Predictive medicine in multiple sclerosis: A systematic review

Julie Havas, Emmanuelle Leray, Fabien Rollot, Romain Casey, Laure Michel, Flora Lejeune, Sandrine Wiertlewski, David Laplaud, Yohann Foucher

Highlights

- A systematic review of developed and/or validated a predictive model for MS.
- Despite finding more than 6,000 studies, 15 articles were retained.
- An over-interpretation of association in terms of prediction in the MS literature.
- A need to integrate good standards in developing and validating predictive models.
- Validated predictive tools for MS management are currently lacking.

Predictive Medicine in Multiple Sclerosis: a Systematic Review

Julie Havas[1], Emmanuelle Leray[2], Fabien Rollot[3,4,5,6], Romain Casey[3,4,5,6], Laure Michel[7], Flora Lejeune[8,9], Sandrine Wiertlewski[8,9], David Laplaud[8,9] and Yohann Foucher[1,9]

Affiliations:
1. SPHERE (methodS in Patient-centered outcomes & HEalth ResEarch) U1246, INSERM, Nantes University, Tours University, Nantes, France.
2. EA 7449 REPERES, Univ. Rennes, EHESP, Rennes, France.
3. Université de Lyon, Université Claude Bernard Lyon 1, F-69000 Lyon, France.
4. Hospices Civils de Lyon, Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, F-69677 Bron, France.
5. Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, F-69003 Lyon, France.
6. EUGENE DEVIC EDMUS Foundation against multiple sclerosis, state-approved foundation, F-69677 Bron, France.
7. Neurology Department, Rennes Clinical Investigation Center 1414, Rennes University Hospital-Rennes University-INSERM, Rennes, France.
8. Centre de Recherche en Transplantation et Immunologie (CRTI), UMR 1064, INSERM, Nantes; Service de Neurologie, CHU Nantes, CIC Inserm 004, France.
9. Nantes University Hospital, Nantes, France.

Corresponding author: David A. Laplaud, INSERM UMR 1064, 30 Bd J. Monnet, 44093 Nantes Cedex 1, France. Tel: +33 240087410; Fax: +33 240087411; E-mail: david.laplaud@univ-nantes.fr

Keywords: Precision medicine; Prognostic tools, Multiple sclerosis, Systematic review.

Total word count: 2560

Running title: Predictive medicine in MS

Disclosure statement: Dr Y. Foucher has received speaking honoraria from Biogen. Dr S. Wiertlewski has received speaking honoraria, travel expense reimbursement for participation in scientific meetings, and has participated in advisory boards in the past years with Biogen, Merck, Novartis, Roche, Sanofi and Teva. Pr D. Laplaud has received funding for travel or speaker honoraria from Biogen, Novartis and Genzyme. He has participated on advisory boards in the past years Biogen-Idec, TEVA Pharma, Novartis and Genzyme. Dr L. Michel received honoraria for consulting from Biogen Idec, Roche, Sanofi Genzyme, Teva, Novartis and Merck Serono. Dr. R. Casey, Dr. F. Lejeune and F. Rollot report no disclosures. Dr E. Leray reports grants from the French National Security Agency of Medicines and Health Products, the EDMUS Foundation and Roche SAS; and personal fees from Novartis, MedDay Pharmaceuticals and Roche SAS, all outside of the submitted work. J.P. holds research funding from the Natural Sciences and Engineering Research Council of Canada and has received consulting fees or fees for service on Data Safety Monitoring Boards from Biogen, the Canadian Study Group on CCSVI, Novartis, and Teva Pharmaceuticals Europe, all outside of the submitted work.

Abbreviations: Multiple sclerosis (MS), interferon beta (IFN-β), medical subject headings (MeSH), systematic reviews and meta-analyses (PRISMA), expanded disability status scale (EDSS), magnetic resonance imaging (MRI), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV).

Abstract

Background. One of the main challenges in multiple sclerosis (MS) is to predict disease progression based on patient characteristics and therapeutic strategies. We therefore performed a systematic review to critically appraise the composite tools available for this purpose.

Methods. We performed electronic database searches in MEDLINE, EMBASE, Web of Science and the Cochrane Library. We included studies in English or French that developed and/or validated a predictive model for MS patients. Two reviewers independently screened articles by title and abstract. Three teams of two reviewers assessed the full text of each relevant study.

Results. Database searches yielded 6,035 studies after deduplication. Among the 42 screened full texts, 15 articles satisfied the eligibility criteria. Of these, six articles examined the development of predictive tools, six articles aimed to validate existing tools and three articles proposed both development and validation. We identified numerous methodological pitfalls, especially the lack of adequate validations in terms of discrimination and calibration. Only two scoring systems were externally validated several times: the Rio and the modified Rio scores. Nevertheless, their accuracies were highly variable, ranging from 65% to 91%.

Conclusions. Overall, there is a lack of validated predictive tools in MS, and further external validation of the existing ones are required. Demonstration of the clinical usefulness is also needed prior to being transferred into clinical practice. Finally, our study illustrates that the MS literature needs to integrate good standards in developing and validating predictive models.

1. Introduction

Multiple sclerosis (MS) is an inflammatory disease that affects the central nervous system. It is the leading cause of non-traumatic neurologic disability in young adults in the USA and Europe (1). MS is a heterogeneous disease with important variability between patients in terms of natural history (2), and this variability is even greater due to the large number of disease-modifying therapies (DMT) (3). However, the expected evolution of the disease, with or without DMT, is essential for guiding informed decisions about initiation, switching, or even cessation of DMT.

One of the main challenges is to predict disease progression based on the patients' characteristics and the therapeutic strategies. In the current era of precision medicine, where genetic and biological parameters may be associated with the disease evolution and the treatment response, this may result in important advances in MS patient treatment. Early identification of suboptimal responder patients could for instance prevent both acute inflammatory injury and the neurodegenerative processes leading to irreversible disabilities and secondary progressive forms. The first and probably most well-known tentative is the Rio scoring system (4), which aims to predict the response to interferon beta (IFN-β) therapy at 1-year post-initiation.

Besides medical decision making, being able to inform patients about their likely disease progression is important. Similar to other chronic diseases, anxiety is a daily concern for MS patients, with a prevalence ranging from 14% to 34% (5). The possible consequences are a reduction in quality of life (6), treatment non-compliance (7), or even exacerbation of disease symptoms (8). For some patients, anxiety is partially due to the absence of information regarding the future of their disease (6). Many patients need better quality information than they initially received. Seventy-five percent of patients reported inadequacies in information they had been offered about MS (9). Besides limiting anxiety, informing a patient of her/his prognosis and corresponding treatment options is of primary importance in a patient-centered vision of care, as this allows joint decisions on further treatments to be made by the patient and neurologist.

The aim of this systematic review is to critically appraise the composite tools available in MS to predict disease evolution. The specific objectives were to identify relevant risk prediction models, to describe the methods used for their development, to investigate their validation, and to discuss their clinical utility.

2. Methods

2.1. Search strategy

This systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (10). The literature search strategies were developed using Medical Subject Headings (MeSH) terms and keywords related to multiple sclerosis, predictive models and validation studies (**Table 1**). We performed electronic database searches in MEDLINE, EMBASE, Web of Science and the Cochrane Library. We also explored the references of the selected articles by using Google Scholar. All sources were reviewed up to the 30th of November 2017.

2.2. Study selection

Two reviewers independently screened articles by title and abstract. Three teams of two reviewers assessed the full text of each potentially relevant study. Where disagreements occurred, the final decision was based on a discussion with another independent reviewer.

We included studies in English or French that developed and/or validated a predictive model for MS. We excluded non-human studies, studies with no original statistics (review articles, reports of registries), studies that did not deal with multiple sclerosis (such as clinically isolated syndrome), medical-economic studies, association studies, studies aiming to develop new methods with no clinical objective, studies which developed or validated non-predictive scores/scales (such as patient reported outcomes), descriptive studies, diagnostic studies, studies with a predictive outcome not related to the disease evolution, and studies where neither the full text nor the summary was available.

We extracted the data using the CHeck list for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (11).

3. Results

3.1. Description of the selected studies

As detailed in **Figure 1**, we identified 6,035 unique articles, among which 5,993 were excluded based on titles and abstracts. Among the 42 screened full texts, 15 articles (4,12–25) that respected the eligibility criteria were included, and all of these dealt with relapsing-remitting MS. Of these, six articles examined the development of prediction tools (4,14–16,19,25), six articles aimed to validate

existing tools (12,20–24), and three articles proposed both development and validation (13,17,18). Seven articles were based on patients treated by IFN-β only, three by several treatments, one by teriflunomide only, one by fingolimod only and two articles included untreated patients. One manuscript did not report information on treatment. The repartition in terms of geography was six from Italy, three from Spain, one from Korea, one from Canada, one from Israel and six from several countries (international studies). In terms of study design, nine articles were based on observational data and five on clinical trials. All articles were published in Neurology journals, except one in multidisciplinary sciences and one in Immunology. The publication year ranged from 1996 to 2017. Seven articles (46.7%) were published since 2013.

3.2. Description of the predictive tools

As illustrated in **Table 2**, nine predictive tools were identified. The Rio score (4), the modified Rio score (13) and the MAGNIMS score (14) were computed at 1-year post IFN-β prescription for predicting the suboptimal response to treatment. These three scoring systems aimed to help the decision for an early switch from IFN-β to second-line therapy for high-risk patients. The Rio score was calculated by using relapse occurrence, Magnetic Resonance Imaging (MRI) activity and EDSS (Expanded Disability Status Scale) progression. In contrast, the modified Rio score (13) and the MAGNIMS score (14) were only based on MRI activity and EDSS progression. The development of the Rio score (4) also differed from the two other scores by ignoring the treatment switch due to lack of efficacy in the definition of the suboptimal response.

The BREMS score (15), the BREMSO score (16) and the tool proposed by Calabrese et al. (17) aimed to predict early onset of the secondary progressive phase. The first score was based on demographic and clinical variables collected during the first-year post-disease onset, while the second score only used data available at disease onset. The authors proposed to use it in observational studies for reducing confounders. The third tool (17) was proposed with no landmark time, i.e. it can be computed at any time of the disease course.

The three other scoring systems (18,19,25) were also proposed with no landmark time. While one can expect that these tools with no landmark should include the disease duration among their predictive variables as a proxy of the disease history, only the model proposed by Weinshenker et al. (19) considered this parameter for predicting the time-to-EDSS 6. Sormani et al. (18) aimed to predict the risk of relapse, while Achiron et al. (25) aimed to predict the neurological disability (a composite outcome from EDSS evolution and relapse occurrence). Among the three tools, only Sormani et al.

(18) proposed a possible utility for identifying MS patients with a high risk of relapse as inclusion criteria in clinical trials.

3.3. Developments: predictive variable selection and modeling

Among the 15 selected articles, nine predictive tools were developed. Six tools were proposed with a statistical selection of the predictive variables (15–19,25), the three others (4,13,14) being scoring systems a priori defined by experts. Among the six articles, the following statistical approaches were used: Markov chain Monte Carlo Bayesian model (15,16), Support Vector Machine (25), logistic model (17,19) and Cox model (18). **Figure 2** presents the distribution of the predictive variables included in the nine predictive tools. The most frequent variables were number of new T2 lesions and relapses. EDSS and age were used in three models. Other predictors less commonly used were gender, duration of the disease, sphincter onset, pure motor onset, moto-sensory onset, sequelae after onset, neurological functional systems, cerebellar cortical volume and cortical lesion volume. The predictive model proposed by Achiron et al. (25) was based on 34 genes with no MRI or other clinical parameters. Importantly, treatments were not included in any model.

3.4. Methodological pitfalls in estimating apparent prognostic capacities

As reported in **Table 2**, the apparent prognostic capacities (estimated from the learning sample) were never compared with other existing prognostic tools. This issue may be because each tool aimed to predict different outcomes or the same outcome with different definitions. When the apparent prognostic capacities are reported in mid- or long-term studies, the corresponding statistical analyses did not appropriately deal with such a time-dependent context. More precisely, while Cox regressions were used for the developments based on right-censored data, it was ignored in estimating the corresponding prognostic capacities by excluding patients with not enough follow-up: indicators such as sensitivity (SE) and specificity (SP) were naively estimated by the corresponding proportions.

3.5. External validations

Three articles (13,17,18) proposed both the development and external validation of a predictive tool. In addition, we identified six studies (12,20–24) with only external validation of previously developed models. Among these nine articles, five proposed a validation of the modified Rio score (13), with accuracies ranging from 65% to 91%. Except in the study by Lattanzi et al. (22) due to small sample

size, the four other articles proposed the same stratification into two groups (score 0-1 versus 2-3) and reported SE from 19% to 96%, SP from 72% to 97%, PPV from 28% to 86% and NPV from 68% to 93%.

Three articles (12,21,23) proposed a validation of the Rio score (4). The accuracies ranged from 62% to 93%. The same stratification (0-1 versus 2-3) resulted in SE ranging from 45% to 98%, SP from 67% to 86%, PPV from 43% to 92% and NPV from 85% to 93%.

One article (20) aimed to validate the predictive capacities of the MAGNIMS score (14), reporting an accuracy of 63%. Other indicators were estimated (SE, SP, PPV and NPV), but the corresponding stratification was not based on the MAGNIMS score alone and was difficult to understand. Bergamashi et al. (24) proposed a validation of the BREMS score (15). Nevertheless, they did not report the accuracy, and the stratification was proposed with irrelevant extreme cuts-offs by using extreme values (5^{th} and 95^{th} percentiles). Sormani et al. (18) proposed both the development and external validation in the same article, but the discriminative capacities were not reported. Calabrese et al. (17) also proposed both the development and external validation. The accuracy equaled 92%, but no rule was reported for the stratification.

3.6. Methodological pitfalls in external validations

No study precisely reported the calibration, for instance, by plotting observed versus predicted probabilities of events. As for the apparent prognostic capacities, the studies mainly reported accuracy, SE, SP, PPV and NPV; but no study considered the right-censoring in the corresponding estimation when necessary. Even more worrisome, the dispersion of these indicators was not reported: we have no idea of the corresponding standard errors or confidence intervals. This is even more important regarding the high range of these values for each scoring system and the small sample sizes of three studies (17,22,23).

4. Discussion

In MS management, the individual evaluation of the expected evolution of disease is important. In order to identify predictive tools potentially useful in clinical practice, we decided to perform this review, including an exhaustive research, a careful selection of studies, and a double-blind data extraction.

Despite finding more than 6,000 studies related to prediction in MS, we retained only 15 articles that were aimed at developing composite predictive tools and/or validating their capacities. One of the main reasons was the over-interpretation of association in terms of prediction (**Figure 1**). It occurred in 2,880 (47.8%) articles (2,217 association/impact studies and 669 studies for evaluating efficacy or safety of drugs). It is quite common to find that factors, defined by authors as prognostic and/or predictive, are in fact only correlated with the outcome. Indeed, the magnitude of odds-ratios (27) or hazard-ratios (28) do not inform on prognostic capacities. Nevertheless, it can lead to misinterpretations concerning the clinical utility of the marker (26).

According to our results, only two scoring systems were externally validated several times: the Rio score (4) and the modified Rio score (13). Compared to the other predictive tools, one can highlight their possible clinical utility by identifying in patients treated by IFN-β for 1 year a stratum at high risk of disability progression, who may benefit from an early switch to second line therapy. Nevertheless, to our knowledge, no study was performed to demonstrate such usefulness.

Our results also highlight that further external validation of the Rio score (4) and the modified Rio score (13) must be performed using large samples and well-validated statistical methods adapted to time-to-event data. Importantly, both calibration and discrimination must be evaluated. Firstly, the calibration aims to evaluate the concordance between observed and predicted probabilities of events. It can be graphically evaluated or statistics can be computed, such as the Hosmer-Lemeshow test (27). Secondly, the discrimination aims to evaluate the separation between individuals who will present events from patients who will not. For instance, indicators such as area under ROC curves (AUC), SE and SP can be used. Importantly, bootstrapping can be used to obtain the corresponding confidence intervals, which were never reported in the articles included in our review.

When necessary for long-term studies, both the calibration and the discrimination analyses must consider right-censoring. For discrimination analyses, Heagerty et al. (28) proposed an estimator of ROC curves in the presence of such incomplete data. The Kaplan-Meier estimator (29) can be used to estimate the observed probabilities in calibration analyses. In contrast, our review highlights the omission of right-censoring in the estimation of prognostic capacities. It consists of removing patients with insufficient follow-up, such a naïve approach being potentially associated with bias and higher variance (30).

One might assume that both the Rio and modified Rio scores may be enriched by other parameters in order to increase their discriminative capacities, as proposed by Sormani et al. (18) One can first study the parameters retained in the other predictive tools, such as age, gender and EDSS (**Figure 2**). Another limitation of the two scoring systems is the lack of update after 1-year post INF-β initiation.

The occurrence of relapse, new MRI results, or EDSS evolution may be useful to compute dynamic predictions and to improve the two time-fixed scoring systems. Recent developments in joint models for longitudinal markers and time-to-event may offer an interesting framework (31,32). In relation to this, our review underlines the importance of avoiding previously identified methodological pitfalls: multiplication of outcomes, different definitions of the same outcome, lack of internal and external validation, poor consideration of incomplete data and small sample size.

5. Conclusions

Validated and clinically useful predictive tools for MS management are currently lacking. Whilst the Rio score (4) and the modified Rio score (13) are the only tools with several external validations, further studies using well-performed external validation and usefulness demonstrations are needed to encourage and justify their use in clinical practice. Our study also demonstrates that the MS literature needs to establish a consensus on the definition, development and validation of predictive models.

6. Acknowledgements

Fabien Rollot and Romain Casey were supported by a grant provided by the French State and handled by the "Agence Nationale de la Recherche," within the framework of the "Investments for the Future" program, under the reference ANR-10-COHO-002 Observatoire Français de la Sclérose en plaques (OFSEP)

7. References

1. Adelman G, Rane SG, Villa KF. The cost burden of multiple sclerosis in the United States: a systematic review of the literature. J Med Econ. 2013;16(5):639–47.

2. Tremlett H, Zhao Y, Rieckmann P, Hutchinson M. New perspectives in the natural history of multiple sclerosis. Neurology. 2010 Jun 15;74(24):2004–15.

3. Wingerchuk DM, Carter JL. Multiple sclerosis: current and emerging disease-modifying therapies and treatment strategies. Mayo Clin Proc. 2014 Feb;89(2):225–40.

4. Río J, Castilló J, Rovira A, Tintoré M, Sastre-Garriga J, Horga A, et al. Measures in the first year of therapy predict the response to interferon beta in MS. Mult Scler. 2009 Jul;15(7):848–53.

5. Garfield AC, Lincoln NB. Factors affecting anxiety in multiple sclerosis. Disabil Rehabil. 2012;34(24):2047–52.

6. Mitchell AJ, Benito-León J, González J-MM, Rivera-Navarro J. Quality of life and its assessment in multiple sclerosis: integrating physical and psychological components of wellbeing. The Lancet Neurology. 2005 Sep 1;4(9):556–66.

7. Mohr DC, Boudewyn AC, Likosky W, Levine E, Goodkin DE. Injectable medication for the treatment of multiple sclerosis: the influence of self-efficacy expectations and injection anxiety on adherence and ability to self-inject. Ann Behav Med. 2001;23(2):125–32.

8. Sherbourne CD, Wells KB, Meredith LS, Jackson CA, Camp P. Comorbid anxiety disorder and the functioning and well-being of chronically ill patients of general medical providers. Arch Gen Psychiatry. 1996 Oct;53(10):889–95.

9. Somerset M, Campbell R, Sharp DJ, Peters TJ. What do people with MS want and expect from health-care services? Health Expectations. 2001 Mar 1;4(1):29–37.

10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336–41.

11. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014 Oct;11(10):e1001744.

12. Romeo M, Martinelli V, Rodegher M, Perego E, Maida S, Sormani MP, et al. Validation of 1-year predictive score of long-term response to interferon-beta in everyday clinical practice multiple sclerosis patients. European journal of neurology. 2015 Jun;22:973–80.

13. Sormani MP, Rio J, Tintore M, Signori A, Li D, Cornelisse P, et al. Scoring treatment response in patients with relapsing multiple sclerosis. Multiple sclerosis (Houndmills, Basingstoke, England). 2013 Apr;19:605–12.

14. Sormani MP, Gasperini C, Romeo M, Rio J, Calabrese M, Cocco E, et al. Assessing response to interferon-β in a multicenter dataset of patients with MS. Neurology. 2016 Jul 12;87(2):134–40.

15. Bergamaschi R, Berzuini C, Romani A, Cosi V. Predicting secondary progression in relapsing-remitting multiple sclerosis: a Bayesian analysis. Journal of the neurological sciences. 2001 Aug 15;189:13–21.

16. Bergamaschi R, Montomoli C, Mallucci G, Lugaresi A, Izquierdo G, Grand'Maison F, et al. BREMSO: a simple score to predict early the natural course of multiple sclerosis. European journal of neurology. 2015 Jun;22:981–9.

17. Calabrese M, Romualdi C, Poretto V, Favaretto A, Morra A, Rinaldi F, et al. The changing clinical course of multiple sclerosis: a matter of gray matter. Annals of neurology. 2013 Jul;74:76–83.

18. Sormani MP, Rovaris M, Comi G, Filippi M. A composite score to predict short-term disease activity in patients with relapsing-remitting MS. Neurology. 2007 Sep 18;69:1230–5.

19. Weinshenker BG, Issa M, Baskerville J. Long-term and short-term outcome of multiple sclerosis: a 3-year follow-up study. Archives of neurology. 1996 Apr;53:353–8.

20. Sormani MP, Truffinet P, Thangavelu K, Rufi P, Simonson C, De Stefano N. Predicting long-term disability outcomes in patients with MS treated with teriflunomide in TEMSO. Neurology: Neuroimmunology and NeuroInflammation. 2017;4.

21. Rio J, Rovira A, Tintore M, Otero-Romero S, Comabella M, Vidal-Jordana A, et al. Disability progression markers over 6-12 years in interferon-beta-treated multiple sclerosis patients. Multiple sclerosis (Houndmills, Basingstoke, England). 2017 Mar 1;1352458517698052.

22. Lattanzi S, Danni M, Cerqua R, Taffi R, Provinciali L, Silvestrini M. Prediction of disability progression in fingolimod-treated patients. Journal of the neurological sciences. 2015 Nov 15;358:432–4.

23. Hyun JW, Kim SH, Jeong IH, Ahn SW, Huh SY, Park MS, et al. Utility of the rio score and modified rio score in korean patients with multiple sclerosis. PLoS ONE. 2015;10:e0129243.

24. Bergamaschi R, Quaglini S, Trojano M, Amato MP, Tavazzi E, Paolicelli D, et al. Early prediction of the long term evolution of multiple sclerosis: the Bayesian Risk Estimate for Multiple Sclerosis (BREMS) score. Journal of neurology, neurosurgery, and psychiatry. 2007 Jul;78:757–9.

25. Achiron A, Gurevich M, Snir Y, Segal E, Mandel M. Zinc-ion binding and cytokine activity regulation pathways predicts outcome in relapsing-remitting multiple sclerosis. Clinical and experimental immunology. 2007 Aug;149:235–42.

26. Dantan E, Combescure C, Lorent M, Ashton-Chess J, Daguin P, Classe J, et al. An original approach was used to better evaluate the capacity of a prognostic marker using published survival curves. J Clin Epidemiol. 2014;67:441–8.

27. Wiley: Applied Logistic Regression, 3rd Edition - David W. Hosmer, Stanley Lemeshow, Rodney X. Sturdivant [Internet]. [cited 2017 Dec 13]. Available from: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470582472.html

28. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44.

29. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958 Jun 1;53(282):457–81.

30. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013 Dec 30;32(30):5381–97.

31. Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. CRC Press; 2012. 279 p.

32. Fournier M-C, Foucher Y, Blanche P, Legendre C, Girerd S, Ladrière M, et al. Dynamic predictions of long-term kidney graft failure: an information tool promoting patient-centred care. Nephrol Dial Transplant. 2019 Mar 11;

8. Tables

Table 1: Database search strategy

Theme	Terms used	Position
Multiple Sclerosis	("multiple sclerosis" OR "disseminated sclerosis") **AND**	MeSH Terms, Title, Abstract, Keyword
Prediction	("predictive value of tests" OR "prediction" OR "predict" OR "predicts" OR "predictive" OR "predicting" OR "predicted" OR "probability" OR "prognosis" OR "prognostication" OR "prognosticate" OR "prognosticates" OR "prognostic" OR "precision medicine" OR "stratified" OR "precision" OR "personalized" OR "personalized" OR "risk assessment" OR "risk") **AND**	MeSH Terms, Title, Abstract, Keyword
Modelling	("models, statistical" OR "model" OR "models" OR "modeling" OR "modelling" OR "equation" OR "equations" OR "regression" OR "algorithm" OR "algorithms" OR "score" OR "scores" OR "scoring" OR "nomograms" OR "nomogram") **AND**	MeSH Terms, Title, Abstract, Keyword
Validation	("prediction" OR "predict" OR "predicts" OR "predictive" OR "predicting" OR "predicted" OR "validation studies as topic" OR "validation" OR "validity" OR "validate" OR "validates" OR "validated" OR "calibration" OR "discrimination" OR "classification" OR "bootstrapping" OR "cross-validation" OR "C-statistic" OR "C-index" OR "ROC curve" OR "ROC" OR "area under curve" OR "AUC" OR "area under curve")	MeSH Terms, Title, Abstract, Keyword

Table 2: Description of the manuscripts related to the Rio and Modified Rio scores.

		THE RIO SCORE (RS) **Rio at al.**(4)			**THE MODIFIED RIO SCORE (MRS)** **Sormani et al.**(13)				
DEVELOPMENT	inclusion criteria	Prospective cohort 2003-2006, RRMS treated with IFN-β, available MRI 12 months after onset treatment			RRMS treated with IFN-β, more than one year of follow-up				
	design, size	prospective cohort, n=222			clinical trial, n=365				
	outcome	treatment failure at 24 months defined by relapse or confirmed disease progression, the latter being defined by EDSS progression ≥ 1 point sustained over at least 6 months and confirmed at the end of the follow-up			time to treatment failure defined by the presence of relapse or confirmed disease progression, the latter being defined by EDSS progression ≥ 1 point when J0 < 6 or progression ≥ 0.5 point when J0 ≥ 6 sustained over at least 6 months or switch to other therapies for lack of efficacy				
	landmark time	12 months after starting treatment			12 months after starting treatment				
	predictors	relapse, EDSS, active lesions			relapse, new T2 lesions				
	tool construction	arbitrary classification evaluated by a logistic model			arbitrary classification evaluated by Cox model				
	utility	to select potential candidates to receive alternative therapeutic approaches that may work better than IFN-β			to select potential candidates to receive alternative therapeutic approaches that may work better than IFN-β				
	limits	confusion between correlation and prediction no internal or external validation			no comparison with the previous Rio Score (4) no consideration of the right-censoring in the study of prognostic capacities				
EXTERNAL VALIDATION		**Romeo et al.**(12)	**Hyun et al.**(23)	**Rio et al.**(21)	**Sormani et al.**(13)	**Romeo et al.**(12)	**Hyun et al.**(23)	**Lattanzi et al.**(22)	**Rio et al.**(21)
	sample	prospective cohort, n=368	retrospective cohort, n=70	prospective cohort, n=233	cohort, n=222	prospective cohort, n=390	retrospective cohort, n=70	retrospective cohort, n=24	prospective cohort, n=233
	differences in outcome definition	disability 1: as in Rio (4) disability 2: EDSS progression ≥ 1.5 points when J0 < 2.5 and 1 point when J0 in 2.5-5.5 sustained over at least 6 months and confirmed at the end of the follow-up	disability: EDSS progression ≥ 1 point when J0 < 6 and 0.5 point when J0 ≥ 6 during the ensuring 2 years of IFN-β	not clearly defined: probably the first event between reaching EDSS at 7.5 or secondary progressive phase	disability 1 as in Rio (4) disability 2: EDSS progression ≥ 1.5 points when J0 < 2.5 and 1 point when J0 in 2.5-5.5 sustained over at least 6 months and confirmed at the end of the follow-up		disability: EDSS progression ≥ 1 points when J0 < 6 and 0.5 point when J0 ≥ 6 during the ensuring 2 years of IFN-β	disability: same as in Sormani et al. (13), but with a cut-off of EDDS at J0 at 5 instead 6	not clearly defined: probably the first event between reaching EDSS at 7.5 or secondary progressive phase
	global performance	accuracy = 62% (disability 1) and 65% (disability 2)	accuracy = 93%	accuracy = 77%	accuracy = 69%	accuracy = 65% (disability 1) and 69% (disability 2)	accuracy = 91%	accuracy = 79%	accuracy = 74%
	prognostic capacities	binary test: 0-1 versus 2-3 SE = 45% (disability 1) and 54% (disability 2) SP = 67% (disability 1) and 68% (disability 2)	binary test: 0-1 versus 2-3 SE = 98%, SP = 75% PPV=92%, VPN=93%	binary test: 0-1 versus 2-3 SE = 40%, SP = 86%, PPV = 43%, NPV = 85%	binary test: 0-1 versus 2-3 SE = 24%, SP = 97% PPV = 83%, NPV = 68%	binary test: 0-1 versus 2-3 SE = 42% (disability 1) and 51% (disability 2) SP = 72% (disability 1) and 72% (disability 2)	binary test: 0-1 versus 2-3 SE = 96%, SP = 75% PPV=86%, VPN=93%	binary test: 0-2 versus 3 SE = 50%, SP = 94% PPV = 80%, NPV = 79%	binary test: 0-1 versus 2-3 SE = 19%, SP = 88%, PPV = 29%, NPV = 81%
	limits/remark	no consideration of the right-censoring in prognostic capacities no calibration no PPV/NPV but can be obtained from survival curves	small sample size no consideration of the right-censoring in prognostic capacities no calibration uncomprehensive analyses in low-risk and high-risk subgroups	the landmark time is not clearly defined (1- or 2-years post-treatment) no consideration of the right-censoring	limits listed above the decision rule changes for validation poorly performed calibration	no PPV/NPV but can be obtained from survival curves no consideration of the right-censoring in prognostic capacities no calibration	small sample size no consideration of the right-censoring in prognostic capacities no calibration irrelevant analyses in low- and high-risk subgroups	small sample size no consideration of the right-censoring in prognostic capacities binary test with a different cut-off patients treated by Fingolimod (different landmark)	no consideration of the right-censoring the landmark time is not clearly defined (1- or 2-years post-treatment)

Table 2 (continued): Description of the manuscripts related to the MAGNIMS, the BREMS and the BREMSO scores.

		THE MAGNIMS SCORE **Sormani et al.**(14)	THE BREMS SCORE **Bergamashi et al.**(15)	THE BREMSO SCORE **Bergamashi et al.**(16)
DEVELOPMENT	inclusion criteria	RRMS treated with IFN-β as their first therapy, assessments of EDSS score, number of relapses and T2 lesions at therapy initiation and after 1 year, at least yearly clinical assessments, including EDSS score and number of relapses, for a minimum of 2 additional years	RRMS with disease duration ≥ 3 years, time between symptoms onset and first examination ≤ 12 months	RRMS patients (2001 McDonald's criteria)
	sample	clinical trial, n=1280	prospective cohort, n=186	prospective cohort, n=14211
	outcome	time to treatment failure defined by the presence of relapses or confirmed disease progression, the latter being defined by EDSS progression > 1 point when J0 < 6 or progression > 0.5 point when J0 > 6 or > 1.5 points when J0=0 sustained over at least 6 months or switch for lack of efficacy	time to onset of secondary progressive phase of the disease, defined by a persistent increase in at least one point in the EDSS level for 6 months	time to onset of secondary progressive phase of the disease as defined in the BREMS study (15) or to major clinical disability (EDS ≥ 6)
	landmark time	12 months after treatment start	12 months after onset of disease	12 months after onset of disease
	predictors	relapse, new T2 lesions	age, sex, sphincter onset, pure motor onset, motor and sensory onset, number of neurological functional systems involved at onset, incomplete recovery after onset	age, sex, sphincter onset, pure motor onset, motor and sensory onset, sequelae after onset, number of involved neurological functional systems at onset, number of sphincter plus motor relapses, EDSS ≥ 4 outside relapse
	methods	Cox model	Markov chain Monte Carlo Bayesian approach	update of the previous model (15)
	utility	to select potential candidates to receive alternative therapeutic approaches that may work better than IFN-β	inclusion criteria in clinical trial to select patients according to their expected disease course pattern surrogate endpoint in clinical trial	stratification of patients with a similar expected evolution to reduce confounders due to the lack of randomization in observational studies
	limits	small number of possible predictors no update of the score after 12 months no consideration of the right-censoring in the study of prognostic capacities different definition of the treatment failure compared the Rio score (4) and modified Rio score (13) no internal or external validation	small sample size no update of the score after 12 months no estimation of the apparent prognostic capacities no validation	no validation no update of the score after 12 months predicted outcome different from the initial BREMS study (15) no consideration of the right-censoring in the study of prognostic capacities
EXTERNAL VALIDATION		**Sormani et al.**(20)	**Bergamashi et al.**(24)	
	sample	clinical trial, n=551	prospective cohort, n=1245	
	differences in outcome definition	the disability worsening with no definition	same as previously (15), except a duration of 12 instead 6 months	
	global performance	accuracy = 63%	no	
	prognostic capacities	SE = 84%, SP = 24%, PPV = 67%, NPV = 45%	binary test: BREMS ≤ 5th percentile (value at 2) versus ≤ 5th percentile: SP = 100%, SE = 8%, PPV = 100%, NPV = 18% binary test: BREMS ≤ 95th percentile (value at -0.63) versus > 95th percentile: SP = 99%, SE = 17%, PPV = 86%, NPV = 83%	
	limits/remark	no calibration patients treated by Teriflunomide (different landmark) the use of a different scoring system compared to the initial proposal (14) reclassification of patients after baseline according the initial proposal (14)	no consideration of the right-censoring in prognostic capacities no calibration no relevance of the proposed extreme values of cut-off	

Table 2 (continued): Description of the manuscripts related to the four other scores with no name.

		Weinshenker et al.(19)	**Sormani et al.**(18)	**Achiron et al.**(25)	**Calabrese et al.**(17)
DEVELOPMENT	inclusion criteria	no precision	RRMS diagnosis for at least 6 months, EDSS score of 0.0 to 5.0, at least one documented relapse in the year before baseline, relapse-free and steroid free in the 30 days prior to baseline, complete clinical and MRI data at baseline, did not have to be treated with disease-modifying agents	RRMS patients treated with IFN-β, never received cytotoxic treatment, free of steroid treatment	RRMS patients (McDonald criteria 2001), at least 5 years of disease duration
	sample	prospective cohort, n=219	clinical trial, n=539	prospective cohort, n=19	prospective cohort, n=334
	outcome	time to reach EDSS at 6	time to first relapse	neurological disability (primary outcome) total number of relapses (secondary outcome)	secondary progressive phase of the disease
	landmark time	No	no	no	No
	predictors	Disease duration, EDSS, follow-up, progression index, other variables from a previous model	previous 2 years relapse, numbers of enhancing lesions	34 genes	age, cortical lesion volume, cerebellar cortical volume
	methods	logistic model	Cox model	Support Vector Machine	logistic model
	utility	not defined	definition of MS patients with high risk of relapse as inclusion criteria in clinical trials	not defined	not defined
	limits	no validation the follow-up is included in the model (conditioning on future)	no evaluation of the apparent discriminative capacities	small sample size training and validation dataset not clearly defined internal validation strategy not clear removal of patients with an intermediate outcome (conditioning on future) no available equation/algorithm/rule	no consideration of time-to-event in this mid-term study the high inter-observer and inter-center variability when heterogeneous scans the white matter variables were not analyzed
EXTERNAL VALIDATION			**Sormani et al.**(18)		**Calabrese et al.**(17)
	sample		clinical trial, n=117		prospective cohort, n=83
	differences in outcome definition		no difference		no difference
	global performances		no		accuracy = 92%
	prognostic capacities		binary test: score ≤ 95th percentile versus ≤ 5th percentile: predictive values can be obtained from the survival curves (up to PPV~85%, NPV~55%)		SE = 84%, SP = 94%
	limits/remark		identical than previously listed the calibration results were poor no evaluation of discriminative capacities		small sample size no calibration no threshold definition for computing SE and SP

9. Figures

Figure 1: Flowchart of the article selection process

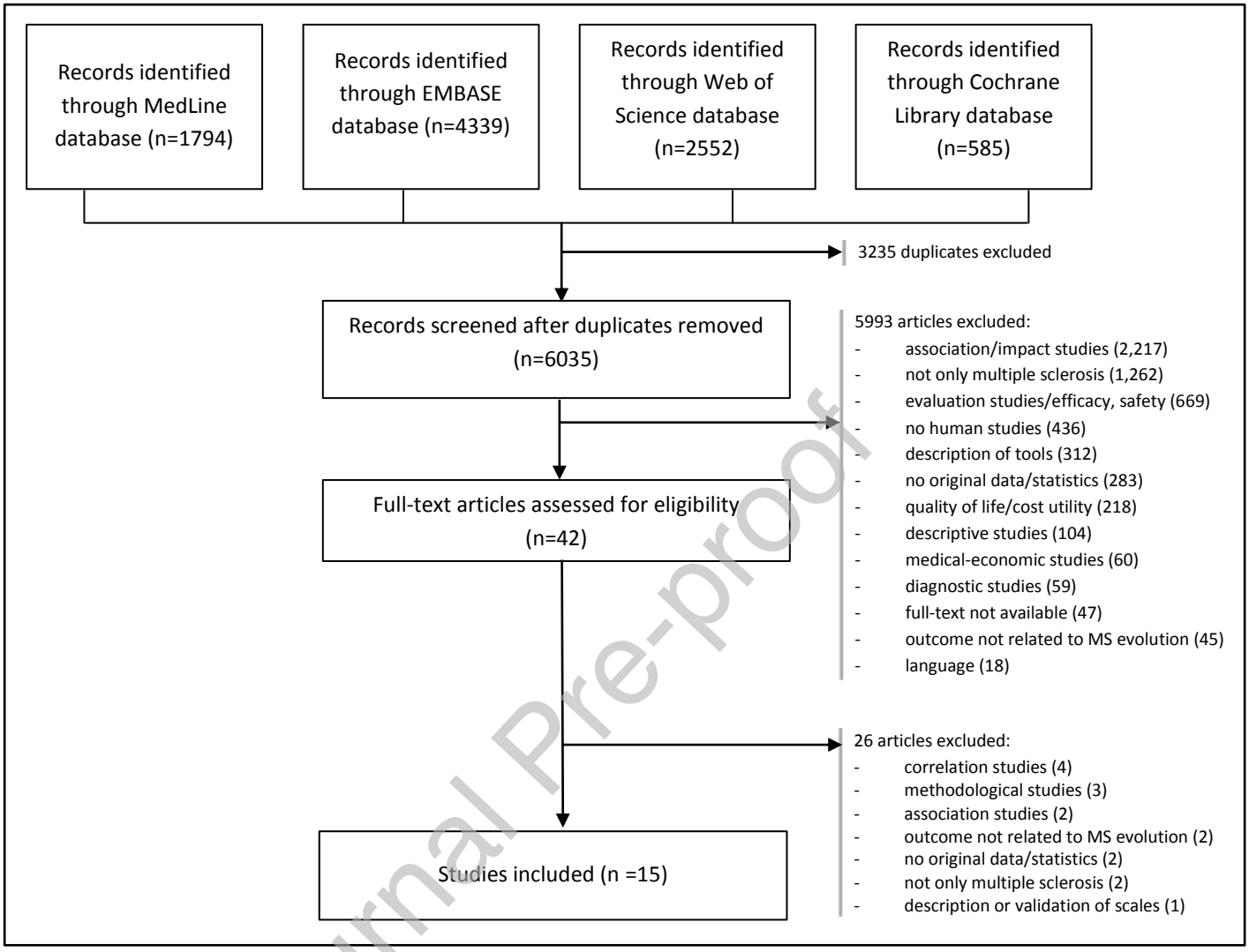

Figure 2: The distribution of the predictors according to the number of tools based on these parameters