were available to the performers/readers of the index test (item 13a), how missing data on the index test and reference standard were handled (item 16), registration number and name of registry (item 28)*.

It is notable that the items concerning the section of test results: *cross tabulation of the index test results by the results of the reference standard (item 23) and estimates of diagnostic accuracy and their precision (item 24),* present low reporting levels (below 50%). It is reported that among common mistakes in reliability analysis in articles that are being published by high impact journals is the assessment of precision by inappropriate statistical tests such as Pearson r, least square and paired t test [35]. During the evaluation progress of a new medical test, is crucial to compare its performance to that of an existing method. When the outcome of the test is a qualitative result (positive-negative), the use of measures like sensitivity/specificity or percent agreement is recommended. For tests that lead to quantitative results, different methods, such as Bland and Altman's limits of agreement (LOA), Pearson correlation (not always appropriate), concordance correlation coefficient (CCC) and intraclass correlation coefficient (ICC) , are indicated [36]. Hence researchers should be instructed to use different statistical tests to assess the precision of their studies.

On the other hand, among the items with the highest reporting are: *scientific and clinical background, including the intended use and clinical role of the index test (item 3), eligibility criteria (item 6), distribution of severity of disease in those with the target condition (item 21a) and sources of funding and other support (item 30)*. The information regarding participants' characteristics is significant because the performance of a test isn't the same among patients with different diseases and thus helps in the generalisability of the results. Also, disclosing notifications about sponsorships of a study permit the reader to judge for potential bias.

Moreover, there are items that have been identified in every study and mostly promote the generalisability of the results: *study objectives and hypotheses (item 4), index test, in sufficient detail to allow replication (item 10a), methods for estimating or comparing measures of diagnostic accuracy (item 14), baseline demographic and clinical characteristics of participants (item 20) and implications for practice, including the intended use and clinical role of the index test (item 27)*.

When we compared the overall quality of reporting in high versus low quality articles it was noted that the two groups were different in terms of STARD adherence and significant differences were spotted in almost a quarter of the investigated items: *identification as a study of diagnostic accuracy using at least one measure of accuracy (item 1), eligibility criteria (item 6), definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory (item 12a), definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory (item 12b), distribution of severity of disease in those with the target condition (item 21a), cross tabulation of the index test results by the results of the reference standard (item 23), estimates of diagnostic accuracy and their precision (item 24)*. In all these items, except eligibility criteria, high quality articles showed better performance. A systematic sampling review that investigated if articles published by journals with high impact factor sufficiently report participants' exclusion criteria, concluded that there is need for better reporting of eligibility criteria [37].

As far abstracts concern, the overall quality of reporting was also moderate (50.80%) and varied a lot across the 11 investigated items. Almost half of them showed extremely poor reporting: *identification as a study of diagnostic accuracy using at least one measure of accuracy (item 1), whether this was a prospective or retrospective study (item 3), eligibility criteria for participants and settings where the data were collected (item 4), estimates of* *diagnostic accuracy and their precision (item 8), registration number and name of registry (item 11)* and the other half showed relatively good reporting quality and in some cases even excellent: *study objectives (item 2), number of participants with and without the target condition included in the analysis (item 7), general interpretation of the results (item 9)*. Similar findings have also been identified in previous studies investigating the adherence of abstracts to "CONSORT checklist"[38,39], highlighting by this, the need for embracement of these guidelines by authors, reviewers and editors.

When we compared the overall quality of reporting in high versus low quality abstracts it came up that the two groups were different in terms of STARD adherence and in only one item: *whether participants formed a consecutive, random, or convenience series (item 5)*, high quality articles showed better performance.

Given that STARD statement has been used since 2003 until today, it is expected to detect improvement in reporting quality during the years. Our study showed that STARD statement hasn't upgraded the reporting quality of articles related to our topic may be because all the eligible articles were published, within a small period of time, the last 5 years. Another finding of our study is that high impact factor is not related τo better reporting, implying by this the necessity for improved reporting in journals either with low or with high impact factor.

Among the 14 journals that the selected articles have been published, 4 journals introduce STARD statement to authors and 6 journals support other reporting guidelines, CONSORT for clinical trials and PRISMA for systematic reviews. One can only assume that since journals indentify the need for unbiased reporting for one study type, it's possible to embrace reporting guidelines for diagnostic accuracy studies as well. Our study detected that 71.4% of the journals endorse STARD statement, indicating that most of the authors are aware of this reporting format.

Our study has some limitations. Firstly, the research was limited in only one database (PubMed) and focused on English language. However, given the fact that before the year 2000 (study's starting point) no studies regarding the topic had been published, we believe that we haven't missed so many studies to alter the findings of our review. Another limitation is that between the selected studies the reference standard was highly heterogeneous, making difficult to evaluate the generalisability of the results. Specifically depending the article, the sNFL levels were compared with those of: CSF NFL levels, MRI data, and scores like EDSS (Expanded Disability Status Scale).

Since today, STARD statement has been used to evaluate the reporting quality of some diagnostic accuracy studies, such as imaging derived parameters (RNFL and ONH) to diagnose glaucoma [40], commercial tests to diagnose Tuberculosis, malaria and human immunodeficiency virus[41], anti CCP antibodies in rheumatoid arthritis[42], but in general, the limited number of studies suggest there is room for more research.

In conclusion, the overall quality of reporting using STARD statement was moderate for both full articles and abstracts, with a large variability in adherence across investigating items, ranging from 0 to 100%. The quality of reporting between high and low quality articles or abstracts was significant different. The introduction of STARD statement hasn't improved the completeness of reporting during the years. The journals seem to publish diagnostic accuracy studies regardless if they suggest the use of STARD statement in the instruction section for authors. Despite the modest adherence even from journals that endorse STARD statement, it is recommended more and more journals to use these reporting guidelines. If authors, reviewers and editors follow with compliance STARD checklist in submitted manuscripts the completeness of reporting and the quality of medical research will be improved.

REFERENCES

1. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. 2018;17(2):162-173. doi:10.1016/S1474-4422(17)30470-2

2. Lycke JN, Karlsson J-E, Andersen O, Rosengren LE. Neurofilament protein in cerebrospinal fluid: a potential marker of activity in multiple sclerosis. *J Neurol Neurosurg Psychiatry*. 1998;64(3):402-404. doi:10.1136/jnnp.64.3.402

3. Varhaug KN, Torkildsen Ø, Myhr K-M, Vedeler CA. Neurofilament Light Chain as a Biomarker in Multiple Sclerosis. *Front Neurol*. 2019;10:338. doi:10.3389/fneur.2019.00338

4. Fialová L, Bartos A, Svarcová J, Zimova D, Kotoucova J, Malbohan I. Serum and cerebrospinal fluid light neurofilaments and antibodies against them in clinically isolated syndrome and multiple sclerosis. *J Neuroimmunol*. 2013;262(1-2):113-120. doi:10.1016/j.jneuroim.2013.06.010

5. Avsar T, Korkmaz D, Tütüncü M, et al. Protein biomarkers for multiple sclerosis: semi-quantitative analysis of cerebrospinal fluid candidate protein biomarkers in different forms of multiple sclerosis. *Mult Scler*. 2012;18(8):1081-1091. doi:10.1177/1352458511433303

6. Martin S-J, McGlasson S, Hunt D, Overell J. Cerebrospinal fluid neurofilament light chain in multiple sclerosis and its subtypes: a meta-analysis of case–control studies. *J Neurol Neurosurg Psychiatry*. 2019;90(9):1059-1067. doi:10.1136/jnnp-2018-319190

7. Piehl F, Kockum I, Khademi M, et al. Plasma neurofilament light chain levels in patients with MS switching from injectable therapies to fingolimod. *Mult Scler J*. 2018;24(8):1046-1054. doi:10.1177/1352458517715132

8. Håkansson I, Tisell A, Cassel P, et al. Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis. *J Neuroinflammation*. 2018;15(1):209. doi:10.1186/s12974-018-1249-7

9. Disanto G, Barro C, Benkert P, et al. Serum Neurofilament light: A biomarker of neuronal damage in multiple sclerosis. *Ann Neurol*. 2017;81(6):857-870. doi:10.1002/ana.24954

10. Novakova L, Zetterberg H, Sundström P, et al. Monitoring disease activity in multiple sclerosis using serum neurofilament light protein. *Neurology*. 2017;89(22):2230-2237.

doi:10.1212/WNL.0000000000004683

11. Kuhle J, Nourbakhsh B, Grant D, et al. Serum neurofilament is associated with progression of brain atrophy and disability in early MS. *Neurology*. 2017;88(9):826-831. doi:10.1212/WNL.0000000000003653

12. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Clin Chem*. 2015;61(12):1446-1452. doi:10.1373/clinchem.2015.246280

13. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799

14. Cohen JF, Korevaar DA, Gatsonis CA, et al. STARD for Abstracts: essential items for reporting diagnostic accuracy studies in journal or conference abstracts. *BMJ*. 2017;358:j3751. doi:10.1136/bmj.j3751

15. What is considered a good impact factor? - LibAnswers. http://mdanderson.libanswers.com/faq/26159. Accessed August 21, 2019.

16. Disanto G, Adiutori R, Dobson R, et al. Serum neurofilament light chain levels are increased in patients with a clinically isolated syndrome. *J Neurol Neurosurg Psychiatry*. 2015;87(2):jnnp-2014-309690. doi:10.1136/jnnp-2014-309690

17. Kuhle J, Barro C, Disanto G, et al. Serum neurofilament light chain in early relapsing remitting MS is increased and correlates with CSF levels and with MRI measures of disease severity. *Mult Scler J*. 2016;22(12):1550-1559. doi:10.1177/1352458515623365

18. Bergman J, Dring A, Zetterberg H, et al. Neurofilament light in CSF and serum is a sensitive marker for axonal white matter injury in MS. *Neurol - Neuroimmunol Neuroinflammation*. 2016;3(5):e271. doi:10.1212/NXI.0000000000000271

19. Al-Temaimi R, AbuBaker J, Al-khairi I, Alroughani R. Remyelination modulators in multiple sclerosis patients. *Exp Mol Pathol*. 2017;103(3):237-241. doi:10.1016/j.yexmp.2017.11.004

20. Varhaug KN, Torkildsen Ø, Myhr K-M, Vedeler CA. Neurofilament Light Chain as a Biomarker in Multiple Sclerosis. *Front Neurol*. 2019;10:338. doi:10.3389/fneur.2019.00338

21. Piehl F, Kockum I, Khademi M, et al. Plasma neurofilament light chain levels in patients with MS switching from injectable therapies to fingolimod. *Mult Scler J*. 2018;24(8):1046-1054. doi:10.1177/1352458517715132

22. Håkansson I, Tisell A, Cassel P, et al. Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis. *J Neuroinflammation*. 2018;15(1):209. doi:10.1186/s12974-018-1249-7

23. Barro C, Benkert P, Disanto G, et al. Serum neurofilament as a predictor of disease worsening and brain and spinal cord atrophy in multiple sclerosis. *Brain*. 2018;141(8):2382-2391. doi:10.1093/brain/awy154

24. Chitnis T, Gonzalez C, Healy BC, et al. Neurofilament light chain serum levels correlate

with 10-year MRI outcomes in multiple sclerosis. *Ann Clin Transl Neurol*. 2018;5(12):1478-1491. doi:10.1002/acn3.638

25. Browne RW, Jakimovski D, Ziliotto N, et al. High-density lipoprotein cholesterol is associated with multiple sclerosis fatigue: A fatigue-metabolism nexus? *J Clin Lipidol*. June 2019. doi:10.1016/j.jacl.2019.06.003

26. Sehr T, Akgün K, Proschmann U, Bucki R, Zendzian-Piotrowska M, Ziemssen T. Early central vs. peripheral immunological and neurobiological effects of fingolimod—a longitudinal study. *J Mol Med*. June 2019. doi:10.1007/s00109-019-01812-x

27. Siller N, Kuhle J, Muthuraman M, et al. Serum neurofilament light chain is a biomarker of acute and chronic neuronal damage in early multiple sclerosis. *Mult Scler J*. 2019;25(5):678-686. doi:10.1177/1352458518765666

28. Akgün K, Kretschmann N, Haase R, et al. Profiling individual clinical responses by high-frequency serum neurofilament assessment in MS. *Neurol - Neuroimmunol Neuroinflammation*. 2019;6(3):e555. doi:10.1212/NXI.0000000000000555

29. Cuello JP, Martínez Ginés ML, Kuhle J, et al. Neurofilament light chain levels in pregnant multiple sclerosis patients: a prospective cohort study. *Eur J Neurol*. May 2019:ene.13965. doi:10.1111/ene.13965

30. Abdelhak A, Huss A, Kassubek J, Tumani H, Otto M. Serum GFAP as a biomarker for disease severity in multiple sclerosis. *Sci Rep*. 2018;8(1):14798. doi:10.1038/s41598-018-33158-8

31. Hyun J-W, Kim Y, Kim G, Kim S-H, Kim HJ. Longitudinal analysis of serum neurofilament light chain: A potential therapeutic monitoring biomarker for multiple sclerosis. *Mult Scler J*. March 2019:135245851984075. doi:10.1177/1352458519840757

32. Dalla Costa G, Martinelli V, Moiola L, et al. Serum neurofilaments increase at progressive multifocal leukoencephalopathy onset in natalizumab-treated multiple sclerosis patients. *Ann Neurol*. 2019;85(4):606-610. doi:10.1002/ana.25437

33. Ferraro D, Guicciardi C, De Biasi S, et al. Plasma Neurofilaments correlate with Disability in Progressive Multiple Sclerosis patients. *Acta Neurol Scand*. July 2019:ane.13152. doi:10.1111/ane.13152

34. Kuhle J, Kropshofer H, Haering DA, et al. Blood neurofilament light chain as a biomarker of MS disease activity and treatment response. *Neurology*. 2019;92(10):e1007-e1015. doi:10.1212/WNL.0000000000007032

35. Sabour S, Abbasnezhad O, Mozaffarian S, Kangavari HN. Accuracy and Precision in Medical Researches ; Common Mistakes and Misinterpretations. 2017;(4):58-60.

36. Morgan CJ, Aban I. Methods for evaluating the agreement between diagnostic tests. *J Nucl Cardiol*. 2016;23(3):511-513. doi:10.1007/s12350-015-0175-7

37. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals. *JAMA*. 2007;297(11):1233. doi:10.1001/jama.297.11.1233

38. Hays M, Andrews M, Wilson R, Callender D, O'Malley PG, Douglas K. Reporting quality

of randomised controlled trial abstracts among high-impact general medical journals: a review and analysis. *BMJ Open*. 2016;6(7):e011082. doi:10.1136/bmjopen-2016-011082

39. Gallo L, Wakeham S, Dunn E, Avram R, Thoma A, Voineskos S. The Reporting Quality of Randomized Controlled Trial Abstracts in Plastic Surgery. *Aesthetic Surg J*. July 2019. doi:10.1093/asj/sjz199

40. Michelessi M, Lucenteforte E, Miele A, et al. Diagnostic accuracy research in glaucoma is still incompletely reported: An application of Standards for Reporting of Diagnostic Accuracy Studies (STARD) 2015. van Wouwe JP, ed. *PLoS One*. 2017;12(12):e0189716. doi:10.1371/journal.pone.0189716

41. Walther S, Schueler S, Tackmann R, Schuetz GM, Schlattmann P, Dewey M. Compliance with STARD Checklist among Studies of Coronary CT Angiography: Systematic Review. *Radiology*. 2014;271(1):74-86. doi:10.1148/radiol.13121720

42. Zintzaras E, Papathanasiou AA, Ziogas DC, Voulgarelis M. The reporting quality of studies investigating the diagnostic accuracy of anti-CCP antibody in rheumatoid arthritis and its impact on diagnostic estimates. *BMC Musculoskelet Disord*. 2012;13(1):113. doi:10.1186/1471-2474-13-113