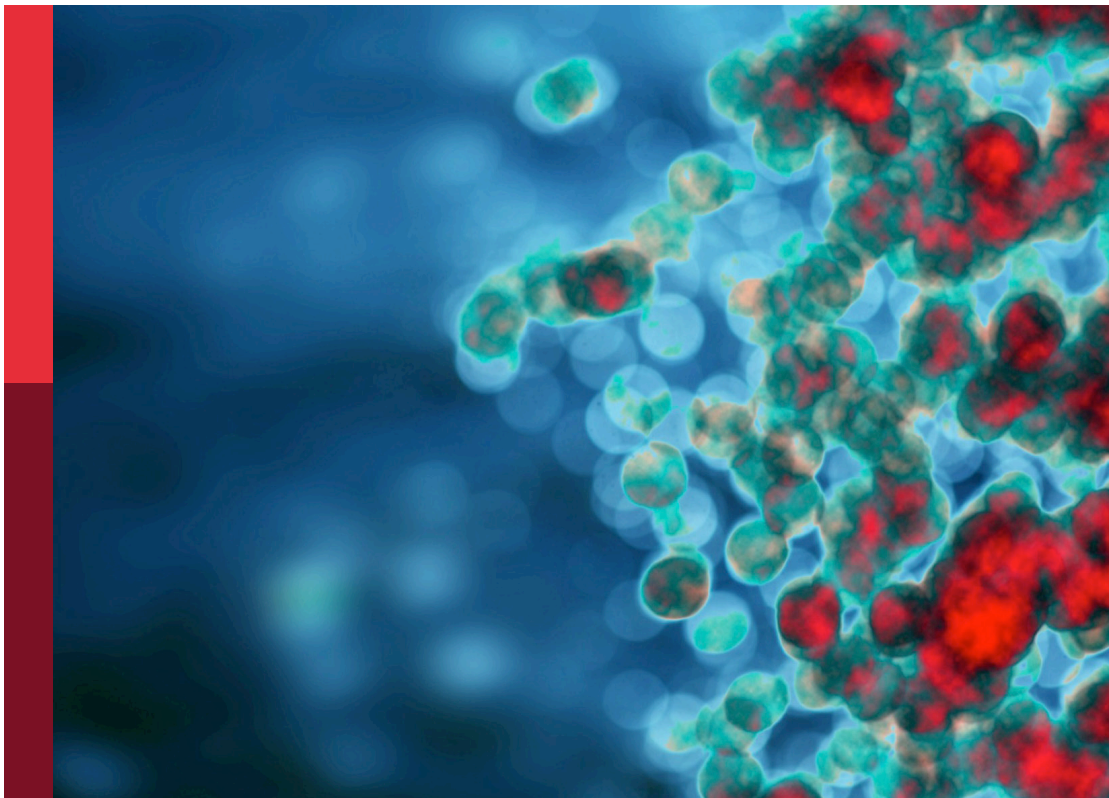
frontiers | Research Topics

Use of big data and artificial intelligence in multiple sclerosis

Edited by
Hans-Peter Hartung, Liesbet M. Peeters,
Giancarlo Comi and Axel Faes

Published in
Frontiers in Immunology
Frontiers in Neurology



ISSN 1664-8714
ISBN 978-2-8325-7054-8
DOI 10.3389/978-2-8325-7054-8

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Use of big data and artificial intelligence in multiple sclerosis

Topic editors

Hans-Peter Hartung — Heinrich Heine University, Germany
Liesbet M. Peeters — University of Hasselt, Belgium
Giancarlo Comi — San Raffaele Hospital (IRCCS), Italy
Axel Faes — University of Hasselt, Belgium

Citation

Hartung, H.-P., Peeters, L. M., Comi, G., Faes, A., eds. (2025). *Use of big data and artificial intelligence in multiple sclerosis*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-7054-8

Table of contents

frontiers | Frontiers in Immunology

TYPE Editorial
PUBLISHED 20 August 2025
DOI 10.3389/fimmu.2025.1679482

OPEN ACCESS

EDITED AND REVIEWED BY
Robert Weissert,
University of Regensburg, Germany

*CORRESPONDENCE
Liesbet M. Peeters
Liesbet.peeters@uhasselt.be

RECEIVED 04 August 2025
ACCEPTED 12 August 2025
PUBLISHED 20 August 2025

CITATION
Peeters LM, Faes A and Hartung H-P (2025) Editorial: Use of big data and artificial intelligence in multiple sclerosis.
Front. Immunol. 16:1679482.
doi: 10.3389/fimmu.2025.1679482

Editorial: Use of big data and artificial intelligence in multiple sclerosis

Liesbet M. Peeters[1,2,3]*, Axel Faes[1,2,3] and Hans-Peter Hartung[4,5,6]

[1]University MS Center (UMSC), Hasselt-Pelt, Belgium, [2]Biomedical Research Center (BIOMED), Hasselt University, Diepenbeek, Belgium, [3]Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium, [4]Department of Neurology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany, [5]Brain and Mind Center, University of Sydney, Sydney, NSW, Australia, [6]Department of Neurology, Palacky University Olomouc, Olomouc, Czechia

KEYWORDS

multiple sclerosis (MS), artificial intelligence, clinical decision support, magnetic resonance imaging, digital health monitoring, biomarkers, trustworthy machine learning

Editorial on the Research Topic
Use of big data and artificial intelligence in multiple sclerosis

Introduction

As health data volume and the sophistication of artificial intelligence (AI) tools grow, their potential to transform the management of complex neuroimmunological conditions like multiple sclerosis (MS) has become increasingly evident (1). MS, a chronic immune-mediated inflammatory disorder of the central nervous system, presents a unique challenge in the health sector due to its multifactorial nature and variable progression patterns. Each patient's journey is marked by distinct symptom trajectories and responses to treatment, demanding personalised approaches in diagnosis, prognosis, and therapeutic interventions. (2, 3).

This Special Topic aims to address the clinical complexity of MS by leveraging data driven insights and innovative health initiatives. The overarching goal is to present the current challenges in MS research and explore recent advances and future trends that can significantly impact patient care. Through a Research Topic of reviews, perspectives, and original research articles, we explore how advanced data techniques and innovative health initiatives are shaping the future of MS research and care.

Inspiring examples to showcase the potential

We kick-start with spotlighting a recently approved European Project, 'Clinical Impact through AI-assisted MS Care' (CLAIMS). Praet et al. explains how this project will develop, validate and seek regulatory approval for an AI-driven clinical decision-support platform, which offers the MS care team a holistic view of the patient through the visualisation of all

relevant patient data and the prognosis on the expected disease trajectories under different treatment regimens. Next to this, two original research contributions further illustrate AI's capacity to enhance MS treatment personalization and early diagnosis. Ilan et al. examine how advanced AI systems can help personalise and diversify treatment regimens, reducing the risk of drug tolerance. Meanwhile, Albuz et al. examine how AI-assessed volumetric measurements of specific brain regions correlate with neuropsychological test outcomes in patients with clinically isolated syndrome, illuminating potential early indicators of MS.

Magnetic resonance imaging and AI in MS

MRI remains central to diagnosing, monitoring, and optimising MS treatment due to its ability to non-invasively visualise both lesional and nonlesional brain pathology. However, the potential of MRI is often constrained in clinical practice by lengthy protocols, challenges in lesion identification, and limited predictive power regarding disability progression. Falet et al. highlight recent AI advances that could enhance MRI's accuracy and broaden its predictive capabilities, improving critical patient outcomes.

Digital tools and AI in MS

The integration of digital monitoring tools, big data, and AI presents new possibilities for real-time tracking of MS symptoms and progression. Dini et al. explore the latest advancements in digital remote monitoring, with devices like wearables and smartphones playing an increasing role in the field. These technologies, coupled with AI analytics, are demonstrating reliability in assessing motor symptoms such as fall risk and gait irregularities, both in clinical settings and through passive, real-life monitoring. While cognitive monitoring is still evolving, AI-driven tools are now beginning to automate neuropsychological test scoring and passive keystroke analysis, setting the stage for continuous, long-term data collection on both motor and cognitive symptoms.

Biomarkers and AI in MS

Expanding the scope to biological markers, Arrambide et al. delve into AI methodologies applied to serum, blood, and cerebrospinal fluid (CSF) biomarkers, outlining key studies, limitations, and future directions. Notably, this systematic review reveals that most research papers on AI applications to biomarker data in MS have been published within the past four years, underscoring that this field is still in its early stages and remains some distance from widespread clinical application.

Future trends

Recognizing the necessity of reliable and interpretable machine learning (ML) in MS, Werthen-Brabants et al. emphasise the need for Trustworthy ML. Given the complex and individualised nature of MS, these authors advocate collaborative efforts among researchers, clinicians, and policymakers to develop ML solutions that are technically robust, clinically relevant, and patient-centred.

Patient-reported outcome measures (PROMs) are vital for capturing the lived experiences of people with MS, providing insights that enrich clinical understanding. However, PROMs are underutilised in both clinical research and routine care. Helme et al. discuss the challenges in scaling PROMs and highlight efforts to integrate health outcomes data across Europe and beyond, noting initiatives like the European Health Data Space (EDHS) that may expand their application.

While the MS community has made substantial progress in leveraging data for research and patient care, several large-scale collaborative efforts across Europe—though not exclusively focused on MS—have the potential to transform the management and application of health data across various diseases, including MS. Peeters highlights key initiatives such as the EHDS, DARWIN-EU, the Observational Health Data Sciences and Informatics (OHDSI), EBRAINS, and ELIXIR. She outlines the challenges that remain in aligning with these initiatives and offers concrete, actionable recommendations to guide the MS research community toward more effective integration and collaboration.

Conclusion

We believe this special topic has opened new perspectives, and gives us some indications of where the field of Big Data and AI in MS is heading. First of all, it testifies that the domain is expanding rapidly. At the same time, however, researchers will have to solve some open issues, such as the need to develop trustworthy, reliable AI models, consistently capture multidimensional longitudinal data, incorporate the patient perspectives and the alignment with evolving regulatory frameworks such as the EHDS. We hope you find this Research Topic as inspiring and impactful to read as it was for us to prepare.

Author contributions

LP: Writing – original draft, Writing – review & editing. AF: Writing – original draft, Writing – review & editing. H-PH: Writing – original draft, Writing – review & editing.

In memoriam

In memoriam of Prof. Giancarlo Comi who helped to launch this project and like many other areas fertilized and promoted the field.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Author disclaimer

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

References

1. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. (2020) 395:1579–86. doi: 10.1016/S0140-6736(20)30226-9

2. Kuhlmann T, Moccia M, Coetzee T, Cohen JA, Correale J, Graves J, et al. Multiple sclerosis progression: time for a new mechanism-driven framework. *Lancet Neurol*. (2023) 22:78–88. doi: 10.1016/s1474-4422(22)00289-7

3. Jakimovski D, Bittner S, Zivadinov R, Morrow SA, Benedict RH, Zipp F, et al. Multiple sclerosis. *Lancet*. (2024) 403:183–202. doi: 10.1016/S0140-6736(23)01473-3

frontiers | Frontiers in Immunology

TYPE Perspective
PUBLISHED 19 August 2024
DOI 10.3389/fimmu.2024.1446748

OPEN ACCESS

EDITED BY
Yolanda Aladro,
European University of Madrid, Spain

REVIEWED BY
Giulia Sancesario,
Santa Lucia Foundation (IRCCS), Italy

*CORRESPONDENCE
Jelle Praet
jelle.praet@icometrix.com

†These authors have contributed equally to this work

RECEIVED 10 June 2024
ACCEPTED 11 July 2024
PUBLISHED 19 August 2024

CITATION
Praet J, Anderhalten L, Comi G, Horakova D, Ziemssen T, Vermersch P, Lukas C, van Leemput K, Steppe M, Aguilera C, Kadas EM, Bertrand A, van Rampelbergh J, de Boer E, Zingler V, Smeets D, Ribbens A and Paul F (2024) A future of AI-driven personalized care for people with multiple sclerosis. *Front. Immunol.* 15:1446748. doi: 10.3389/fimmu.2024.1446748

A future of AI-driven personalized care for people with multiple sclerosis

Jelle Praet[1*†], Lina Anderhalten[2†], Giancarlo Comi[3,4], Dana Horakova[5], Tjalf Ziemssen[6], Patrick Vermersch[7], Carsten Lukas[8], Koen van Leemput[9,10,11], Marjan Steppe[12], Cristina Aguilera[13], Ella Maria Kadas[14], Alexis Bertrand[15], Jean van Rampelbergh[16], Erik de Boer[17], Vera Zingler[18], Dirk Smeets[1], Annemie Ribbens[1†] and Friedemann Paul[2,19,20,21,22†] for CLAIMS consortium

[1]icometrix NV, Leuven, Belgium, [2]Experimental and Clinical Research Center (ECRC), A Cooperation Between the Max Delbrück Center for Molecular Medicine in the Helmholtz Association and Charité - Universitätsmedizin Berlin, Berlin, Germany, [3]Department of Neurorehabilitative Sciences, Casa di Cura Igea, Italy, [4]Department of Neurology, Vita-Salute San Raffaele University-Ospedale San Raffaele, Milan, Italy, [5]Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czechia, [6]Center of Clinical Neuroscience, Department of Neurology, University Clinic Carl Gustav Carus, TU Dresden, Dresden, Germany, [7]Univ. Lille, InsermU1172 LilNCog, CHU Lille, FHU Precise, Lille, France, [8]Institute of Neuroradiology, St. Josef Hospital, Ruhr-University Bochum, Bochum, Germany, [9]Athinoula A. Martinos Center, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, United States, [10]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland, [11]Department of Computer Science, Aalto University, Espoo, Finland, [12]European Charcot Foundation, Brussels, Belgium, [13]SYNAPSE Research Management Partners, Madrid, Spain, [14]Nocturne GmbH, Berlin, Germany, [15]AB Science, Clinical Development, Paris, France, [16]Imcyse SA, Liège, Belgium, [17]Bristol-Myers Squibb Company Corp, Princeton, NJ, United States, [18]F. Hoffmann-La Roche Ltd., Product Development Medical Affairs, Neuroscience, Basel, Switzerland, [19]Experimental and Clinical Research Center (ECRC), Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany, [20]Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany, [21]Neuroscience Clinical Research Center (NCRC), Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany, [22]Department of Neurology with Experimental Neurology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

Multiple sclerosis (MS) is a devastating immune-mediated disorder of the central nervous system resulting in progressive disability accumulation. As there is no cure available yet for MS, the primary therapeutic objective is to reduce relapses and to slow down disability progression as early as possible during the disease to maintain and/or improve health-related quality of life. However, optimizing treatment for people with MS (pwMS) is complex and challenging due to the many factors involved and in particular, the high degree of clinical and sub-clinical heterogeneity in disease progression among pwMS. In this paper, we discuss these many different challenges complicating treatment optimization for pwMS as well as how a shift towards a more pro-active, data-driven and personalized medicine approach could potentially improve patient outcomes for pwMS. We describe how the 'Clinical Impact through AI-assisted MS Care' (CLAIMS) project serves as a recent example of how to realize such a shift

towards personalized treatment optimization for pwMS through the development of a platform that offers a holistic view of all relevant patient data and biomarkers, and then using this data to enable AI-supported prognostic modelling.

KEYWORDS

multiple sclerosis, personalized medicine, disease progression, prognosis, diagnosis, AI, data

1 The heterogeneous disease course of multiple sclerosis

Multiple sclerosis (MS) is a devastating immune-mediated disorder of the central nervous system (CNS) resulting in progressive disability accumulation in most individuals affected (1, 2). MS imposes a significant burden on patients, affecting all aspects of their life, and additionally, it poses a significant challenge to society as with growing disability, indirect expenses (productivity losses associated with sick absence, inability to work, and early retirement) and care costs rise substantially (3).

The classical view on MS describes different clinical subtypes, with relapsing-remitting MS (RRMS) being the most common form, occurring in 85% of patients (National MS Society). Patients with RRMS experience neurological exacerbation (relapses) as well as intermittent periods of remission in which they remain clinically stable. Relapses can either recover completely or leave persistent clinical disability, referred to as Relapse Associated Worsening (RAW). Among these patients, approximately two-thirds progress to secondary-progressive MS (SPMS) (4). In contrast to RRMS, the disease course of patients with SPMS or primary-progressive MS (PPMS, 15% of MS patients) is mainly driven by a gradual worsening of disability in the absence of relapse activity (5).

Recent research has challenged this classical view of distinct MS subtypes, as they may not sufficiently account for the large spectrum of multifaceted clinical phenotypes and disease courses as well as sub-clinical disease variability (6). This disease heterogeneity is further complicated by a high prevalence of comorbidities and multi-pharmacy in MS. Data from the NARCOMS registry suggested that, at the time of MS diagnosis, 35% of MS patients suffer physical comorbidities while 18% reported a psychiatric comorbidity (7, 8). Additionally, accumulation of clinical disability independent of acute inflammatory relapses - commonly referred to as Progression Independent of Relapse Activity (PIRA) (9) - was found to occur in any of the classical MS subtypes, including RRMS, and at any stage of the disease (10, 11). Most importantly, in a substantial proportion of people with MS (pwMS), PIRA occurs already very early on, and this is associated with worse long-term outcomes (2). Recent studies have also shown that PIRA gradually becomes the dominant driver of disability worsening as the disease progresses (9).

While new insights into PIRA continue to be unraveled, exact criteria of how to define, assess, and monitor PIRA are still lacking. Several definitions have been put forward, but these focus mainly only on measuring disability worsening by means of the Expanded Disability Status Scale (EDSS) and Confirmed Disability Worsening (CDW) (2). Relying solely on EDSS or CDW to describe PIRA, however, seems to be insufficient as (i) there are heterogeneous symptoms and disease aspects contributing to disability worsening and MS severity, and (ii) this omits sub-clinical processes such as compartmentalized inflammation, chronically active (smouldering) lesions, diffuse normal-appearing matter damage (12, 13), as well as brain (14) and spinal cord atrophy (15, 16). Such processes seem to represent relevant substrates of (silent/smouldering) disease progression even during early stages and to contribute to enhanced long-term disability worsening in pwMS (17). In this regard, the topographical disease model proposed by Krieger et al. may facilitate the interpretation of the clinical course revision, providing a unified visualization across phenotypes, while providing insights in the interplay between the distinct processes of relapse activity and progression, and accounting for latent variables such as relapse localization, frequency, severity, recovery and progression rate (18). Additionally, this model was recently validated in terms of brain MRI markers (19). Aligning with this model, individuals deemed neurologically normal in early MS (e.g., with an EDSS score of 0) demonstrated subtle deficits in high-challenging motor tasks (20) and often have fatigue (21) and cognitive impairments (22). The former was also shown to correlate with imaging markers of disease burden and brain reserve, challenging traditional severity definitions and underscoring the importance of looking beyond standard clinical measures such as the EDSS (20).

2 A changing landscape in treatment strategies

The heterogeneity in disease progression among individuals with MS (both clinically and sub-clinically) contributes to a high

diversity in treatment responses across pwMS (23). As there is no cure available yet for MS, the primary therapeutic objective is to slow down disability progression and to reduce relapses as early as possible during the disease to maintain and/or improve the health-related quality of life (24).

To this end, all regulatory-approved disease-modifying treatments (DMT) have shown their worth in preventing relapses during the few years of the clinical trial in which their efficacy was evaluated. However, the impact on the long-term accumulation of disability and chronic subtle disease processes was often limited as even the most effective DMTs available were only able to mitigate the short-term risk of disability progression by 30-42% (25). A recent review from Gasperini et al. emphasizes how dire the situation really is, indicating that only 30- 40% of patients receiving a DMT remain stable over a period of 5 to 7 years, and only up to 10% over a period of 7 to 10 years after initiating DMT (26).

Despite the approval of ±20 different DMTs by the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) (27, 28), concerns about side effects and efficacy might discourage many pwMS from initiating a high-efficacy DMT therapy (29, 30), an issue further aggravated by therapeutic inertia (31). Additionally, those who do receive a DMT usually start with one of the less effective but well-established therapies due to their minimal side effects (32). Traditionally, it's only when these well-established DMTs fail to prevent relapses and disability progression, that the treatment is escalated to a higher-efficacy treatment, which usually is more expensive, might have more pronounced side effects, and is potentially more challenging to administer (oral and injectables versus infusions) (33). However, multiple studies support the observation that reducing the accrual of neurological damage in the initial stages of the disease potentially improves overall clinical outcomes throughout the patient's lifespan when employing early intervention with higher efficacy DMT (34–38). Additionally, DMTs were shown to be more efficacious, and side effects less likely to occur in younger patients (39). Taken together, these studies question the traditional treatment escalation paradigm which is therefore nowadays considered outdated by most physicians. Instead, current thinking emphasizes the potential advantages of early initiation of high-efficacy DMTs, indicating the need for and the significance of an early MS diagnosis, proactive monitoring to detect disease activity early, and shared decision-making as crucial elements in patient care (32, 40).

Additionally, given the shortcomings of current DMTs to halt long-term disability accumulation, a next generation of DMTs might focus more on the silent progression of the disease. A first novel category of DMTs in this regard are potentially the Bruton tyrosine kinase inhibitors. This new class of drugs might become the first to target both acute inflammatory relapses as well chronic inflammatory processes in the CNS thought to drive disability accumulation (41). In this context, especially the early recognition of individuals prone to developing PIRA will be essential. A better understanding of PIRA and RAW as well as their interplay, combined with data-driven prognosis, will enhance the selection of current and future DMTs and allow to treat patients beyond just relapse activity. Nevertheless, certain variables pose challenges to the trajectory of precision medicine and treatment optimization on an individual level. While there are guidelines on the use of DMTs in MS (24), these are all based on expert judgment and differ across countries, even within the EU (28, 37). This variance extends to therapy selection post-diagnosis or during follow-ups, driven by perceived levels of clinical and subclinical disease activity and progression.

3 Precision medicine enables treatment optimization

Accumulating evidence suggests that the reactive treatment of lesion activity is insufficient, negatively impacting long-term patient outcomes (42). In the complex landscape of MS treatment, an increasing acknowledgment of disease heterogeneity and underlying disease mechanisms underscores the imperative for a paradigm shift toward proactive, data-driven precision medicine (43). However, despite its promise, such data-driven approaches come hand in hand with substantial challenges.

The understanding of the complex and heterogeneous underlying neuropathology of MS is still limited. The adoption of precision medicine in MS is further complicated by the chronic nature of the disease, exhibiting variable courses over time. Consequently, given the longitudinal disease aspect, one must account for the fact that data might be incomplete at times, particularly in routine practice. In addition, the influence of comorbidities adds another layer of complexity (44). Various biomarkers are deemed relevant for their role in identifying diverse MS aspects and patterns of progression in MS, aiding diagnosis, prognosis, and treatment selection (45). However, they might not capture the full complexity of MS and their interpretation requires a nuanced understanding of the disease context. Moreover, the heterogeneous nature of MS challenges the development of universally applicable biomarkers and complicates the tracking of different treatment effects on an individual basis (46).

Notably, with a variety of treatment options being available (27, 28), emerging biomarkers, including liquid and imaging markers, have shown potential in monitoring treatment efficacy (45, 47). However, the validation, availability, and implementation of biomarker assessments in real-world clinical practice is often still missing as this differs significantly from their application in clinical trials. Moreover, biomarkers that demonstrate both sensitivity and specificity in the context of progressive MS are still lacking (47). While early diagnosis and prognosis modelling are pivotal for timely and effective treatment initiation, the ability to clearly define and disentangle disability accumulation attributed to RAW or PIRA will be key to optimizing individual treatment over the course of the disease.

Advancements in artificial intelligence (AI) can offer enhanced and data-driven support by considering longitudinal data on multiple biomarkers simultaneously and subtyping patients more accurately. In particular, this can include biomarkers more related to PIRA such as motor dysfunction beyond EDSS (2, 48), optical

coherence tomography (49–51), magnetic resonance imaging markers predictive of disability worsening such as brain atrophy (14), slowly expanding lesions and paramagnetic rim lesions (52–54) and cognitive impairment (55–57), as well as subjective markers [i.e. patient-reported outcomes (PROs) such as quality of live (58, 59)]. We believe that a holistic overview of the patient will be crucial to avoid overlooking relevant information, including both existing and new biomarkers as our disease understanding evolves further (Figure 1).

Such transformative approaches hold the potential to significantly enhance treatment strategies and extend the adjusted quality of life years for individuals with MS. Nevertheless, the current landscape is still fragmented, often focusing on singular aspects or biomarkers rather than adopting a more holistic and

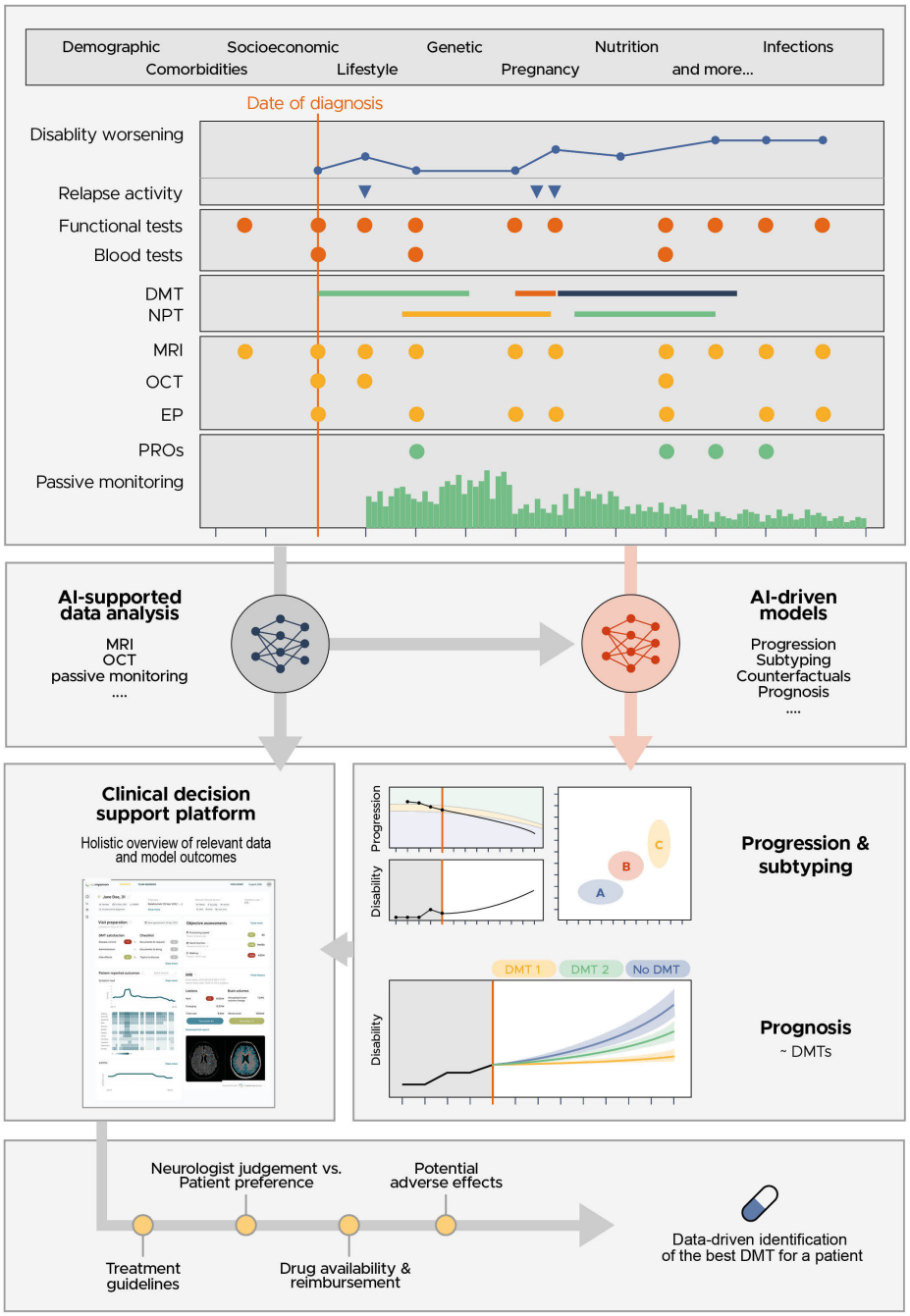


FIGURE 1
A clinical decision support tool should be capable of visualizing the very heterogenous MS patient data, the AI-supported analysis of this data and the outcome of prognostic models using this data, enabling a data-driven discussion between the neurologist and patient to identify the best DMT for the patient.

comprehensive approach. Data strategies to reduce the level of heterogeneity, particularly improving data harmonization by means of a common data model, are wishful to guarantee standardization in clinical decision making (60). However, the implementation of such initiatives is still in the early stages. Care pathways for pwMS are also not commonly standardized and while some diagnostic and treatment guidelines and recommendations are available (1, 61, 62), the assessment of relevant outcomes may not always be sufficiently covered and integrated into the routine clinical workflow (63). A modular-integrative framework of digital patient pathways for MS management and treatment is needed, which should incorporate AI, data harmonization and review relevant research concerning the use of pathways in healthcare (64, 65). Although initial evidence of acting upon AI-driven MRI biomarkers has indicated to improve patient outcome (66), the evaluation of impact in real-world practice and evidence on whether acting upon data-driven models and biomarkers truly improves the quality of life for patients with MS are crucial components that demand more attention in the pursuit of effective precision medicine strategies for MS.

4 Clinical impact through AI-assisted MS care

A data-driven and personalized clinical decision support tool is urgently needed for MS, to prevent and slow down disease progression more efficiently via optimizing treatment. The EU-funded 'Clinical Impact through AI-assisted MS Care' (CLAIMS, www.claims.ms) project aims to address this need. The project will develop, validate and seek regulatory approval for an AI-driven clinical decision-support platform, which offers the MS care team a holistic view of the patient through the visualization of all relevant patient data and the prognosis on the expected disease trajectories under different treatment regimens.

Initially, the project focusses on the development and optimization of these prognostic models via the use of retrospectively collected clinical routine data in combination with clinical trial data. A detailed description of this retrospective multi-center observational study (called RECLAIM) is accessible via ClincialTrials.gov. This study aims to collect and harmonize both clinical and subclinical data and store it in a central database on a secure cloud environment. Data harmonization will be following the common data model proposed in Parciak et al. (67), but kept to the minimum necessary as we aim to stay as close as possible to the real-world clinical setting and to ensure the clinical relevance.

The combination of real-world with clinical trial data is an important aspect of the study. Clinical trial data is very homogeneous and highly curated, making it an ideal dataset to develop AI-driven prognostic models. For instance, MRI scans obtained in clinical trials adhere to a standardized protocol, include all necessary sequences, and ensure follow-up scans within a specific timeframe. In contrast, MRI scans acquired in a real-world setting frequently don't meet these requirements (68, 69). As the CLAIMS project aims to create AI-based prediction models applicable in real-world clinical settings, it is crucial to also incorporate routine care data in the development and validation phases. By combining both types of data, we aim to achieve an extensive dataset that leverages the strengths of both types of data ensuring applicability in a routine clinical care setting where confounding factors (e.g., comorbidities), low quality data and missing data are common (70, 71).

The focus will be on modelling disease progression. Disease progression models often have strong assumptions about the monotonicity of disease progression processes, the missingness model and associated completeness of the data, the longitudinal regularity of the observations, and homoscedastic noise characteristics of the measurements. Due to the different MS subtypes, and relapse and recurrence events, many of these assumptions do not hold in a MS setting. Furthermore, when using clinical observational data, data points are missing-not-at-random, both because patients often miss their appointments, but also because certain examinations (clinical assessments, MRI, etc) are performed as a function of patient presentation. Tackling this requires us to explore applicability of advanced and appropriate models of data imputation, and from generative models that explicitly model the causal relationships of the observations.

Contrary to clinical research trials where patients are assigned to a treatment or placebo arm at random, in an observational setting, DMTs are given to patients according to guideline recommendations and patient presentation. Observational data is thus biased by these guidelines, and appropriate measures are needed to control for this bias. Causal inference mechanisms via counterfactuals allows one to model such observational data and predict what the potential outcome would have been under a counterfactual treatment. By disentangling causes and effects, one gains a clearer understanding of the underlying biological or pathological markers that are predictive of the observed effect and outcome. This enables a more grounded clustering of patients (e.g., what are the patient characteristics that predict drug efficacy), providing an explanation of the optimal therapeutic inference (e.g., what is the biological reason why a certain drug is optimal for a specific patient). While some of these challenges have been addressed in highly controlled randomized clinical research environments, solving them using an observational experimental setup would allow one to exploit large amounts of data while ensuring the models remain accurate when deployed in a real-world environment where the aforementioned problems exist. Observational studies using real-world data allow for more heterogeneous and comprehensive cohorts, thereby elevating external validity and supplying valuable insights to guide treatment approaches (69).

At the time of writing this paper, the first version of the CLAIMS platform was already available, building upon a regulatory cleared AI solution for brain MRI quantification, a patient app for pwMS and a regulatory cleared AI solution for optical coherence tomography (OCT) quantification (72–75), but without the prognostic models (Figure 2). The complete clinical decision support platform, including the prognostic models, will be included in prospective clinical trial (called PROCLAIM), designed to obtain regulatory approval, and bringing it to the market as soon as possible. Meanwhile, the platform will be iteratively improved as

10.3389/fimmu.2024.1446748

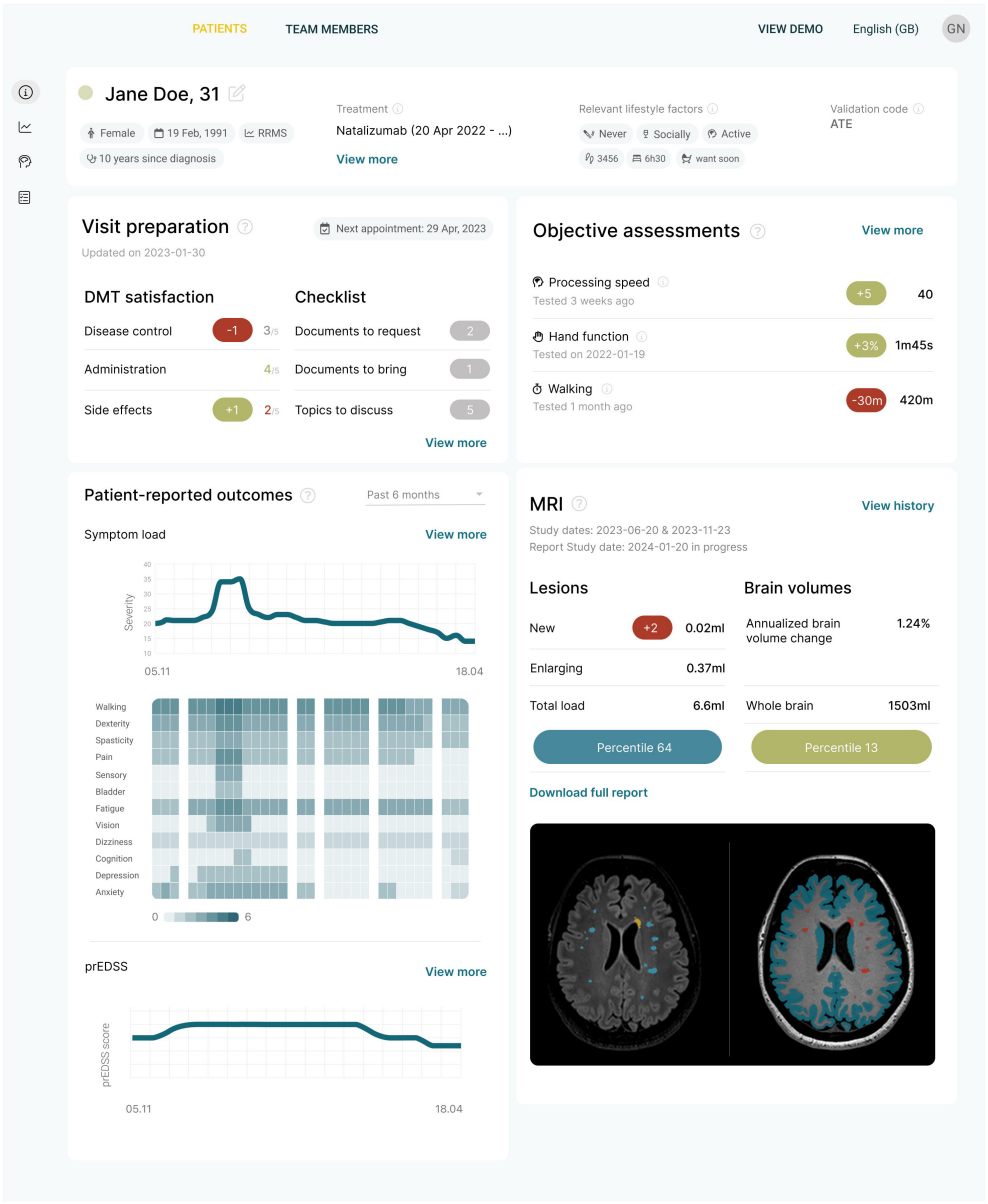


FIGURE 2
The first iteration of the clinical decision support platform being developed in the CLAIMS project. It offers a concise overview of the most important data for making a clinical decision.

new biomarker data becomes available and models are further refined. This iterative approach ensures that the CLAIMS project achieves true clinical impact for patients sooner rather than later.

5 Digital health and how this support prognosis

The CLAIMS project is exploring an additional avenue for the identification of promising markers of disease progression by capturing digital biomarkers using digital health tools. A first set of digital health tools includes AI solutions tailored for the quantification of brain MRI scans (74, 75). Notable advancement of these tools' accuracy, in combination with rigorous technological, workflow, clinical and even initial health economic validation makes that this solution steadily gains recognition as standard of care. In the United States, this trend towards embracing AI-based brain MRI quantification is further exemplified by the recent provision of two new Current Procedural Terminology (CPT) codes. Evidence has shown that by using such a solution, disease activity can be detected up to 3 years earlier with a potentially significant impact on treatment decisions (66).

Patient apps, another major trend in the digital health tools, could enhance the early detection of disease progression in pwMS and allow monitoring disease progression in between visits with their treating physician. This can be achieved by monitoring symptoms and disability progression through capturing patient-reported outcomes (PROs), through passive monitoring of various

markers (activity, sleep, vital signs, …) or through the digital administration of tests assessing for example cognition, vision, mobility, etc. (76, 77). In addition, these tools can play an important role in increasing and monitoring medication adherence, improving a patient's lifestyle through creating awareness, and to educate and empower patients in managing their disease better. As such, disease monitoring via digital health tools provides a dynamic, more continuous, and more nuanced understanding of disease progression.

Development of such tools poses a socio-technical challenge. Any tool which aims to obtain regulatory clearance for use in a clinical setting will need to obtain sufficient technical and clinical evidence, which is often a long and laborious process. A bigger challenge, however, is patient adoption and thereafter adherence in using the tools. Concerns on data security and privacy need to be adequately addressed and simultaneously, it needs to be very clear to patients that they will benefit from enhanced care and personalized interventions driven by a more holistic understanding and monitoring of their health status and disease progression. CLAIMS aims to address this by empowering and educating patients on the need to better monitor their disease. In this light, the patient app used in CLAIMS is positioned as a companion app, available to support the patient as needed, focusing on topics of interest to the patient, rather than mandating the app usage. Actively involving patients and capturing their feedback on the app utilization, whether via real-world usage or within a clinical study setting, will contribute valuable insights, allowing to further refine the tools and ultimately, the clinical decision support platform.

Besides patient adherence, integration into routine clinical workflows poses another challenge. To address this, the clinical decision support platform in CLAIMS aims to keep the steps of platform adoption to a bare minimum. It aggregates all of a patient's data, including data from the patient app, from the AI-driven MRI analysis and from the AI-driven OCT-analysis. While the full datasets and analyses will be available via this platform, the main dashboard focusses on providing a holistic overview of all clinically actionable measures and markers. While this is rather straightforward for subjective and episodic data such as with questionnaires or simple tests captured via the patient app, this will be harder to achieve for data from passive monitoring. The latter is known to generate large longitudinal datasets where AI algorithms are needed to identify subtle patterns and disease subtypes, and to predict trajectories.

Patient-reported outcomes (PROs) represent a unique occasion to involve patients using digital health tools and measure the impact of health care on outcomes that hold utmost significance to pwMS. However, the variety of PRO measures available and the absence of standards across different healthcare centers and countries present a considerable challenge (58). The recently established initiative 'Patient-Reported Outcomes for Multiple Sclerosis' (PROMS), consisting of an interdisciplinary, international network of different stakeholders, addresses the challenge of creating PRO measures that meet the diverse needs of all parties involved to enhance the influence of both scientific research and patient perspectives on the lives of pwMS (59). In this context, digital health tools enable meaningful assessments, but patient satisfaction can influence assessment compliance and indirectly affect outcome measures. To assess patient satisfaction with digital tools, patient-reported and expert-reported experience measures (PREM) should be collected in parallel (78).

6 The road ahead

As our understanding of MS increases, it becomes evident that we should go beyond making treatment decisions solely based on relapses, EDSS progression and lesion activity and move towards proactively treating pwMS for the best possible prognostic outcome. A focus on maintaining/improving health-related quality of life and slowing down disease progression and disability worsening - also independent of relapse activity - has sprouted a clear need for data-driven and personalized clinical decision support tools in MS. Such tools are crucial to administer the right drug to the right patient at the right time to preserve long-term neurological function while minimizing side effects. However, such solutions require well validated biomarkers and models that clearly link to the specificity of the disease course and outcome at individual patient level and can be easily implemented along the clinical care path of the patient.

The CLAIMS project aims to develop such a data-driven and personalized clinical decision support tool while addressing the posed challenges. Biomarker validation and model building will be performed in the retrospective RECLAIM study using both real world data and data from clinical trials. Subsequently, the prospective PROCLAIM study will evaluate the envisioned platform in daily clinical routine, evaluating feasibility and impact on patient care pathways and patient outcome. As such the project will generate a platform for daily clinical routine that provides a holistic view of each patient including existing and novel biomarker assessments to better monitor relapse related disability worsening and progression independent of relapse activity. Driven by deep-learning-based disease subtyping and progression models, the platform will allow the estimation of individual disease trajectories and as such contribute to the urgent need of a more pro-active and data-driven precision medicine in MS care.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

JP: Conceptualization, Funding acquisition, Project administration, Visualization, Writing – original draft, Writing – review & editing. LA: Conceptualization, Project administration, Writing – original draft, Writing – review & editing. GC: Conceptualization, Writing – review

& editing. DH: Conceptualization, Writing – review & editing. TZ: Conceptualization, Writing – review & editing. PV: Conceptualization, Writing – review & editing. CL: Conceptualization, Writing – review & editing. KV: Conceptualization, Writing – review & editing. MS: Conceptualization, Writing – review & editing. CA: Conceptualization, Funding acquisition, Project administration, Writing – review & editing. EK: Conceptualization, Writing – review & editing. AB: Conceptualization, Writing – review & editing. JV: Conceptualization, Writing – review & editing. ED: Conceptualization, Writing – review & editing. VZ: Conceptualization, Writing – review & editing. DS: Conceptualization, Funding acquisition, Writing – review & editing. AR: Conceptualization, Funding acquisition, Project administration, Visualization, Writing – original draft, Writing – review & editing. FP: Conceptualization, Funding acquisition, Project administration, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The CLAIMS project is supported by the Innovative Health Initiative Joint Undertaking (JU) under grant agreement No 101112153. The JU receives support from the European Union's Horizon Europe research and innovation program and COCIR, EFPIA, EuropaBio, MedTech Europe, Vaccines Europe, AB Science SA and Icometrix NV. This work was partially supported by an ITEA grant (20030 HeKDisco, HBC.2021.0500) from Flanders Innovation and Entrepreneurship. DH has received support from the Charles University Cooperation Program in Neuroscience, from the National Institute for Neurological Research (Program EXCELES, ID Project No. LX22NPO5107), from the European Union –Next Generation EU, and from the General University Hospital in Prague (project MH CZ-RVO-VFN64165).

Conflict of interest

JP is a shareholder of icometrix NV. AR is a shareholder of icometrix NV. DS is a shareholder of icometrix NV. MS has no relevant or material financial interests that relate to the research described in this paper. However, she is employed as Operational Director of The European Charcot Foundation. TZ reports scientific advisory board and/or consulting for Biogen, Roche, Novartis, Celgene, and Merck; compensation for serving on speaker's bureaus for Roche, Novartis, Merck, Sanofi, Celgene, and Biogen; and research support from Biogen, Novartis, Merck, and Sanofi. VZ is a shareholder of F. Hoffmann-La Roche Ltd PV reports honorarium, contributions to meeting from Biogen, Sanofi-Genzyme, Novartis, Teva, Merck, Roche, Imcyse, AB Science, Janssen, Ad Scientiam and BMS-Celgene. Research support from Novartis, Sanofi-Genzyme and Merck. EK is a shareholder of Nocturne GmbH. DH received compensation for travel, speaker honoraria and consultant fees from Biogen Idec, Novartis, Merck, Bayer, Sanofi Genzyme, Roche, and Teva, as well as support for research activities from Biogen Idec. JR is a shareholder of Imcyse SA. GC has received consulting and speaking fees from Novartis, Sanofi, Janssen, Bristol Myers Squibb, Roche and Rewind. FP provided research support to Neurosciences Clinical Research Center, German Ministry for Education and Research (BMBF), Deutsche Forschungsgemeinschaft (DFG), Einstein Foundation, Guthy Jackson Charitable Foundation, EU FP7 Framework Program, Biogen, Genzyme, Merck Serono, Novartis, Bayer, Roche, Parexel and Almirall, received honoraria for lectures, presentations, speakers from Guthy Jackson Foundation, Bayer, Biogen, Merck Serono, Sanofi Genzyme, Novartis, Viela Bio, Roche, UCB, Mitsubishi Tanabe and Celgene, in addition, received compensation for serving on a scientific advisory board of Celgene, Roche, UCB and Merck, and is an Academic Editor of PLos One and Associate Editor of Neurology® Neuroimmunology & Neuroinflammation, all unrelated to this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* (2018) 17:162–73. doi: 10.1016/s1474-4422(17)30470-2

2. Tur C, Carbonell-Mirabent P, Cobo-Calvo A, Otero-Romero S, Arrambide G, Midaglia L, et al. Association of early progression independent of relapse activity with long-term disability after a first demyelinating event in multiple sclerosis. *JAMA Neurol.* (2023) 80:151. doi: 10.1001/jamaneurol.2022.4655

3. Kobelt G, Berg J, Lindgren P, Jönsson B. Costs and quality of life in multiple sclerosis in Europe: method of assessment and analysis. *Eur J Health Econ.* (2006) 7:5–13. doi: 10.1007/s10198-006-0365-y

4. Tutuncu M, Tang J, Zeid NA, Kale N, Crusan DJ, Atkinson EJ, et al. Onset of progressive phase is an age-dependent clinical milestone in multiple sclerosis. *Multiple Sclerosis.* (2013) 19:188–98. doi: 10.1177/1352458512451510

5. Lublin FD, Reingold SC, Cohen JA, Cutter GR, Sørensen PS, Thompson AJ, et al. Defining the clinical course of multiple sclerosis. *Neurology.* (2014) 83:278–86. doi: 10.1212/wnl.0000000000000560

6. Kuhlmann T, Moccia M, Coetzee T, Cohen JA, Correale J, Graves J, et al. Multiple sclerosis progression: time for a new mechanism-driven framework. *Lancet Neurol.* (2023) 22:78–88. doi: 10.1016/s1474-4422(22)00289-7

7. Marrie RA, Horwitz RI, Cutter G, Tyry T, Vollmer T. Association between comorbidity and clinical characteristics of MS. *Acta Neurol Scand.* (2011) 124:135–41. doi: 10.1111/j.1600-0404.2010.01436.x

8. Marrie RA, Cohen JA, Stuve O, Trojano M, Sørensen PS, Reingold S, et al. A systematic review of the incidence and prevalence of comorbidity in multiple sclerosis: Overview. *Multiple Sclerosis.* (2015) 21:263–81. doi: 10.1177/1352458514564491

9. Lublin FD, Häring DA, Ganjgahi H, Ocampo A, Hatami F, Čuklina E, et al. How patients with multiple sclerosis acquire disability. *Brain*. (2022) 145:3147–61. doi: 10.1093/brain/awac016

10. Kappos L, Wolinsky JS, Giovannoni G, Arnold DL, Wang Q, Bernasconi C, et al. Contribution of relapse-independent progression vs relapse-associated worsening to overall confirmed disability accumulation in typical relapsing multiple sclerosis in a pooled analysis of 2 randomized clinical trials. *JAMA Neurol*. (2020) 77:1132. doi: 10.1001/jamaneurol.2020.1568

11. Portaccio E, Bellinvia A, Fonderico M, Pastò L, Razzolini L, Totaro R, et al. Progression is independent of relapse activity in early multiple sclerosis: a real-life cohort study. *Brain*. (2022) 145:2796–805. doi: 10.1093/brain/awac111

12. Lassmann H. Targets of therapy in progressive MS. *Multiple Sclerosis*. (2017) 23:1593–9. doi: 10.1177/1352458517729455

13. Lassmann H. Pathogenic mechanisms associated with different clinical courses of multiple sclerosis. *Front Immunol*. (2019) 9:3116. doi: 10.3389/fimmu.2018.03116

14. Cagol A, Schaedelin S, Barakovic M, Benkert P, Todea RA, Rahmanzadeh R, et al. Association of brain atrophy with disease progression independent of relapse activity in patients with relapsing multiple sclerosis. *JAMA Neurol*. (2022) 79:682. doi: 10.1001/jamaneurol.2022.1025

15. Bischof A, Papinutto N, Keshavan A, Rajesh A, Kirkish G, Zhang X, et al. Spinal cord atrophy predicts progressive disease in relapsing multiple sclerosis. *Ann Neurol*. (2022) 91:268–81. doi: 10.1002/ana.26281

16. Cagol A, Benkert P, Melie-Garcia L, Schaedelin SA, Leber S, Tsagkas C, et al. Association of spinal cord atrophy and brain paramagnetic rim lesions with progression independent of relapse activity in people with MS. *Neurology*. (2024) 102:e207768. doi: 10.1212/wnl.0000000000207768

17. UCA San Francisco MS-EPIC Team, Bruce ACC, Hollenbach JA, Bove R, Kirkish G, Sacco S, et al. Silent progression in disease activity–free relapsing multiple sclerosis. *Ann Neurol*. (2019) 85:653–66. doi: 10.1002/ana.25463

18. Krieger SC, Cook K, De Nino S, Fletcher M. The topographical model of multiple sclerosis. *Neurology® Neuroimmunol Neuroinflamm*. (2016) 3:e279. doi: 10.1212/nxi.0000000000000279

19. Krieger SC, Billiet T, Maes C, Barros N, Ribbens A, Wang C, et al. MSMilan2023 – paper poster - session 1' (2023). *Multiple Sclerosis*. (2023) 29:137–393. doi: 10.1177/13524585231196192

20. Krieger SC, Antoine A, Sumowski JF. EDSS 0 is not normal: Multiple sclerosis disease burden below the clinical threshold. *Multiple Sclerosis*. (2022) 28:2299–303. doi: 10.1177/13524585221108297

21. Runia TF, Jafari N, Siepman DAM, Hintzen RQ. Fatigue at time of CIS is an independent predictor of a subsequent diagnosis of multiple sclerosis. *J Neurol Neurosurg Psychiatry*. (2015) 86:543–6. doi: 10.1136/jnnp-2014-308374

22. Paul F. Pathology and MRI: exploring cognitive impairment in MS. *Acta Neurol Scand*. (2016) 134:24–33. doi: 10.1111/ane.12649

23. Kalincik T, Manouchehrinia A, Sobisek L, Jokubaitis V, Spelman T, Horakova D, et al. Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. *Brain*. (2017) 140:2426–43. doi: 10.1093/brain/awx185

24. Montalban X, Gold R, Thompson AJ, Otero-Romero S, Amato MP, Chandraratna D, et al. ECTRIMS/EAN Guideline on the pharmacological treatment of people with multiple sclerosis. *Multiple Sclerosis J*. (2018) 24:96–120. doi: 10.1177/1352458517751049

25. Wingerchuk DM, Carter JL. Multiple sclerosis: current and emerging disease-modifying therapies and treatment strategies. *Mayo Clin Proc*. (2014) 89:225–40. doi: 10.1016/j.mayocp.2013.11.002

26. Gasperini C, Prosperini L, Tintoré M, Sormani MP, Filippi M, Rio J, et al. Unraveling treatment response in multiple sclerosis. *Neurology*. (2019) 92:180–92. doi: 10.1212/wnl.0000000000006810

27. Amin M, Hersh CM. Updates and advances in multiple sclerosis neurotherapeutics. *Neurodegenerative Dis Manage*. (2023) 13:47–70. doi: 10.2217/nmt-2021-0058

28. Bayas A, Christ M, Faissner S, Klehmet J, Pul R, Skripuletz T, et al. Disease-modifying therapies for relapsing/active secondary progressive multiple sclerosis – a review of population-specific evidence from randomized clinical trials. *Ther Adv Neurol Disord*. (2023) 16:175628642211468. doi: 10.1177/17562864221146836

29. Visser LH, Van Der Zande A. Reasons patients give to use or not to use immunomodulating agents for multiple sclerosis. *Eur J Neurol*. (2011) 18:1343–9. doi: 10.1111/j.1468-1331.2011.03411.x

30. Jokubaitis VG, Spelman T, Lechner-Scott J, Barnett M, Shaw C, Vucic S, et al. The Australian Multiple Sclerosis (MS) Immunotherapy Study: A prospective, multicentre study of drug utilisation using the MSBase platform. *PloS One*. (2013) 8: e59694. doi: 10.1371/journal.pone.0059694

31. Saposnik G, Andhavarapu S, de la Maza SS, Castillo-Triviño T, Borges M, Barón BP, et al. Delayed cognitive processing and treatment status quo bias in early-stage multiple sclerosis. *Multiple Sclerosis Related Disord*. (2022) 68:104138. doi: 10.1016/j.msard.2022.104138

32. Giovannoni G, Butzkueven H, Dhib-Jalbut S, Hobart J, Kobelt G, Pepper G, et al. Brain health: time matters in multiple sclerosis. *Multiple Sclerosis Related Disord*. (2016) 9:S5–S48. doi: 10.1016/j.msard.2016.07.003

33. Inojosa H, Proschmann U, Akgün K, Ziemssen T. The need for a strategic therapeutic approach: multiple sclerosis in check. *Ther Adv Chronic Dis*. (2022) 13:204062232110630. doi: 10.1177/20406223211063032

34. Harding K, Williams O, Willis M, Hrastelj J, Rimmer A, Joseph F, et al. Clinical outcomes of escalation vs early intensive disease-modifying therapy in patients with multiple sclerosis. *JAMA Neurol*. (2019) 76:536. doi: 10.1001/jamaneurol.2018.4905

35. Buron MD, Chalmer TA, Sellebjerg F, Barzinji I, Bech D, Christensen JR, et al. Initial high-efficacy disease-modifying therapy in multiple sclerosis. *Neurology*. (2020) 95:e1041–51. doi: 10.1212/wnl.0000000000010135

36. Simpson A, Mowry EM, Newsome SD. Early aggressive treatment approaches for multiple sclerosis. *Curr Treat Options Neurol*. (2021) 23:19. doi: 10.1007/s11940-021-00677-1

37. Spelman T, Magyari M, Piehl F, Svenningsson A, Rasmussen PV, Kant M, et al. Treatment escalation vs immediate initiation of highly effective treatment for patients with relapsing-remitting multiple sclerosis. *JAMA Neurol*. (2021) 78:1197. doi: 10.1001/jamaneurol.2021.2738

38. Freeman L, Longbrake EE, Coyle PK, Hendin B, Vollmer T. High-efficacy therapies for treatment-naïve individuals with relapsing–remitting multiple sclerosis. *CNS Drugs*. (2022) 36:1285–99. doi: 10.1007/s40263-022-00965-7

39. Weideman AM, Tapia-Maltos MA, Johnson K, Greenwood M, Bielekova B. Meta-analysis of the age-dependent efficacy of multiple sclerosis treatments. *Front Neurol*. (2017) 8:577. doi: 10.3389/fneur.2017.00577

40. He A, Merkel B, Brown JWL, Ryerson LZ, Kister I, Malpas CB, et al. Timing of high-efficacy therapy for multiple sclerosis: a retrospective observational cohort study. *Lancet Neurol*. (2020) 19:307–16. doi: 10.1016/s1474-4422(20)30067-3

41. Krämer J, Bar-Or A, Turner TJ, Wiendl H. Bruton tyrosine kinase inhibitors for multiple sclerosis. *Nat Rev Neurol*. (2023) 19:289–304. doi: 10.1038/s41582-023-00800-7

42. Hult KJ. Measuring the potential health impact of personalized medicine: evidence from multiple sclerosis treatments. In: *Economic Dimensions of Personalized and Precision Medicine* (2019) (Chicago, USA: University of Chicago Press). p. 185–216. doi: 10.7208/chicago/9780226611235.003.0008

43. Van Wijmeersch B, Hartung HP, Vermersch P, Pugliatti M, Pozzilli C, Grigoriadis N, et al. Using personalized prognosis in the treatment of relapsing multiple sclerosis: A practical guide. *Front Immunol*. (2022) 13:991291. doi: 10.3389/fimmu.2022.991291

44. Marrie RA, Fisk JD, Fitzgerald K, Kowalec K, Maxwell C, Rotstein D, et al. Etiology, effects and management of comorbidities in multiple sclerosis: recent advances. *Front Immunol*. (2023) 14:1197195. doi: 10.3389/fimmu.2023.1197195

45. Yang J, Hamade M, Wu Q, Wang Q, Axtell R, Giri S, et al. Current and future biomarkers in multiple sclerosis. *Int J Mol Sci*. (2022) 23:5877. doi: 10.3390/ijms23115877

46. Voigt I, Inojosa H, Wenk J, Akgün K, Ziemssen T. Building a monitoring matrix for the management of multiple sclerosis. *Autoimmun Rev*. (2023) 22:103358. doi: 10.1016/j.autrev.2023.103358

47. Gill AJ, Schorr EM, Gadani SP, Calabresi PA. Emerging imaging and liquid biomarkers in multiple sclerosis. *Eur J Immunol*. (2023) 53:2250228. doi: 10.1002/eji.202250228

48. Trentzsch K, Schumann P, Śliwiński G, Bartscht P, Haase R, Schriefer D, et al. Using machine learning algorithms for identifying gait parameters suitable to evaluate subtle changes in gait in people with multiple sclerosis. *Brain Sci*. (2021) 11:1049. doi: 10.3390/brainsci11081049

49. Guerrieri S, Comi G, Leocani L. Optical coherence tomography and visual evoked potentials as prognostic and monitoring tools in progressive multiple sclerosis. *Front Neurosci*. (2021) 15:692599. doi: 10.3389/fnins.2021.692599

50. Paul F, Calabresi PA, Barkhof F, Green AJ, Kardon R, Sastre-Garriga J, et al. Optical coherence tomography in multiple sclerosis: A 3-year prospective multicenter study. *Ann Clin Trans Neurol*. (2021) 8:2235–51. doi: 10.1002/acn3.51473

51. Graves JS. Identifying multiple sclerosis activity. *Neurology*. (2022) 99:269–70. doi: 10.1212/wnl.0000000000200903

52. Absinta M, Sati P, Masuzzo F, Nair G, Sethi V, Kolb H, et al. Association of chronic active multiple sclerosis lesions with disability in *vivo*. *JAMA Neurol*. (2019) 76:1474. doi: 10.1001/jamaneurol.2019.2399

53. Blindenbacher N, Brunner E, Asseyer S, Scheel M, Siebert N, Rasche L, et al. Evaluation of the 'ring sign' and the 'core sign' as a magnetic resonance imaging marker of disease activity and progression in clinically isolated syndrome and early multiple sclerosis. *Multiple Sclerosis J - Exp Trans Clin*. (2020) 6:205521732091548. doi: 10.1177/2055217320915480

54. Preziosa P, Pagani E, Meani A, Moiola L, Rodegher M, Filippi M, et al. Slowly expanding lesions predict 9-Year multiple sclerosis disease progression. *Neurology® Neuroimmunol Neuroinflamm*. (2022) 9:e1139. doi: 10.1212/nxi.0000000000001139

55. Oreja-Guevara C, Blanco TA, Ruiz LB, Pérez MAH, Meca-Lallana V, Ramió-Torrentà L. Cognitive dysfunctions and assessments in multiple sclerosis. *Front Neurol*. (2019) 10:581. doi: 10.3389/fneur.2019.00581

56. Podda J, Ponzio M, Pedullà L, Bragadin MM, Battaglia MA, Zaratin P, et al. Predominant cognitive phenotypes in multiple sclerosis: Insights from patient-centered outcomes. *Multiple Sclerosis Related Disord*. (2021) 51:102919. doi: 10.1016/j.msard.2021.102919

57. Carotenuto A, Costabile T, Pontillo G, Moccia M, Falco F, Petracca M, et al. Cognitive trajectories in multiple sclerosis: a long-term follow-up study. *Neurol Sci*. (2021) 43:1215–22. doi: 10.1007/s10072-021-05356-2

58. Brichetto G, Zaratin P. Measuring outcomes that matter most to people with multiple sclerosis: the role of patient-reported outcomes. *′ Curr Opin Neurol*. (2020) 33:295–9. doi: 10.1097/wco.0000000000000821

59. Zaratin P, Vermersch P, Amato MP, Brichetto G, Coetzee T, Cutter G, et al. The agenda of the global patient reported outcomes for multiple sclerosis (PROMS) initiative: Progresses and open questions. *Multiple Sclerosis Related Disord*. (2022) 61:103757. doi: 10.1016/j.msard.2022.103757

60. Peeters LM, Parciak T, Kalra D, Moreau Y, Kasilingam E, van Galen P, et al. Multiple Sclerosis Data Alliance – A global multi-stakeholder collaboration to scale-up real world data research. *Multiple Sclerosis Related Disord*. (2021) 47:102634. doi: 10.1016/j.msard.2020.102634

61. Yamout B, Sahraian M, Bohlega S, Al-Jumah M, Goueider R, Dahdaleh M, et al. Consensus recommendations for the diagnosis and treatment of multiple sclerosis: 2019 revisions to the MENACTRIMS guidelines. *Multiple Sclerosis Related Disord*. (2020) 37:101459. doi: 10.1016/j.msard.2019.101459

62. Wattjes MP, Ciccarelli O, Reich DS, Banwell B, de Stefano N, Enzinger C, et al. 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol*. (2021) 20:653–70. doi: 10.1016/s1474-4422(21)00095-8

63. Tur C, Moccia M, Barkhof F, Chataway J, Sastre-Garriga J, Thompson AJ, et al. Assessing treatment outcomes in multiple sclerosis trials and in the clinical setting. *Nat Rev Neurol*. (2018) 14:75–93. doi: 10.1038/nrneurol.2017.171

64. Voigt I, Benedict M, Susky M, Scheplitz T, Frankowitz S, Kern R, et al. A digital patient portal for patients with multiple sclerosis. *Front Neurol*. (2020) 11:400. doi: 10.3389/fneur.2020.00400

65. Wenk J, Voigt I, Inojosa H, Schlieter H, Ziemssen T. Building digital patient pathways for the management and treatment of multiple sclerosis. *Front Immunol*. (2024) 15:1356436. doi: 10.3389/fimmu.2024.1356436

66. Sima DM, Esposito G, Van Hecke W, Ribbens A, Nagels G, Smeets D. Health economic impact of software-assisted brain MRI on therapeutic decision-making and outcomes of relapsing-remitting multiple sclerosis patients—A microsimulation study. *Brain Sci*. (2021) 11:1570. doi: 10.3390/brainsci11121570

67. Parciak T, Geys L, Helme A, van der Mei I, Hillert J, Schmidt H, et al. Introducing a core dataset for real-world data in multiple sclerosis registries and cohorts: Recommendations from a global task force. *Multiple Sclerosis*. (2023) 30:396–418. doi: 10.1177/13524585231216004

68. Sastre-Garriga J, Pareto D, Battaglini M, Rocca MA, Ciccarelli O, Enzinger C, et al. MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. *Nat Rev Neurol*. (2020) 16:171–82. doi: 10.1038/s41582-020-0314-x

69. Aboseif A, Roos I, Krieger SC, Kalincik T, Hersh CM. Leveraging Real-World evidence and observational studies in treating multiple sclerosis. *Neurol Clinics*. (2024) 42:203–27. doi: 10.1016/j.ncl.2023.06.003

70. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J Multidiscip Healthcare*. (2018) 11:295–304. doi: 10.2147/jmdh.s160029

71. Vercruyssen S, Brys A, Verheijen M, Steach B, Van Vlierberge E, Sima DM, et al. Abstracts from the 34th annual meeting of the consortium of multiple sclerosis centers. *Int J MS Care*. (2020) 22:1–116. doi: 10.7224/1537-2073-22.s2.1

72. Yadav SK, Motamedi S, Oberwahrenbrock T, Oertel FC, Polthier K, Paul F, et al. CuBe: parametric modeling of 3D foveal shape using cubic Bézier. *Biomed Optics Express*. (2017) 8:4181. doi: 10.1364/boe.8.004181

73. Yadav SK, Kadas EM. Optic nerve head three-dimensional shape analysis. *J Biomed Optics*. (2018) 23:1. doi: 10.1117/1.jbo.23.10.106004

74. Rakić M, Vercruyssen S, Van Eyndhoven S, de la Rosa E, Jain S, Van Huffel S, et al. 'icobrain ms 5.1: Combining unsupervised and supervised approaches for improving the detection of multiple sclerosis lesions. *NeuroImage Clin*. (2021) 31:102707. doi: 10.1016/j.nicl.2021.102707

75. Van Hecke W, Costers L, Descamps A, Ribbens A, Nagels G, Smeets D, et al. A novel digital care management platform to monitor clinical and subclinical disease activity in multiple sclerosis. *Brain Sci*. (2021) 11:1171. doi: 10.3390/brainsci11091171

76. Dillenseger A, Weidemann ML, Trentzsch K, Inojosa H, Haase R, Schriefer D, et al. Digital biomarkers in multiple sclerosis. *Brain Sci*. (2021) 11:1519. doi: 10.3390/brainsci11111519

77. Voigt I, Inojosa H, Dillenseger A, Haase R, Akgün K, Ziemssen T. Digital twins for multiple sclerosis. *Front Immunol*. (2021) 12:669811. doi: 10.3389/fimmu.2021.669811

78. Scholz M, Haase R, Trentzsch K, Stölzer-Hutsch H, Ziemssen T. Improving digital patient care: lessons learned from patient-reported and expert-reported experience measures for the clinical practice of multidimensional walking assessment. *Brain Sci*. (2021) 11:786. doi: 10.3390/brainsci11060786

frontiers | Frontiers in Immunology

TYPE Review
PUBLISHED 18 October 2024
DOI 10.3389/fimmu.2024.1459502

OPEN ACCESS

EDITED BY
Axel Faes,
University of Hasselt, Belgium

REVIEWED BY
Veronica Popescu,
University of Hasselt, Belgium
Eleftheria Kodosaki,
University College London, United Kingdom
Aram Zabeti,
University of Cincinnati Gardner Neuroscience Institute, United States

*CORRESPONDENCE
Georgina Arrambide
garrambide@cem-cat.org
Carmen Tur
ctur@cem-cat.org

RECEIVED 04 July 2024
ACCEPTED 30 September 2024
PUBLISHED 18 October 2024

CITATION
Arrambide G, Comabella M and Tur C (2024) Big data and artificial intelligence applied to blood and CSF fluid biomarkers in multiple sclerosis. *Front. Immunol.* 15:1459502. doi: 10.3389/fimmu.2024.1459502

Big data and artificial intelligence applied to blood and CSF fluid biomarkers in multiple sclerosis

Georgina Arrambide*, Manuel Comabella and Carmen Tur*

Multiple Sclerosis Centre of Catalonia (Cemcat), Department of Neurology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain

Artificial intelligence (AI) has meant a turning point in data analysis, allowing predictions of unseen outcomes with precedented levels of accuracy. In multiple sclerosis (MS), a chronic inflammatory-demyelinating condition of the central nervous system with a complex pathogenesis and potentially devastating consequences, AI-based models have shown promising preliminary results, especially when using neuroimaging data as model input or predictor variables. The application of AI-based methodologies to serum/blood and CSF biomarkers has been less explored, according to the literature, despite its great potential. In this review, we aimed to investigate and summarise the recent advances in AI methods applied to body fluid biomarkers in MS, highlighting the key features of the most representative studies, while illustrating their limitations and future directions.

KEYWORDS

multiple scleorsis (MS), fluid biomarkers, demyelinating, machine learning and AI, deep learning

Introduction

Artificial intelligence (AI) techniques have proved very useful for the diagnosis and prognostication of several conditions around the world (1), including multiple sclerosis (MS) (2). AI methods used in medical research, including MS research, may include machine learning (ML) and deep learning (DL) analyses. Typically, while ML analyses are based on tabulated data as input to the model, DL models use raw data – typically images – as input to the model. Model outputs depend on the type of task that is needed, e.g., a given diagnosis (instead of another one), a certain disability milestone, or the presence of MRI activity in people who are receiving a given drug.

Multiple sclerosis (MS) is a chronic inflammatory-demyelinating condition of the central nervous system (CNS) with heterogeneous genetic and environmental risk factors (3). Disease diagnosis and monitoring strongly rely on routine clinical assessments and the use of conventional brain and spinal cord magnetic resonance imaging (MRI) as a biomarker. A biological marker, or biomarker, is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic

processes or pharmacologic responses to a therapeutic intervention (4). Besides MRI, body fluid biomarkers can also provide additional, independent data on MS. AI applications in MS can potentially help us better support the diagnosis, find markers for prognosis, facilitate accurate monitoring, and eventually understand the mechanisms of the disease. Focusing on these main challenges, this review aims to summarise the recent advances in AI applied to blood, serum and CSF biomarkers in MS, highlighting the key features of the most representative studies (Figure 1) (5). This review also aims to illustrate its limitations and future directions.

Search strategy

We performed a search in PubMed based on the following criteria: (i) search terms: ((multiple sclerosis) or demyelination or (demyelinating disease)) AND ((artificial intelligence) or (deep learning) or (machine learning)) AND (biomarkers OR markers OR (biological markers) OR (fluid biomarkers) OR (body fluid biomarkers)); (ii) language of publication: English; (iv) type of paper: original research. For the purpose of this narrative review, we have focused on three aspects: (i) diagnosis & differential diagnosis; (ii) prediction of clinical outcome; (iii) understanding of pathogenic mechanisms. Thus, after the first literature search, we manually selected the papers if they were included in one of these three categories. Papers not clearly included in any of these categories were not considered in the review. Thus, we did not include papers whose main focus was methodological or animal research, and papers related to fluid biomarkers other than blood, serum and CSF. We also excluded review papers, editorials, and case reports. The PubMed search yielded 206 articles, published between 1996 (and especially between 2009) and 2024, both included (Figure 2). After excluding those not meeting our inclusion criteria, we revised 29 papers for their inclusion in this narrative review (Figure 2). Most of these papers have been published between 2019 and 2024 (Figure 3).

Once all papers were selected, they were divided into MS diagnosis and differential diagnosis (N=6), prediction of disease evolution (N=14), and understanding mechanisms of damage in MS (N=9). Of note, for some papers we found a degree of overlap and the decision to include them into one or another category depended on the main objectives described by the authors.

MS diagnosis and differential diagnosis

The diagnosis of MS relies on integrating clinical, MRI, and laboratory findings and excluding alternative diagnoses, especially in the presence of red flags. Indeed, the diagnosis of MS is not devoid of challenges: other conditions may mimic MS, clinically or radiologically (6). In these circumstances, the use of AI algorithms may be useful (Table 1), especially in body fluid biomarker discovery studies such as those done with "omics" technology.

AI has been implemented to identify genetic susceptibility biomarkers. Pasella et al. (7) used decision trees (DT) to create a predictive tool assessing the likelihood of MS including alleles responsible for human leukocyte antigen (HLA) class I molecules and killer immunoglobulin-like receptor (KIR) genes, responsible for natural killer (NK) lymphocyte receptors. They studied 299 persons with MS (PwMS) and 619 healthy controls (HC). The algorithm accurately identified 80.94% of PwMS and 71.08% HC in the training set and 73.24% and 66.07%, respectively, in the validation set. Guo et al. (8) used Support Vector Machine (SVM) to identify gene expression profiles on the transcriptome of peripheral blood mononuclear cells (PBMC) from 26 PwMS and 18 subjects with other neurological diseases (OND). This approach

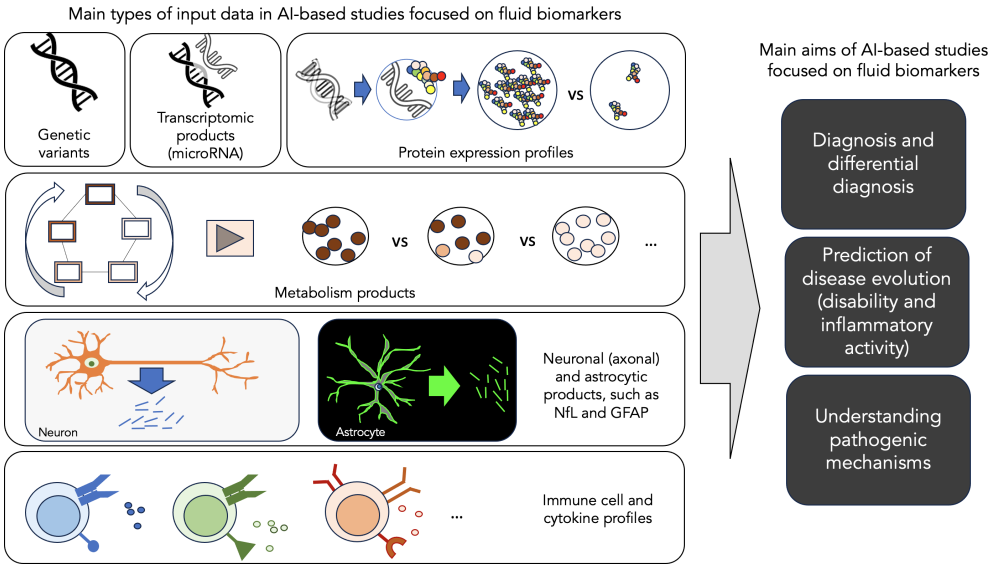


FIGURE 1
Main aims of AI-based studies focused on fluid biomarkers. This figure illustrates the main types of input data and the main aims of AI-based studies focused on fluid biomarker data in MS.

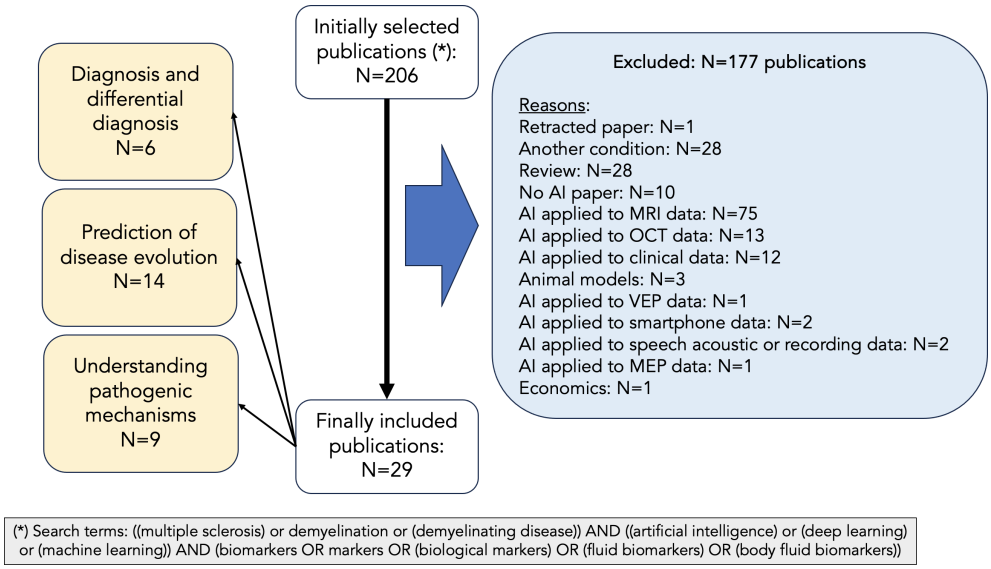


FIGURE 2
PRISMA chart describing article selection. We have followed a systematic approach for selecting the papers to be considered in our manuscript. After performing a PubMed search with the following terms: (multiple sclerosis or demyelination or demyelinating disease) AND (artificial intelligence or deep learning or machine learning) AND (biomarkers or markers or biological markers or fluid biomarkers or body fluid biomarkers), 206 records were obtained. Of those, only 29 were considered for this review after excluding those not meeting our inclusion criteria.

identified 8 genes differentially expressed between groups with 86% accuracy in the validation study. These genes involved the protein kinase cascade, inactivation of mitogen-activated protein kinases (MAPK), and regulation of signal transduction and apoptosis.

The metabolomes of cells and tissues include lipids, amino acids, sugars and other molecules (9). Andersen et al. (10) used random forests (RF) to identify blood-based metabolite profiles that could discriminate between 12 male PwMS and 13 male controls. The top 6 candidate metabolites informative for MS, defined as having an area under the receiver operating characteristic (ROC) curve (ROC-AUC) >80%, participate in glutathione metabolism, fatty acid metabolism and oxidation, cellular membrane composition, and transient receptor potential channel signalling. Whilst metabolomics focuses on hydrophilic molecules, lipidomics has emerged as an independent "omics" due to its complexity (9). Lötsch et al. (11) used unsupervised ML to compare 43 lipid mediators in serum from 102 PwMS and 301 HC. The analyses showed 98% accuracy to differentiate PwMS from HC. Then, the authors used supervised ML implemented as RF and computed ABC analysis-based feature selection, to create a classifier. This approach identified 8 lipid biomarkers differentially expressed in PwMS with ≥95% accuracy in training and test datasets.

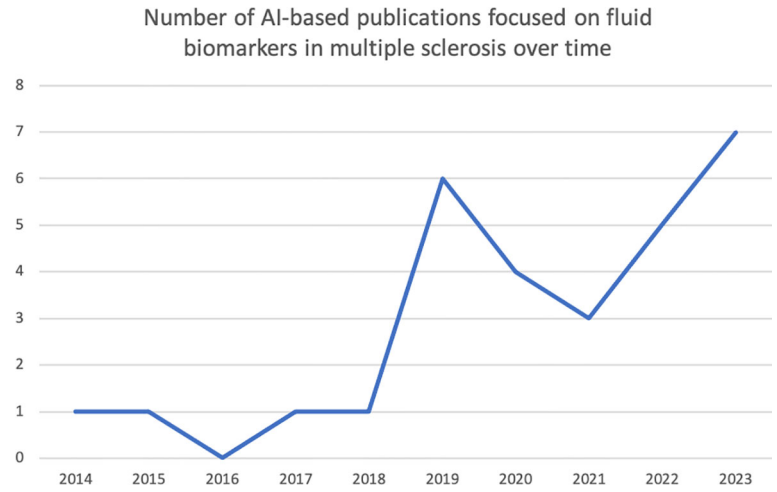


FIGURE 3
Distribution of the research papers on AI applied to biomarker data in MS over time. This histogram shows the number of research articles (of those 29 selected) published per year. It is to be noted that most of the papers have been published in the last 4 years.

TABLE 1 Summary of selected studies focused on diagnosis and differential diagnosis.

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Pasella et al., Front Neuroinform. 2023 [ref (7)]	MS: n=299 (RRMS n=218, PPMS n=81) Healthy controls: n=619	0	Alleles responsible for HLA class I molecules and KIR genes, obtained from PBMC	DT	Genotyping for alleles at *HLA-A*, *-B*, *-C*, and *-DRB1* loci. Primers specific to 11KIR genes: *IR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL5*, *KIR3DL1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5*, *KIR3DS1*	MS vs non-MS	identified 80.94% of MS patients in the training set and 73.24% in the validation set. Identified 71.08% of healthy controls in the training set and 66.07% in the validation set	Immunogenetic risk factors, specifically alleles responsible for HLA class I molecules and KIR genes, responsible for natural killer lymphocyte receptors
Guo et al., PLoS One. 2014 [ref (8)]	MS: n=26 OND: n=18	0	27336 probe sets obtained from gene expression profiles from the Array Express Database. Samples obtained from PBMC	SVM, ROC algorithm, Boruta algorithm	8 genes differentially expressed between MS and OND	MS vs OND	AUC 0.711-0.852. Accuracy of 86% in validation study	The 8 differentially expressed genes in MS vs OND were related to the protein kinase cascade, inactivation of MAPK, and regulation of signal transduction and apoptosis
Andersen et al., Mult Scler Relat Disord. 2019 [ref (10)]	Male subjects with MS: n=12 Male controls: n=13	0	Serum metabolites (lipid and amino acid profiles)	RF	12 metabolites	MS vs controls	6 metabolites with AUCs >80%: pyroglutamate, laurate, acylcarnitine C14:1, N-methylmaleimide, and 2 phosphatidylcholines (PC ae 40:5, PC ae 42:5)	Identified metabolites participate in glutathione metabolism, fatty acid metabolism and oxidation, cellular membrane composition, and transient receptor potential channel signalling. Their gene expression association suggested enrichment for pathways associated with apoptosis and mitochondrial dysfunction.
Lötsch et al., Sci Rep. 2018 (ref [11)]	MS: n=102 Healthy controls: n=301	0	43 lipid mediators from serum samples: ceramides (@)	Self-organising maps of neural networks, swarm intelligence and Minimum Curvilinear Embedding. In a second step, RF and computed ABC analysis-based feature selection	Classifier with 8 lipid biomarkers (GluCerC16, LPA20:4, HETE15S, LacCerC24:1, C16Sphinganine, biopterin, and endocannabinoids PEA and OEA)	MS vs healthy controls	98% accuracy for the 43 lipid mediators; classifier with ≥95% accuracy in training and test data sets	Most lipid mediator concentrations were reduced in MS. Exceptions were the ceramide LacCerC24:1 and the sphingolipid C16Sphinganin, found at higher concentrations in MS Cer16 and Cer24 might amplify cytokine-induced cell death of myelin-producing oligodendrocytes. HETE15S was shown to be regulated in CSF of MS patients. Enhanced activity of autotaxin was observed in serum samples of MS patients. PEA and OEA have been found in RRMS and SPMS. Neopterin is an activation marker of the innate immune system with increased levels in autoimmune diseases including the CSF of MS patients

(Continued)

10.3389/fimmu.2024.1459502

TABLE 1 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Probert., et al. Front Immunol. 2021 [ref (12)]	MS with +OCB: n=41 Non-MS controls with +OCB: n=64 (*)	0	Metabolites and proteins in CSF	Multivariate OPLS-DA	8 metabolites significantly decreased in MS: 4 (myo-inositol, isoleucine, leucine, glutamine) had higher specificity than OCB for MS diagnosis. 9 biomarkers outperformed OCB as predictor of MS (CCN5, CDCC80, NTN1, vWF, DKK4, SOST, ERBB3, IGL4, and IGKV1-5). All significantly decreased in MS vs non-MS except for IGL4 and IGKV1-5, which were increased.	MS vs non-MS	The combination of CCN5, vWF, GFAP, and OCB status provided the best overall diagnostic properties (sensitivity 89%, specificity 92%, accuracy 91%) compared to OCB status	Integrative metabolomics and proteomic enrichment analysis revealed upregulated JAK-STAT and glycolysis pathways in MS, consistent with an increased inflammatory response and altered energy metabolism.
Gaetani et al., Int J Mol Sci. 2023 (ref [13)]	+OCB RRMS: n=58; -OCB RRMS: n=24; OND: n=36 (&)	0	Quantification of 92 immune activation CSF proteins	Hierarchical clustering to profile CSF proteins. Binomial and multinomial LASSO regressions to differentiate patient groups	92 tested proteins minus 45 with a call rate <85%, age, sex, NfL	MS vs OND; +OCB RRMS vs OND; -OCB RRMS vs OND	All: CD5 (AUC 0.87) and IL-12B (AUC 0.81). +OCB RRMS vs OND: IL-12B, CX3CL1, FGF-19, CST5, and MCP-1 (91% sensitivity, 94% specificity in the training set; 81% and 95%, respectively, in the validation set) -OCB RRMS vs OND: CX3CL1, CD5, CCL4, and OPG as well as NfL (87% sensitivity, 80% specificity in the training set; 56% and 48% in the validation set)	CD5 may act as a receptor in regulating T cell proliferation. IL-12B promotes differentiation of T cells into T helper 1 (Th1) cells. CX3CL1 increases IFN-γ and TNF-α gene expression and IFN-γ secretion by CD4+ T cells. FGF signalling may regulate inflammation and myelination in MS since an abundance. CST5 has shown potential as a relapse marker. MCP-1 may be involved in the recruitment of monocytes/macrophages and activated lymphocytes. CCL4 is involved in the disruption of the blood-brain barrier. OPG suppresses mRNA expression of CCL20, a chemokine involved in Th17 cell recruitment with anti-inflammatory effects
Martynova et al., Mediators Inflamm. 2020 [ref (14)]	MS: n=101 (RRMS n=49, SPMS n=31, PPMS n=21) and Non-MS subjects: serum n=101 and CSF n=25 ($)		45 leucocyte-activation regulatory cytokines measured in serum and CSF	k-Nearest Neighbour, DT, XGB, Gaussian Naïve Bayes and RF	22 cytokines altered in CSF and 20 in serum, 10 commonly affected in both (IL-1α IL-4, IL-18, CCL7, CCL27, CSF, IFN-γ, LIF, M-CSF, and TNF-α). Three	MS vs non-MS	Diagnostic accuracy: ≥92% when any randomly selected 5 of any cytokines were used. The highest accuracy, 99%, obtained when including CCL27, IFN-γ, and IL-4	CCL27 could trigger T memory cells to produce IL-4 and IFN-γ. Interleukins and chemokines affected in serum and CSF could direct leukocyte migration targeting Th1 cells.

(Continued)

10.3389/fimmu.2024.1459502

10.3389/fimmu.2024.1459502

TABLE 1 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
					independent datasets: cytokines affected both in CSF and serum, only in CSF and only in serum			

(*) Epilepsy (n=5), functional neurological disorder (n=12), gait disorder (n=1), meningitis (n=2), motor paresis (n=3), movement disorder (n=2), MG (n=2), neuralgic amytrophy (n=1), neuroinfection (n=3), normal pressure hydrocephalus (n=1), polyneuropathy (n=5), polyradiculitis (n=2), primary headache disorder (n=13), sensory disturbance (n=8), SLE (n=1), visual disturbance (n=1), white matter lesions/leukoencephalopathy (n=2); (&): headache: n=16; psychiatric disorders: n=13; mononeuropathy: n=4), dysmetabolic polyneuropathy (n=3); ($): tension type headache, residual encephalopathy, unspecified demyelinating disease of the CNS, cerebrovascular diseases, PML, migraine with aura; (@): Cer16:0, Cer18:0, Cer18:1, Cer20:0, Cer24:0, Cer24:1, GluCerC16:0, GluCerC24:1, LacCerC16:0, LacCerC24:0, LacCerC24:0); lyosophosphatidic acids (LPA16:0, LPA18:0, LPA18:1, LPA18:2, LPA18:3, LPA20:4); sphingolipids (sphinganine, sphingosine, S1P, SA1P C16Sphinganine, C18Sphinganine, C24Sphinganine, C24:1Sphinganine); prostaglandins (PGD2, PGF1α, PGE2, TXB2); dihydroxyeicosatrienoic acids (DHET5.6, DHET11.12, DHET14.15); hydroxyeicosatetraenoic acids (HETE 5 S, HETE_12S, HETE_15S, HETE_20S); endocannabinoids (AEA, OEA, PEA, 2-AG) and pterins (biopterin, neopterin); Abbreviations (in alphabetical order): AUC, area under the curve; CCL, chemokin (C-C motif) ligand; CCN5, connective tissue growth factor/cysteine-rich protein/nephroblastoma overexpressed-5; CD, cluster of differentiation; CDCC80, coiled-coil domain-containing protein 80; CSF, cerebrospinal fluid; CST5, cystatin D; CX3CL, chemokine (C-X3-C motif) ligand 1; DKK4, dickkopf-related protein 4; DT, decision trees; ERBB3, receptor tyrosine-protein kinase erbB-3; FGF, fibroblast growth factor; GFAP, glial fibrillary acidic protein; HLA, human leukocyte antigen; IFN, interferon; IGKV1-5, immunoglobulin kappa variable 1-5; IGL4, insulin growth factor-like family member 4; IL, interleukin; JAK-STAT, Janus kinase/signal transduction and transcription activation; KIR, killer immunoglobulin-like receptor; LASSO, least absolute shrinkage and selection operator regression; LIF, leukemia inhibitory factor; MAPK, mitogen-activated protein kinases; MCP, monocyte chemoattractant protein; M-CSF, macrophage colony-stimulating factor; MG, myasthenia gravis; MS, multiple sclerosis; NfL, neurofilament light chain; NTN1, netrin-1; OCB, oligoclonal bands; OND, other neurological diseases; OPG, osteoprotegerin; OPLS-DA, orthogonal partial least squares discriminant analysis; PBMC, peripheral blood mononuclear cells; PPMS, primary progressive multiple sclerosis; RRMS, relapsing remitting multiple sclerosis; ROC, receiver operating characteristic curve; RF, random forests; SLE, systemic lupus erythematosus; SOST, sclerostin; SPMS, secondary progressive multiple sclerosis; SVM, support vector machine; Th, T helper cells; TNF, tumor necrosis factor; vWF, von Willebrand factor; XGB, Extreme Gradient Boosting.

Other studies have focused on CSF biomarkers. Probert et al. (12) used ML to profile metabolites and proteins in CSF samples from 41 PwMS and positive IgG oligoclonal bands (+OCB) and 64 patients with OND and +OCB. Multivariate orthogonal partial least squares discriminant analyses (OPLS-DA) showed that combining connective tissue growth factor/Cysteine-rich protein/Nephroblastoma overexpressed-5 (CCN5), von Willebrand Factor (vWF), glial fibrillary acidic protein (GFAP), and OCB provided the best diagnostic properties to discriminate MS from OND (89% sensitivity, 92% specificity, 91% accuracy). Gaetani et al. (13) used hierarchical clustering to profile 92 immune activation CSF proteins in +OCB relapsing-remitting MS (RRMS) (n=58), -OCB RRMS (n=24), and OND (n=36). Next, they used binomial and multinomial least absolute shrinkage and selection operator (LASSO) regressions to differentiate among these groups. Cluster of differentiation 5 (CD5) (ROC-AUC 0.87) and interleukin 12B (IL-12B) (ROC-AUC 0.81) were the best MS vs OND predictors. The model that best differentiated +OCB RRMS from OND included IL-12B and 4 other proteins (sensitivity 91% and 81%, specificity 94% and 95% in the training and validation sets, respectively). The model that best differentiated -OCB RRMS from OND included CD5, 3 other immune activation proteins as well as NfL, assessed additionally (sensitivity 87% and 56%, specificity 80% and 48% in the training and validation sets, respectively).

One study assessed proteins in both CSF and serum. Martynova et al. (14) used five ML models to study differences in 45 leucocyte-activation regulatory cytokines, measured in serum and CSF of 101 PwMS and in 101 serum and 25 CSF samples from non-MS subjects. Twenty-two cytokines were altered in CSF and 20 in serum, of which 10 were commonly affected. Next, three independent datasets including cytokines affected in CSF and serum, only in CSF, and only in serum were used as input to ML models to predict MS. Diagnostic accuracy was ≥92% when any randomly selected five of any cytokines were used.

Prediction of MS evolution

The high heterogeneity of MS in terms of disease evolution means that the prognostication in clinical practice is extremely difficult. Although the presence of a high number of inflammatory-demyelinating lesions in the brain (15), and the presence of infratentorial (16), cortical (17), spinal cord (18), lesions at the time of the first attack are well-known predictors of a worse clinical evolution, these associations are only meaningful at a group level. That is, the prediction of the disease at the individual level based on these known predictors is still far from optimal. For that reason, over the years, a number of authors have aimed at predicting MS evolution based on these factors but through the development of AI models, with a much greater potential – at least theoretically – than classical statistical models. In spite of this, though, the ability to currently build (and publish) AI models to predict disease evolution based on MRI and clinical data is still limited. This limited ability becomes evident especially when a model built in a given cohort is applied in a completely unseen, independent, validation cohort,

showing a much lower accuracy than expected (much lower than that of the original cohort). This possibly suggests that the variability across people with MS is probably larger than what we thought and that mismatches between accuracies in original (training and testing) cohorts and external validation cohorts may be due to an overfitting of the data by the model in the original cohorts. Additionally, this may also suggest that other aspects apart from MRI and clinical data may be playing a role in the evolution of the disease. Over the last 10 but especially over the last 5 years, some studies using AI models applied to biomarker data to explain concurrent and future disease evolution have started to emerge (Table 2).

Regarding the studies that have focused on the concurrent prediction of clinical outcomes, in 2019, Flauzino et al. (19), published a study where 122 people with MS were tested on several serum biomarkers to predict concurrent disability status. These biomarkers, which were related to the immune-inflammatory response, lipid and protein metabolic pathways, and oxidative stress, were able to predict which patients had an Expanded Disability Status Scale (EDSS) (20) score above or below 3.0 with high accuracy (Area under the ROC curve = 0.842). These results suggest that Immune inflammatory, metabolic and oxidative stress pathways may play a key role in disability accumulation in MS and deserve further research. In another interesting study focused on concurrent prediction, Brummer and colleagues (21) showed how serum neurofilament light (NfL) levels could improve our ability to detect cognitive dysfunction, especially when added to MRI predictors such as grey matter volume. The authors of this study not only built a ML model with high predictive accuracy, but also validated the ML model in an external cohort, supporting the generalisability of the model (21). Finally, we highlight the paper from Jackson and colleagues (22), where ML models based on random forest regression were built to predict a multi-dimensional score of disease severity using genetic variants previously identified as related to MS severity. Interestingly, the results, which could be validated in an external cohort, showed that the 19 most predictive genetic variants were located in 12 genes associated with immune cell regulation, complement activation and functions of neurons (22). This supports the robustness of the results while providing important insights on the mechanisms of progression in MS.

Regarding the studies with a longitudinal design, there is a high variability in terms of the length of the prediction period, ranging from 6 months to 11 years, and in terms of the nature of the predictor data, i.e., the input of the ML model. For instance, there are studies which have used genetic data, focusing on the presence of certain genetic variants or single nucleotide polymorphisms (SNPs) (23, 24). Other studies have focused instead on the presence of certain epigenetic mechanisms, such as DNA methylation (25), and on certain gene expression profiles (26, 27). Also, a few studies have demonstrated the ability of (immune) cellular profiles to predict clinical outcome (23). Finally, there are studies which have based their predictions on the presence of specific serum and CSF proteins and metabolites (28, 29). In relation to the output data, i.e., the outcome of the ML model, most studies focus on disability progression measures (19, 21–23, 25, 28, 30, 31), although some of them have chosen acute activity (generally MRI activity) outcomes (24, 26, 27, 32) and one focused on the development of anti-drug neutralising antibodies (33), known to reduce the effectiveness of the disease-modifying drug (33).

In relation to the studies which have used SNP data to predict future outcome, the article by Andorra et al. (23) is of special interest. In this study, not only SNPs located in Human Leukocyte Antigen (HLA) and non-HLA genes were considered as predictors, but also data on immune cell populations, proteomics, brain MRI, and optic coherence tomography (OCT) data. In this study, whose results were validated in an external cohort, the authors predicted the development of confirmed disability accumulation on different disability outcomes after 2 years of follow-up, with high sensitivity (23).

Among the studies with longest predictive periods, there is the paper by Uphaus et al. (28), which used NfL data to predict 6-year development of relapse-free progression and transition from RRMS to SPMS with high accuracies, especially for the former outcome and especially when combined with age and T2 lesion volume (28). More recently, Everest et al. (31) published a paper where CSF proteomics data was used to predict unfavourable evolutions over an 8-year follow-up period (on average) with very high accuracies. In this paper, which included an external validation analysis, the authors propose several novel candidate CSF protein biomarkers with a promising future in disease prediction modelling (31). Finally, Campagna et al. (25) exploited the DNA methylation profiles of 235 women with MS to predict disease severity over an 11-year period, again with high accuracy. Although this model was not externally validated in an independent cohort, the length of its prediction and the nature of the biomarker used make it especially relevant. Interestingly, those genes with greater levels of methylation seemed to be related to neuronal structure and function (25).

Investigation of disease mechanisms

The pathophysiological processes in MS are not completely understood and are believed to be highly heterogeneous across people and disease stages. Fluid biomarker studies using AI to understand pathogenetic mechanisms could contribute to a greater characterisation of MS by expanding the concept of classical phenotypes (Table 3).

PBMCs can bear specific dysregulation in genes at different stages of MS. Acquaviva et al. (34) analysed transcriptomic profiles of PBMCs from individuals with CIS (n=57), RRMS (n=108), SPMS (n=26), PPMS (n=35), OND (n=27), and HC (n=60), divided into training (n=224) and validation (n=89) datasets. They defined classifiers (MS vs non-MS, relapsing vs progressive MS) using nested cross-validation in the training dataset. Then they used ward DT-based algorithms [RF, functional trees (FTs) and adaptive boosting applied to FT (ADAboost-FT) to evaluate their performance in the validation dataset. ADAboost-FT generated the best model to differentiate MS from non-MS (94.3% sensitivity, 87.5% precision). Identified transcripts in MS were related to interferon signalling, chromatin remodelling, and apoptosis. The

TABLE 2 Summary of selected studies focused on prediction of disease course: relapses and disability accumulation.

Reference	Training and testing cohort, N	Independent validation cohort, N	Follow-up time (study design)	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
*Cross-sectional prediction**									
Flauzino et al., Metab Brain Dis. 2019 [ref (19)]	122 patients with MS, i.e., RRMS, N=103; PPMS, N=3; SPMS, N=16	0	NA	Serum biomarkers including immune-inflammation, metabolic, and nitro-oxidative stress features	Multilayer perceptron neural network	Immune inflammatory (Th17/Treg ratio), metabolic (LDL/HDL ratio, uric acid, homocysteine) and oxidative stress (lipid hydro-peroxides, carbonyl protein, AOPP, NO metabolites) biomarkers, together with age, sex, disease duration, body mass index, and presence of metabolic syndrome	Disability status based on EDSS score: i) ≥3.0 vs <3.0 (binary outcome) ii) as a continuous outcome	ROC AUC = 0.842	Immune inflammatory, metabolic and oxidative stress pathways play a key role in disability accumulation in MS
Jackson et al., Ann Hum Genet. 2020 [ref (22)]	205	94	NA	113 genetic variants previously identified as related to MS severity	Random forest regression	19 genetic variants (GeM-MSS model)	MS-DSS, a score defined through a statistical model which takes into account CNS damage and demographic features [ref (46)]	GeM-MSS RMSE (error) = 0.464	The 19 genetic variants included in the GeM-MSS are related to 12 genes associated with immune cell regulation, complement activation and functions of neurons
Brummer et al., Brain Commun. 2022 [ref (21)]	152 patients with early MS	101 early MS	NA	Serum NfL	Support vector regression	Serum NfL, lesion volume, grey matter volume	Cognitive status based on SDMT score (continuous outcome)	Accuracy = 90.8%, greater than the accuracy of the models with individual predictors	The combination of blood and imaging measures improves the accuracy of predicting cognitive impairment
Zhu et al., Brain Commun. 2023 [ref (30)]	431	0	NA	19 serum protein biomarkers: APLP1, CCL20, CD6, CDCP1, CNTN2, CXCL9, CXCL13, FLRT2, GFAP, MOG, NfL, OPG, OPN, PRTG, SERPINA9, TNFSF10A, TNFSF13B, VCAN	LASSO, Random forest, Extreme Gradient Boosting, Support Vector Machines, stacking ensemble learning	7 clinical factors (age at sample collection, sex, race/ethnicity, disease subtype, disease duration, DMT, and time interval between sample collection and closest PRO assessment) and 19 serum protein biomarkers	Disability status based on PDDS score: ≥4 vs <4 (binary outcome) PDDS score: as categorical variable	ROC AUC = up to 0.91 (for LASSO prediction of PDDS using combined clinical and biomarker profiles as input)	Combined (clinical + biomarkers) models: the best LASSO better than other ML approaches Serum multi-protein biomarker profiles: better than single-protein (e.g., NfL or GFAP) models

(Continued)

10.3389/fimmu.2024.1459502

TABLE 2 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Follow-up time (study design)	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Longitudinal prediction (**)									
Ebrahimkhani et al., Mol Neurobiol. 2020 [ref (32)]	29 RRMS patients who were about to start on fingolimod	0	0.5 years (6 months); however, the study does not focus on future but concurrent prediction (i.e., disease activity and microRNA dysregulation occur over the same period of time)	Exosome miRNAs	Random forest	Out of all micro-RNAs, 15 were selected for being dysregulated between active and non-active patients, 6 months after fingolimod onset. Of those, 11 were selected for having ROC AUC 95%CI above 0.50. Then, out of a total of 2037 combinations of these 11 microRNAs, 3 combinations ($) were chosen for their highest accuracy	Disease activity vs no activity, based on MRI, i.e., presence of gadolinium-enhancing lesions (binary outcome)	Prediction accuracy (of combined microRNAs) = 0.92	microRNA signatures are noninvasive biomarkers which may help predict treatment response in the future
Baranzini et al., Mult Scler. 2015 [ref (27)]	155 RRMS on beta-interferon treatment	0	0,77 years (40 weeks)	Gene expression profiles at treatment onset or over the follow-up (i.e., *induction ratios* of gene expressions after treatment onset)	Random forest	Triplet (3-gene) expression profiles (several triplet combinations were assessed)	Disease activity free on treatment (presence of clinical and/or MRI activity) vs suboptimal response (binary outcome)	Predictive accuracy = 0.59-0.68 ROC AUC = up to 0.63	Future (IFNb) treatment response may be predicted with gene expression profiles at treatment onset or over the first weeks after that, using models of machine learning
Waddington et al., Front Immunol. 2020 [ref (33)]	89 patients with RRMS/ first demyelinating attack who were about to start on beta-interferon treatment	0	1 year	156 serum metabolites (see paper for full details)	Random forest, support vector machine, and LASSO logistic regression (K-nearest neighbour and decision trees also tested for comparison)	60 and 59 serum metabolites (out of 156) at baseline (before IFNb onset) and after 3 months, respectively; the remaining 96 and 97 metabolites, respectively, were excluded because of a strong correlation between them and the finally chosen 60 and 59 ones	ADA positive, i.e., i) bAbs+ & nAbs+ or ii) bAbs- but nAbs+ and titer ≥ 320 U/mL, within 12 months of starting treatment, vs ADA negative (binary outcome)	Classification accuracy (baseline) = 0.695-0.854 Classification accuracy (3 months after IFNb onset) = 0.712-0.863	ADA status may be predicted through serum metabolites
Herman et al., iScience. 2023 [ref (29)]	123	56	1 year	498 CSF metabolites	Elastic-net regularized classifier model In addition,	CSF metabolites: out of 498, 15 metabolites are selected	MS phenotype: PMS vs RRMS (binary outcome)	ROC AUC = 0.93, better than any of the single	This study provides confidence in individual patient prediction

(Continued)

10.3389/fimmu.2024.1459502

TABLE 2 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Follow-up time (study design)	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Longitudinal prediction (**)									
					conformal prediction analyses provides confidence in individual patient predictions			metabolic features in isolation	(=0.88), which can help with patient monitoring
Andorra et al., J Neurol. 2023 [ref (23)]	322	271	2 years	Genomics: MS-associated (HLA and non-HLA) SNPs; Cytomics: levels of effector and regulatory T cells, B cells, and NK cells; Phospho-proteomics: 25 kinases participating in pathways associated with MS	Random forest	Brain MRI, OCT, and multiomics (genotyping, cytomics and phospho-proteomics) from PBMC	CDA on different scales (EDSS, T25WT, 9HPT, SDMT, SL25, HCVA) vs no-CDA (binary outcomes); NEDA vs no-NEDA (binary outcome); MSSS, ARMSS, onset of DMT, escalation from low- to high-efficacy DMT (continuous outcomes)	ROC AUC = from 0.50 (T25WT-CDA) to 0.81 (SL25-CDA); Balanced accuracies = from 0.5 (9HPT or T25WT) to 0.69 (starting therapy) Sensitivities = almost all between 0.82 and 0.94 PPVs = almost all between 0.8 and 0.9	Models provided better sensitivities and PPVs than accuracies or AUC; Models including imaging & genetics or omics slightly improved model performance (with respect to models with clinical predictors only) and only in 50% of the times
Ferrè et al., J Pers Med. 2023 [ref (24)]	304 patients on fingolimod treatment	77 patients on fingolimod treatment	2 years	Genetic data	Random forest	123 SNPs (*genetic model*), clinical data (*clinical model*), or both (*combined model*)	NEDA vs no-NEDA (binary outcome)	ROC AUC genetic model = 0.65 ROC AUC combined (genetic and clinical) model = 0.71	ML models integrating clinical and genetic data can help predict disease evolution in pwMS on fingolimod
Fagone et al., Mol Med Rep. 2019 [ref (26)]	12 patients with RRMS who were about to start on natalizumab	0	3 years	Whole–genome expression data from CD 4+ T cells (assessed before natalizumab onset)	UnCorrelated Shrunken Centroid Algorithm (¢)	Genetic expression of 17 genes related to CD4+ T cells	Disease activity or not, based on presence (vs absence) of relapses over the whole follow-up of 3 years (binary outcome)	Accuracy = 0.892	Gene expression profiles may help design personalised therapeutic strategies for patients with MS

(Continued)

10.3389/fimmu.2024.1459502

TABLE 2 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Follow-up time (study design)	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Longitudinal prediction (**)									
Uphaus et al., EBioMedicine 2021 [ref (28)]	196 patients with RRMS/ first demyelinating attack	204 RRMS/ first demyelinating attack	Median: 6 (IQR 4.3-7.5) years	Serum NfL	Support vector machine	Serum NfL levels at baseline and ratio NfL follow-up/baseline +/- age & T2 lesion number at baseline	Relapse-free progression (binary outcome); Transition to SPMS (binary outcome)	For relapse-free progression: ROC AUC = 0.811 (NfL + age & T2 lesion number) For SPMS transition: ROC AUC = 0.651	Serum NfL levels may help predict future relapse-free progression in clinical practice, together with age and T2 lesions at baseline
Everest et al., PLoS One. 2023 [ref (31)]	94	40	Mean: 8.2 ± 2.2 years	CSF proteomics data: 151 differentially expressed CSF proteins, including C3bCfb, A2M, ATF7, PRBP, Haptoglobin, PDS5B, Myosin, CD36, and ApoA1 (ref (47)]	Genetic algorithm (Holland J. Adaptation in natural and artificial systems. University of Michigan Press, 1975)	CSF proteomics data	Disease severity status (binary outcome) based on ARMSS score on last follow-up: ≥5 (unfavourable group) vs <5 (favourable)	Rule 1 (to select ARMSS≥5): ROC AUC = 86.34% Rule 2 (to select ARMSS<5): ROC AUC = 73.26%	Novel candidate CSF protein biomarkers are proposed, to be validated in larger samples
Campagna et al., Clin Epigenetics. 2022 [ref (25)]	235 female patients with RMS	0	Median: 11.13 (IQR 9.49; 12.59) years	DNA methylation data assessed through Illumina methylation EPIC array	Elastic-net regression and logistic regression	Clinical data (age and symptoms), DNA methylation data of genes related to neuronal structure and function	Disease severity status (binary outcome) based on ARMSS score: mild vs severe (i.e., median ARMSS score below or above 20th or 80th percentile, respectively, of the cohort)	Methylation model ROC AUC = 0.91 (vs clinical model ROC AUC = 0.74)	Whole-blood methylation can predict disease severity in RMS and seems to affect genes related to neuronal structure and function

(*) Articles shown in chronological order; (**) Articles shown based on length of follow-up; (¢) UC SC; http://home.cc.umanitoba.ca/~psgendb/birchhomedir/BIRC HDE V/doc/MeV/manual/usc.html; ($) Combination 1: miR-432-5p and miR-485-5p; combination 2: miR-432-5p, -485-5p, -375; combination 3: miR-432-5p, −485-5p, −134-5p; Abbreviations (in alphabetical order): 9HPT, 9-hole peg test; A2M, alpha-2-macroglobulin; ADA, anti-drug antibodies; AOPP, Advanced oxidation protein products; APLP1, amyloid beta precursor like protein 1; ApoA1, apolipoprotein A1; ARMSS, age-related MS severity scale; ATF7, cyclic AMP-dependent transcription factor ATF-7; AUC, area under the ROC curve; bAbs, IFNb-binding antibodies; C3bCfb, chain F, crystal structure of complement C3b in complex with factor B; CCL20, chemokine (C-C motif) ligand 20; CD6, cluster of differentiation 6; CDA, confirmed disability accumulation; CDCP1, CUB-domain-containing protein 1; CNTN2, contactin-2; CXCL13, chemokine (C-X-C motif) ligand 13; CXCL9, chemokine (C-X-C motif) ligand 9; DMT, disease modifying treatment; EDSS, Expanded Disability Status Scale; FLRT2, fibronectin leucine-rich transmembrane protein 2; GFAP, glial fibrillary acidic protein; HCVA, high contrast vision; IFNb, interferon beta; IL12B, interleukin-12 subunit beta; IQR, interquartile range; LASSO, Least Absolute Shrinkage and Selection Operator; miRNA, microRNA, which are small, non-coding RNA molecules; MOG, myelin oligodendrocyte glycoprotein; MS, multiple sclerosis; MS-DSS, MS disease severity scale, defined thanks to a statistical model [ref (46)] which takes into account, the amount of CNS-tissue destruction measured by Combinatorial MRI scale of CNS tissue destruction (COMRIS-CTD) [ref (43)], and demographic data; MSSS, multiple sclerosis severity scale; Myosin, human skeletal mRNA for myosin heavy chain light meromyosin region; N0, sample size of the training and testing cohort; N1, sample size of the validation cohort; NA, not applicable; nAbs, IFNb-neutralising antibodies; NEDA, no evidence of disease activity; NfL, neurofilament light chain; OPG, osteoprotegerin; OPN, osteopontin; PBMC, peripheral blood mononuclear cells; PDDS, patient-determined disease steps; PDS5B, human androgen-induced prostate proliferative shutoff associated protein (AS3); PMS, progressive MS; PPV, positive predictive value; PRBP, plasma retinol binding protein; PRO, patient-reported outcome; PRTG, protogenin; RRMS, relapsing-remitting MS; SDMT, Symbol Digit Modality Test; SERPINA9, serpin family A member 9; SL25, 2.5% low contrast visual acuity; SNPs, single nucleotide polymorphisms; T25WT, timed 25 feet walking test; TNFSF10A, tumor necrosis factor ligand superfamily member 10; TNFSF13B, tumor necrosis factor ligand superfamily member 13B; VCAN, versican.

10.3389/fimmu.2024.1459502

TABLE 3 Summary of selected studies focused on disease mechanisms.

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
Acquaviva et al., Cell Rep Med. 2020 [ref (34)]	313 subjects: CIS (n=57), RRMS (n=108), SPMS (n=26), PPMS (n=35), OND (n=27) Healthy subjects (n=60)	0	Transcriptomic profiles of PBMCs	Training set: nested cross-validation Validation set: ward DT-based algorithms (RF, FTs and ADAboost-FT)	Raw and processed microarray data from the GEO database, age, sex	MS classifiers: MS vs non-MS Relapsing vs progressive MS	MS vs non-MS: on 139 probes, 94.3% sensitivity and 87.5% precision. Relapsing vs progressive MS: 222 probes, 83.3% sensitivity and 93.8% precision. PPMS vs RRMS: 266 probes, 90% sensitivity and 90% precision. SPMS vs RRMS: 201 probes, 87.5% sensitivity and 100% precision	Identified transcripts in MS vs non-MS: related to interferon signalling, chromatin remodelling and apoptosis. Identified transcripts in relapsing vs progressive MS: related to cell cycle and T cell activation for both progressive forms; protein ubiquitination, cell migration, and fatty acid metabolism for PPMS; and regulation of GTPase activity, locomotor behaviour, and blood coagulation in the SPMS signature.
Sun et al., Front Genet. 2022 [ref (36)]	miRNA-MS associations from the disease-related miRNA from the HMDD. MS-related miRNAs as positive samples, and randomly selected associations with n times the number of positive samples from unlabelled miRNAs associations as negative samples, where n∈(2,10,20,30,40,50)	0	MS-related miRNAs	CNN vs DT, SVM, logistic regression, and GaussianNB	miRNAs	Top 10 predicted miRNAs: hsa-miR-605-5p, hsa-miR-15b-5p, hsa-miR-16-5p, hsa-miR-17-5p, hsa-miR-181a-5p, hsa-miR-181b-5p, hsa-miR-181c-5p, hsa-miR-18a-3p, hsa-miR-195-5p, and hsa-miR-196a-5p.	ROC-AUC 0.87 with CNN	Some of the miRNAs were differentially expressed in RRMS or related to Th17 cell differentiation; one of them (miR-16-5p) decreased in PBMCs after initiation of therapy with interferon β
Lötsch et al., Int J Mol Sci. 2017 [ref (38)]	MS: n=102 Healthy subjects: n=301	0	3 types of lipid biomarkers in serum: eicosanoids: n=11; ceramides: n=10; and lysophosphatidic acids: n=6	ESOM combined with the U*-matrix visualisation technique	Eicosanoids, ceramides and lysophosphatidic acids	Data structures in eicosanoid and ceramide serum concentrations	Ecosanoid concentrations: sensitivity 54%, specificity 100%, accuracy 77%. Ceramid concentrations: sensitivity 89.2%, specificity 100%, accuracy 94.6%.	Lipid metabolism has been suggested to play a critical role in the pathophysiology of MS, influencing inflammation, neurodegeneration, myelin damage, and repair processes
Mezzaroba et al., Mol Neurobiol. 2020 [ref (39)]	MS: n=174 (CIS n=5; RRMS n=144, SPMS n=20, PPMS n=5) Controls: n=182	0	Plasma levels of TNF-α, sTNFR1, sTNFR2, adiponectin,	NNA and RBF/SVM	TNF-α, sTNFR1, sTNFR2, adiponectin, hydroperoxides,	MS vs controls	Low concentrations of four antioxidants (zinc, adiponectin, TRAP and SH groups)	Lower concentrations of all four antioxidants (zinc, adiponectin, TRAP and SH groups) were predictive of MS when compared to controls. TRAP and adiponectin were the

(Continued)

10.3389/fimmu.2024.1459502

TABLE 3 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
			hydroperoxides, AOPP, nitric oxide metabolites, TRAP, SH groups, and serum levels of zinc		AOPP, nitric oxide metabolites, TRAP, SH groups, and zinc		combined with increased sTNFR2: 98.7% sensitivity, 91.7% specificity, AUC-ROC 0.990. SVM analysis (validation): 93.51% training accuracy, 92.03% validation accuracy. NNA training: sensitivity 98.2%, specificity83.3%, AUC-ROC 0.997	most important predictors, followed by zinc and sTNFR2
Goyal et al., Front Neurol. 2019 [ref (40)]	MS: n=910 Healthy volunteers/ controls: n=199		Serum cytokines: IL-1β, IL-2, IL-4, IL-8, IL-10, IL-13, IFN-γ, and TNF-α	SVM, DT, RF and neural networks	IL-1β, IL-2, IL-4, IL-8, IL-10, IL-13, IFN-γ, and TNF-α, age, sex, disease duration, EDSS and MSSS (cytokines for MS vs non-MS, and cytokines and other variables for relapsing vs non-relapsing MS)	MS vs non-MS Relapsing vs non-relapsing MS	MS vs non-MS: RF model: sensitivity 75.6%, 85.7% specificity, 90.91% accuracy, ROC-AUC 0.957 Relapsing vs non-relapsing MS: the RF model had the highest accuracy (70%). In the validation set, the RF model was the best discriminator	Cytokines play an important role in the differentiation of Th cells and recruitment of auto-reactive T and B cells in MS
Seitz et al., Ther Adv Neurol Disord. 2021 [ref (42)]	Early MS: n=156: n=110 with no history of ON n=46 with prior history of ON	0	sNfL levels	SVM	sNfL age, sex, disease duration, EDSS	OCT: OPL volume and atrophy	SVM: sNfL levels 75.7% accurate at predicting OPL volume (training 75.9%, testing 76.2%). Longitudinal analysis of sNfL and OPL in ON eyes: sNfL levels 72.1% accurate at predicting OPL atrophy (training 72.5%, testing 71.8%)	NfL was predominantly expressed in the RNFL, GCIPL and OPL in comparison to other layers (murine retina). The findings suggest NfL and OPL associations may be due mostly to inflammation leading to axonal damage
Kosa et al., Nat Commun. 2022 [ref (43)]	MS: n=227 Healthy subjects: n=24		1305 proteins in CSF	RF	Proteins in CSF, age, sex	MS severity: CombiWISE-based MS-DSS at baseline	Training: baseline MS-DSS: 75 unique biomarkers explaining	Identification of 7 patient clusters differing in CSF concentration of proteins from four protein modules (1. Myeloid lineage/TNF; 2.

(Continued)

10.3389/fimmu.2024.1459502

TABLE 3 Continued

Reference	Training and testing cohort, N	Independent validation cohort, N	Biomarker profiles	AI method: algorithms	Model input	Model output	Model performance	Comments
						and follow-up, and BVD severity outcome	62% variance. MS-DSS on follow-up, 34 unique biomarkers and 35 for BVD explaining 60% variance. Validation: CSF-based MS-DSS at baseline predicted 17% variance, 26% of MS-DSS at follow-up, 22% of BVD severity model	CNS repair; 3. Complement/coagulation; and 4. Adaptive immunity and CNS stress). Cluster 2: predominance of males with progressive MS, relatively low expression in the CNS repair module and high expression in the myeloid lineage/TNF and complement/coagulation modules. These patients had a higher MS severity. Clusters 3 and 4 relatively enriched for female subjects. Cluster 3: high expression of adaptive immunity and CNS module proteins and enriched with relapsing MS subjects. Cluster 4: relatively high expression of all protein modules except for complement/coagulation, with a relatively low MS severity
Gross et al., Brain. 2021 [ref (44)]	Autoimmune neuroinflammatory diseases: n=282 (relapsing MS n=196, NMOSD n=15, Susac syndrome n=14, AE n=57) Degenerative diseases: n=93 (amyotrophic lateral sclerosis n=52, mild Alzheimer´s Disease n=41) Vascular conditions: n=97 Non-inflammatory controls: n=74 (with somatoform disorders or who donated CSF during the course of spinal anesthesia). Total n=546	Additional subjects: n=231 (neuroinflammatory diseases: n=32; neurodegenerative diseases: n=156; neurovascular diseases: n=8; non-inflammatory controls: n=35)	CSF analysis with multiparameter flow cytometry to identify 34 CSF and blood biomarkers after assessing for collinearity	Feature selection with dimensionality reduction and unsupervised cluster analyses	34 CSF and blood features	Neuroinflammatory processes vs other conditions: cells/ml, monocytes, NK cells, and B cells in CSF and CD56dim NK cells in peripheral blood. MS vs other neuroinflammatory disorders: CSF plasma cells and intrathecal IgG synthesis	Neuroinflammatory diseases vs others: 70% sensitivity, 81% specificity, 76% accuracy,ROC-AUC 85% MS vs other neuroinflammatory disorders: Accuracy vs: NMOSD: 87.3%; Susac Syndrome: 95.3%; A E: 89.4%. ROC-AUC vs: NMOSD: 91.5; Susac Syndrome: 90.7; AE: 82.7	MS vs other autoimmune diseases: besides parameters such as intrathecal plasma cells concomitant with IgG synthesis, the analyses identified intrathecal IgA and IgM synthesis. There were other disease-specific parameters, such as alterations in circulating peripheral blood CD56bright NKcells and intrathecal lactate concentrations in NMOSD; circulating CD4+ and CD8+ T cells in Susac Syndrome; and circulating and intrathecal lymphocytes, intrathecal NK T cells, monocytes, and CD14+CD16+ monocytes in AE.

ADAboost-FT, adaptive boosting applied to functional trees; AE, autoimmune encephalitis; AOPP, advanced oxidation protein products; BVD, brain volume deficit; CD, cluster of differentiation; CIS, clinically isolated syndrome; CNN, convolutional neural network; CombiWISE, combinatorial weight-adjusted disability score; CSF, cerebrospinal fluid; DT, decision tree; EDSS, Expandid Disability Status Scale; ESOM, emergent self-organising feature maps; FT, functional trees; GaussianNB, Gaussian Naïve Bayes; GCIPL, macular ganglion cell-inner plexiform layer; GEO, gene expression omnibus data repository; CNS, central nervous system; GTPase, guanosine triphosphate enzyme; HMDD, Human microRNA Disease Database; IFN, interferon; IL, interleukin; miRNA, microRNA; MS, multiple sclerosis; MS-DSS, Multiple Sclerosis Disease Severity Score; MSSS, Multiple Sclerosis Severity Score; NK, natural killer; NMOSD, neuromyelitis optica spectrum disorders; NNA, neural network analysis; OCT, optical coherence tomography; ON, optic neuritis; OND, other neurological diseases; OPL, outer plexiform layer; PBMCs, peripheral blood mononuclear cells; PPMS, primary progressive multiple sclerosis; RBF/SVM, support vector machine with radial basis function; RF, random forests; RNFL, retinal nerve fiber layer; ROC-AUC, receiver-operating characteristic curve-area under the curve; RRMS, relapsing remitting multiple sclerosis; SH, sulphydryl; sNfL, neurofilament light chain in serum; SPMS, secondary progressive multiple sclerosis; sTNFR, soluble tumour necrosis factor receptor; SVM, support vector machine; Th, T helper cells; TNF, tumour necrosis factor; TRAP, total radical-trapping antioxidant parameter.

10.3389/fimmu.2024.1459502

relapsing vs progressive MS classifier showed 83.3% sensitivity and 93.8% precision. Associated biological themes included cell cycle and T cell activation for both progressive forms; protein ubiquitination, cell migration, and fatty acid metabolism for PPMS; and GTPase activity regulation, locomotor behaviour, and blood coagulation in SPMS.

MicroRNAs (miRNAs) play critical roles in post-transcriptomal gene expression regulation. In MS, miRNAs have been implicated in various aspects of the disease's pathophysiology (35). Sun et al. (36) proposed a convolutional neural network (CNN)-based model to identify MS-related miRNAs and compared it to other existing methods: DT, SVM, logistic regression, and Gaussian Naïve Bayes. Using the miRNA-MS associations from the Human microRNA Disease Database (HMDD), the CNN model showed the highest ROC-AUC (0.87). Some of the top 10 predicted miRNAs were differentially expressed in RRMS or related to Th17 cell differentiation, whereas another one decreased after initiation of therapy with interferon β.

Lipid metabolism may influence inflammation, neurodegeneration, myelin damage, and repair processes in MS (37). Lötsch et al. (38) used unsupervised ML implemented as emergent self-organising feature maps (ESOM) combined with the U*-matrix visualisation technique to analyse eicosanoids, ceramides, and lysophosphatidic acids in serum of 102 PwMS and 301 HC, to find distance and density-based structures. Clear data structures were observed in eicosanoid and ceramide concentrations. Whereas the classification of MS vs HC yielded a moderate performance with eicosanoids (54% sensitivity, 100% specificity, 77% accuracy) the structures emerging with ceramides resulted in a high performance (89.2% sensitivity, 100% specificity, 94.6% accuracy).

An imbalance of oxidant and antioxidant molecules has been implicated in demyelination and axonal damage in MS. Mezzaroba et al. (39) used supervised ML (neural network analysis [NNA] and SVM with radial basis function [RBF/SVM]) to evaluate discriminatory patterns in plasma of 9 oxidants and antioxidants and zinc serum levels, in 174 PwMS and 182 controls. The combination of low levels of four antioxidants and increased levels of one oxidant yielded the best prediction for MS (sensitivity 98.7%, specificity 91.7%, AUC-ROC 0.990). The SVM analyses obtained 93.51% training and 92.03% validation accuracies *(39)*.

Cytokines play an important role in Th cell differentiation and recruitment of auto-reactive T and B cells in MS. Goyal et al. (40) used four ML models (SVM, DT, RF, and neural networks) to identify serum cytokines predictive of MS. They also assessed the cytokines with age, sex, disease duration, EDSS, and MSSS to classify MS into remitting and non-remitting MS. They used 910 serum samples from PwMS and 199 from HC (total n=1109). Of these, 900 were included in the training set and 209 in the testing set. RF was the model that best predicted MS (sensitivity 75.6%, specificity 85.7%, accuracy 90.91%, ROC-AUC 0.957) and also had the highest accuracy (70%) to differentiate relapsing from non-relapsing MS. In the validation set, the RF model was again the best discriminator (40).

Neurofilament light chain (NfL) is a biomarker of axonal damage in MS (41). Seitz et al. (42) used SVM analysis to test for associations between baseline serum NfL (sNfL) and different retinal thickness measures in 156 early MS patients: 110 with no history of optic neuritis (ON) and 46 with ON. After adjusting for age, sex, disease duration, and EDSS, a significant correlation was found only between high sNfL levels and low outer plexiform layer (OPL) volume in patients with a history of ON. Follow-up OCTs available for 38 subjects with a mean (SD) follow-up of 2.1 (1.4) years showed baseline sNfL correlated with absolute OPL atrophy in ON. sNfL levels predicted OPL volume with 75.9% training and 76.2% testing accuracies. In the longitudinal analysis, sNfL predicted OPL atrophy with 72.5% training and 71.8% testing accuracies.

Other studies have focused on CSF biomarkers. Kosa et al. (43) used RF to search for biomarkers among 1305 proteins in CSF of 227 PwMS to build models predictive of disease severity. To differentiate natural aging and sex effects from MS-related mechanisms they used data from 24 HC. MS severity was assessed using the combinatorial weight-adjusted disability score (CombiWISE)-based MS Disease Severity Score (MS-DSS) measured at baseline and follow-up, and the brain volume deficit (BVD) severity outcome, based on linear regression models of brain parenchymal fraction and age, calculated from MRIs performed within 3 months of CSF collection. Initial analyses demonstrated positive associations of coagulation and complement cascades and negative associations for NOTCH signalling and neuron recognition categories with MS severity. After adjusting for age and sex, the model selected 75 biomarkers explaining 62% of variance for baseline MS-DSS. For follow-up MS-DSS, 34 biomarkers were selected and 35 for BVD explaining 60% of variance. The effect sizes decreased to 17%, 26%, and 22% of variance in the validation cohort (n=98). Using unsupervised cluster analyses, the authors identified seven patient clusters differing in CSF protein concentrations from four protein modules. Of note, one cluster had a predominance of men with progressive MS, a relatively low expression in the CNS repair module and high expression in the myeloid lineage/TNF and complement/coagulation modules. These patients had a higher MS severity.

Cellular characterisation in blood and CSF can help differentiate between CNS disorders and clarify their pathophysiological processes. Gross et al. (44) combined feature selection with dimensionality reduction and unsupervised cluster analyses to investigate parameters altered across autoimmune neuroinflammatory diseases [RRMS n=196, neuromyelitis optica spectrum disorders (NMOSD) n=15, Susac syndrome n=14, autoimmune encephalitis (AE) n=57], other CNS conditions (neurodegenerative n=93, vascular n=97), and non-inflammatory controls (n=74) (total n=546). The validation cohort included 231 additional subjects (neuroinflammatory n=32, neurodegenerative n=156, neurovascular n=8, non-inflammatory controls n=35). Exploratory analyses identified four CSF parameters and one peripheral blood parameter that together discriminated neuroinflammatory diseases from other groups (70% sensitivity,

81% specificity, 76% accuracy, ROC-AUC of 85%). When aiming to differentiate MS from other neuroinflammatory diseases, CSF plasma cells and intrathecal IgG synthesis alone were sufficient to distinguish RRMS from other neuroinflammatory diseases with high accuracy and ROC-AUC (NMOSD: 87.3% and 91.5%; Susac syndrome: 95.3% and 90.7%; AE: 89.4% and 82.7%). Finally, the authors compared cell profiles in RIS, CIS and early RRMS (≤36 months from disease onset) vs late RRMS (>36 months). Alterations in the proportions of CD56dim NK cells and biomarkers of intrathecal inflammation gradually increased during disease evolution. When splitting RRMS based on inflammatory activity, minor effects were shown in most intrathecal parameters, whereas changes in peripheral and intrathecal CD4+CD8+ T cells and intrathecal plasma cells were more pronounced.

Limitations of AI-based research in MS fluid biomarkers

AI-based studies using fluid biomarkers in MS offer promising results. However, these studies have limitations which are worth being mentioned. In general, all these studies still have relatively small sample sizes, which, together with the lack of external validation analyses in many of them, limit the generalisability of the results. Also, despite the low number of studies published so far, there is a large methodological variability, which, at times, is not explained in detail, making it very difficult to replicate the analyses done (Tables 1–3). These limitations are common to all AI-based studies that harness biomarker data to improve the diagnosis, predict or understand the disease, thus hampering the application of all these models to clinical practice.

In relation to the specific limitations of those studies focused on diagnosis, the number and types of diseases which have been compared with are limited. Furthermore, many of the tests (biomarkers) used by the authors are not available in routine clinical practice. These aspects reduce the utility of these models in practice, at least in the short term, suggesting the need for more research.

Regarding the studies focused on prediction of disease evolution, apart from the general limitations abovementioned, many of them have cross-sectional designs or, if they have a longitudinal design, there is a relatively short follow-up time in most of the cases. Also, very often, the effect of treatment is not taken into account. Furthermore, most studies were not adjusted for important demographic, clinical and technical aspects, such as race, ethnicity, disease duration, brain volume, and the interval between sampling and relapses or their treatment. Finally, despite the developments in AI-based models in MS which use raw neuroimaging and deep learning techniques to predict clinical outcome, the integration of these into AI-based models which use fluid biomarkers (or the other way around) is still lacking. Little is known about the complementary roles of both types of predictors and the potential synergies between them. However, it is highly likely that only when both are used together in comprehensive models, a real impact on the clinical management of MS can be achieved. Such integration requires, though, intensive methodological research which will hopefully bear fruit in the near future.

Lastly, regarding the limitations of the studies focused on understanding disease mechanisms, many of them are far too focused on certain paths or predictors, therefore not allowing us to explain or understand the whole picture. Also, very importantly, the fact that many of these biomarkers, paths, or predictors, may explain the same variance of a given outcome measure but we are not aware of that – because typically one study tends to focus on a given path – implies that many of the associations found may be reflecting mere epiphenomena rather than causally related events. Whereas this might be less relevant for building predictive models, for those studies which aim at understanding the disease through AI, this may be deleterious.

Conclusions and future directions

The application of AI-based methodologies to tackle key challenges in MS is exponentially increasing. However, in this context, the number of studies published in the literature focusing on the use of fluid biomarker data is still small. Most of these publications are focused on serum biomarkers, genetic variants, and gene expression profiles as predictors. Of note, only half of them have included an external validation analysis of the developed AI model, thus hampering a full interpretation of the results and their potential generalisability.

Importantly, after the assessment of the papers published so far, it may be said that the research on AI applied to biomarker data is still quite in its early days and that we are still far from clinical applications. So far, AI methodologies have been very useful for biomarker discovery in MS, but the large heterogeneity of methods and results suggests that we may need many years of research before prototypes can be launched to help healthcare professionals and patients in the clinic.

Along the same lines, even though many studies reported much higher accuracy levels when fluid biomarker, MRI, and clinical data were combined as predictors of diagnosis or disease evolution, large studies combining the most important types of predictor acquired in the clinic are lacking. Only when these take place and are replicated in large independent cohorts will we be able to comprehend their full potential and start considering that a change in patient management thanks to the introduction of those AI-based models is possible. Of note, for these models to be useful in the clinic, they need to use, as input data (predictors), routinely-acquired biomarkers, including laboratory, imaging, and clinical data. On the other hand, it is possible that a branch of AI-based research in MS, i.e., that focused on understanding the pathogenic mechanisms and those processes underlying disability accumulation, continues to exist with the use of less common (non-routinely acquired) biomarkers. This research is also important and will surely bring to light crucial knowledge on the disease, essential for its ultimate eradication. A final conclusion is that all studies carried out so far confirm the leading role of inflammatory pathways in MS.

Future directions include the development of larger studies with validation in independent datasets. Also, future directions should aim at the design of longitudinal studies with longer follow-ups (for those mainly focused on future prediction), hopefully accounting

for the complex effects of disease-modifying treatments and other dynamic data, as well as the integration of fluid biomarkers, neuroimaging, optical coherence tomography (OCT) imaging, and clinical predictor data to build robust and powerful models.

Furthermore, forthcoming research endeavours must transition from the current exploratory phase of AI-based methodologies applied to biomarker data in MS to a more translational stage. This shift necessitates thorough evaluation of the clinical utility of the constructed AI models. For that, the future lies in creating guidelines for AI-based analyses to improve the comparability across studies, to shed light on the steps needed to go from discovery to clinical practice implementation, and to evaluate utility of AI-based algorithms in practice. Additionally, we should be able to learn from AI-based investigations on other neurodegenerative diseases (45) to overcome the challenges surrounding these types of studies.

As a final consideration, it is imperative to recognise that addressing ethical and inequality concerns surrounding AI-based analyses is just as crucial as resolving technical challenges. With the exponential growth of AI studies, maintaining research integrity in AI research demands not only initial attention but also ongoing evolution, keeping pace with the rapid advancement of science to meet the needs and expectations of us all.

Author contributions

GA: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. MC: Writing – original draft, Writing – review & editing. CT: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. CT is currently being funded by the Miguel Servet contract, awarded by the Instituto de Salud Carlos III (ISCIII), Spanish Ministry of Science and Innovation (award number: CP23/00117). She has also received research support from the ISCIII through the FORTALECE grant (FORT23/00034).

Conflict of interest

GA has received compensation for consulting services, speaking honoraria or participation in advisory boards from Merck, Roche, and Horizon Therapeutics; and travel support for scientific meetings from Novartis, Roche, ECTRIMS and EAN. She serves as editor for Europe of the Multiple Sclerosis Journal – Experimental, Translational and Clinical journal; and as a member of the editorial and scientific committee of Acta Neurológica Colombiana. She is a member of the International Women in Multiple Sclerosis iWiMS network executive committee, of the European Biomarkers in Multiple Sclerosis BioMS-eu steering committee, and of the MOGAD Eugene Devic European Network MEDEN steering group.

MC has received compensation for consulting services and speaking honoraria from Bayer Schering Pharma, Merck Serono, Biogen-Idec, Teva Pharmaceuticals, Sanofi-Aventis, Genzyme, and Novartis.

CT is currently being funded by a Miguel Servet contract, awarded by the Instituto de Salud Carlos III ISCIII, Ministerio de Ciencia e Innovación de España CP23/00117. She has also received a 2020 Junior Leader La Caixa Fellowship fellowship code: LCF/BQ/PI20/11760008, awarded by "la Caixa" Foundation ID 100010434, a 2021 Merck's Award for the Investigation in MS, awarded by Fundación Merck Salud Spain, a 2021 Research Grant PI21/01860 awarded by the ISCIII, Ministerio de Ciencia e Innovación de España, and a FORTALECE research grant FORT23/00034 also by the ISCIII, Ministerio de Ciencia e Innovación de España. In 2015, she received an ECTRIMS Post-doctoral Research Fellowship and has received funding from the UK MS Society. She is a member of the Editorial Board of Neurology Journal and Multiple Sclerosis Journal. She has also received honoraria from Roche, Sanofi, Bristol-Myers Squibb, and Novartis and is a steering committee member of the O'HAND trial and of the Consensus group on Follow-on DMTs.

Publisher's note

References

1. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. *Lancet*. (2020) 395:1579–86. doi: 10.1016/S0140-6736(20)30226-9

2. La Rosa F, Wynen M, Al-Louzi O, Beck ES, Huelnhagen T, Maggi P, et al. Cortical lesions, central vein sign, and paramagnetic rim lesions in multiple sclerosis: Emerging machine learning techniques and future avenues. *NeuroImage Clin*. (2022) 36:103205. doi: 10.1016/j.nicl.2022.103205

3. Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. *N Engl J Med*. (2018) 378:169–80. doi: 10.1056/NEJMra1401483

4. Tumani H, Hartung H-P, Hemmer B, Teunissen C, Deisenhammer F, Giovannoni G, et al. Cerebrospinal fluid biomarkers in multiple sclerosis. *Neurobiol Dis*. (2009) 35:117–27. doi: 10.1016/j.nbd.2009.04.010

5. McCulloch W, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. (1943) 5:115–133.

6. Solomon AJ, Arrambide G, Brownlee WJ, Flanagan EP, Amato MP, Amezcua L, et al. Differential diagnosis of suspected multiple sclerosis: an updated consensus approach. *Lancet Neurol*. (2023) 22:750–68. doi: 10.1016/S1474-4422(23)00148-5

7. Pasella M, Pisano F, Cannas B, Fanni A, Cocco E, Frau J, et al. Decision trees to evaluate the risk of developing multiple sclerosis. *Front Neuroinform*. (2023) 17:1248632. doi: 10.3389/fninf.2023.1248632

8. Guo P, Zhang Q, Zhu Z, Huang Z, Li K. Mining gene expression data of multiple sclerosis. *PloS One*. (2014) 9:e100052. doi: 10.1371/journal.pone.0100052

9. Wang R, Li B, Lam SM, Shui G. Integration of lipidomics and metabolomics for in-depth understanding of cellular mechanism and disease progression. *J Genet Genomics*. (2020) 47:69–83. doi: 10.1016/j.jgg.2019.11.009

10. Andersen SL, Briggs FBS, Winnike JH, Natanzon Y, Maichle S, Knagge KJ, et al. Metabolome-based signature of disease pathology in MS. *Mult Scler Relat Disord*. (2019) 31:12–21. doi: 10.1016/j.msard.2019.03.006

11. Lötsch J, Schiffmann S, Schmitz K, Brunkhorst R, Lerch F, Ferreiros N, et al. Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy. *Sci Rep*. (2018) 8:14884. doi: 10.1038/s41598-018-33077-8

12. Probert F, Yeo T, Zhou Y, Sealey M, Arora S, Palace J, et al. Determination of CSF GFAP, CCN5, and vWF levels enhances the diagnostic accuracy of clinically defined MS from non-MS patients with CSF oligoclonal bands. *Front Immunol*. (2021) 12:811351. doi: 10.3389/fimmu.2021.811351

13. Gaetani L, Bellomo G, Di Sabatino E, Sperandei S, Mancini A, Blennow K, et al. The immune signature of CSF in multiple sclerosis with and without oligoclonal bands: A machine learning approach to proximity extension assay analysis. *Int J Mol Sci*. (2023) 25(1):139. doi: 10.3390/ijms25010139

14. Martynova E, Goyal M, Johri S, Kumar V, Khaibullin T, Rizvanov AA, et al. Serum and cerebrospinal fluid cytokine biomarkers for diagnosis of multiple sclerosis. *Mediators Inflammation*. (2020) 2020:2727042. doi: 10.1155/2020/2727042

15. Tintore M, Rovira À, Río J, Otero-Romero S, Arrambide G, Tur C, et al. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. *Brain*. (2015) 138:1863–74. doi: 10.1093/brain/awv105

16. Chung KK, Altmann D, Barkhof F, Miszkiel KA, Brex PA, O'Riordan J, et al. A 30-Year Clinical and Magnetic Resonance imaging observational study of multiple sclerosis and clinically isolated syndromes. *Ann Neurol*. (2020) 87(1):63–74. doi: 10.1002/ana.25637

17. Scalfari A, Romualdi C, Nicholas RS, Mattoscio M, Magliozzi R, Morra A, et al. The cortical damage, early relapses, and onset of the progressive phase in multiple sclerosis. *Neurology*. (2018) 90:e2107–18. doi: 10.1212/WNL.0000000000005685

18. Brownlee WJ, Altmann DR, Prados F, Miszkiel KA, Eshaghi A, Gandini Wheeler-Kingshott CAM, et al. Early imaging predictors of long-term outcomes in relapse-onset multiple sclerosis. *Brain*. (2019) 142:2276–87. doi: 10.1093/brain/awz156

19. Flauzino T, Simão ANC, de Carvalho Jennings Pereira WL, Alfieri DF, Oliveira SR, Kallaur AP, et al. Disability in multiple sclerosis is associated with age and inflammatory, metabolic and oxidative/nitrosative stress biomarkers: results of multivariate and machine learning procedures. *Metab Brain Dis*. (2019) 34:1401–13. doi: 10.1007/s11011-019-00456-7

20. Kurtzke JF. Rating neurologic impairment in multiple sclerosis. *Neurology*. (1983) 33:1444–4. doi: 10.1212/WNL.33.11.1444

21. Brummer T, Muthuraman M, Steffen F, Uphaus T, Minch L, Person M, et al. Improved prediction of early cognitive impairment in multiple sclerosis combining blood and imaging biomarkers. *Brain Commun*. (2022) 4:fcac153. doi: 10.1093/braincomms/fcac153

22. Jackson KC, Sun K, Barbour C, Hernandez D, Kosa P, Tanigawa M, et al. Genetic model of MS severity predicts future accumulation of disability. *Ann Hum Genet*. (2020) 84:1–10. doi: 10.1111/ahg.12342

23. Andorra M, Freire A, Zubizarreta I, de Rosbo NK, Bos SD, Rinas M, et al. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. *J Neurol*. (2024) 271(3):1133–49. doi: 10.1007/s00415-023-12132-z

24. Ferrè L, Clarelli F, Pignolet B, Mascia E, Frasca M, Santoro S, et al. Combining clinical and genetic data to predict response to fingolimod treatment in relapsing remitting Multiple Sclerosis patients: A precision medicine approach. *J Pers Med*. (2023) 13(1):122. doi: 10.3390/jpm13010122. Epub ahead of print.

25. Campagna MP, Xavier A, Lea RA, Stankovich J, Maltby VE, Butzkueven H, et al. Whole-blood methylation signatures are associated with and accurately classify multiple sclerosis disease severity. *Clin Epigenet*. (2022) 14:194. doi: 10.1186/s13148-022-01397-2

26. Fagone P, Mazzon E, Mammana S, Di Marco R, Spinasanta F, Basile MS, et al. Identification of CD4+ T cell biomarkers for predicting the response of patients with relapsing–remitting multiple sclerosis to natalizumab treatment. *Mol Med Rep*. (2019) 20:678–84. doi: 10.3892/mmr.2019.10283

27. Baranzini SE, Madireddy LR, Cromer A, D'Antonio M, Lehr L, Beelke M, et al. Prognostic biomarkers of IFNb therapy in multiple sclerosis patients. *Mult Scler*. (2015) 21:894–904. doi: 10.1177/1352458514555786

28. Uphaus T, Steffen F, Muthuraman M, Ripfel N, Fleischer V, Groppa S, et al. NfL predicts relapse-free progression in a longitudinal multiple sclerosis cohort study. *EBioMedicine*. (2021) 72:103590. doi: 10.1016/j.ebiom.2021.103590

29. Herman S, Arvidsson McShane S, Zjukovskaja C, Khoonsari PE, Svenningsson A, Burman J, et al. Disease phenotype prediction in multiple sclerosis. *iScience*. (2023) 26:106906. doi: 10.1016/j.isci.2023.106906

30. Zhu W, Chen C, Zhang L, Hoyt T, Walker E, Venkatesh S, et al. Association between serum multi-protein biomarker profile and real-world disability in multiple sclerosis. *Brain Commun*. (2023) 6(1):fcad300. doi: 10.1093/braincomms/fcad300

31. Everest E, Uygunoglu U, Tutuncu M, Bulbul A, Onat UI, Unal M, et al. Prospective outcome analysis of multiple sclerosis cases reveals candidate prognostic cerebrospinal fluid markers. *PloS One*. (2023) 18:e0287463. doi: 10.1371/journal.pone.0287463

32. Ebrahimkhani S, Beadnall HN, Wang C, Suter CM, Barnett MH, Buckland ME, et al. Serum exosome MicroRNAs predict multiple sclerosis disease activity after fingolimod treatment. *Mol Neurobiol*. (2020) 57:1245–58. doi: 10.1007/s12035-019-01792-6

33. Waddington KE, Papadaki A, Coelewij L, Adriani M, Nytrova P, Kubala Havrdova E, et al. Using serum metabolomics to predict development of anti-drug antibodies in multiple sclerosis patients treated with IFNβ. *Front Immunol*. (2020) 11:1527. doi: 10.3389/fimmu.2020.01527

34. Acquaviva M, Menon R, Di Dario M, Dalla Costa G, Romeo M, Sangalli F, et al. Inferring multiple sclerosis stages from the blood transcriptome via machine learning. *Cell Rep Med*. (2020) 1:100053. doi: 10.1016/j.xcrm.2020.100053

35. Junker A, Hohlfeld R, Meinl E. The emerging role of microRNAs in multiple sclerosis. *Nat Rev Neurol*. (2011) 7:56–9. doi: 10.1038/nrneurol.2010.179

36. Sun X, Ren X, Zhang J, Nie Y, Hu S, Yang X, et al. Discovering miRNAs associated with Multiple Sclerosis based on network representation learning and deep learning methods. *Front Genet*. (2022) 13:899340. doi: 10.3389/fgene.2022.899340

37. Lorincz B, Jury EC, Vrablik M, Ramanathan M, Uher T. The role of cholesterol metabolism in multiple sclerosis: From molecular pathophysiology to radiological and clinical disease activity. *Autoimmun Rev*. (2022) 21:103088. doi: 10.1016/j.autrev.2022.103088

38. Lötsch J, Thrun M, Lerch F, Brunkhorst R, Schiffmann S, Thomas D, et al. Machine-learned data structures of lipid marker serum concentrations in multiple sclerosis patients differ from those in healthy subjects. *Int J Mol Sci*. (2017) 18(6):1217. doi: 10.3390/ijms18061217

39. Mezzaroba L, Simão ANC, Oliveira SR, Flauzino T, Alfieri DF, de Carvalho Jennings Pereira WL, et al. Antioxidant and anti-inflammatory diagnostic biomarkers in multiple sclerosis: A machine learning study. *Mol Neurobiol*. (2020) 57:2167–78. doi: 10.1007/s12035-019-01856-7

40. Goyal M, Khanna D, Rana PS, Khaibullin T, Martynova E, Rizvanov AA, et al. Computational intelligence technique for prediction of multiple sclerosis based on serum cytokines. *Front Neurol*. (2019) 10:781. doi: 10.3389/fneur.2019.00781

41. Khalil M, Teunissen CE, Lehmann S, Otto M, Piehl F, Ziemssen T, et al. Neurofilaments as biomarkers in neurological disorders — towards clinical application. *Nat Rev Neurol*. (2024) 20(5):269–87. doi: 10.1038/s41582-024-00955-x

42. Seitz CB, Steffen F, Muthuraman M, Uphaus T, Krämer J, Meuth SG, et al. Serum neurofilament levels reflect outer retinal layer changes in multiple sclerosis. *Ther Adv Neurol Disord*. (2021) 14:17562864211003478. doi: 10.1177/17562864211003478

43. Kosa P, Barbour C, Varosanec M, Wichman A, Sandford M, Greenwood M, et al. Molecular models of multiple sclerosis severity identify heterogeneity of pathogenic mechanisms. *Nat Commun*. (2022) 13:7670. doi: 10.1038/s41467-022-35357-4

44. Gross CC, Schulte-Mecklenbeck A, Madireddy L, Pawlitzki M, Strippel C, Räuber S, et al. Classification of neurological diseases using multi-dimensional CSF analysis. *Brain*. (2021) 144:2625–34. doi: 10.1093/brain/awab147

45. Kaur A, Mittal M, Bhatti JS, Thareja S, Singh S. A systematic literature review on the significance of deep learning and machine learning in predicting Alzheimer's disease. *Artif Intell Med*. (2024) 154:102928. doi: 10.1016/j.artmed.2024.102928

46. Weideman AM, Barbour C, Tapia-Maltos MA, Tran T, Jackson K, Kosa P, et al. New multiple sclerosis disease severity scale predicts future accumulation of disability. *Front Neurol*. (2017) 8:598. doi: 10.3389/fneur.2017.00598

47. Avsar T, Durası İM, Uygunoğlu U, Tütüncü M, Demirci NO, Saip S, et al. CSF proteomics identifies specific and shared pathways for multiple sclerosis clinical subtypes. *PloS One*. (2015) 10:e0122045. doi: 10.1371/journal.pone.0122045

frontiers | Frontiers in Immunology

TYPE Perspective
PUBLISHED 22 October 2024
DOI 10.3389/fimmu.2024.1461361

OPEN ACCESS

EDITED BY
Eugenio Pucci,
AST Fermo Marche Region Health System, Italy

REVIEWED BY
Gianluigi Mancardi,
University of Genoa, Italy
Adamantios Koumpis,
University Hospital of Cologne, Germany

*CORRESPONDENCE
Liesbet M. Peeters
liesbet.peeters@uhasselt.be

RECEIVED 08 July 2024
ACCEPTED 02 October 2024
PUBLISHED 22 October 2024

CITATION
Peeters LM (2024) The arisal of data spaces: why I am excited and worried.
Front. Immunol. 15:1461361.
doi: 10.3389/fimmu.2024.1461361

The arisal of data spaces: why I am excited and worried

Liesbet M. Peeters[1,2,3]*

[1]University MS Center (UMSC), Hasselt-Pelt, Belgium, [2]Biomedical Research Institute (BIOMED), Hasselt University, Diepenbeek, Belgium, [3]Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium

This paper explores the significant role of real-world data (RWD) in advancing our understanding and management of Multiple Sclerosis (MS). RWD has proven invaluable in MS research and care, offering insights from larger and diverse patient populations. A key focus of the paper is the European Health Data Space (EHDS), a significant development that promises to change how healthcare data is managed across Europe. This initiative is particularly relevant to the MS community. The paper highlights various data initiatives, discussing their importance for those affected by MS. Despite the potential benefits, there are challenges and concerns, especially about ensuring that the growth of various data platforms remains beneficial for MS patients. The paper suggests practical actions for the global MS community to consider, aimed at optimizing the use of RWD. The emphasis of this discussion is on the secondary use of health data, particularly in the European context. The content is based on the author's own experiences and interpretations, offering a personal yet informed view on using RWD to improve MS research and patient care.

KEYWORDS

real-world data, European Health Data Space, secondary use of health data, collaborative research, data interoperability

Introduction

The multiple sclerosis (MS) community is fortunate to have a longstanding and successful legacy of using real-world data (RWD, Table 1) to address complex clinical problems. RWD often reflects larger and more representative populations and therefore is specifically fit-for-purpose to investigate for example disease behavior in a real-world setting, validation of outcome measures, comparative effectiveness and long-term safety of therapies. Additionally, RWD plays a crucial role in enhancing patient advocacy by informing policies on employment, reimbursement of treatments and access to healthcare services, as well as supporting routine healthcare practices. A growing number of real-world MS databases and registries produce long-term outcome data from large cohorts of people with MS (1–3).

The heterogeneity in MS management across Europe, combined with the variability in data collection methods (different formats and data acquisition software systems used

TABLE 1 Glossary - for the purpose of this paper, the following definitions of concepts and terminologies are introduced as follows:.

- **Data space**: Comprehensive term that captures various dimensions of data handling, from its storage and organization to its processing, access and analytical use.
- **Real-World Data (RWD)**: Pragmatically defined as any data that is gathered in the context of standard care as opposed to data gathered in an experimental setting such as a randomized clinical trial. Examples include registry data and data collected and stored using electronic health records (EHR). Real-world-evidence (RWE) is defined as any evidence generated using RWD.
- **Core dataset**: Set of variables that represent the common denominator across different initiatives and their accompanying (minimal) datasets.
- **Common Data Model (CDM)**: Standardized representation of content, independent from a purpose or research question, combined with a defined common infrastructure. Its purpose is to enable collaborative analyses by providing a defined framework and structure.
- **Primary use of health data**: When health data is used to deliver health care to the individual from whom it is collected. For example: an MRI measurement taken for the purpose of diagnosing MS.
- **Secondary (re-)use of health data:** When (existing) health data, originally collected for a specific primary purpose, is used for alternative objectives or research that differs from the initiative reason for data collection. For example: data originally collected for patient care and treatment optimization is re-used to inform regulatory policies and decisions, potentially leading to improved treatment guidelines and enhanced patient safety in the MS patient community.
- **Patient registries:** Organized systems that use observational methods to collect uniform data on a population defined by a particular disease, condition or exposure, and that is followed over time.
- **Big data:** large datasets which may be complex, multi-dimensional, unstructured and heterogeneous, which are accumulating rapidly and which may be analyzed computationally to reveal patterns, trends, and associations (e.g. RWD (such as electronic health records, insurance claims data and data from patient registries), genomics, clinical trials, spontaneous adverse drug reaction reports, social media and wearable devices).

across various data sources and MS registries), presents significant challenges. These differences can impact the interpretation of RWD at scale. Despite these challenges, the research community has realized that combining data from diverse sources across the globe presents significant opportunities for advancing our understanding of MS. To manage the challenges associated with heterogeneity, strategies such as incorporating detailed information about the origin and specification of the source data, ensuring use of high-quality data, involving domain experts in interpreting results, and investing in data harmonization strategies are essential. These approaches have enabled the research community to turn these challenges into opportunities, as seen in initiatives like the Big Multiple Sclerosis Data Network (BMSD - bigmsdata.org) and the COVID-19 in MS Global Data Sharing initiative (GDSI).

BMSD is the largest real-world MS data network and brings together leading MS registries and databases to allow joint analyses of very large merged or federated sets of structured clinical data. It was initiated in 2014 and currently consists of the national MS registries of the Czech Republic (4), Denmark (5), France (6), Italy (7) and Sweden (8) as well as the international MSBase (9). The total number of MS patients in BMSD amounts to over 250,000. In recent years, the BMSD has led on several studies, yielding critical data-driven insights into MS treatment and progression. For example, they uncovered significant patterns in treatment management strategies (10) and disability progression in secondary progressive MS (11). GDSI was project led by the MS Data Alliance and MS International Federation in collaboration with a multitude of global partners (12). In March 2020, as COVID-19 spread, the demand for data on its impact on people with MS surged. Within months, 19 global partners shared data on over 10,000 people with MS, which helped update global advice for MS patients regarding COVID-19 (13–15).

While the MS community has made significant strides in utilizing RWD for research and patient care, several existing and emerging large-scale collaborative efforts across Europe – though not specific for MS – are set to profoundly impact how RWD is managed and utilized across various disease, including MS. In the following paragraphs, several of these key initiatives will be highlighted and explained in detail, focusing on their objectives, relevance to the MS community, and the potential benefits of engaging with them. These 'highlighted initiatives' represent transformative efforts that are shaping the future of healthcare data. However, while they offer exciting possibilities, they also present unique challenges. The subsequent discussion will explore these challenges and offer actionable recommendations to help the MS community effectively navigate this evolving landscape, mitigate risks, and maximize the opportunities these initiatives provide.

Highlighted initiative 1: The European Health Data Space (EHDS) – a revolutionary legislative framework

The EHDS is set to revolutionize healthcare management across a wide spectrum of stakeholders. Europe has been making continuous efforts aiming at enhancing the harmonization and integration of health data, which is needed in order to be able to create a digitized and connected healthcare system, as foreseen in the EHDS regulation. The EHDS proposal aspires to (i) support individuals to take control of their own health data, (ii) support the use of health data for better healthcare delivery, better research, innovation and policy making and (iii) enables the EU to make full use of the potential offered by a safe and secure exchange, use and reuse of health data (16). Two projects, while differing in focus, collectively aspire to enhance the concrete implementation of the EHDS: TEHDAS and HealthData@EU. TEHDAS (Towards The European Health Data Space - tehdas.eu), running from February 2021 to July 2023, focused on developing principles for the secondary use of health data, emphasizing dialogue and engagement across stakeholders, and establishing governance models for cross-border cooperation. This project involved 25 European countries and numerous stakeholders in discussions about health data usage for research and policymaking. In contrast, the HealthData@EU Pilot (ehds2pilot.eu), launched in October 2022, is building a pilot infrastructure for the EHDS, focusing on infrastructure development, testing, and evaluation. Involving 17 partners, this project aims to connect data platforms, develop services for research project support, and provide guidelines for data standards and security.

Highlighted initiative 2: DARWIN-EU – an initiative by the European Medicine Agency (EMA)

The EMA has gained significant interest in the use of RWD to assess the benefit-risk of medicines across their lifecycle and to monitor the safety of medicine, specifically post-authorisation. A post-authorisation safety study (PASS) is a study that is carried out after a medicine has been authorized to obtain further information on a medicine's safety, or to measure the effectiveness of risk-management measures. Figure 1 highlights some of the key activities of EMA and/or the Heads of Medicine Agencies (HMA) with respective timelines.

The initiative for patient registries, launched in September 2015, aspired to explore ways of expanding the use of patient registries by introducing and supporting a systematic and standardized approach to their contribution to the benefit-risk evaluation of medicines (17). Within the scope of this initiative, two workshops of specific interest were hosted and summarized in extensive reports: (i) A more general disease-agnostic workshop on patient registries (2016) to better understand the barriers and facilitators to collaboration between stakeholders. The workshop report provides recommendations on actions to improve stakeholder collaboration and optimize the use of registries to support regulatory decision-making (18): (ii) An MS specific workshop aiming to reach consensus on implementable MS specific recommendations for advancing the systematic use of MS registries to support regulatory evaluations. Similar workshops were hosted for other disease registries such as for example haemophilia (19), cystic fibrosis (20) and cancer (21).

After a short period of public consultation, the guideline on registry-based studies was published in 2021. This guideline addresses the methodological, regulatory and operational aspects involved in using registry-based studies to support regulatory decision-making. It aims to help with defining study populations and designing study protocols. It provides guidance on data collection, data quality management and data analyses to achieve high quality evidence (22). Meta-data catalogues offering descriptive statistics will further support data quality assessment, and evolving guidelines on data quality criteria will continue to improve and standardize this process.

The HMA-EMA Joint Big Data Taskforce Phase II report (23) suggests how the European regulatory network can use Big Data to improve public health and innovation. The first and top priority activity formulated is to deliver a sustainable platform to access and analyze healthcare data from across the EU (Data Analysis and Real World Interrogation Network - DARWIN - darwin-eu.org). Other priority recommendations include to establish a framework for data quality and to enhance data discoverability by strengthening the current European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (EnCePP) resources databases (24) in line with the 'Good Practice Guide for the use of the Metadata Catalogue of RWD sources' (25).

Highlighted initiative 3: the Observational Medical Outcomes Partnership (OMOP) – driving data harmonization

The freely available OMOP (Observational Medical Outcomes Partnership) common data model (CDM) refers to the open community standardized data model, which is designed to

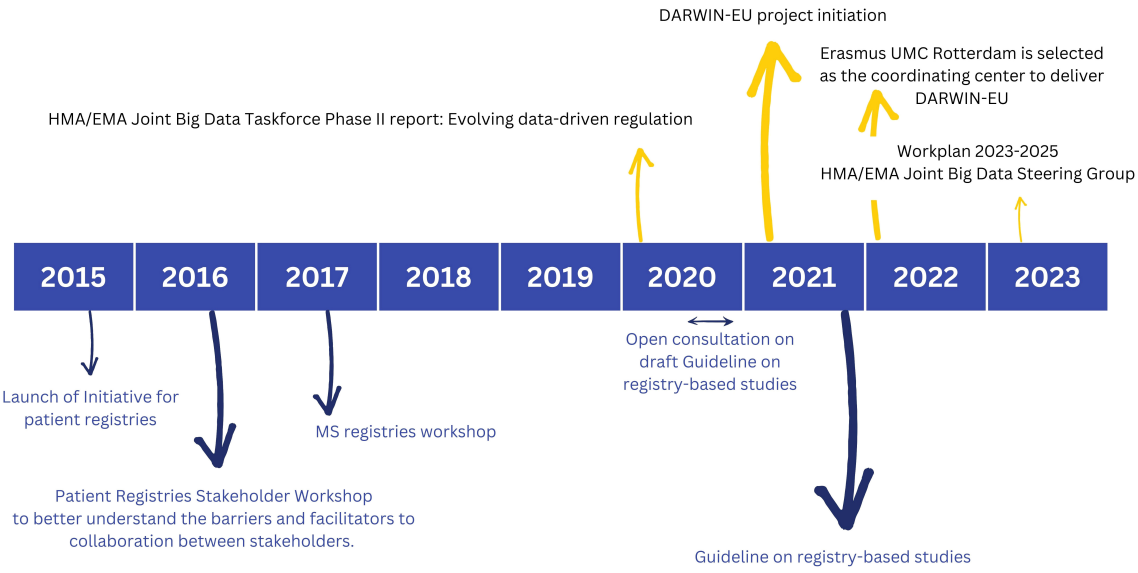


FIGURE 1
Highlighted activities of the European Medicine Agency (EMA) and/or the Heads of Medicine Agencies (HMA) with respective timelines focusing on 'patient registries' (bottom) and 'big data' (top). Reference documents to learn more include: Initiative for patient registries - Strategy and Mandate of the Cross-Committee Task Force (17); Guideline on registry-based studies (22); HMA/EMA Joint Big Data Taskforce Phase II report (23); Report - Patient Registry Workshop (18); Report on Multiple Sclerosis (MS) Registries (42): Work Plan 2023-2025 HMA/EMA Joint Big Data Steering Group (51).

integrate and harmonize healthcare data from various sources, such as electronic health records (EHRs), claims databases, and other observational databases (26, 27). The OMOP CDM is a patient-centric relational database with several standardized tables, distinguished in domains like condition, procedures, drug usage, measurements or observations. Some of the key standard terminologies used in the OMOP common data model include SNOMED CT ((28) - snomed.org) and LOINC ((29) - loinc.org). The large community behind the OMOP CDM is consolidated in the Observational Health Data Sciences and Informatics community (OHDSI - ohdsi.org). Some OHDSI tools of specific interest include HADES, a set of open source R-packages for large-scale analytics (30) and ATLAS, which facilitates the design and execution of analyses (31). The 2023 annual report on ohdsi.org highlighted impressive numbers: over 3,700 collaborators from 83 countries, a data network of 543 databases from 49 countries, and more than 956 million patient records, covering about 12% of the global population.

Several large-scale collaborative RWD initiatives have adopted the OMOP CDM. Some examples include PIONEER focusing on prostate cancer [prostate-pioneer.eu; (32)], the European Reference Network for Rare Adult Solid Cancers [EURACAN; euracan.eu; (33)], and HONEUR with a specific focus on hematology [portal.honeur.org; (34)]. The European Health Data and Evidence Network [EHDEN; ehden.eu; (35)] deserves special attention, since it managed to establish the largest European federated RWD network. The EHDEN network currently consists of 187 Data Partners in 29 countries across the European region, with greater than 850 million anonymous health records.

Highlighted initiatives 4: European Research Data infrastructures: EBRAINS focusing on brain-related research data and ELIXIR for life sciences (-omics) data

Complementing these efforts are European research data infrastructures like EBRAINS (ebrains.eu) and ELIXIR (elixir-europe.org), which enhance research data handling and analysis for brain-related and life sciences (-omics) data, respectively. ELIXIR unifies bioinformatics resources and life science data for easier mining and reuse. This distributed digital infrastructure connects scientists from 23 countries (>250 research institutes), offering services like data deposition databases, data analysis, management, and compute services. ELIXIR also operates a vibrant training network through the TeSS Training Portal (36), registering over 1,200 training materials and training more than 19,000 people between September 2015 and March 2019 (37). ELIXIR played a leading role in the beyond one million genome project (b1mg-project.eu) that recently ended. During the COVID-19 pandemic, ELIXIR provided a range of services to study COVID-19 (38).

EBRAINS offers a digital infrastructure to boost collaborative brain research in neuroscience, brain health, and brain-related technology. Emerging from the Human Brain Project (HBP) (2013–2023), a European Flagship project with a €607 million investment, it involved over 500 researchers from 19 countries and 155 institutions. The HBP developed 160+ digital tools for multi-scale brain research and facilitated extensive collaboration among research teams (39). Some highlighted examples of potentially interesting tools and services include the Knowledge Graph - multi-modal metadata platform, the Medical Informatics Platform (MIP) - enabling access and analyses of anonymized medical data (40) and The Virtual Brain - a reference tool for full-brain simulation (41).

Discussion

There is great promise for the MS community in aligning closely and promptly with the EHDS legislation and engaging with emerging large-scale data initiatives that are not specific to MS

The EHDS is about to be implemented and is expected to have as significant and far-reaching impact. A proactive approach, which includes early investigation of alignment and synergy, would enable the MS community to understand the potential risks and challenges associated with this new legislation from the start. This foresight would allow for more effective long-term planning, the ability to anticipate future trends, and the development of risk management strategies to navigate anticipated changes in the regulatory environment. Moreover, collaborations with data initiatives not specific to MS not only pave the way for valuable partnerships and networking opportunities, but they also offer significant opportunities to explore new research questions and enhance existing studies with complementary insights.

Nevertheless, the path forward is marked by numerous, significant challenges that need to be addressed

Although I am a firm advocate for the EHDS and the collaboration with the previously mentioned data initiatives, I must highlight a series of challenges and lingering questions. These will be summarized in the following section, underlining the complexities we still need to navigate:

- *How will the implementation of the EHDS impact the current utilization of MS registries and other RWD sources?* As previously emphasized, MS registries and other RWD sources are vital for addressing pressing clinical questions related to MS. Currently, there is significant variation in the governance principles applied within the existing and emerging registries and RWD sources, which complicates collaborative efforts (2, 42). Given the uncertainty regarding how the EHDS will influence the conduct of large-scale, multi-centric studies using data from different member states, it is yet to be determined whether the EHDS will simplify or further complicate these collaborations.

- *The EHDS primarily focuses on Europe, while other continents are advancing parallel initiatives within their regions, such as the Sentinel Initiative* (43)) *and the Framework for the FDA's Real-World Evidence Program* (44). *This raises the question: Can we expect alignment between these initiatives to address clinical challenges on a global scale?* Investigating phenomena like silent progression, pediatric MS, early detection of MS onset in at-risk individuals (referred to as prodrome), and conducting large-scale epidemiological studies, such as the Atlas of MS (45), requires a wealth of high-quality data. Global collaboration is crucial to tackle these complex questions, especially considering the global prevalence and incidence rates of MS. An estimated 2.8 million people worldwide live with MS, equating to 35.9 per 100,000 population, with a pooled incidence rate of 2.1 per 100,000 persons/year (45). For instance, the COVID-19 in MS global data sharing initiative brought together data from 19 partners but compiled 'only' 10,000 patient records (13–15). Similarly, the BMSD network, with the potential of over 250,000 patient records, experiences a significant reduction in numbers when specific inclusion criteria are applied (11).
- *How can we ensure that the disease-agnostic recommendations, services, and tools are not only fit-for-purpose but also implementable for addressing MS-related questions, given that their straightforward application to the MS community is evidently not as feasible as assumed?* A prime example is the OMOP CDM, which, despite its broad application, is currently not entirely suitable for MS registry data. This statement is based on the experiences of my research group and in line with the documented experience from pulmonary hypertension databases (46). The underlying problem and probably the main reason for the different mapping designs is the observational character of MS RWD sources that are not connected to an electronic health record and filled with clinical data from there. Furthermore, a significant gap exists between guidelines formulated by EMA and their practical application, as highlighted by two key reports – the EMA Report on MS Registries (18) and the EMA Guideline on Registry-Based Studies (22). These documents, while authoritative, lack the necessary detail, have little or no focus on patient's input or patient relevant outcome measures and have not been checked sufficiently for real-world and sustainable implementation. For example, the discussion about financial sustainability is insufficiently incorporated into these reference documents. Despite the aforementioned challenges, there are notable examples of successful collaborations. The German MS registry and the MS DataConnect Cohort of the University MS Center in Belgium are part of the federated data network of EHDEN (35). In the MultipleMS consortium (multiplems.eu), linked to the International Multiple Sclerosis Genetics Consortium (47), and the COVID-19 in MS global data sharing initiative (12), the ELIXIR community has played a key role in supporting the technical architectures for data storage, management, and sharing in these large-scale collaborative efforts.

In a continuously changing and complex environment, it is essential to prioritize pragmatic actions.

To this end, a set of concrete, actionable suggestions for the MS community are formulated (see also Figure 2).

- Suggested action 1: Building upon the strong foundation of collaboration established within the MS community to further enhance our collaborative efforts. As we move toward formulating detailed and implementable global recommendations for data collection, it is clear that the responsibility for this initiative will continue to rest with the MS community. Recently, a global multi-stakeholder task force defined a core dataset for MS to guide emerging registries in their dataset definitions and speed-up and support harmonization across registries and RWD MS initiatives. A regular revision of the current Core DataSet is anticipated, especially in regards to the currently excluded variables or pragmatic choices of values (48). Dataset variables needing a dedicated set of data elements (e.g. in the area of patient-reported outcomes or pharmacovigilance) are also not included. The latter is anticipated to be driven by leading networks like BMSD or PROMS initiative focusing on these specific topics. Another interesting activity to enhance multi-stakeholder collaboration is to regularly organize large-scale multi-stakeholder engagement meetings (18, 49).
- Suggested action 2: Investigate the potential of existing and emerging data spaces to address some urgent and critical questions formulated by the MS community, adhering to the principle of 'learning by doing.' Specific pilot projects could be established and carried out to assess the suitability of current recommendations for data standardization, interoperability, infrastructure, and governance in the MS context. Following these pilot projects, identifying areas for potential synergy and proposing necessary adjustments will be crucial. An innovative approach could involve organizing a study-a-thon in collaboration with OHDSI and/or EHDEN. A study-a-thon is a focused, multi-day research event that generates reliable evidence on a specific medical topic across different countries and health systems. It gathers multidisciplinary teams to expedite scientific contributions without sacrificing the quality of research, facilitated through a reproducible process (50). This method could effectively showcase the advantages of collaborating with these networks within a limited timeframe. Concurrently, the MS Data Alliance is investigating how the OMOP CDM can be tailored to address the challenges previously identified. This research is specifically focused on the feasibility of automatically converting the MS Data Alliance Core Dataset (48) to the OMOP CDM, with the results expected to be publicly and freely available to the MS community soon.
- Suggested action 3: Team-up with other disease areas to co-2create recommendations to ensure that the EHDS encapsulates disease-specific requirements. The challenges highlighted earlier in this paper, while focusing on MS, are

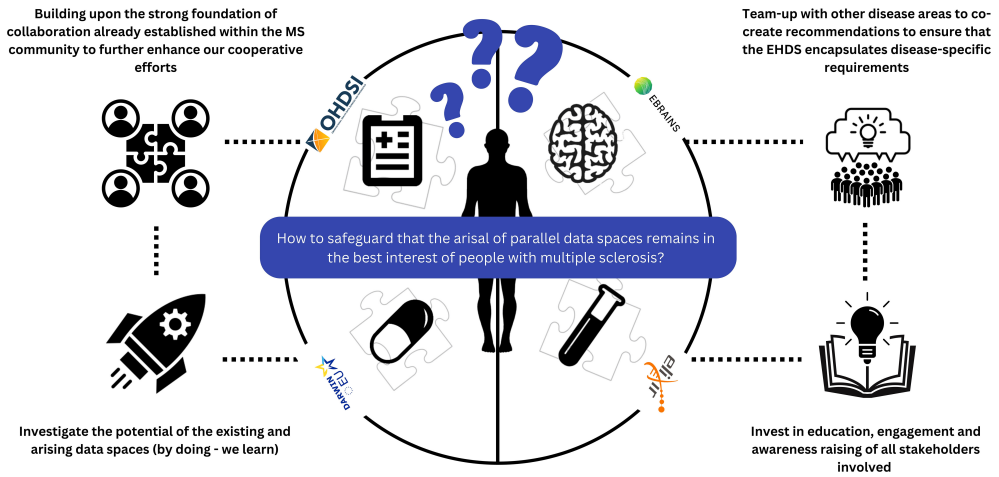


FIGURE 2
Summary overview of suggested action points towards the MS community to safeguard that the arisal of parallel data spaces remains in the best interest of people with multiple sclerosis.

not unique to it. Similar issues are encountered by communities studying chronic diseases that require long-term, high-dimensional follow-up. Particularly relevant are those groups already actively engaged in EHDS discussions, such as those focused on cystic fibrosis, cancer and diabetes (20, 21, 49)). A practical first step would be to co-create a joint statement, consolidating a unified response to the EHDS proposal and addressing the identified challenges.

- Suggested action 4: Invest in education, engagement and awareness raising of all stakeholders involved to ensure proper understanding related to the EHDS as well as general data science principles. Stakeholders include regulators, clinicians, researchers, industry, and people with MS, all of whom are equally important. The level of being informed about how to contribute to the RWD ecosystem as well as experience in actively participating in large-scale RWD collaborative initiatives differs between stakeholders and individuals. Being limited informed and/or having limited experience leads to reduced active participation in initiatives that aim to address the urgent needs within the ecosystem. People with MS (or broader citizens) can actively contribute by co-creating legislation — deciding what is acceptable, how, and for what health data can be used - as well as helping to define priorities in the global research agenda.

Conclusion

Rapid advances in artificial intelligence (AI) and the growing health data volume are expected to significantly impact the health sector. AI has already shown promise in helping to improve diagnostic performances, workflow and cost-effectiveness. AI has the potential to speed-up the complex process of data management and –analysis, specifically with the recent developments in the field of generative AI (e.g. ChatGPT). As we stand at the intersection of immense potential and complex challenges, there is both a reason for excitement and a cause for concern. By coming together – researchers, clinicians, patients, policymakers, and other stakeholders – we can harness the full potential of RWD while navigating its complexities. This is a journey that we must embark on together, informed by diverse perspectives and united by a common goal: to revolutionize MS care and research for the betterment of people affected by MS worldwide. Let this paper be the catalyst for that collaborative journey.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

LP: Writing – original draft, Writing – review & editing, Conceptualization, Visualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors acknowledge the assistance of ChatGPT4, an AI language model developed by OpenAI, for its support in structuring and refining the content of this paper.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Trojano M, Tintore M, Montalban X, Hillert J, Kalincik T, Iaffaldano P, et al. Treatment decisions in multiple sclerosis — insights from real-world observational studies. *Nat Rev Neurol*. (2017) 13:105–18. doi: 10.1038/nrneurol.2016.188

2. Geys L, Parciak T, Pirmani A, McBurney R, Schmidt H, Malbaša T, et al. The multiple sclerosis data alliance catalogue. *Int J MS Care*. (2021) 23:261–8. doi: 10.7224/1537-2073.2021-006

3. Cohen JA, Trojano M, Mowry EM, Uitdehaag BM, Reingold SC, Marrie RA. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Mult Scler J*. (2020) 26:23–37. doi: 10.1177/1352458519892555

4. Stastna D, Drahota J, Lauer M, Mazouchova A, Menkyova I, Adamkova J, et al. The Czech National MS Registry (ReMuS): Data trends in multiple sclerosis patients whose first disease-modifying therapies were initiated from 2013 to 2021. *BioMed Pap*. (2023) 168(3):262–70. doi: 10.5507/bp.2023.015.html

5. Koch-Henriksen N, Stenager E, Brønnum-Hansen H. Studies based on the Danish multiple sclerosis registry. *Scand J Public Health*. (2011) 39:180–4. doi: 10.1177/1403494811405097

6. Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud DA, et al. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. *Mult Scler J*. (2020) 26:118–22. doi: 10.1177/1352458518815602

7. on behalf of the Italian Multiple Sclerosis Register Centers Group, Trojano M, Bergamaschi R, MP A, Comi G, Ghezzi A, et al. The Italian multiple sclerosis register. *Neurol Sci*. (2019) 40:155–65. doi: 10.1007/s10072-018-3610-0

8. Hillert J, Stawiarz L. The Swedish MS registry – clinical support tool and scientific resource. *Acta Neurol Scand*. (2015) 132:11–9. doi: 10.1111/ane.2015.132.issue-S199

9. Kalincik T, Butzkueven H. The MSBase registry: Informing clinical practice. *Mult Scler J*. (2019) 25:1828–34. doi: 10.1177/1352458519848965

10. Hillert J, Magyari M, Soelberg Sørensen P, Butzkueven H, van der Welt A, Vukusic S, et al. Treatment switching and discontinuation over 20 years in the big multiple sclerosis data network. *Front Neurol*. (2021) 12:647811. doi: 10.3389/fneur.2021.647811

11. Signori A, Lorscheider J, Vukusic S, Trojano M, Iaffaldano P, Hillert J, et al. Heterogeneity on long-term disability trajectories in patients with secondary progressive MS: a latent class analysis from Big MS Data network. *J Neurol Neurosurg Psychiatry*. (2023) 94:23–30. doi: 10.1136/jnnp-2022-329987

12. Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. *Mult Scler J*. (2020) 26:1157–62. doi: 10.1177/1352458520941485

13. Simpson-Yap S, Pirmani A, Kalincik T, De Brouwer E, Geys L, Parciak T, et al. Updated results of the COVID-19 in MS global data sharing initiative: anti-CD20 and other risk factors associated with COVID-19 severity. *Neurol Neuroimmunol Neuroinflamm*. (2022) 9:e200021. doi: 10.1212/NXI.0000000000200021

14. Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of disease-modifying therapies with COVID-19 severity in multiple sclerosis. *Neurology*. (2021) 97:1870–85. doi: 10.1212/WNL.0000000000012753

15. Simpson-Yap S, Pirmani A, De Brouwer E, Peeters LM, Geys L, Parciak T, et al. Severity of COVID19 infection among patients with multiple sclerosis treated with interferon-β. *Mult Scler Relat Disord*. (2022) 66:104072. doi: 10.1016/j.msard.2022.104072

16. Proposal for a regulation - The European Health Data Space - European Commission. Available online at: https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en (Accessed on July 8, 2024).

17. EMA. *Patient registry initiative-strategy and mandate of the cross-committee task force*. EMA. Initiative London (2017).

18. EMA. *Patient registries workshop, 28 October 2016-observations and recommendations arising from the workshop*. London: EMA (2017).

19. EMA. Report on haemophilia registries workshop 8 june 2018(2018). Available online at: https://www.ema.europa.eu/en/documents/report/report-haemophilia-registries-workshop_en.pdf (Accessed on July 8, 2024).

20. EMA. *Report on Cystic Fibrosis Registries - Workshop 14 June 2017*. London: EMA. (2017).

21. EMA. Report of the workshop on the use of registries in the monitoring of cancer therapies based on tumours' genetic and molecular features - 29 November 2019(2020). Available online at: https://www.ema.europa.eu/system/files/documents/report/report-workshop-registries_en.pdf (Accessed on July 8, 2024).

22. *Guideline on registry-based studies - Scientific guideline*. European Medicines Agency. Available at: https://www.ema.europa.eu/en/guideline-registry-based-studies-scientific-guideline.

23. Taskforce H. Phase II report:'evolving data-driven regulation. *Eur Med Agency*. (2019).

24. Plueschke K, Jonker C, Strassmann V, Kurz X. Collection of data on adverse events related to medicinal products: A survey among registries in the ENCePP resources database. *Drug Saf.=*. (2022) 45:747–54. doi: 10.1007/s40264-022-01188-x

25. EMA EMA. Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources(2022). Available online at: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf (Accessed on July 8, 2024).

26. OMOP common data model. Available online at: https://ohdsi.github.io/CommonDataModel/ (Accessed on July 8, 2024).

27. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc*. (2015) 22:553–64. doi: 10.1093/jamia/ocu023

28. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. *Proc Conf Am Med Inform Assoc AMIA Fall Symp*. (1997), 640–4.

29. Drenkhahn C, Ingenerf J. The LOINC content model and its limitations of usage in the laboratory domain. *Stud Health Technol Inform*. (2020) 270:437–42. doi: 10.3233/SHTI200198

30. OHDSI/Hades. Observational health data sciences and informatics(2024). Available online at: https://github.com/OHDSI/Hades (Accessed on July 8, 2024).

31. OHDSI/Atlas. Observational health data sciences and informatics(2024). Available online at: https://github.com/OHDSI/Atlas (Accessed on July 8, 2024).

32. Omar MI, Roobol MJ, Ribal MJ, Abbott T, Agapow PM, Araujo S, et al. Introducing PIONEER: a project to harness big data in prostate cancer research. *Nat Rev Urol*. (2020) 17:351–62. doi: 10.1038/s41585-020-0324-x

33. Blay JY, Casali P, Bouvier C, Dehais C, Galloway I, Gietema J, et al. European Reference Network for rare adult solid cancers, statement and integration to health care systems of member states: a position paper of the ERN EURACAN. *ESMO Open*. (2021) 6:100174. doi: 10.1016/j.esmoop.2021.100174

34. Bardenheuer K, Van Speybroeck M, Hague C, Nikai E, Price M. Haematology Outcomes Network in Europe (HONEUR)—A collaborative, interdisciplinary platform to harness the potential of real-world data in hematology. *Eur J Haematol*. (2022) 109:138–45. doi: 10.1111/ejh.v109.2

35. Voss EA, Blacketer C, Van Sandijk S, Moinat M, Kallfelz M, Van Speybroeck M, et al. European Health Data & Evidence Network—learnings from building out a standardized international health data network. *J Am Med Inform Assoc*. (2023) 31:209–19. doi: 10.1093/jamia/ocad214

36. Beard N, Bacall F, Nenadic A, Thurston M, Goble CA, Sansone SA, et al. TeSS: a platform for discovering life-science training opportunities. *Bioinformatics*. (2020) 36:3290–1. doi: 10.1093/bioinformatics/btaa047

37. Harrow J, Drysdale R, Smith A, Repo S, Lanfear J, Blomberg N. ELIXIR: providing a sustainable infrastructure for life science data at European scale. *Bioinformatics*. (2021) 37:2506–11. doi: 10.1093/bioinformatics/btab481

38. Blomberg N, Lauer KB. Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. *Eur J Hum Genet*. (2020) 28:719–23. doi: 10.1038/s41431-020-0637-5

39. Lorents A, Colin ME, Bjerke IE, Nougaret S, Montelisciani L, Diaz M, et al. Human brain project partnering projects meeting: status quo and outlook. *eneuro*. (2023) 10:ENEURO.0091–23.2023. doi: 10.1523/ENEURO.0091-23.2023

40. Redolfi A, De Francesco S, Palesi F, Galluzzi S, Muscio C, Castellazzi G, et al. Medical informatics platform (MIP): A pilot study across clinical Italian cohorts. *Front Neurol*. (2020) 11:1021. doi: 10.3389/fneur.2020.01021

41. Jirsa V, Wang H, Triebkorn P, Hashemi M, Jha J, Gonzalez-Martinez J, et al. Personalised virtual brain models in epilepsy. *Lancet Neurol*. (2023) 22:443–54. doi: 10.1016/S1474-4422(23)00008-X

42. EMA. Report on multiple sclerosis registries - workshop 7 July 2017(2017). Available online at: https://www.ema.europa.eu/system/files/documents/report/wc500236644_en.pdf (Accessed on July 8, 2024).

43. Brown JS, Mendelsohn AB, Nam YH, Maro JC, Cocoros NM, Rodriguez-Watson C, et al. The US Food and Drug Administration Sentinel System: a national resource for a learning health system. *J Am Med Inform Assoc JAMIA*. (2022) 29:2191–200. doi: 10.1093/jamia/ocac153

44. Schurman B. *The framework for FDA's real-world evidence program*. (2019).

45. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler J*. (2020) 26:1816–21. doi: 10.1177/1352458520970841

46. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol*. (2021) 21:238. doi: 10.1186/s12874-021-01434-3

47. International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet*. (2013) 45:1353–60. doi: 10.1038/ng.2770

48. Parciak T, Geys L, Helme A, van der Mei I, Hillert J, Schmidt H, et al. Introducing a core dataset for real-world data in multiple sclerosis registries and cohorts: Recommendations from a global task force. *Mult Scler J*. (2023) 30:13524585231216004. doi: 10.1177/13524585231216004

49. Hogervorst MA, Møllebæk M, Vreman RA, Lu TA, Wang J, De Bruin ML, et al. Perspectives on how to build bridges between regulation, health technology assessment and clinical guideline development: a qualitative focus group study with European experts. *BMJ Open*. (2023) 13:e072309. doi: 10.1136/bmjopen-2023-072309

50. Hughes N, Rijnbeek PR, van Bochove K, Duarte-Salles T, Steinbeisser C, Vizcaya D, et al. Evaluating a novel approach to stimulate open science collaborations: a case series of "study-a-thon" events within the OHDSI and European IMI communities. *JAMIA Open*. (2022) 5:ooac100. doi: 10.1093/jamiaopen/ooac100

51. EMA. *Big Data Workplan 2023-2025 - HMA/EMA joint Big Data Steering Group* (2024). Available online at: https://www.ema.europa.eu/en/documents/work-programme/workplan-2023-2025-hma-ema-joint-big-data-steering-group_en.pdf (Accessed on July 8, 2024).

frontiers | Frontiers in Neurology

TYPE Original Research
PUBLISHED 11 December 2024
DOI 10.3389/fneur.2024.1500140

OPEN ACCESS

EDITED BY
Hans-Peter Hartung,
Heinrich Heine University, Germany

REVIEWED BY
Reza Rahmanzadeh,
TheUltra.ai, Switzerland
Alessio Signori,
University of Genoa, Italy

*CORRESPONDENCE
Ibrahim Acir
✉ iacir33@gmail.com

RECEIVED 22 September 2024
ACCEPTED 29 November 2024
PUBLISHED 11 December 2024

CITATION
Albuz Ö, Acir I, Haşimoğlu O, Suskun M, Hocaoğlu E and Yayla V (2024) Cranial volume measurement with artificial intelligence and cognitive scales in patients with clinically isolated syndrome.
Front. Neurol. 15:1500140.
doi: 10.3389/fneur.2024.1500140

Cranial volume measurement with artificial intelligence and cognitive scales in patients with clinically isolated syndrome

Özlem Albuz[1], Ibrahim Acir[1]*, Ozan Haşimoğlu[2], Melis Suskun[1], Elif Hocaoğlu[1] and Vildan Yayla[1]

[1]Bakırköy Dr. Sadi Konuk Eğitim ve Araştırma Hastanesi, Istanbul, Türkiye, [2]Basaksehir Cam and Sakura City Hospital, Istanbul, Türkiye

Objective: We aimed to investigate the relationship between volumetric measurements of specific brain regions which were measured with artificial intelligence (AI) and various neuropsychological tests in patients with clinically isolated syndrome.

Materials and methods: A total of 28 patients diagnosed with CIS were included in the study. The patients were administered Öktem Verbal Memory Processes Test, Symbol Digit Modalities Test (SDMT), Backward-Forward Digit Span Test, Stroop Test, Trail Making Test, Controlled Oral Word Association Test (COWAT), Brief Visuospatial Memory Test, Judgement of Line Orientation Test, Beck Depression Scale, Beck Anxiety Scale and Fatigue Severity Scale. Artificial intelligence assisted BrainLab Elements™ Atlas-Based Automatic Segmentation program was used for calculating volumes. The measured volumes were compared with the reference database. In addition, neuropsychological test performances and volumetric measurements of the patients were compared.

Results: Of the patients included in the study, 78.6% were female and 21.4% were male, with an average age of 33 years. Verbal Memory Processes Test, SDMT, Backward-Forward Digit Span, JLOT, and Stroop Test showed significant correlations with multiple anatomical regions, particularly the anterior thalamic nucleus, which was associated with the highest number of cognitive tests. The JLOT exhibited the strongest correlation with six different brain regions ($p < 0.001$).

Conclusion: The Judgement of Line Orientation and Stroop Tests, correlated with multiple brain regions, especially the anterior thalamic nucleus, underscoring the importance of these tests in assessing cognitive function in CIS.

KEYWORDS

multiple sclerosis, clinically isolated syndrome, artificial intelligence, BrainLab, brain volume analysis

Introduction

Clinically isolated syndrome (CIS) is defined as one of the subtypes of multiple sclerosis (MS) according to the 2017 McDonald MS criteria. It is a monophasic clinical episode suggestive of a focal or multifocal, inflammatory demyelinating event in the central nervous system, lasting at least 24 h, with or without subsequent improvement, not accompanied by infection or fever, and including symptoms resembling a typical MS relapse (1). Although

almost any neurological finding may be the first clinical episode in patients with CIS, somatosensory findings, optic neuritis, transverse myelitis, brainstem syndrome, and cognitive involvement are most commonly observed (2, 3). Cognitive impairment was first mentioned by Charcot in 1877 as "slowness in the perception of MS patients." Cognitive impairment has been reported to be approximately 34–65% (4).

Brain tissue loss (atrophy) is thought to reflect neuroaxonal damage. Volumetric measurements are performed with fully automatic segmentation software over 3D T1-weighted sequences to evaluate atrophy. Atrophy starts in the early period of the disease, and it is known to be strongly associated with cognitive impairment (5).

Cognitive impairments observed in MS include impairments in information processing efficiency and speed, attention maintenance and complex attention, working memory, learning, problem-solving, language and visuospatial memory, long-term memory, abstract thinking, verbal fluency, and executive functions (6). The characteristics of cognitive impairment in the CIS group are similar to those of MS, and information processing speed and verbal memory are most commonly affected. It has been suggested that cognitive dysfunction observed in patients with CIS may predict the transformation of the disease into MS and the disability that occurs over time (7, 8).

The possibility of establishing a correlation between radiological images and cognitive impairment in MS is very important, and many studies have been conducted on this subject. In studies, cognitive impairment was found to be associated with T2 lesion load, neocortical gray matter, volume loss in the thalamus, hippocampus, and corpus callosum on MR imaging (6, 9–11).

Our primary aim encompassed a comprehensive inquiry into the intricate interplay between the volumetric measurements derived from distinct cerebral regions in CIS patients and a diverse array of neuropsychological tests, delving into the nuanced associations and potential implications within this multifaceted relationship.

Materials and method

In this study, a total of 28 patients comprising 6 males and 22 females diagnosed with CIS, and who were under observation at the demyelinating diseases outpatient clinic between February–June 2023, were assessed. Inclusion criteria stipulated that patients must have been diagnosed with clinically isolated syndrome, be 18 years of age or older, be proficient in Turkish, and exhibit normal laboratory test results concerning cognitive function. Exclusion criteria encompassed substance abuse, recent acute exacerbations or corticosteroid use within 4 weeks before clinical and MR imaging tests, presence of central nervous system diseases, significant affective disorders or severe psychiatric illnesses, utilization of psychostimulant or psychotropic drugs affecting cognitive functions, alcohol or substance dependence, as well as a history of attention deficit-hyperactivity disorder and learning disabilities.

Patients underwent cranial MR imaging with a slice interval of 1 mm. The imaging was conducted in the supine position utilizing a 1.5 Tesla magnetic field strength (Siemens Magnetom Amira) device equipped with an 8-channel head coil, adhering to the MS acquisition protocol. All images were acquired using the same device and included Turbo spin echo T1 (TR 1,060 ms, TE Shortest ms, slice thickness 1 mm with no gaps, matrix 252 × 240 pixels) and T2w (TR 2,500 ms, TE: shortest 260 ms, slice thickness 1 mm with no gaps, matrix 252 × 252 pixels) sequences. The radiological images were converted to the appropriate format and transferred to the BrainLab Elements™ Atlas-Based Automatic Segmentation program, where the volumes of the patients were evaluated by a certified neurosurgeon trained in volume measurement. In this system, the most accurate boundaries of the grey matter and basal ganglia were automatically identified by comparing the voxel parameters of the patient with the parameters in the atlas averages through artificial intelligence. Subsequently, after the fusion of the T2w and T1 MR images of the patients in the BrainLab Elements program, all grey matter and basal nuclei were automatically segmented separately in the object segmentation module, and their boundaries and volumes were calculated. The boundaries were cross-checked on the T2w image, and any inaccuracies in segmentations were rectified. The volume values obtained were then juxtaposed with the average volume values in the MNI PD25 and ICBM152 standard human brain database (12), and the variance for each anatomical region was recorded. The measured volumes included the amygdala, capsule externa, capsule interna, nucleus caudatus, cerebellum, nucleus dentatus, fornix, globus pallidus, hypothalamus, nucleus accumbens, basal nucleus of Meynert, nucleus ruber, optic nerve, pedunculopontine nucleus, putamen, substantia nigra, anterior thalamic nucleus, zona incerta, and ventricle volumes, which were subsequently compared to the reference database using the BrainLab Elements™ Atlas-Based Automatic Segmentation program. The measured volumes of the patients were compared with the reference database (topographic volume-standardization atlas of the human brain) (Figures 1, 2) (13).

Öktem Verbal Memory Processes Test, Paced Auditory Serial Addition Test (PASAT), Symbol Digit Modalities Test (SDMT), Backward-Forward Digit Span Test, Stroop Test, Trail Making Test, Controlled Oral Word Association Test (COWAT), Brief Visuospatial Memory Test (BVMT-R), Judgment of Line Orientation Test (JLOT), Beck Depression Scale, Beck Anxiety Scale, and Fatigue Severity Scale (FSS) neuropsychological tests were administered, which lasted approximately 90 min within 2 weeks following MRI. The PASAT test was only administered to one person due to communication and cooperation difficulties between the patients and the administrator, as well as the challenges in administering the test. Therefore, this test was excluded from the study.

All statistical analyses were performed using IBM SPSS Statistics version 29.0. Descriptive statistics were expressed as mean ± standard deviation (mean ± SD) or median (25th–75th percentile) values for continuous variables and as numbers (n) and percentage (%) for categorical variables. The comparison between categorical variables was conducted using the chi-square test or Fisher's exact test. The determination of normal distribution was based on the number of observations in the groups, histograms, and the Shapiro–Wilk test. The Mann–Whitney U test was employed to compare continuous variables that were not normally distributed between two groups. If normal distribution was confirmed, Student's t-test was utilized. The linear relationship between two continuous variables was assessed using Pearson or Spearman correlation coefficients, and their significance was analyzed based on the presence or absence of normal distribution. Correlation coefficients falling between 0 and ± 0.3 were interpreted as indicating no correlation, while coefficients between 0.3 and 0.5 suggested a weak correlation in a positive (or negative)

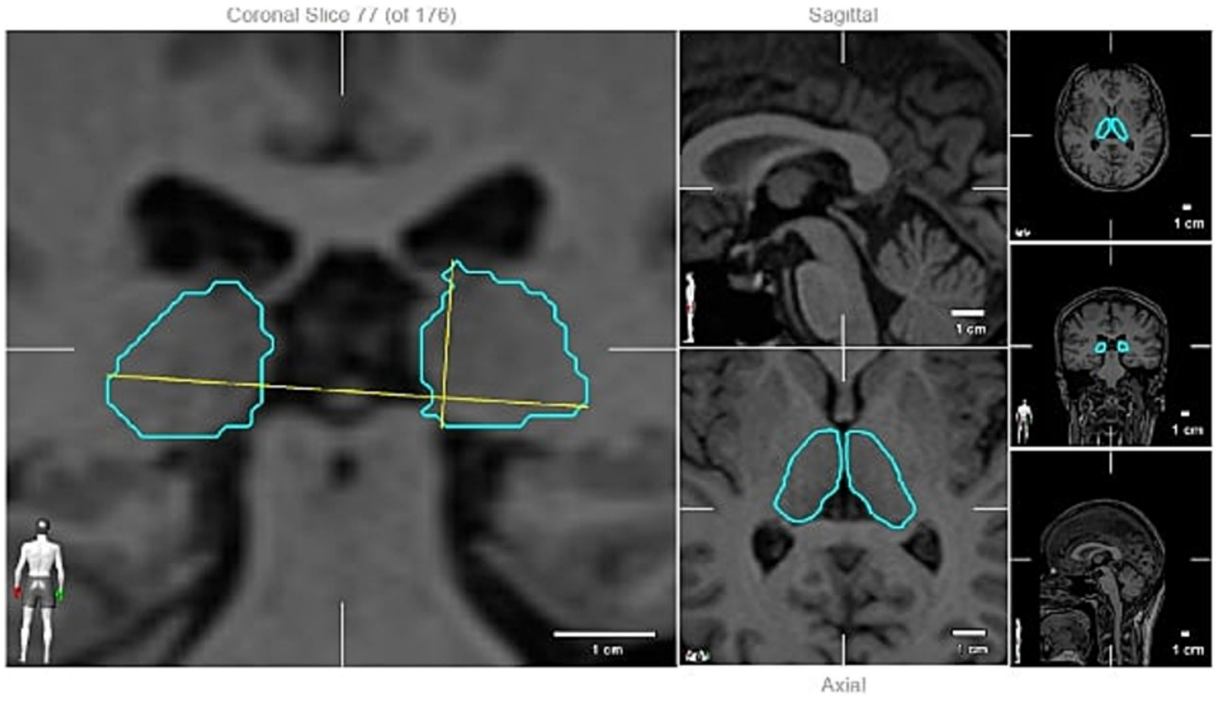


FIGURE 1
Thalamic volume measuring (an example).

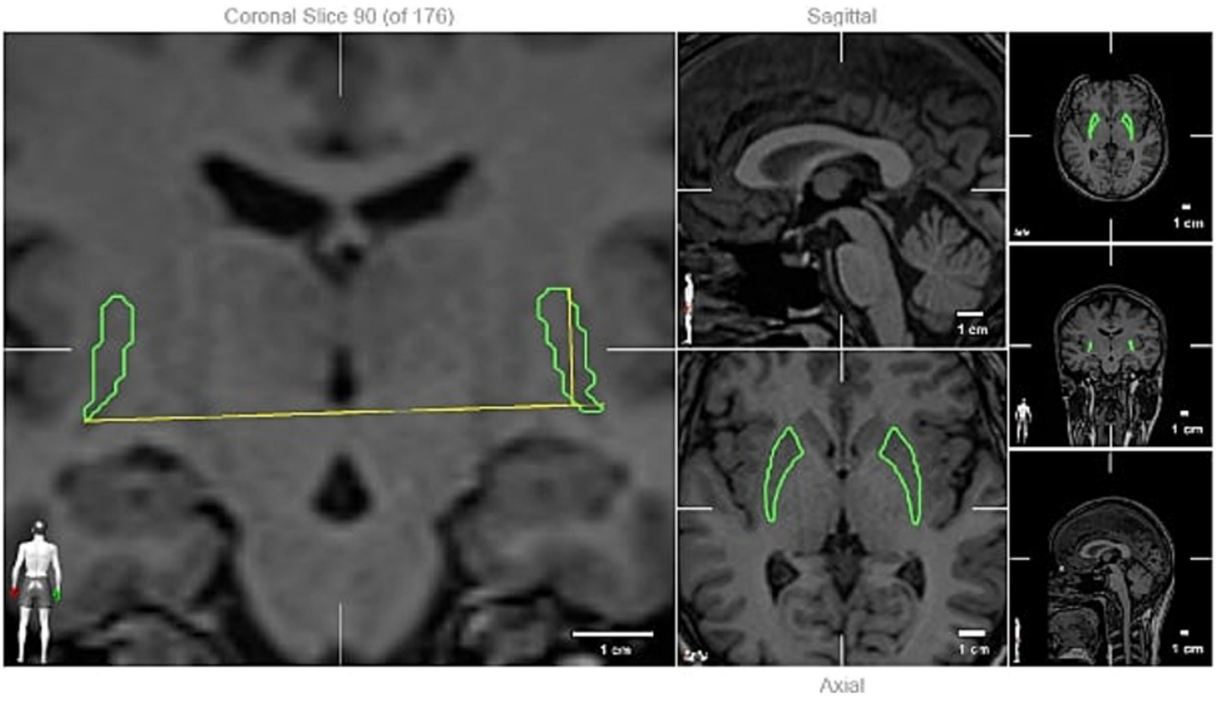


FIGURE 2
Putamen volume measuring (an example).

direction. Coefficients ranging from 0.5 to 0.7 indicated a moderate correlation in a positive (or negative) direction, while coefficients exceeding 0.7 were indicative of a strong correlation (positive or negative). In cases where the influence of a third variable was considered, partial correlation coefficients were calculated. Two-way *p*-values less than 0.05 were considered statistically significant.

10.3389/fneur.2024.1500140

Results

Demographic and clinical characteristics of the patients who participated in our study are summarised in Table 1. A total of 22 (78.6%) of the patients were female, while 6 (21.4%) were male. The ages of all patients ranged between 17 and 51 years, with a mean age of 33.0 years. Half of the patients (50%) had less than 8 years of education, while the other half had more than 8 years of education. Clinical attacks manifested as optic neuritis in 15 (53.6%) patients, brainstem symptoms in 4 (14.3%) patients, sensory symptoms in 8 (28.6%) patients, and cerebellar symptoms in 1 (3.57%) patient. Diabetes mellitus (DM) was present in 2 patients, hypertension (HT) was present in 2 patients, and hypothyroidism was present in 1 patient. However, thyroid function tests were within normal limits in all patients according to laboratory tests (Table 1).

When the volumetric examinations of the patients were compared according to gender, a statistically significant difference was found between the two groups in cerebellum, hypothalamus, nucleus accumbens, periquaductal grey matter and subthalamic nucleus volumes ($p < 0.05$) (Table 2).

The measured volumes of the patients were compared with the volumes of amygdala, basal ganglia (caudate + putamen + globus pallidus), capsule interna, nucleus caudatus, cerebellum, thalamus, globus pallidus, putamen and ventricle in the reference database. Mean ± SD values and statistical comparisons are shown in Table 3. Amygdala, basal ganglia caudate + putamen + globus pallidus, capsule interna, nucleus caudatus, thalamus, globus pallidus putamen and cerebellum were found to be significantly different from the population mean in the sample group with clinically isolated syndrome ($p < 0.001$).

When examining the correlation between cognitive tests and anatomical regions, no significant correlation was found with the COWAT, BVMT-R, Beck Depression Scale, Beck Anxiety Scale and FSS tests. However, significant correlations were observed with the Öktem Verbal Memory Processes Test, SDMT, Backward-Forward Digit Span Test, JLOT and Stroop Tests. The JLOT was the test that showed correlations with the most anatomical locations (6 anatomical regions). The anterior thalamic nucleus was identified as the anatomical region that correlated with the highest number of cognitive tests. The statistically significant results of the correlation analyses between the cognitive tests and anatomical region volumetric measurements of the patients are shown in Table 4.

The regression analysis revealed distinct patterns in the relationship between age, sex, tracking test performance, and volumetric measurements. For the tracking test, the model demonstrated strong explanatory power, accounting for 54% of the variance in performance. Age emerged as a significant predictor, with increasing age associated with longer tracking times ($B = 6.361$, $p < 0.001$). The standardized coefficient ($\beta = 0.738$) confirmed age as the most influential factor. In contrast, sex had no statistically significant effect on tracking test performance ($p = 0.795$). The model was statistically significant overall ($F = 14.678$, $p < 0.001$), emphasizing the role of age in predicting tracking performance.

The analysis of brain region volumes, including the amygdala, thalamus, capsula interna, putamen, globus pallidus, and nucleus caudatus, showed limited explanatory power. For the amygdala, the model accounted for only 8.3% of the variance, with neither age ($p = 0.688$) nor sex ($p = 0.176$) significantly influencing its volume. Similarly, the capsula interna volume model explained 11.2% of the variance, with age showing no significant effect ($p = 0.843$) and sex being marginally non-significant ($p = 0.095$), suggesting a potential relationship that may require further investigation.

For the thalamus, the model explained 7.6% of the variance, with neither age ($p = 0.397$) nor sex ($p = 0.309$) demonstrating statistical significance. The putamen model performed poorly, explaining only 1.5% of the variance, with both age ($p = 0.545$) and sex ($p = 0.980$)

TABLE 1 Clinical and demographic characteristics of patients with clinically isolated syndrome included in the study.

	[All] *N* = 28
Gender	
Woman	22 (78.6%)
Male	6 (21.4%)
Age	33.0 [17–51]
Education status	
<8 years	14 (50%)
>8 years	14 (50%)
Marital status	
Married	17 (60.7%)
Single/divorced	11 (39.3%)
Profession	
Not working	16 (57.1%)
Labourer, civil servant, other	12 (42.9%)
BMI	26.5 [19.6; 38.1]
Smoking	10 (35.7%)
Alcohol use	2 (7.14%)
Presence of comorbidities	
DM	2 (7.14%)
HT	3 (10.7%)
Hypothyroidism	1 (3.57%)
Other	3 (10.71%)
First attack pattern	
Optic neuritis	15 (53.6%)
Brain stem	4 (14.3%)
Sensory	8 (28.6%)
Cerebellar	1 (3.57%)

TABLE 2 Comparison of volumetric measurements according to gender.

	Female	Male	*p* overall	*N*
	N = 22	*N* = 6		
Cerebellum	121 (13.6)	136 (4.32)	0.010	28
Hypothalamus	1.25 (0.14)	1.45 (0.09)	0.003	28
Nucleus accumbens	0.94 (0.12)	1.09 (0.14)	0.022	28
Periacuaductal grey matter	0.24 (0.05)	0.31 (0.04)	0.008	28
Subthalamic nucleus	0.18 (0.02)	0.20 (0.01)	0.039	28

TABLE 3 Comparison of patient volumes with population averages according to topographic volume-standardisation atlas of the human brain database.

	Mean ± SD (patient)	Mean ± SD (atlas) (ATLAS)	*t* value	*p*
Amigdala	2.87 ± 0.29	3.12 ± 0.47	−4.6187	**<0.001**
Basal ganglia (caudate + putamen + globus pallidus)	19.41 ± 1.20	22.12 ± 2.98	−7.1675	**<0.001**
Capsula interna	9.11 ± 1.04	10.62 ± 1.55	−7.6597	**<0.001**
Caudate nucleus	7.22 ± 10.90	7.78 ± 1.32	−3.3053	**0.003**
Cerebellum	116.73 ± 12.62	124 ± 13.8	−2.901	**0.007**
Ventricle	23.3 ± 5.52	21.18 ± 16.71	1.999	0.06
Thalamus	11.1 ± 1.33	14.61 ± 1.46	−13.89	**<0.001**
Globus pallidus	3.07 ± 0.59	3.69 ± 0.38	−8.5068	**<0.001**
Putamen	8.51 ± 0.93	11.26 ± 1.66	−15.64	**<0.001**

Bold values: highly significant.

TABLE 4 Correlation analysis between cognitive tests and anatomical region volumetric measurements.

Cognitive test	Anatomic region	Correlations	*p*
Öktem Verbal Memory Processes Test	Subtalamic nucleus	−0.421	0.026
SDMT	Acumbal nucleus	0.376	0.048
SDMT	Anterior thalamic nucleus	0.482	0.009
Trail Making Test	Internal capsule	−0.463	0.013
Trail Making Test	Acumbens nucleus	−0.501	0.007
Trail Making Test	Meynert's basal nucleus	−0.389	0.040
Trail Making Test	Putamen	−0.435	0.021
Trail Making Test	Talamus	−0.486	0.009
Backward-Forward Digit Span Test	Anterior thalamic nucleus	0.374	0.050
Judgement of Line Orientation Test	Capsule interna	0.515	0.005
Judgement of Line Orientation Test	Dentate nucleus	0.477	0.010
Judgement of Line Orientation Test	Globus pallidus	0.436	0.020
Judgement of Line Orientation Test	Acumbens nucleus	0.541	0.003
Judgement of Line Orientation Test	Anterior talamic nucleus	0.453	0.016
Judgement of Line Orientation Test	Talamus	0.409	0.031
Stroop Test	Capsule interna	−0.413	0.040
Stroop Test	Anterior talamic nucleus	−0.545	0.005
Stroop Test	Nucleus caudatus	−0.400	0.047

failing to show significant effects. Similarly, the globus pallidus and nucleus caudatus models explained 6.9 and 6.3% of the variance, respectively, with no significant contributions from age or sex for either region.

Discussion

Clinical isolated syndrome is a single episode of inflammatory demyelination of the central nervous system suggestive of MS. The main mechanism in the pathophysiology of the disease is thought to involve multifocal inflammation, demyelination, oligodendrocyte loss, reactive gliosis, and axonal degeneration (14). In our study, our primary objective was to assess whether atrophy was present by comparing the measured volumes in specific brain regions of patients with clinically isolated syndrome with those in the reference database (topographic volume-standardization atlas of the human brain). Our secondary objective was to evaluate the correlation between the volumes measured in specific brain regions and the results of various cognition tests assessing different cognitive functions, aiming to determine which cognitive performance is most accurately predicted by volume parameters. Previous research has primarily emphasized the role of subcortical structures like the thalamus and basal ganglia in tasks related to executive functions and memory. However, this study expands the scope by examining a more comprehensive set of cognitive tasks, including visuospatial memory, information processing, and working memory, and their associations with specific brain regions in patients with CIS.

Cognitive impairment, often overlooked in daily practice but with a detrimental impact on the daily life activities of patients, is

frequently observed in MS. Studies have shown that the prevalence of cognitive impairment ranges from 40 to 65% and may manifest as early as the initial stages of the disease, including during the CIS period (9, 15). It is understood that demyelinating plaques in the periventricular white matter, axonal loss, and neocortical atrophy play crucial roles in the pathophysiology of cognitive impairment. Zipoli et al. (7) identified cognitive impairment in a significant proportion of patients with CIS and concluded that this had prognostic value in predicting conversion to MS. The pattern of cognitive impairment observed in patients with CIS closely resembles that observed in patients with MS, characterized by reduced information processing speed, impaired working memory, executive functions, and attention deficits (15, 16).

In a study that divided MS patients into 3 clusters according to disability status and compared regional volumes with a healthy control group, the volumes of the thalamus, hypothalamus, putamen, and nucleus caudatus were found to be significantly different. It was thought that the ventral diencephalon underwent early degeneration during the course of MS (17). Similarly, in another study aimed at evaluating the relationship between subcortical grey matter and cognition in RRMS patients, atrophy was most prominent in the nucleus caudatus, globus pallidus, and thalamus (18). Furthermore, a study conducted in patients with CIS revealed atrophy in the thalamus, hypothalamus, putamen, nucleus caudatus, and cerebellum compared to the control group (19). In a longitudinal study with a 1-year follow-up MR imaging of RRMS and CIS patients, it was observed that atrophy developed in the grey matter, including the thalamus, nucleus caudatus, putamen, and brainstem. Deep grey matter volume, especially the thalamus volume, was predictive of cognitive performance and disability progression (20). When we compared the volumes measured in our study with the reference database, we found that the volumes of the amygdala, basal ganglia (nucleus caudatus + putamen + globus pallidus), capsule interna, nucleus caudatus, thalamus, globus pallidus, and putamen were significantly different in our patients. This result aligns with findings from other studies and suggests the development of degeneration and secondary atrophy during the clinically isolated syndrome period. Additionally, one of the unique and robust aspects of our study is the utilization of the artificial intelligence-supported BrainLab measurement method, which enables more precise and accurate measurements compared to the measurement methods commonly used in the literature.

The thalamus plays an important role in cognitive functions including attention, information processing speed and memory (21). Neurodegeneration of thalamic nuclei and connections which develops due to inflammation and cytotoxic damage leads to cognitive impairment. Many studies have concluded that thalamic atrophy develops in the early period of the disease and is a strong indicator of cognitive deficits (20, 22). In a study conducted in RRMS patients, thalamus was found to be associated with visuospatial memory (23). In another study conducted in MS patients, SDMT performance was found to be mostly associated with the thalamus and putamen and it was argued that the thalamus plays an important role in information processing efficiency (24). In a different study, thalamus volume was found to be associated with trail making test, Judgement of Line Orientation Test and SDMT performance and it was concluded that it played an important role in memory, working memory and information processing speed (25). In a study conducted by Houtchens et al. (26) in MS patients, it was suggested that thalamus volume was a significant biomarker for information processing speed and visuospatial memory. In a study conducted in patients with CIS, atrophy of the thalamus, putamen and nucleus caudatus was found and it was concluded that thalamic atrophy was an indicator in cognitive evaluation (19). In our study, a significant atrophy was found in the thalamus volume in patients with CIS compared to the reference database. Our study supports that thalamic atrophy develops even in the early period of MS and even in patients with CIS, as in other studies. The fact that a different method was used in our study instead of the commonly used measurement methods in the literature and the results were found to be similar with other studies indicates that there is a correlation between the results of the measurement methods. In addition, there was a correlation between thalamus volume and the tracking test and Judgement of Line Orientation Test.

It has been shown in many studies that the anterior thalamic nucleus plays an important role in learning and memory (27). In a study evaluating the anterior thalamic nuclei in mice, it was shown that they have roles in different stages of memory (28). In another study, a decrease in episodic memory processes, information processing speed, directed attention, working memory and executive functions performance was observed in correlation with age-related decrease in anterior thalamic volume and secondary atrophy (29). In a 3-year follow-up study in MS patients, the anterior thalamic nucleus was found to be more atrophic in patients with cognitive deterioration than in cognitively preserved patients (30). In a cross-sectional study conducted in MS patients, a relationship was found between cognitive deterioration and focal atrophy of the anterior thalamic nucleus (31). In a study examining all nuclei of the thalamus in detail, SDMT performance was found to be correlated with the volume of the left ventral anterior nucleus (32). In our study, there was a correlation between anterior thalamic nucleus volume and SDMT, Backward-Forward Digit Span Test, Stroop and Judgement of Line Orientation Test performance. The positive correlations observed with the SDMT and Backward-Forward Digit Span Test suggest that this region is actively involved in tasks requiring working memory and information processing speed. In contrast, the negative correlation with the Stroop Test indicates that while the anterior thalamic nucleus is engaged in cognitive control and attention tasks, its activity may decrease as performance on inhibitory control tasks improves. This dual role highlights the complexity of the anterior thalamic nucleus in modulating different aspects of cognition, particularly in tasks that require both rapid information processing and cognitive inhibition. These results provide a nuanced understanding of the anterior thalamic nucleus' contributions to cognitive functions, especially in patients with cognitive impairments.

The nucleus accumbens is known as the centre of reward and pleasure. It plays a modulatory role in the flow of information between the amygdala, basal ganglia, mesolimbic and dopaminergic regions and the prefrontal cortex. The nucleus accumbens is believed to be associated with the cognitive impairment seen in Alzheimer's disease. It is thought that dopaminergic system changes frequently observed in Alzheimer's patients are associated with impaired memory performance and reward processing dysfunctions (33). In a study conducted on mice, it was observed that the nucleus accumbens has an important role in mesocorticolimbic dopamine function and cognition (34). In our study, a statistically significant correlation was

found between nucleus accumbens volume and SDMT, trail making and Judgement of Line Orientation Test. Based on this, we can say that nucleus accumbens volume predicts working memory, information processing speed, executive functions and visuospatial memory performance. In our research, we did not find any studies on the relationship between nucleus accumbens volume and cognition tests in patients with CIS. We think that comprehensive studies should be conducted on this subject and these findings are one of the unique aspects of our study.

The capsulae interna coordinates cognitive, motor and sensory pathways. Fibre tracts in the anterior crus are associated with emotion, cognition, decision making and motivation (35). In a study evaluating motor and cognitive disorders with diffusion tensor imaging (DTI) in MS patients, a significant correlation was found between capsular interna DTI metrics and 9-hole peg test and PASAT performance (36). In our study, capsular interna volume was found to be atrophic according to the reference database and a significant correlation was found between capsular interna volume and stroop, trail making and Judgement of Line Orientation Test. This finding suggests that, similar to MS, capsular interna volume plays a role in working memory, information processing speed, executive functions and visuospatial functions. It was thought that cognitive functions were affected in patients with CIS before conversion to MS and that the change in capsular volume could explain this.

The cholinergic neuron population in the basal nucleus of Meynert's nucleus is involved in learning, long-term memory, control and maintenance of attention. Its degeneration causes various neuropsychiatric disorders. The association between the accumulation of Lewy bodies in the nucleus of Meynert and dementia and the favourable results obtained in dementia with DBS treatment applied to the nucleus of Meynert are proof of this. In the correlation study of BICAMS and volumetric measurement in MS patients, a significant relationship was found between them and predicted cognitive change in follow-up. In addition, the volume of Meynert's nucleus was found to be associated with lower SDMT score (37). In our study, a significant correlation was found between the performance of the tracking test and the volume of Meynert's basal nucleus and it was thought to be predictive of working memory, information processing speed and executive functions. In our research, we could not find any study in this direction in patients with CIS. Therefore, we think that comprehensive studies should be conducted on this subject and these findings are one of the unique aspects of our study. In addition, we believe that large-scale double-blind controlled studies are needed to evaluate the effect of early initiation of cholinesterase inhibitor treatment on the protection of patients from cognitive impairment.

Our findings, particularly the significant correlation between the Judgement of Line Orientation Test and six distinct anatomical regions, as well as the association of the anterior thalamic nucleus with working memory and information processing speed, can provide valuable insights for managing CIS patients. These correlations suggest that the anterior thalamic nucleus plays a critical role in multiple cognitive domains, especially those related to visuospatial processing, working memory, and rapid cognitive functioning. For CIS patients, who often experience early neurological symptoms that may precede multiple sclerosis, assessing cognitive functions through specific tests like the Judgement of Line Orientation Test and evaluating the integrity of the anterior thalamic nucleus may offer a more targeted approach for early intervention. For instance, using the Judgement of Line Orientation Test can help assess visuospatial abilities, a domain that may be disrupted in CIS due to early thalamic or parietal lobe involvement. Furthermore, the strong correlation of the anterior thalamic nucleus with working memory and processing speed highlights the importance of monitoring these cognitive functions in CIS patients, as deficits in these areas may signal more extensive brain involvement or the transition to MS. By incorporating these specific tests into routine clinical assessments for CIS patients, healthcare providers can better identify early cognitive changes, tailor cognitive rehabilitation strategies, and potentially intervene earlier in the disease course.

Our study highlights the differential predictive power of age and gender on various brain region volumes and cognitive functions. While age emerged as a significant predictor for tracking test performance, it showed no substantial impact on the volumes of key subcortical structures such as the thalamus, amygdala, and putamen. Similarly, gender demonstrated borderline significance for some regions, such as the capsula interna, but was not a robust predictor overall. These results suggest that volumetric changes in certain brain regions may occur independently of these demographic factors, aligning with the growing understanding that intrinsic disease processes in CIS play a dominant role in neurodegeneration.

There is no comprehensive study of this type in the literature that examines various cognitive functions, cranial volumetric measurements and their correlation in patients with CIS. The strengths of this study are that a homogeneous group was formed, a larger number of anatomical regions that had not been evaluated before were evaluated compared to other studies, more precise and accurate volume measurements were provided by using artificial intelligence with the BrainLab Elements™ Atlas-Based Automatic Segmentation programme, and a large number of neuropsychological tests covering the main cognitive functions were used. The limitations of our study are that, it is a single-centre study, cross-sectional evaluation and we did not estimate pre-disease intelligence. The number of CIS patients included in the study is relatively lower compared to MS patients. Additionally, for volumetric analysis to be performed, MR imaging needs to be acquired using consistent techniques and sequences, which further limited the number of eligible patients. This is one of the reasons for the small sample size, which presents a limitation in terms of the generalizability of the results. However, despite this limitation, careful and reliable analyses were conducted using the BrainLab Elements™ Atlas-Based Automatic Segmentation program. In addition, the fact that we did not include the anatomical locations of demyelinating lesions in our analyses can be counted as another factor. Future longitudinal studies are needed to determine the usefulness and predictive value of volumetric measurements and cognitive functions in determining the risk of conversion to MS in patients with CIS.

Conclusion

In conclusion, our study highlights the significant role of the anterior thalamic nucleus in various cognitive functions, particularly in working memory, information processing speed, and visuospatial tasks in patients with CIS. The Judgement of Line Orientation Test emerged as a key tool for assessing visuospatial abilities, demonstrating strong correlations with multiple brain regions in patients with CIS.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Bakırköy Dr. Sadi Konuk Clinical Research Ethical Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ÖA: Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. IA: Conceptualization, Data curation, Writing – original draft, Writing – review & editing. OH: Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing. MS: Data curation, Writing – original draft, Writing – review & editing. EH: Methodology, Software, Writing – original draft, Writing – review & editing. VY: Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Bradley WG, Daroff RB, Fenichel GM, Jankovic J. (2008) Neurology in clinical practice. pp.1584–1612, *5*, Philadelphia, PA, Elsevier/Saunders.

2. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol.* (2011) 69:292–302. doi: 10.1002/ana.22366

3. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* (2018) 17:162–73. doi: 10.1016/S1474-4422(17)30470-2

4. Nocentini U, Pasqualetti P, Bonavita S, Buccafusca M, De Caro MF, Farina D, et al. Cognitive dysfunction in patients with relapsing-remitting multiple sclerosis. *Mult Scler.* (2006) 12:77–87. doi: 10.1191/135248506ms1227oa

5. Radue EW, Barkhof F, Kappos L, Sprenger T, Häring DA, De Vera A, et al. Correlation between brain volume loss and clinical and MRI outcomes in multiple sclerosis. *Neurology*. (2015) 84:784–93. doi: 10.1212/WNL.0000000000001281

6. Bagert B, Camplair P, Bourdette D. Cognitive dysfunction in multiple sclerosis natural history, pathophysiology and management. *CNS Drugs*. 16:445–55. doi: 10.2165/00023210-200216070-00002

7. Zipoli V, Goretti B, Hakiki B, Siracusa G, Sorbi S, Portaccio E, et al. Cognitive impairment predicts conversion to multiple sclerosis in clinically isolated syndromes. *Mult Scler*. (2010) 16:62–7. doi: 10.1177/1352458509350311

8. Deloire M, Ruet A, Hamel D, Bonnet M, Brochet B. Early cognitive impairment in multiple sclerosis predicts disability outcome several years later. *Mult Scler*. (2010) 16:581–7. doi: 10.1177/1352458510362819

9. Amato MP, Zipoli V, Portaccio E. Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. *J Neurol Sci.* (2006) 245:41–6. doi: 10.1016/j.jns.2005.08.019

10. Chiaravalloti ND, DeLuca J. Cognitive impairment in multiple sclerosis. *Lancet Neurol.* (2008) 7:1139–51. doi: 10.1016/S1474-4422(08)70259-X

11. Amato MP, Portaccio E, Goretti B, Zipoli V, Battaglini M, Letizia Bartolozzi M, et al. Association of neocortical volume changes with cognitive deterioration in relapsing-remitting multiple sclerosis. *Arch Neurol.* 64:1157–61. doi: 10.1001/archneur.64.8.1157

12. Xiao Y, Lau JC, Anderson T, DeKraker J, Collins DL, Peters T, et al. An accurate registration of the BigBrain dataset with the MNI PD25 and ICBM152 atlases. *Sci Data*. (2019) 6:210. doi: 10.1038/s41597-019-0217-0

13. Akeret K, van Niftrik CHB, Sebök M, Muscas G, Visser T, Staartjes VE, et al. Topographic volume-standardization atlas of the human brain. *Brain Struct Funct*. (2021) 226:1699–711. doi: 10.1007/s00429-021-02280-1

14. Dutta R, Trapp BD. Mechanisms of neuronal dysfunction and degeneration in multiple sclerosis. *Prog Neurobiol*. (2011) 93:1–12. doi: 10.1016/j.pneurobio.2010.09.005

15. Feuillet L, Reuter F, Audoin B, Malikova I, Barrau K, Ali Cherif A, et al. Early cognitive impairment in patients with clinically isolated syndrome suggestive of multiple sclerosis. *Mult Scler*. (2007) 13:124–7. doi: 10.1177/1352458506071196

16. Anhoque CF, Domingues SCA, Teixeira AL, Domingues RB. Prejuízo cognitivo na síndrome clínica isolada: Uma revisão sistemática. *Dement Neuropsychol*. (2010) 4:86–90. doi: 10.1590/S1980-57642010DN40200002

17. Fujimori J, Fujihara K, Ogawa R, Baba T, Wattjes M, Nakashima I. Patterns of regional brain volume loss in multiple sclerosis: a cluster analysis. *J Neurol*. (2020) 267:395–405. doi: 10.1007/s00415-019-09595-4

18. Schoonheim MM, Popescu V, Lopes FCR, Wiebenga OT, Vrenken H, Douw L, et al. Subcortical atrophy and cognition: sex effects in multiple sclerosis. *Neurology*. (2012) 79:1754–61. doi: 10.1212/WNL.0b013e3182703f46

19. Henry RG, Shieh M, Okuda DT, Evangelista A, Gorno-Tempini ML, Pelletier D. Regional grey matter atrophy in clinically isolated syndromes at presentation. *J Neurol Neurosurg Psychiatry*. (2008) 79:1236–44. doi: 10.1136/jnnp.2007.134825

20. Radetz A, Koirala N, Krämer J, Johnen A, Fleischer V, Gonzalez-Escamilla G, et al. Gray matter integrity predicts white matter network reorganization in multiple sclerosis. *Hum Brain Mapp*. (2020) 41:917–27. doi: 10.1002/hbm.24849

21. Amin M, Ontaneda D. Thalamic injury and cognition in multiple sclerosis. *Front Neurol*. (2021) 11:623914. doi: 10.3389/fneur.2020.623914

22. Štecková T, Hluštík P, Sládková V, Odstrčil F, Mareš J, Kaňovský P. Thalamic atrophy and cognitive impairment in clinically isolated syndrome and multiple sclerosis. *J Neurol Sci*. (2014) 342:62–8. doi: 10.1016/j.jns.2014.04.026

23. Andravizou A, Siokas V, Artemiadis A, Bakirtzis C, Aloizou AM, Grigoriadis N, et al. Clinically reliable cognitive decline in relapsing remitting multiple sclerosis: is it the tip of the iceberg? *Neurol Res*. (2020) 42:575–86. doi: 10.1080/01616412.2020.1761175

24. Batista S, Zivadinov R, Hoogs M, Bergsland N, Heininen-Brown M, Dwyer MG, et al. Basal ganglia, thalamus and neocortical atrophy predicting slowed cognitive processing in multiple sclerosis. *J Neurol*. (2012) 259:139–46. doi: 10.1007/s00415-011-6147-1

25. Matías-Guiu JA, Cortés-Martínez A, Montero P, Pytel V, Moreno-Ramos T, Jorquera M, et al. Identification of cortical and subcortical correlates of cognitive performance in multiple sclerosis using voxel-based morphometry. *Front Neurol*. (2018) 9:920. doi: 10.3389/fneur.2018.00920

26. Houtchens MK, Benedict RHB, Killiany R, Sharma J, Jaisani Z, Singh B, et al. Thalamic atrophy and cognition in multiple sclerosis. *Neurology*. (2007) 69:1213–23. doi: 10.1212/01.wnl.0000276992.17011.b5

27. Van der Werf YD, Jolles J, Witter MP, Uylings HB. Contributions of thalamic nuclei to declarative memory functioning. *Cortex*. 39:1047–62. doi: 10.1016/s0010-9452(08)70877-3

28. Safari V, Nategh M, Dargahi L, Zibaii ME, Khodagholi F, Rafiei S, et al. Individual subnuclei of the rat anterior thalamic nuclei differently affect spatial memory and passive avoidance tasks. *Neuroscience*. (2020) 444:19–32. doi: 10.1016/j.neuroscience.2020.07.046

29. Fama R, Sullivan EV. Thalamic structures and associated cognitive functions: relations with age and aging. *Neurosci Biobehav Rev*. (2015) 54:29–37. doi: 10.1016/j.neubiorev.2015.03.008

30. Bergsland N, Zivadinov R, Dwyer MG, Weinstock-Guttman B, Benedict RHB. Localized atrophy of the thalamus and slowed cognitive processing speed in MS patients. *Mult Scler*. (2016) 22:1327–36. doi: 10.1177/1352458515616204

31. Bisecco A, Rocca MA, Pagani E, Mancini L, Enzinger C, Gallo A, et al. Connectivity-based parcellation of the thalamus in multiple sclerosis and its implications for cognitive impairment: a multicenter study. *Hum Brain Mapp*. (2015) 36:2809–25. doi: 10.1002/hbm.22809

32. Trufanov A, Bisaga G, Skulyabin D, Temniy A, Poplyak M, Chakchir O, et al. Thalamic nuclei degeneration in multiple sclerosis. *J Clin Neurosci*. (2021) 89:375–80. doi: 10.1016/j.jocn.2021.05.043

33. Nobili A, Latagliata EC, Viscomi MT, Cavallucci V, Cutuli D, Giacovazzo G, et al. Dopamine neuronal loss contributes to memory and reward dysfunction in a model of Alzheimer's disease. *Nat Commun*. (2017) 8:8. doi: 10.1038/ncomms14727

34. Laplante F, Zhang ZW, Huppé-Gourgues F, Dufresne MM, Vaucher E, Sullivan RM. Cholinergic depletion in nucleus accumbens impairs mesocortical dopamine activation and cognitive function in rats. *Neuropharmacology*. (2012) 63:1075–84. doi: 10.1016/j.neuropharm.2012.07.033

35. Safadi Z, Grisot G, Jbabdi S, Behrens TE, Heilbronner SR, McLaughlin NCR, et al. Functional segmentation of the anterior limb of the internal capsule: linking white matter abnormalities to specific connections. *J Neurosci*. (2018) 38:2106–17. doi: 10.1523/JNEUROSCI.2335-17.2017

36. Sbardella E, Petsas N, Tona F, Prosperini L, Raz E, Pace G, et al. Assessing the correlation between grey and white matter damage with motor and cognitive impairment in multiple sclerosis patients. *PLoS One*. (2013) 8:e63250. doi: 10.1371/journal.pone.0063250

37. Hildesheim FE, Benedict RHB, Zivadinov R, Dwyer MG, Fuchs T, Jakimovski D, et al. Nucleus basalis of Meynert damage and cognition in patients with multiple sclerosis. *J Neurol*. (2021) 268:4796–808. doi: 10.1007/s00415-021-10594-7

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 31 January 2025
DOI 10.3389/fimmu.2025.1532660

OPEN ACCESS

EDITED BY
Hans-Peter Hartung,
Heinrich Heine University, Germany

REVIEWED BY
Reza Rahmanzadeh,
TheUltra.ai, Switzerland
Ioannis Nikolaidis,
Hippokration General Hospital, Greece

*CORRESPONDENCE
Vinzenz Fleischer
vinzenz.fleischer@unimedizin-mainz.de

†These authors have contributed equally to this work

RECEIVED 22 November 2024
ACCEPTED 16 January 2025
PUBLISHED 31 January 2025

CITATION
Fleischer V, Brummer T, Muthuraman M, Steffen F, Heldt M, Protopapa M, Schraad M, Gonzalez-Escamilla G, Groppa S, Bittner S and Zipp F (2025) Biomarker combinations from different modalities predict early disability accumulation in multiple sclerosis. *Front. Immunol.* 16:1532660. doi: 10.3389/fimmu.2025.1532660

Biomarker combinations from different modalities predict early disability accumulation in multiple sclerosis

Vinzenz Fleischer[1*†], Tobias Brummer[1†], Muthuraman Muthuraman[1,2], Falk Steffen[1], Milena Heldt[1], Maria Protopapa[1], Muriel Schraad[1], Gabriel Gonzalez-Escamilla[1], Sergiu Groppa[1], Stefan Bittner[1†] and Frauke Zipp[1†]

[1]Department of Neurology, Focus Program Translational Neuroscience (FTN), Rhine Main Neuroscience Network (rmn²), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany, [2]Department of Neurology, Section of Neural Engineering with Signal Analytics and Artificial Intelligence, University Hospital Würzburg, Würzburg, Germany

Objective: Establishing biomarkers to predict multiple sclerosis (MS) disability accrual has been challenging using a single biomarker approach, likely due to the complex interplay of neuroinflammation and neurodegeneration. Here, we aimed to investigate the prognostic value of single and multimodal biomarker combinations to predict four-year disability progression in patients with MS.

Methods: In total, 111 MS patients were followed up for four years to track disability accumulation based on the Expanded Disability Status Scale (EDSS). Three clinically relevant modalities (MRI, OCT and blood serum) served as sources of potential predictors for disease worsening. Two key measures from each modality were determined and related to subsequent disability progression: lesion volume (LV), gray matter volume (GMV), retinal nerve fiber layer, ganglion cell-inner plexiform layer, serum neurofilament light chain (sNfL) and serum glial fibrillary acidic protein. First, receiver operator characteristic (ROC) analyses were performed to identify the discriminative power of individual biomarkers and their combinations. Second, we applied structural equation modeling (SEM) to the single biomarkers in order to determine their causal inter-relationships.

Results: Baseline GMV on its own allowed identification of subsequent EDSS progression based on ROC analysis. All other individual baseline biomarkers were unable to discriminate between progressive and non-progressive patients on their own. When comparing all possible biomarker combinations, the tripartite combination of MRI, OCT and blood biomarkers achieved the highest discriminative accuracy. Finally, predictive causal modeling identified that LV mediates significant parts of the effect of GMV and sNfL on disability progression.

Conclusion: Multimodal biomarkers, i.e. different major surrogates for pathology derived from MRI, OCT and blood, inform about different parts of the disease pathology leading to clinical progression.

KEYWORDS

multiple sclerosis, biomarker, magnetic resonance imaging, neurofilament, optical coherence tomography, disease progression, prediction, structural equation modeling

Introduction

In multiple sclerosis (MS), disability progression is closely related to neuroaxonal degeneration (1, 2). Therefore, identifying and quantifying axonal damage is an essential step towards improved clinical decision-making and prognostication. Currently, magnetic resonance imaging (MRI) is the most established non-invasive modality for diagnosing, evaluating treatment effectiveness, and monitoring disease progression in patients with MS. In particular, conventional structural MRI metrics, like T2-hyperintense lesion volume (LV) and gray matter volume (GMV), have been proven to be reproducible and well-validated in reflecting disease activity and progression, respectively (3, 4). However, recent technical advances, such as single molecule array (SiMoA) and easily accessible optical coherence tomography (OCT), have enabled additional non-invasive measurements of neurodegeneration-related biomarkers with increasing clinical application (5, 6). Therefore, blood-based biomarkers such as serum neurofilament light chain (sNfL) and serum glial fibrillary acidic protein (sGFAP), as well as measures of retinal thickness (retinal nerve fiber layer (RNFL), ganglion cell inner plexiform layer (GCIPL)) have gained significant interest for diagnostic purposes and are expected to be applied in clinical routine soon.

Nevertheless, all biomarkers have certain limitations due to the nature of their respective modalities: MRI is most effective at detecting focal white matter lesions in the brain and spinal cord, but lesions in gray matter structures can only be reliably visualized with rather high field strengths (7). Additionally, conventional MRI is functionally "blind" to what is known as "normal-appearing white matter" (NAWM). Blood biomarkers of neuronal (sNfL) or glial (sGFAP) damage can be influenced by different factors such as age, blood volume, genetics, and other medical conditions such as impaired renal function (8–10). Additionally, measures of retinal thickness may not always accurately reflect the presence and extent of inflammation or damage in the brain and spinal cord, as they may be affected by factors such as pupil dilation, eye movements, and the presence of cataracts or other eye conditions, which can impact the accuracy of the results (6, 11). Furthermore, the spatial resolution is limited, as OCT captures only a small part of the central nervous system (CNS). Thus, the concept of "one biomarker" indicating the existence of an underlying disease-specific process remains a utopia in predicting disease progression. However, individual challenges may be overcome by combining biomarkers from different modalities that ideally also represent multiple aspects of MS pathology.

Utilizing multiple biomarkers from different modalities has already been demonstrated in other neurological disorders such as Alzheimer's disease, where a combination of positron emission tomography (PET)-imaging and cerebrospinal fluid (CSF) biomarkers has enabled a more precise diagnostic evaluation (12, 13). In people with MS, initial efforts have shown that multimodal biomarkers can predict neuropsychological parameters such as cognitive impairment (14). However, it is unclear which biomarker combinations offer the best discriminative accuracy for disease progression of MS. The combination of several biomarkers altogether, by means of predictive modeling, may be able to compile large amounts of multimodal data, in order to attain solid conclusions and decision making in MS monitoring.

Thus, the aim of this study was to investigate the prognostic value of individual biomarkers (MRI, OCT and blood), as well as their combinations in predicting four-year disease activity and progression in MS. To test this, we determined LV and GMV from MRI, RNFL and GCIPL from OCT and sGFAP and sNfL from blood within a cohort of 111 MS patients who were clinically followed up for four years.

Methods

Participants

In total, out of 141 MS patients that were retrospectively screened for this project, 111 MS patients that underwent a comprehensive and detailed clinical assessment were finally included in the analysis (Figure 1). The selected cohort included MS patients with MRI (T2-hyperintense LV and GMV), blood (sNfL and sGFAP), and OCT (RNFL and GCIPL) measurements at the outpatient clinic of the Department of Neurology, at the University Medical Center Mainz (Germany) (Table 1). All included patients had relapsing-remitting multiple sclerosis (RRMS) as diagnosed according to the 2017 revised McDonald diagnostic criteria (15). The mean (± standard deviation) disease duration of all patients at study inclusion was 3.15 ± 4.26 years.

10.3389/fimmu.2025.1532660

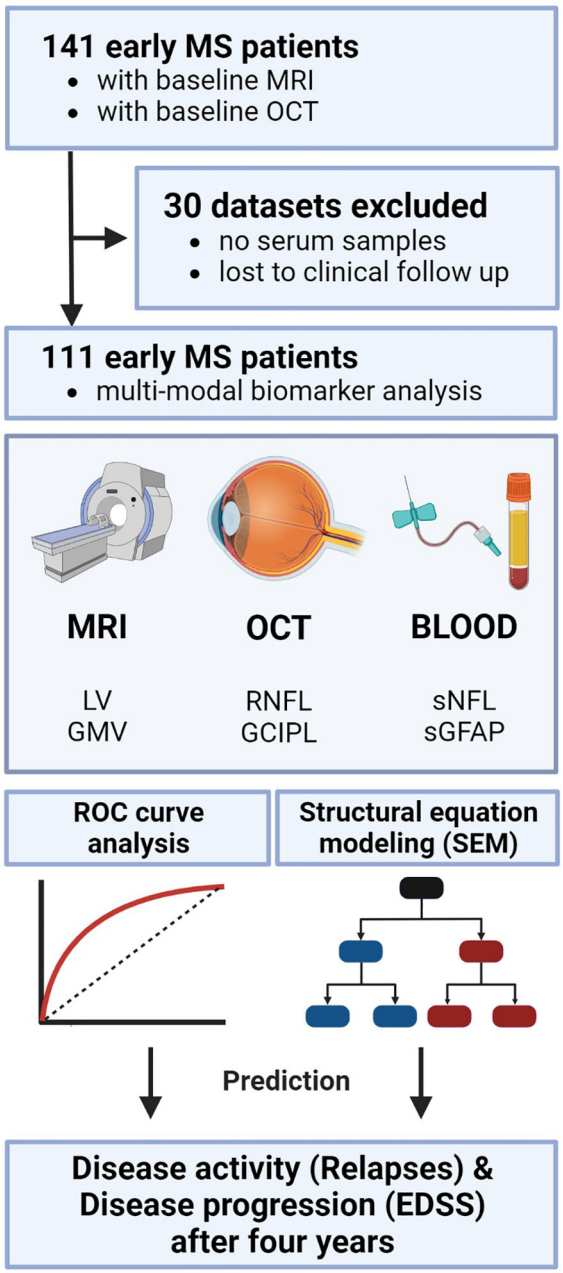


FIGURE 1
Study analysis design. Study protocol and design including the three modalities each with two biomarkers as potential predictors. Statistically, ROC analysis was performed to evaluate the discriminative power of single and combined biomarkers. Subsequently, SEM was applied to test the causal inter-relationships between the variables. EDSS, expanded disability status scale; GCIPL, ganglion cell-inner plexiform layer; GMV, gray matter volume; OCT, optical coherence tomography; RNFL, retinal nerve fiber layer; ROC, receiver operator characteristics; SEM, structural equation modeling; sGFAP, serum glial fibrillary acidic protein; sNfL, serum neurofilament light; LV, lesion volume.

All diagnostic baseline measurements were performed within 6 months of study inclusion. An experienced neurologist clinically assessed patients and their Expanded Disability Status Scale (EDSS) score at study entry and follow up visit (3.74 ± 1.25 years), along with clinical relapse history over the study period. EDSS progression was defined as an increase of ≥ 1 point in the EDSS score for a baseline score of ≥ 1.5 or a 1.5 point increase for a baseline score of 0 (16). A clinical relapse was defined as a monophasic clinical episode with new neurological symptoms, lasting more than 24 h and in the absence of fever or infection (15). The annualized relapse rates (ARR) were calculated by dividing the total number of all observed relapses by the total number of patient-years. All measurements were performed at least 30 days after a clinical relapse and/or a high-dose corticosteroid treatment.

TABLE 1 Basic characteristics. Demographic and clinical data of the included MS patients as well as MRI, OCT and blood biomarkers at baseline.

Demographics	MS patients (n = 111)
Age [years] mean ± SD	34.8 ± 9.67
Sex [female] (percent)	79 (71)
Disease duration [years] mean ± SD	3.15 ± 4.26
Disease-modifying treatment	
None (percent)	18 (16)
Mild to moderate efficacy (percent)	69 (62)
High efficacy (percent)	24 (22)
Clinical measures	
Baseline EDSS median (25^{th}; 75^{th} percentile)	1.0 (0.0; 2.0)
Follow up EDSS median (25^{th}; 75^{th} percentile)	1.5 (0.0; 2.5)
Patients with EDSS progression (percent)	46 (41.4)
Relapses over 4 years mean ± SD	0.76 ± 1.17
Annualized relapse rate mean ± SD	0.21 ± 0.33
Time to follow up [years] mean ± SD	3.74 ± 1.25
Patients with history of optic neuritis (percent)	33 (30)
MRI measures	
LV [ml] mean ± SD	5.97 ± 9.57
GMV [fraction] mean ± SD	0.43 ± 0.03
OCT measures	
RNFL [mm^3] mean ± SD	0.21 ± 0.02
GCIPL [mm^3] mean ± SD	0.76 ± 0.1
Blood measures	
sNfL [z-score] mean ± SD	0.115 ± 2.21
sGFAP [pg/ml] mean ± SD	121.2 ± 43.8

Mild to moderate efficacy = interferons, glatiramer acetate, teriflunomide, dimethyl fumarate. High efficacy = natalizumab, anti-CD20 monoclonal antibodies, sphingosine-1-phosphate receptor modulators, alemtuzumab.
EDSS, extended disability status scale; GCIPL, ganglion cell-inner plexiform layer; GMV, gray matter volume; LV, lesion volume; MRI, magnetic resonance imaging; OCT, optical coherence tomography; RNFL, retinal nerve fiber layer; SD, standard deviation; sGFAP, serum glial fibrillary acidic protein; sNfL, serum neurofilament light.

sNfL and sGFAP measurements

Serum samples were collected by attending physicians at the University Medical Center Mainz. Samples were processed at room temperature within 2 hours. Serum samples were spun at 2000xg at room temperature for 10 minutes, aliquoted in polypropylene tubes and stored at −80°C. sNfL and sGFAP concentrations were measured as previously described (10, 14). In brief, sNfL and sGFAP levels were determined using the highly sensitive single molecule array (SiMoA) technology (17). Samples were measured in duplicates by SiMoA HD-1 (Quanterix, USA) using NF-Light Advantage kits according to the manufacturer's instructions. The mean inter-assay and intra-assay coefficient of variation was less than 10%. Measurements were performed in a blinded fashion without information about clinical data.

MRI data acquisition

MRI data acquisition was performed as previously described (14). In brief, structural MRI was performed on a 3-Tesla MRI scanner (Magnetom Tim Trio, Siemens, Germany) with a 32-channel receive-only head coil. In all patients, imaging was performed using a sagittal 3D T1-weighted magnetization-prepared rapid gradient echo (MP-RAGE) sequence (TE/TI/TR = 2.52/900/1900 ms, flip angle = 9°, field of view = 256 × 256 mm2, matrix size = 256 × 256, slab thickness = 192 mm, voxel size = 1 × 1 × 1 mm^3) and a sagittal 3D T2-weighted fluid-attenuated inversion recovery (FLAIR) sequence (TE/TI/TR = 388/1800/5000 ms, echo-train length = 848, field of view = 256 × 256 mm^2, matrix size = 256 × 256, slab thickness = 192 mm, voxel size = 1 × 1 × 1 mm^3). A clinician scientist blinded to the patient data excluded major anatomical abnormalities based on the subject's T1-weighted and FLAIR images of the whole brain.

Quantification of white matter LV and GMV

The quantification of WM (white matter) volume, lesion volume and GMV was performed as previously described (14). Using voxel-based morphometry (VBM) analysis in the Statistical Parametric Mapping (SPM8) software, the GM and WM volumes were calculated. The volumes of WM lesions were assessed using the cross-sectional lesion growth algorithm of the lesion segmentation toolbox (18) included in the SPM8 software. 3D FLAIR images were co-registered to 3D T1-weighted images and bias corrected. After partial volume estimation, lesion segmentation was performed with 20 different initial threshold values for the lesion growth algorithm (18). By comparing manually and automatically estimated lesion maps, the optimal threshold (κ value, dependent on image contrast) was determined, and average values were calculated for each patient. A uniform κ value of 0.1 was applied in all patients in order to automatically estimate lesion volume and filling of 3D T1-weighted images. Subsequently, the filled 3D T1-weighted images and the native 3D T1-weighted images were segmented into GM, WM, and CSF and then normalized to the Montreal Neurological Institute (MNI) space. The quality of the segmentations was visually inspected to increase reliability.

OCT: image acquisition and scanning protocol

The analysis was performed as previously described (19, 20). In brief, the Advised Protocol for OCT Study Terminology and Elements (APOSTEL) recommendations were followed (21) including a quality control for the raw OCT scans complying with the OSCAR-IB criteria (22). MS patients with accompanying diseases potentially affecting the optic nerve or other ocular disease were excluded in advance. Hence, none of the patients had a history of glaucoma, retinopathy or other neurological disorders (besides RRMS). An experienced operator performed OCT image acquisition following a unified standard acquisition protocol using a spectral domain OCT (Heidelberg Spectralis, Heidelberg Engineering, Germany) with Heidelberg Eye Explorer software (HEYEX, version 1.10.2.0). The measurements were acquired in a shaded room at ambient light without pupillary dilation. Intra-retinal layers of the macula were gauged by a standardized scan comprising 61 vertical or horizontal B-scans while focusing on the fovea at a scanning angle of 30° × 25° and a resolution of 768 × 496 pixels. Automatic real time was set to nine at high-speed scanning mode. Confocal scanning laser ophthalmoscopy was performed in parallel and revealed no evidence of pathology. No further fundoscopic imaging was carried out. To account for inter-eye within-patient dependencies, we calculated the mean of both eyes in patients with no history of optic neuritis; in patients with a history of unilateral optic neuritis, we only used the OCT scan of the non-affected eye. Hence, the main statistical analysis was performed at a per-patient level. All B-scans were automatically segmented (followed by manual correction by a trained rater) using segmentation beta-software (Spectralis Viewing Module version 6.9.5.0) of the Heidelberg Eye Explorer (version 1.10.2.0) provided by the manufacturer. The segmentation lines were the following retinal layers: RNFL, GCIPL, inner nuclear layer, outer plexiform layer and outer nuclear layer. The mean volume of the individual retinal layers was computed in an area of a radius of 3.45 mm around the fovea including the fovea using the Early Treatment of Diabetic Retinopathy Study (ETDRS) grid. Lastly, RNFL and GCIPL were finally selected as primary estimate for neuroaxonal damage of the retina, as both have been associated with brain atrophy and disability worsening (23, 24).

Statistics

Statistical analysis was performed using SPSS 23 (SPSS, Chicago, IL, USA), MedCalc (Version 20.115) and GraphPad Prism 9 software. Summary statistics are presented as mean ± standard deviation (SD), or median (25^{th} and 75^{th} percentile), or number (percentage), where applicable. To create a combined variable for each biomarker combination, a binary logistic

regression model for each combination (corrected for sex, age, disease duration and disease-modifying treatment) was estimated in order to get the predicted probability from each model. Then, we used this probability as the test variable in the subsequent receiver operating characteristic (ROC) procedure (14).

A ROC analysis was performed to calculate the predictive discriminating values for each biomarker and the combinations. This statistical method is preferentially used to make a series of discriminations into two different states based on a specific diagnostic variable. Here, the presence or absence of relapses or EDSS worsening, served as binary classifiers. Every value of that discriminating variable is used as a cut-off with calculation of the corresponding sensitivity and specificity.

Structural equation modeling

The analysis was performed as previously described (25) using the SEM toolbox for MATLAB (version 13a; Mathworks, Natick, MA, USA). SEM represents a statistical technique that is used to test and estimate structural relationships between variables in a model. By structural, we mean that we incorporate causal assumptions as part of the model. Hence, SEM represents a multivariate technique that is able test complex relationships among multiple variables simultaneously, and estimate the strength and direction of these relationships. In our model, we explored the association between multimodal biomarkers and the clinical outcomes (clinical relapses and EDSS progression). We used the Maximum likelihood method of estimation to fit the models. In order to adjust the models for a large sample size, we used the Root Mean Square Error of Approximation (RMSEA) index, which improves precision without increasing bias (26). The RMSEA index estimates lack of fit in a model compared to a perfect model and therefore should be low. In all models, the Invariant under a Constant Scaling (ICS) and ICS factor (ICSF) criteria should be close to zero, indicating that models were appropriate for analysis. Finally, based on the Akaike Information Criterion (AIC) the quality of each model relative to other models was estimated, with smaller values signifying a better fit of the model. The strength of associations between the variables in the models was quantified by standardized coefficients (s), ranging from 0 (no association) to 1 (very strong association). To correct for potential confounders the models were adjusted for sex, age, disease duration and disease-modifying treatment (DMT). P-values less than 0.05 were considered statistically significant.

Results

Patient characteristics

All demographics and clinical characteristics of the investigated cohort are summarized in Table 1. In total, 141 early MS patients with baseline MRI and OCT were selected. Thirty patients were excluded from the final analysis because either there was no serum sample available or they were lost to clinical follow-up (Figure 1). The mean follow-up time in our longitudinal cohort of 111 patients was 3.74 ± 1.25 years. The mean age ± SD was 34.8 ± 9.67 years; 79 patients (71.0%) were female and 32 (29.0%) were male. The mean disease duration at study inclusion was 3.15 ± 4.26 years. All patients had a relapsing-remitting disease course (RRMS) according to the 2017 revised McDonald criteria (15). At the time of inclusion, 18 patients (16%) were not receiving any DMT, 69 (62%) were receiving a mild to moderate efficacy DMT, and 24 (22%) were receiving a high efficacy DMT. The median baseline disability, quantified with EDSS, was 1.0 (25th and 75th percentile: 0.0–2.0). Overall, 46 patients (41.4%) experienced EDSS progression during the observation period. The mean ARR was 0.21 ± 0.33; 33 (30%) patients had a history of optic neuritis. The results from blood biomarker, MRI, and OCT measurements are also summarized in Table 1.

Predictive discrimination model

An overall ROC analysis was performed to determine the predictive discriminating value of the individual and combined measures to distinguish MS patients with and without disease activity (determined through the presence or absence of relapses during this time) and with and without disability progression (determined through the presence or absence of EDSS worsening over four years). Resulting values with AUC, standard error, 95% confidence interval and p-values are presented in detail in Figures 2A and 3A.

In general, none of the individual biomarkers were able to predict the occurrence of clinical relapses within the 4-year observation period (AUC-range: 0.523 – 0.602). All p-values for testing AUC = 0.5 vs. AUC ≠ 0.5 were greater than 0.05 and were hence not significantly different from a random classifier (Figure 2B). Only LV showed a trend towards significance (AUC = 0.602; p = 0.060). In the ROC analysis based on the presence or absence of EDSS progression, GMV was the only single biomarker to show significant predictive capability for EDSS progression on its own (AUC = 0.614, SE = 0.054; p = 0.035), whereas all other single biomarkers did not (AUC-range = 0.502 - 0.596) (Figure 3B).

When we combined biomarkers within their respective modality, MRI markers (LV + GMV) were able to predict both relapses (AUC = 0.631, SE = 0.054; p = 0.015) and EDSS progression over the four-year period (AUC = 0.621, SE = 0.055; p = 0.026). Combined blood biomarkers (sNfL + sGFAP) were only able to predict EDSS progression (AUC = 0.632, SE = 0.059; p = 0.025), while combined OCT measures (RNFL + GCIPL) were unable to predict either clinical relapses (AUC = 0.599, SE = 0.054; p = 0.069) or EDSS progression (AUC = 0.507, SE = 0.058, p = 0.906) (Figures 2C, 3C).

However, all combinations of two biomarker modalities significantly predicted clinical relapses (AUC range = 0.636 – 0.643) and EDSS progression (AUC range = 0.631 – 0.699) (Figures 2D, 3D). The best prediction for EDSS progression using two modalities was achieved with a combination of MRI and blood biomarkers (AUC = 0.699, SE = 0.055; p < 0.001).

Most notably, the combination of all six biomarkers achieved the highest AUC for discriminating MS patients with clinical

10.3389/fimmu.2025.1532660

A ROC curve results

Biomarkers		AUC	SE	95% CI	*p-value*
Single biomarkers	GMV	0.523	0.056	0.426 – 0.619	*0.675*
	LV	0.602	0.054	0.505 – 0.694	*0.060*
	RNFL	0.537	0.055	0.440 – 0.632	*0.503*
	GCIPL	0.590	0.055	0.492 – 0.682	*0.100*
	sNFL	0.559	0.055	0.462 – 0.653	*0.220*
	sGFAP	0.570	0.062	0.462 – 0.673	*0.259*
Combined biomarkers (one modality)	MRI (GMV & LV)	0.631	0.054	0.534 – 0.721	***0.015***
	OCT (RNFL & GCIPL)	0.599	0.055	0.502 – 0.691	*0.069*
	BLOOD (sNFL & sGFAP)	0.604	0.060	0.496 – 0.705	*0.086*
Combined biomarkers (two modalities)	MRI & OCT	0.643	0.054	0.547 – 0.732	***0.008***
	MRI & BLOOD	0.636	0.059	0.528 – 0.734	***0.021***
	BLOOD & OCT	0.637	0.059	0.530 – 0.735	***0.020***
Combined biomarkers (three modalities)	MRI & OCT & BLOOD	0.678	0.057	0.572 – 0.772	***0.002***

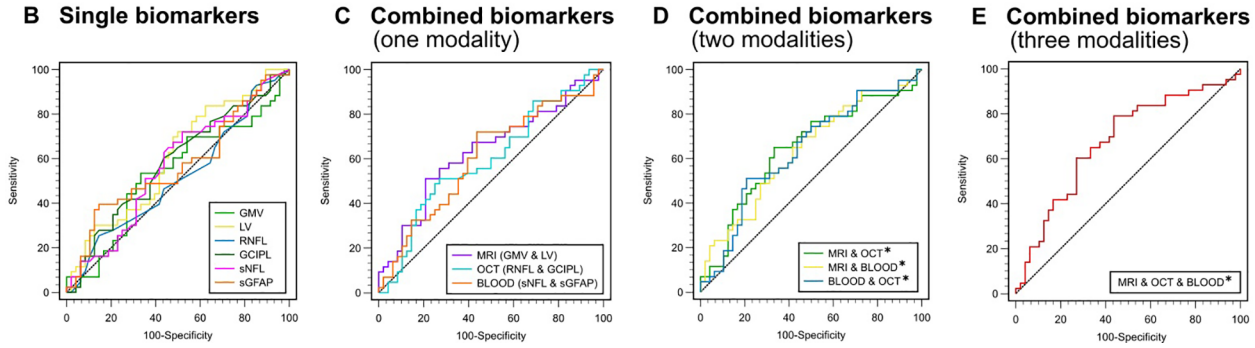


FIGURE 2
ROC analysis for the discrimination between the presence or absence of relapse activity **(A)** Color-coded table depicting the ROC analysis for individual and combinations of biomarkers. AUC, p-value and 95%-CI for the prediction of clinical relapses (yes/no). **(B)** ROC curves for single biomarkers. **(C)** ROC curves for combined biomarkers within one modality. **(D)** ROC curves for combined biomarkers within two modalities. **(E)** ROC curve for combined biomarkers of all three modalities (GMV + LV, RNFL + GCIPL and sNfL + sGFAP). AUC, area under the curve; CI, confidence interval; GCIPL, ganglion cell-inner plexiform layer; GMV, gray matter volume; LV, lesion volume; OCT, optical coherence tomography; RNFL, retinal nerve fiber layer; ROC, receiver operator characteristics; SE, standard error; sGFAP, serum glial fibrillary acidic protein; sNfL, serum neurofilament light.

relapse activity from those without (AUC = 0.678, SE = 0.057; p = 0.002) and for discriminating progressive from non-progressive MS patients (AUC = 0.706, SE = 0.055; p < 0.001) (Figures 2E, 3E). Overall, these results demonstrate that the predictive capability of single biomarkers remains limited except for GMV, whereas combining multimodal biomarkers stepwise improves their accuracy in prediction of both relapse activity and disease progression within early multiple sclerosis.

MRI and blood biomarkers influence disease activity and progression

In order to create a prediction model analyzing complex relationships among multiple variables, we next applied SEM to assess the causal relationship of the most promising biomarker combinations determined in the ROC approach, namely MRI (LV + GMV) and blood (sNfL + sGFAP) biomarkers. In addition to the ROC analysis, SEM allows us to test a model for its compatibility with the data in its entirety simultaneously. In the predictive modeling approach, the RMSEA index for the models was below 0.03 and the AIC comparing the models varied between 0.006 and 0.019. The obtained fit indices in the SEM analysis implied a good fit of the constructed models to the observed data, providing robust relations between the variables. Within the SEM model quantifying the pathways, the input variables (GMV, sNfL, sGFAP and LV) predicted both ARR and EDSS progression. Our model with resultant standardized coefficients (s) identified that GMV (s = 0.58; p < 0.01) and sNfL (s = 0.63; p < 0.01) significantly predict ARR and EDSS progression through lesion volume as mediator (ARR [s = 0.59; p < 0.01] and EDSS [s = 0.73; p < 0.001]) (Figure 4). Taken together, LV mediates the path between GMV and sNfL on the one side, and ARR and EDSS progression on the other side.

Discussion

Here, we present a longitudinal study utilizing a classification model and a multivariate analysis technique to predict both disease activity and progression in patients with early MS based on multimodal biomarker combinations. In our discrimination model, the triple combination of MRI (LV and GMV), OCT (RNFL and GCIPL) and blood biomarkers (sNfL and sGFAP) achieved the best performance in predicting disability progression

10.3389/fimmu.2025.1532660

A ROC curve results

Biomarkers		AUC	SE	95% CI	*p-value*
Single biomarkers	GMV	0.614	0.054	0.517 – 0.705	***0.035***
	LV	0.572	0.057	0.475 – 0.666	*0.201*
	RNFL	0.502	0.057	0.406 – 0.599	*0.970*
	GCIPL	0.512	0.058	0.416 – 0.608	*0.815*
	sNFL	0.596	0.054	0.499 – 0.688	*0.075*
	sGFAP	0.540	0.064	0.432 – 0.645	*0.529*
Combined biomarkers (one modality)	MRI (GMV & LV)	0.621	0.055	0.534 – 0.712	***0.026***
	OCT (RNFL & GCIPL)	0.507	0.058	0.410 – 0.603	*0.906*
	BLOOD (sNFL & sGFAP)	0.632	0.059	0.525 – 0.731	***0.025***
Combined biomarkers (two modalities)	MRI & OCT	0.631	0.054	0.535 – 0.721	***0.014***
	MRI & BLOOD	0.699	0.055	0.594 – 0.790	***<0.001***
	BLOOD & OCT	0.646	0.059	0.539 – 0.744	***0.013***
Combined biomarkers (three modalities)	MRI & OCT & BLOOD	0.706	0.055	0.601 – 0.797	***<0.001***

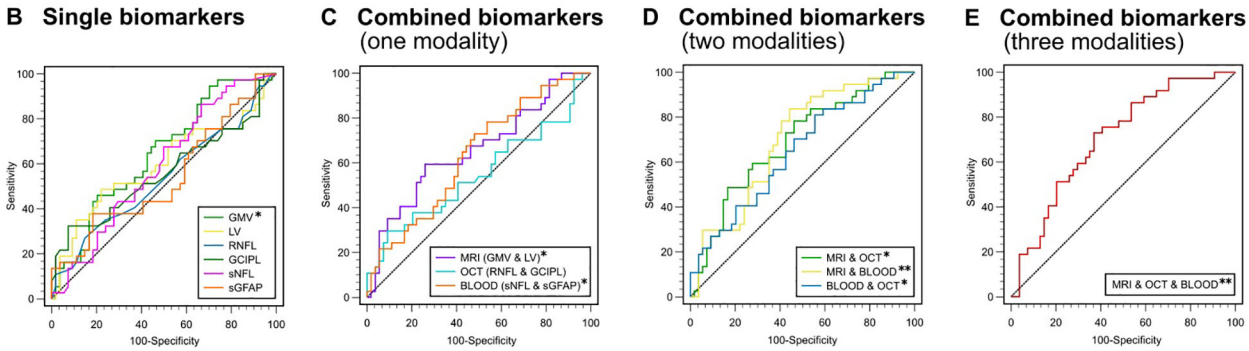


FIGURE 3
ROC analysis for the discrimination between the presence or absence of EDSS progression. **(A)** Color-coded table depicting the ROC analysis for individual and combinations of biomarkers. AUC, p-value and 95%-CI for the prediction of EDSS progression (yes/no). **(B)** ROC curves for single biomarkers. **(C)** ROC curves for combined biomarkers within one modality. **(D)** ROC curves for combined biomarkers within two modalities. **(E)** ROC curve for combined biomarkers of all three modalities (GMV + LV, RNFL + GCIPL and sNfL + sGFAP). AUC, area under the curve; CI, confidence interval; GCIPL, ganglion cell-inner plexiform layer; GMV, gray matter volume; OCT, optical coherence tomography; RNFL, retinal nerve fiber layer; ROC, receiver operator characteristics; SE, standard error; sGFAP, serum glial fibrillary acidic protein; sNfL, serum neurofilament light; LV, lesion volume.

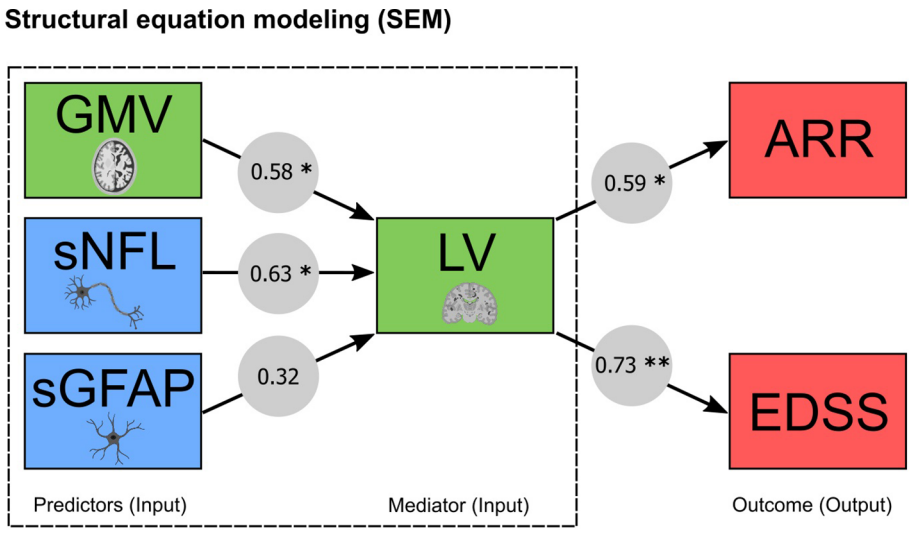


FIGURE 4
MRI and blood biomarkers and their capability to predict clinical outcomes through structural equation modeling (SEM). Predictive modeling of MRI (GMV and LV) and blood (sNfL and sGFAP) biomarkers. Arrows denote the relationship between the variables expressed as standardized coefficients, which are shown for each path (* significant at $p < 0.01$; ** significant at $p < 0.001$). ARR, annualized relapse rate; EDSS, expanded disability status scale; GMV, gray matter volume; sGFAP, serum glial fibrillary acidic protein; sNfL, serum neurofilament light; LV, lesion volume.

as well as disease activity within the upcoming four years. Our subsequently constructed SEM model established sNfL, GMV and LV as viable predictors of both disease activity and progression. Beyond that, the model further indicated that LV significantly mediates the effect of sNfL and GMV on future disease activity and progression over the study period. Thereby, our multi-biomarker approach highlights the importance of accounting for LV (neuroinflammation) when implementing cross-modal biomarkers in predicting clinical outcomes in MS.

Our findings align well with the current understanding of the pathophysiology in early, inflammation-driven MS, where disease activity (T2-hyperintense LV) drives ongoing neuroaxonal degeneration (sNfL and GMV) and clinical disability progression (27). Although each biomarker has been found to predict certain aspects of MS pathology individually (6, 17, 28–30), they all have their own individual strengths and weaknesses. In line with this, the predictive ability of each biomarker in our ROC analyses was limited when used on its own, but gained an incremental value when applied in combination with other biomarkers. Importantly, combining biomarkers from different modalities, such as MRI and blood biomarkers, resulted in a significant improvement in predicting both relapse activity and disease progression. This implies that certain biomarkers might be able to compensate for the limitations of others. For example, blood biomarkers have been found to be poor predictors of fatigue in MS (14, 31), while imaging of deep gray matter and brainstem structures have shown strong associations with measures of fatigue (25). Additionally, blood biomarkers provide a holistic view of cellular damage across the entire neuroaxis with high temporal resolution but lack of spatial resolution (5, 8), while conventional MRI markers provide great spatial resolution but are naturally "blind" for slightly injured tissue such as NAWM. Therefore, using both imaging and blood biomarkers can provide a more comprehensive understanding of disease progression in MS, as they can offer complementary information of different aspects of the disease process. Furthermore, the integration of potentially latent variables via observed variables in the characterization of cross-modal biomarkers may help to identify patients at risk of disease progression, and therefore aid therapeutic decision-making. Appropriate biomarkers may even been chosen according to a patient's individual symptoms and signs, which could allow for the creation of more personalized treatment plans. Accordingly, a recent study found predictors with mid- to high-accuracy for several disability outcomes in MS by combining clinical and imaging with omics information (32). This machine learning study particularly identified algorithms for predicting the escalation of therapy from first-line to high-efficacy treatment.

A plethora of different blood biomarker candidates has been evaluated in clinical and pre-clinical studies on neuroinflammation (33). However, sNfL and more recently sGFAP have shown the greatest prognostic potential in MS (14, 33), therefore, we preselected those biomarkers for our study. There are several surrogate markers of neurodegeneration in MR imaging, such as brain parenchyma fraction, total brain volume, and GMV (34). We decided to primarily include GMV in our analyses since it is widely used and has a strong association with neurodegeneration and cognitive impairment (29, 34). However, as models and algorithms become more complex and advanced, it makes sense to include more biomarkers in order to further improve predictive accuracies. In MS, OCT has been used to detect thinning of retinal layers; this loss of retinal nerve fibers may be indicative of underlying neurodegeneration (6). However, in our early MS cohort, inclusion of OCT did not show a remarkable additive effect in predicting disease progression or relapse rates. This may have several reasons: first, changes in the eyes of our early MS cohort may be subtle and not always be detectable with OCT. Furthermore, although OCT has a good resolution for damage to the visual system, namely the retina and the layers immediately beneath it, as well as the optical radiation, it may not provide sufficient information on neurodegeneration in other regions of the CNS, such as infratentorial structures (6, 11). Additionally, previous studies have shown RNFL to be a significantly variable measure, especially when considering non-optic neuritis eyes (35–37). In line with this, in our cohort, only 33 patients had a history of prior optic neuritis and in order to look at neurodegeneration in MS in general, we only included OCT results from eyes without prior optic neuritis in our analyses. This may have limited the predictive capability of our OCT results; however, both GCIPL and RNFL are well-established markers and have been associated with disease progression even when applied for non-optic neuritis eyes (38).

Our study also has some limitations: First, we investigated a real-world cohort. Hence, the time point for measuring all biomarkers showed some ranges. However, a real-world cohort has the advantage of resembling a more realistic clinical situation and may therefore suffer less from a selection bias (39). Second, longer follow-up observations are warranted. Third, total GM atrophy is related to disability in MS (29, 40), but also regional GM atrophy e.g. thalamic volume plays a key role for clinical progression (41). Finally, also changes within the NAWM are relevant for disease worsening in MS (42, 43). Hence, further studies are needed to incorporate more specific and advanced MRI-derived markers into such multimodal approaches.

Altogether, the combination of multimodal biomarkers (LV, GMV, RNFL, GCIPL, sNfL, sGFAP) that represent different parts of the disease pathology offer advantages in predicting upcoming disability accumulation in MS. In addition, predictive modeling specifically revealed that total lesion volume is a substantial mediator of the prognostic properties of gray matter and neurofilament on future progression indicating the significance of overall cerebral lesion load in fostering neuronal loss and subsequent disability. Validation and replication of multimodal biomarkers identified so far will be required for generating the evidence to be applied in personalized health care for people with MS.

Data availability statement

The datasets presented in this article are not readily available because restrictions apply to the availability of these data, which were used under license for the current study and are therefore not publicly available. The raw data used in preparation of the figures

and tables will be shared in anonymized format upon reasonable request by a qualified investigator for purposes of replicating procedures and results. Requests to access the datasets should be directed to vinzenz.fleischer@unimedizin-mainz.de.

Ethics statement

The study was approved by the local ethics committee (numbers: 2018-13622, 837.019.10); written informed consent was obtained from all patients. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

VF: Conceptualization, Formal analysis, Supervision, Writing – original draft. TB: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft. MM: Formal analysis, Methodology, Writing – review & editing. FS: Data curation, Formal analysis, Methodology, Writing – review & editing. MH: Data curation, Formal analysis, Writing – review & editing. MP: Data curation, Writing – review & editing. MS: Data curation, Writing – review & editing. GG: Data curation, Methodology, Writing – review & editing. SG: Formal analysis, Resources, Writing – review & editing. SB: Data curation, Resources, Writing – review & editing. FZ: Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the German Research Foundation (DFG; CRC TRR128 (project number: 213904703) to VF, MM, SG, SB, FZ and CRC TRR 355/1 (project number: 490846870) to SB) and the Hermann and Lilly Schilling foundation (SB). TB is supported by the Clinician Scientist Fellowship "TransMed Jumpstart Program: 2019_A72", which is supported by the Else Kröner Fresenius Foundation.

Acknowledgments

The authors thank Dr. Cheryl Ernest for proofreading the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Barro C, Benkert P, Disanto G, Tsagkas C, Amann M, Naegelin Y, et al. Serum neurofilament as a predictor of disease worsening and brain and spinal cord atrophy in multiple sclerosis. *Brain*. (2018) 141:2382–91. doi: 10.1093/brain/awy154

2. Sormani MP, Kappos L, Radue EW, Cohen J, Barkhof F, Sprenger T, et al. Defining brain volume cutoffs to identify clinically relevant atrophy in RRMS. *Mult Scler*. (2017) 23:656–64. doi: 10.1177/1352458516659550

3. Brex PA, Ciccarelli O, O'Riordan JI, Sailer M, Thompson AJ, Miller DH. A longitudinal study of abnormalities on MRI and disability from multiple sclerosis. *N Engl J Med*. (2002) 346:158–64. doi: 10.1056/NEJMoa011341

4. O'Riordan JI, Thompson AJ, Kingsley DP, MacManus DG, Kendall BE, Rudge P, et al. The prognostic value of brain MRI in clinically isolated syndromes of the CNS. A 10-year follow-up. *Brain*. (1998) 121:495–503. doi: 10.1093/brain/121.3.495

5. Bittner S, Oh J, Havrdova EK, Tintore M, Zipp F. The potential of serum neurofilament as biomarker for multiple sclerosis. *Brain*. (2021) 144:2954–63. doi: 10.1093/brain/awab241

6. Petzold A, Balcer LJ, Calabresi PA, Costello F, Frohman TC, Frohman EM, et al. Retinal layer segmentation in multiple sclerosis: a systematic review and meta-analysis. *Lancet Neurol*. (2017) 16:797–812. doi: 10.1016/S1474-4422(17)30278-8

7. Madsen MAJ, Wiggermann V, Bramow S, Christensen JR, Sellebjerg F, Siebner HR. Imaging cortical multiple sclerosis lesions with ultra-high field MRI. *NeuroImage Clin*. (2021) 32:102847. doi: 10.1016/j.nicl.2021.102847

8. Benkert P, Meier S, Schaedelin S, Manouchehrinia A, Yaldizli O, Maceski A, et al. Serum neurofilament light chain for individual prognostication of disease activity in people with multiple sclerosis: a retrospective modelling and validation study. *Lancet Neurol*. (2022) 21:246–57. doi: 10.1016/S1474-4422(22)00009-6

9. Akamine S, Marutani N, Kanayama D, Gotoh S, Maruyama R, Yanagida K, et al. Renal function is associated with blood neurofilament light chain level in older adults. *Sci Rep*. (2020) 10:20350. doi: 10.1038/s41598-020-76990-7

10. Yalachkov Y, Schafer JH, Jakob J, Friedauer L, Steffen F, Bittner S, et al. Effect of estimated blood volume and body mass index on GFAP and nfL levels in the serum and

CSF of patients with multiple sclerosis. *Neurol Neuroimmunol Neuroinflamm*. (2023) 10:e200045. doi: 10.1212/NXI.0000000000200045

11. Costello F, Van Stavern GP. Should optical coherence tomography be used to manage patients with multiple sclerosis? *J Neuroophthalmol*. (2012) 32:363–71. doi: 10.1097/WNO.0b013e318261f7e7

12. Scheltens P, Blennow K, Breteler MM, de Strooper B, Frisoni GB, Salloway S, et al. Alzheimer's disease. *Lancet*. (2016) 388:505–17. doi: 10.1016/S0140-6736(15)01124-1

13. Zhao A, Li Y, Yan Y, Qiu Y, Li B, Xu W, et al. Increased prediction value of biomarker combinations for the conversion of mild cognitive impairment to Alzheimer's dementia. *Transl Neurodegener*. (2020) 9:30. doi: 10.1186/s40035-020-00210-5

14. Brummer T, Muthuraman M, Steffen F, Uphaus T, Minch L, Person M, et al. Improved prediction of early cognitive impairment in multiple sclerosis combining blood and imaging biomarkers. *Brain Commun*. (2022) 4:fcac153. doi: 10.1093/braincomms/fcac153

15. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. (2018) 17:162–73. doi: 10.1016/S1474-4422(17)30470-2

16. Kalincik T, Cutter G, Spelman T, Jokubaitis V, Havrdova E, Horakova D, et al. Defining reliable disability outcomes in multiple sclerosis. *Brain: J Neurol*. (2015) 138:3287–98. doi: 10.1093/brain/awv258

17. Bittner S, Steffen F, Uphaus T, Muthuraman M, Fleischer V, Salmen A, et al. Clinical implications of serum neurofilament in newly diagnosed MS patients: A longitudinal multicentre cohort study. *EBioMedicine*. (2020) 56:102807. doi: 10.1016/j.ebiom.2020.102807

18. Schmidt P, Gaser C, Arsic M, Buck D, Forschler A, Berthele A, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *Neuroimage*. (2012) 59:3774–83. doi: 10.1016/j.neuroimage.2011.11.032

19. Seitz CB, Droby A, Zaubitzer L, Kramer J, Paradis M, Klotz L, et al. Discriminative power of intra-retinal layers in early multiple sclerosis using 3D OCT imaging. *J Neurol*. (2018) 265:2284–94. doi: 10.1007/s00415-018-8988-3

20. Seitz CB, Steffen F, Muthuraman M, Uphaus T, Kramer J, Meuth SG, et al. Serum neurofilament levels reflect outer retinal layer changes in multiple sclerosis. *Ther Adv Neurol Disord*. (2021) 14:17562864211003478. doi: 10.1177/17562864211003478

21. Cruz-Herranz A, Balk LJ, Oberwahrenbrock T, Saidha S, Martinez-Lapiscina EH, Lagreze WA, et al. The APOSTEL recommendations for reporting quantitative optical coherence tomography studies. *Neurology*. (2016) 86:2303–9. doi: 10.1212/WNL.0000000000002774

22. Schippling S, Balk LJ, Costello F, Albrecht P, Balcer L, Calabresi PA, et al. Quality control for retinal OCT in multiple sclerosis: validation of the OSCAR-IB criteria. *Mult Scler*. (2015) 21:163–70. doi: 10.1177/1352458514538110

23. Martinez-Lapiscina EH, Arnow S, Wilson JA, Saidha S, Preiningerova JL, Oberwahrenbrock T, et al. Retinal thickness measured with optical coherence tomography and risk of disability worsening in multiple sclerosis: a cohort study. *Lancet Neurol*. (2016) 15:574–84. doi: 10.1016/S1474-4422(16)00068-5

24. Cagol A, Fuertes NC, Stoessel M, Barakovic M, Schaedelin S, D'Souza M, et al. Optical coherence tomography reflects clinically relevant gray matter damage in patients with multiple sclerosis. *J Neurol*. (2023) 270:2139–48. doi: 10.1007/s00415-022-11535-8

25. Fleischer V, Ciolac D, Gonzalez-Escamilla G, Grothe M, Strauss S, Molina Galindo LS, et al. Subcortical volumes as early predictors of fatigue in multiple sclerosis. *Ann Neurol*. (2022) 91:192–202. doi: 10.1002/ana.26290

26. Kelley K, Lai K. Accuracy in parameter estimation for the root mean square error of approximation: sample size planning for narrow confidence intervals. *Multivariate Behav Res*. (2011) 46:1–32. doi: 10.1080/00273171.2011.543027

27. Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. *Lancet*. (2018) 391:1622–36. doi: 10.1016/S0140-6736(18)30481-1

28. Uphaus T, Steffen F, Muthuraman M, Ripfel N, Fleischer V, Groppa S, et al. NfL predicts relapse-free progression in a longitudinal multiple sclerosis cohort study. *EBioMedicine*. (2021) 72:103590. doi: 10.1016/j.ebiom.2021.103590

29. Fisher E, Lee JC, Nakamura K, Rudick RA. Gray matter atrophy in multiple sclerosis: a longitudinal study. *Ann Neurol*. (2008) 64:255–65. doi: 10.1002/ana.21436

30. Oship D, Jakimovski D, Bergsland N, Horakova D, Uher T, Vaneckova M, et al. Assessment of T2 lesion-based disease activity volume outcomes in predicting disease progression in multiple sclerosis over 10 years. *Mult Scler Relat Disord*. (2022) 67:104187. doi: 10.1016/j.msard.2022.104187

31. Aktas O, Renner A, Huss A, Filser M, Baetge S, Stute N, et al. Serum neurofilament light chain: No clear relation to cognition and neuropsychiatric symptoms in stable MS. *Neurol Neuroimmunol Neuroinflamm*. (2020) 7:e885. doi: 10.1212/NXI.0000000000000885

32. Andorra M, Freire A, Zubizarreta I, de Rosbo NK, Bos SD, Rinas M, et al. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. *J Neurol*. (2024) 271:1133–49. doi: 10.1007/s00415-023-12132-z

33. Yang J, Hamade M, Wu Q, Wang Q, Axtell R, Giri S, et al. Current and future biomarkers in multiple sclerosis. *Int J Mol Sci*. (2022) 23:5877. doi: 10.3390/ijms23115877

34. Sastre-Garriga J, Pareto D, Battaglini M, Rocca MA, Ciccarelli O, Enzinger C, et al. MAGNIMS consensus recommendations on the use of brain and spinal cord atrophy measures in clinical practice. *Nat Rev Neurol*. (2020) 16:171–82. doi: 10.1038/s41582-020-0314-x

35. Knier B, Berthele A, Buck D, Schmidt P, Zimmer C, Muhlau M, et al. Optical coherence tomography indicates disease activity prior to clinical onset of central nervous system demyelination. *Mult Scler*. (2016) 22:893–900. doi: 10.1177/1352458515604496

36. Albrecht P, Ringelstein M, Muller AK, Keser N, Dietlein T, Lappas A, et al. Degeneration of retinal layers in multiple sclerosis subtypes quantified by optical coherence tomography. *Mult Scler*. (2012) 18:1422–9. doi: 10.1177/1352458512439237

37. Chua J, Bostan M, Li C, Sim YC, Bujor I, Wong D, et al. A multi-regression approach to improve optical coherence tomography diagnostic accuracy in multiple sclerosis patients without previous optic neuritis. *NeuroImage Clin*. (2022) 34:103010. doi: 10.1016/j.nicl.2022.103010

38. Dreyer-Alster S, Gal A, Achiron A. Optical coherence tomography is associated with cognitive impairment in multiple sclerosis. *J Neuroophthalmol*. (2022) 42:e14–21. doi: 10.1097/WNO.0000000000001326

39. Chodankar D. Introduction to real-world evidence studies. *Perspect Clin Res*. (2021) 12:171–4. doi: 10.4103/picr.picr_62_21

40. Filippi M, Preziosa P, Copetti M, Riccitelli G, Horsfield MA, Martinelli V, et al. Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology*. (2013) 81:1759–67. doi: 10.1212/01.wnl.0000435551.90824.d0

41. Eshaghi A, Prados F, Brownlee WJ, Altmann DR, Tur C, Cardoso MJ, et al. Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Ann Neurol*. (2018) 83:210–22. doi: 10.1002/ana.25145

42. Fleischer V, Kolb R, Groppa S, Zipp F, Klose U, Groger A. Metabolic patterns in chronic multiple sclerosis lesions and normal-appearing white matter: intraindividual comparison by using 2D MR spectroscopic imaging. *Radiology*. (2016) 281:536–43. doi: 10.1148/radiol.2016151654

43. Llufriu S, Kornak J, Ratiney H, Oh J, Brenneman D, Cree BA, et al. Magnetic resonance spectroscopy markers of disease progression in multiple sclerosis. *JAMA Neurol*. (2014) 71:840–7. doi: 10.1001/jamaneurol.2014.895

frontiers | Frontiers in Immunology

TYPE Perspective
PUBLISHED 17 February 2025
DOI 10.3389/fimmu.2025.1487709

OPEN ACCESS

EDITED BY
Joachim Havla,
Ludwig Maximilian University of Munich, Germany

REVIEWED BY
Sun Xin,
Hebei North University, China
Fady Albashiti,
LMU Munich University Hospital, Germany

*CORRESPONDENCE
Paola Zaratin
paola.zaratin@aism.it

RECEIVED 28 August 2024
ACCEPTED 28 January 2025
PUBLISHED 17 February 2025

CITATION
Helme A, Kalra D, Brichetto G, Peryer G, Vermersch P, Weiland H, White A and Zaratin P (2025) Artificial intelligence and science of patient input: a perspective from people with multiple sclerosis. *Front. Immunol.* 16:1487709. doi: 10.3389/fimmu.2025.1487709

Artificial intelligence and science of patient input: a perspective from people with multiple sclerosis

Anne Helme[1], Dipak Kalra[2], Giampaolo Brichetto[3], Guy Peryer[4], Patrick Vermersch[5], Helga Weiland[6], Angela White[7] and Paola Zaratin[3]*

[1]Multiple Sclerosis International Federation, London, United Kingdom, [2]Dept. Medical Informatics & Statistics, The European Institute for Innovation through Health Data, Ghent University Hospital, Gent, Belgium, [3]Research Department, Italian Multiple Sclerosis Foundation, Genoa, Italy, [4]Multiple Sclerosis Society UK, London, United Kingdom, [5]Univ. Lille, Inserm U1172 LilNCog, Centre Hospitalier Universitaire de Lille (CHU) Lille, Fédératif Hospitalo-Universitaire (FHU) Precise, Lille, France, [6]Multiple Sclerosis South Africa, Western Cape, South Africa, [7]National Multiple Sclerosis Society, New York, NY, United States

Artificial intelligence (AI) can play a vital role in achieving a shift towards predictive, preventive, and personalized medicine, provided we are guided by the science with and of patient input. Patient-reported outcome measures (PROMs) represent a unique opportunity to capture experiential knowledge from people living with health conditions and make it scientifically relevant for all other stakeholders. Despite this, there is limited uptake of the use of standardized outcomes including PROMs within the research and healthcare system. This perspective article discusses the challenges of using PROMs at scale, with a focus on multiple sclerosis. AI approaches can enable learning health systems that improve the quality of care by examining the care health systems presently give, as well as accelerating research and innovation. However, we argue that it is crucial that advances in AI – whether relating to research, clinical practice or health systems policy – are not developed in isolation and implemented 'to' people, but in collaboration 'with' them. This implementation of science with patient input, which is at the heart of the Global PROs for Multiple Sclerosis (PROMS) Initiative, will ensure that we maximize the potential benefits of AI for people with MS, whilst avoiding unintended consequences.

KEYWORDS

artificial intelligence, patient reported outcomes, health outcomes, multiple sclerosis, ethics

1 Introduction

There is an increasing demand for a shift towards predictive, preventive, and personalized medicine (1, 2) and artificial intelligence (AI) can play a vital role in achieving this. Multiple sclerosis (MS), an autoimmune condition affecting nearly 3 million people across the world (3), is very heterogeneous, affecting people's lives in

different ways. A single treatment or care approach will not be suitable for every individual. The presentation and course of MS reflect myriad factors that can be difficult to capture in a comprehensive manner. So, whilst MS is not itself particularly rare, once people with MS (pwMS) are sub-divided into groups requiring different treatment and care services, and who have different priorities when it comes to health outcomes, everyone becomes part of a rare group. Determining the right approach to treatment and care needs to take into account all of the variability that exists within that person's life: their sex, age, environment, access to care, economic resources, comorbidities and many other factors. AI-based solutions may be necessary to support the capture and use of these complex data, so that health outcomes can be optimized for everyone.

2 Health outcomes that matter to people with MS

Health outcomes reflect information about the impact on people from health and care interventions. Leveraging patient experiential knowledge and make it scientifically measurable via Patient Generated Health Data (PGHD) is a critical part of the humanisation of health in line with Value-Based Health Care EU pillars (4–6). PGHD include patient reported outcome measures (PROMs), patient-reported experience measures (PREMs - people's perspectives of their experience while receiving care) or Patient Preferences and Acceptability for Innovative health interventions (PPI). Among these, PROMs provide a patient perspective on the impact that a disease (and its treatment) has on their physical, functional, and psychological status without interpretation from anyone else. There is no unique definition of PROs: "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else" in accordance to the Food and Drug Administration (FDA) (7) or "any outcome evaluated directly by the patient him/herself and based on patient's perception of a disease and its treatment(s)" in accordance to the European Medicines Agency (EMA) (8). The FDA definition of PROs designates both active and passive information as PROs, while the EMA definition seems to restrict PROs to active reports only. AI could help to incorporate PROMs reflecting different functional domains alongside other research and clinical data if relevant PROMs for the target population and adequate infrastructure for collecting PROs are available.

The Global Patient Reported Outcomes for MS (PROMS) Initiative launched on 12 September 2019 at the 35th Congress of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS). It is jointly led by the European Charcot Foundation (ECF) and the Multiple Sclerosis International Federation (MSIF) with the Italian MS Society acting as lead agency for and on behalf of the global MSIF movement (9, 10). The strategic intent of the PROMS Initiative is to engage people with MS in developing and prioritizing PROMs that give us a picture of their status today and changes over time. At present, clinical and care measurements are snapshots of individual functional domains and pwMS are frustrated that functional domains and corresponding interrelationships that matter most to them are not addressed by currently available PROMs (11). Within this framework, applying AI to PROMs can be a catalyst for a renewed humanism from research to care, but this vision will only be achieved by furthering the optimal engagement of pwMS (12).

3 The route to a unified view on PROMs for MS

Challenges with capturing, measuring and using PROs have been recently described by the PROMS Initiative (5) and are summarized below:

i. reaching consensus on relevant PROMs for specific and targeted populations (i.e. acknowledging there cannot be a 'one-size-fits-all' approach for PROMs), which have been validated and can be used within and across countries for accurate comparisons;
ii. developing practical and usable tools (e.g. apps, wearables, other devices) to enable the routine capture of multiple changing outcomes over time, which requires acceptability and therefore a user-friendly and useful solution for collecting the information (13, 14);
iii. translating subjective impressions from PRO questionnaires (such as Likert scales) into valid numerical data, and determining what threshold constitutes a meaningful change for different individuals (15);
iv. calibrating changes in outcomes over time against the types and costs of health and care interventions that have created those outcomes. This can help target health spending most effectively (i.e. assessing value), without leading to unintended consequences such as restriction of access to care, support, disability status or benefits.

Commonly used PROMs in the MS field include the MS Impact Scale-29 (16), Multiple Sclerosis Quality of Life-54 (17), Patient Determined Disease Steps (18), SymptoMScreen (19) among others. At the current time, PROMs are mainly used as a correlate with classical metrics (in the case of MS, such as the Expanded Disability Status Scale (EDSS), Timed 25-foot Walk (T25W) and others). PROMs are used as confirmation of these classical metrics, rather than adding their own specific and unique value.

As mentioned earlier, pwMS are frustrated that currently available measures do not capture the experiences that have the greatest impact on their daily lives. In addition, PROMs also need to be measured formally so they can be collected consistently and compared over time for the same person and between people (20). There are many initiatives and resources focused on the creation and standardization of health outcome measures, including PROMs, for example the International Consortium for Health Outcomes Measurement (ICHOM) (21), the Patient-Reported Outcomes Measurement and Information System (PROMIS) (22), and the Core Outcome Measures in Effectiveness Trials (COMET) initiative (23). PROMOPROMS is an initiative focused on PROMs that matter most to people with MS and the implementation of

these in clinical practice (24), and a recent global survey of pwMS identified the functional domains that have the greatest impact on their lives (25). Identifying distinct clusters of PwMS who share symptom patterns across functional domains and experiential knowledge, along with their interdependencies, will pave the way for a personalized application of PROMs from clinical trials to clinical practice and vice versa.

Despite this, there is limited uptake of the use of PROMs within the research and healthcare system. Without a significant body of evidence, health systems are poorly placed to learn, potentially ineffective interventions are sustained and health system budgets are wasted (26). The opportunity is also lost for PROMs to be used directly by people and their clinicians (27). The application of AI to PROMs data can support learning health systems, but a renewed humanism from research to care will only be achieved if researchers and the clinical community works effectively alongside people with MS.

The ALAMEDA project (28) made progress towards AI-enabled prediction, prevention and intervention. ALAMEDA is a Horizon 2020 EU-funded project aiming to make use of AI to reduce the costs of treating disorders such as MS, Parkinson's, and stroke, hence easing the burden on healthcare systems. In a pilot study carried out by the Italian MS Foundation (FISM), wearable technology and smartphone apps enabled the longitudinal collection of continuous digital-health data and electronic PRO data from pwMS across domains including mobility, sleep, mental and cognitive ability, emotional status and quality of life. This data supported the development and testing of AI algorithms with the aim of detecting and predicting relevant changes in disease progression.

In particular, the MS pilot focuses on key aspects such as the use of predictive systems to improve decision support systems for multiple sclerosis and the use of wearable technology (from sensors to electronic patient reported outcomes) in MS. The end goal of the MS pilot study was to test AI/machine-learning based algorithms that are able to predict the risk of developing a relapse in MS. Therefore, a characteristic research interest of the MS study is to explore the use of combined PRO and wearable-provided data as input for relapse prediction algorithms (29).

Crucial to the success of the ALAMEDA project is the use of MULTI-ACT guidelines (30) to engage relevant and representative stakeholders, including pwMS. Through co-design with pwMS, preferences and opinions about devices, frequency of measurement and potential barriers and facilitators for adhering to long-term patient-reported data collection were identified. In addition, pwMS were also involved in identifying and prioritizing suitable endpoints that might act as signs of a forthcoming relapse. All these factors helped shape the final protocol for the ALAMEDA MS pilot study (29).

4 The potential for AI to improve health outcomes for people with MS

The use of AI within healthcare systems is not yet standardized or routine, and more research is needed into its cost-effectiveness. It includes interventions used by healthcare professionals such as AI-assisted clinical decision support systems, as well as those used by individuals, such as chatbots that provide health information and smartphones with AI-related applications. Applying AI technology to the analysis and use of health data – particularly when it has been patient generated or patient-reported - has the potential to improve prognosis, prevent and treat disease progression and improve lives, through taking a personalized approach to diagnosis, treatment and care (31, 32).

The role of AI in healthcare spans all clinical conditions and is widely studied, for example in the oncology field recent studies have examined whether machine learning models include PRO data, and how AI could impact the doctor-patient relationship (33, 34). In the field of MS, an example of a decision support system in development is 'Clinical impact through AI-assisted MS care' (CLAIMS), an AI-driven clinical decision-support platform that aims to model expected disease trajectories depending on treatment regimen (35). A review by Inojosa et al. (36) explores the opportunities for using large language models as a form of AI in MS management.

Crucially, the involvement of AI in research and healthcare must be guided by the science *with* and *of* patient input. The power of science *with* patient input relies on an innovative framework used to engage patients (10, 30), while the science *of* patient input relies on patient-generated health data (PGHD). Among PGHD, PROMs represent a unique opportunity to capture experiential knowledge from people living with health conditions and make it scientifically relevant for all other stakeholders – the mission of the Global PROMS Initiative (10).

With the advent of the European Health Data Space (EHDS), all EU member states will be required to focus on the quality and interoperability of priority health data items (37). The EHDS will enable large, enriched datasets encompassing information from the whole of the EU. Where standardized PROMs are in use for certain health conditions, collected in a clinical setting and stored in people's medical records, these too will be available. The scale and complexity of data within the EHDS will necessitate the use of AI to interrogate these large datasets, combining clinical and PRO data to develop meaningful insights. AI will be instrumental in enabling greater use of PROMs in value-based healthcare decisions, such as those made by national health technology agencies, leading to improved delivery of healthcare across the region and better outcomes for individuals.

As set out in the framework by Rivera et al. (31), patient reported outcomes could be used as an input to an AI model, they could be an output predicted by the model, or an outcome in terms of the evaluation of the AI intervention. Within a healthcare setting, PRO measures may be used to monitor symptoms, monitor adherence to treatment, measure response to treatment, or determine when someone needs a clinical review. Using PROs in an AI or learning system could enable clinical decision making to incorporate the consideration of a person's wellbeing, beyond overall survival or delayed progression of disease.

An example of how combining PROMs and AI could provide benefits for pwMS is through using AI approaches to interrogate

individual-level data captured from multiple sources. PROs might be captured passively (e.g. via a smartphone enabled with technology such as a step-counter, accelerometer, altimeter etc) or input actively from a person recording their symptoms, feelings, use of medications and lifestyle factors such as diet and exercise. Added to this might be daily temperature or atmospheric pressure readings. PwMS report that fatigue is a huge challenge to daily living. Patterns uncovered by AI interrogation of complex patient-reported data over time could provide insights into which factors increase or decrease levels of fatigue. These factors could be environmental or aspects that can be influenced by the person through lifestyle changes or self-management. Importantly, if the AI model identifies consistent changes in data patterns over time, this might signal an underlying change in the condition, such as progression of MS, prompting referral to a healthcare professional.

5 Challenges with using AI in MS healthcare: perspective from people with MS

The increasing use of digital technology that deploys AI poses several challenges, including representativeness, data privacy, health equity and consent (38). When developing models or interventions involving AI and PRO data, an essential consideration is that the data used to develop and train AI systems needs to be representative of the population in which the AI approach will be implemented. If models are developed on a specific, limited population of people with a particular condition, there may be issues when applying them to people with different demographic backgrounds (39), which could lead to misdiagnosis or incorrect management. This is especially true for complex conditions such as MS, which can present very differently across individuals, especially when considered in the context of multimorbidity and on a global basis. In addition, a common symptom of MS is cognitive dysfunction. If a person is not able to provide PRO data that accurately reflects their condition, because the questionnaire is too complex for example, then the resulting dataset on which an AI model is trained may not reflect the real needs of the population.

Health interventions that involve AI will only make it successfully into the clinic if they are fully acceptable by people with health conditions and their clinicians and care providers. Trust and honest communication are crucial components of the interaction between a healthcare professional and a person with MS. Whilst there may be improvements to health outcomes from AI in terms of clinical decision making – and the latest AI technology developed by Google has even been shown to conduct sophisticated diagnostic conversations (40) - there could be a risk that overreliance on AI algorithms reduces a clinician's ability to relate to people they are caring for as individuals. People want to see that their healthcare professional is also drawing on their experience and intuition as part of the decision-making process. Artificial intelligence might complement the role of healthcare professionals, but should not replace them.

A study comparing responses to frequently asked questions showed that people with MS rated those written by ChatGPT as higher in empathy compared by those written by a neurologist (41). Yet some people will find it hard to trust decisions that are purely an output of an AI system and any errors caused by use of such technology will have a profound impact on the relationship between a person and their clinician. McCradden et al. (42) argue that where health settings use AI-based predictions, these should not be prioritized above patient experiential knowledge. To enhance trust, people should be made aware when AI or algorithms are being used in decision-making relating to their healthcare. There needs to be transparency in terms of the data and instruments upon which AI and its underlying algorithms are based as well as any unconscious biases that may be inherent in both programming and interpretation. To help overcome barriers to uptake of AI health technologies, clinical trials of the technology should be co-designed with people with lived experience, and use relevant PROMs as a trial endpoint (43).

MS is a condition present across the globe. AI should not just improve outcomes for people with MS in well-resourced settings, and it is clear that AI has the potential to both improve and decrease health equity (44–46). In terms of MS healthcare, remote monitoring and digital technology that deploys AI algorithms could help fulfil a need caused by a lack of specialist healthcare professionals in some settings. If AI can improve the accuracy and speed of diagnosis, allowing for earlier intervention and personalized care plans, this should reduce the variation in care experienced by pwMS, both within and between countries. Yet the benefits of AI-assisted technology may not be available to everyone. The accessibility and costs of the technology – including any supporting infrastructure, personnel or regulatory requirements needed to integrate AI systems into the current system - may provide a barrier for lower socioeconomic populations (47) or countries where MS is relatively rare. A lack of use of the technology in these settings can contribute to a negative feedback loop, whereby the continual refinement and updating of the AI algorithms are based on a limited population, becoming increasingly less representative of the diversity of people with MS across the world.

A critically important consideration relates to privacy and security of personal health data. Whether in a clinical or research setting, the use of AI is likely to involve the collection and analysing of sensitive information. Also, personal health data may have social, cultural, and religious implications in communities that are less familiar with or accepting of health conditions such as MS. It is essential that safeguards are in place for handling, storing and using this type of data securely. People must have a clear understanding of the purpose for which their data might be used and give consent for their data to be used in this way. A focus on consent is even more important for people who may be experiencing cognitive dysfunction. It is important to remember, too, that data generated by and collected with AI and/or algorithms may produce consequences outside of health systems, including decisions regarding pensions, disability payments, and other services. For people with MS who rely on access to treatment, therapy, and other forms of support, there is a constant concern about the potential

that this support could be restricted based on incorrect interpretation of personal data, whether by human or AI decision-making.

6 Discussion

How can we maximize the potential benefits of AI for people with MS, whilst avoiding unintended consequences? As mentioned earlier, this requires science with patient input, which is at the heart of the Global PROMS Initiative. Advances in AI – whether relating to research, clinical practice or health systems policy - should not be developed in isolation and implemented 'to' people, but in collaboration 'with' them. Underlying this, communication and transparency is key. Encouragingly, these considerations are reflected in the recent WHO guidance on the "Ethics and governance of artificial intelligence for health: guidance on large multi-modal models." (48)

Quality of life is defined differently for everyone with MS and cannot be viewed purely clinically. AI algorithms cannot replace the emotional and psychological understanding of an individual and their expectations in relation to their wellbeing. The clinical interaction should always be 'personal', and it is important to guard against anything that reduces people to data points. There is a need for future research to determine whether AI in complement with standard of care has a beneficial impact on outcomes such as disability and quality of life.

As a community of people with MS, we urge that the use of AI in patient care proceeds with caution as well as anticipation. For care to maximize quality of life, it must be holistic, encompassing emotional, psychological and social as well as physical aspects. Any benefits from AI must not come at the expense of damage to the relationship between clinicians and the people they care for, widening health inequity, or worsening health and social outcomes for people with MS.

Crucially, the Global PROMS Initiative will help ensure that people with MS are involved in the development of PROMs for MS from research through to global implementation. They will have space to raise ethical questions in relation to the growing use of AI as it applies to large, patient-reported datasets. They can prompt other members of this multi-stakeholder initiative to move away from thinking of people with MS as data points, and consider the impact of any recommendations on all aspects of the life of a person with MS. Only by working collaboratively in this way will we ensure that future advances in AI safeguard individuals and be acceptable to the whole community.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

AH: Writing – original draft, Writing – review & editing. DK: Conceptualization, Writing – original draft, Writing – review & editing. GB: Writing – original draft, Writing – review & editing. GP: Writing – review & editing, Writing – original draft. PV: Writing – review & editing. HW: Conceptualization, Writing – original draft, Writing – review & editing. AW: Writing – review & editing, Writing – original draft. PZ: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank all the members of the PROMS Initiative Engagement Coordination Team for their insight during many discussions on this and related topics.

Conflict of interest

AH is an employee of the MS International Federation, which receives income from a wide range of sources, including healthcare and other companies, individuals, member organizations, campaigns, foundations, and trusts. During the past 5 years, MSIF received funding from the following companies: Bristol Myers Squibb, Sanofi, Merck, Viatris formerly Mylan, Novartis, Biogen, and Roche-all of which is publicly disclosed. PV has received honorarium and contributions to meeting from Biogen, Sanofi-Genzyme, Novartis, Teva, Merck, Roche, Imcyse, AB Science, Janssen, Ad Scientiam and BMS, and research support from Novartis, Sanofi-Genzyme and Merck.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Beccia F, Hoxhaj I, Castagna C, Strohäker T, Cadeddu C, Ricciardi W, et al. An overview of Personalized Medicine landscape and policies in the European Union. *Eur J Public Health*. (2022) 32:844–51. doi: 10.1093/eurpub/ckac103

2. EU4Health programme 2021-2027 – a vision for a healthier European Union - European Commission (2024). Available online at: https://health.ec.europa.eu/funding/eu4health-programme-2021-2027-vision-healthier-european-union_en (Accessed February 27, 2024).

3. Number of people with MS | Atlas of MS. Available online at: https://www.atlasofms.org/map/united-kingdom/epidemiology/number-of-people-with-ms/ (Accessed September 25, 2024).

4. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project healthDesign. *JMIR Hum Factors*. (2016) 3:e26. doi: 10.2196/humanfactors.5919

5. Zaratin P, Samadzadeh S, Seferoğlu M, Ricigliano V, dos Santos Silva J, Tunc A, et al. The global patient-reported outcomes for multiple sclerosis initiative: bridging the gap between clinical research and care – updates at the 2023 plenary event. *Front Neurol*. (2024) 15:1407257/full. doi: 10.3389/fneur.2024.1407257/full

6. European Commission: Directorate-General for Health and Food Safety. *Defining value in 'value-based healthcare' – Report of the Expert Panel on Effective Ways of Investing in Health (EXPH)*. Luxembourg: Publications Office (2019).

7. FDA. *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims*. US: US Food and Drug Administration (2009).

8. European Medicines Agency. Draft reflection paper on the use of patient reported outcome (PRO) measures in oncology studies (2014). Available online at: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-use-patient-reported-outcome-pro-measures-oncology-studies_en.pdf (Accessed January 09, 2025).

9. proms-initiative.org. Patient reported outcome for multiple sclerosis. Available online at: https://proms-initiative.org/ (Accessed February 27, 2024).

10. Zaratin P, Vermersch P, Amato MP, Brichetto G, Coetzee T, Cutter G, et al. The agenda of the global patient reported outcomes for multiple sclerosis (PROMS) initiative: Progresses and open questions. *Mult Scler Relat Disord*. (2022) 61:103757. doi: 10.1016/j.msard.2022.103757

11. Bharadia T, Vandercappellen J, Chitnis T, Eelen P, Bauer B, Brichetto G, et al. Patient-reported outcome measures in MS: Do development processes and patient involvement support valid quantification of clinically important variables? *Mult Scler J - Exp Transl Clin*. (2022) 8:20552173221105642. doi: 10.1177/20552173221105642

12. McCradden MD, Kirsch RE. Patient wisdom should be incorporated into health AI to avoid algorithmic paternalism. *Nat Med*. (2023) 29:765–6. doi: 10.1038/s41591-023-02224-8

13. Alsulami S, Konstantinidis S, Wharrad H. Use of wearables among Multiple Sclerosis patients and healthcare Professionals: A scoping review. *Int J Med Inf*. (2024) 184:105376. doi: 10.1016/j.ijmedinf.2024.105376

14. Lavelle G, Norris M, Flemming J, Harper J, Bradley J, Johnston H, et al. Validity and acceptability of wearable devices for monitoring step-count and activity minutes among people with multiple sclerosis. *Front Rehabil Sci*. (2022) 2:737384/full. doi: 10.3389/fresc.2021.737384/full

15. Cella D, Nolla K, Peipert JD. The challenge of using patient reported outcome measures in clinical practice: how do we get there? *J Patient-Rep Outcomes*. (2024) 8:35. doi: 10.1186/s41687-024-00711-1

16. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain J Neurol*. (2001) 124:962–73. doi: 10.1093/brain/124.5.962

17. Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil*. (1995) 4:187–206. doi: 10.1007/BF02260859

18. Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. *BMC Neurol*. (2013) 13:37. doi: 10.1186/1471-2377-13-37

19. Green R, Kalina J, Ford R, Pandey K, Kister I. SymptoMScreen: A tool for rapid assessment of symptom severity in MS across multiple domains. *Appl Neuropsychol Adult*. (2017) 24:183–9. doi: 10.1080/23279095.2015.1125905

20. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. *N Engl J Med*. (2016) 374:504–6. doi: 10.1056/NEJMp1511701

21. International consortium for health outcomes measurement. Available online at: https://www.ichom.org (Accessed February 27, 2024).

22. Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, et al. The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health Qual Life Outcomes*. (2013) 11:210. doi: 10.1186/1477-7525-11-210

23. COMET initiative | Home. Available online at: https://comet-initiative.org/ (Accessed February 27, 2024).

24. Brichetto G, Zaratin P. Measuring outcomes that matter most to people with multiple sclerosis: the role of patient-reported outcomes. *Curr Opin Neurol*. (2020) 33:295–9. doi: 10.1097/WCO.0000000000000821

25. Brichetto G, Helme A, Ghirotto L, Iorio G, Belscott L, Peryer G. Frequency and impact of symptoms experienced by people with MS - Results from the global PROMS Initiative survey. ECTRIMS 2024 Late Breaking Poster. P891/4083. *Multiple Sclerosis J*. (2024) 30(3_suppl):1148–211. doi: 10.1177/13524585241269220

26. OECD. *Tackling Wasteful Spending on Health*. Paris: Organisation for Economic Co-operation and Development (2017). Available at: https://www.oecd-ilibrary.org/social-issues-migration-health/tackling-wasteful-spending-on-health_9789264266414-en (Accessed February 27, 2024).

27. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf*. (2014) 23:508–18. doi: 10.1136/bmjqs-2013-002524

28. alamedaproject.eu. Alameda project . Available online at: https://alamedaproject.eu/ (Accessed January 03, 2025).

29. Sorici A, Băjenaru L, Mocanu IG, Florea AM, Tsakanikas P, Ribigan AC, et al. Monitoring and predicting health status in neurological patients: the ALAMEDA data collection protocol. *Healthc Basel Switz*. (2023) 11:2656. doi: 10.3390/healthcare11192656

30. Zaratin P, Bertorello D, Guglielmino R, Devigili D, Brichetto G, Tageo V, et al. The MULTI-ACT model: the path forward for participatory and anticipatory governance in health research and care. *Health Res Policy Syst*. (2022) 20:22. doi: 10.1186/s12961-022-00825-2

31. Rivera SC, Liu X, Hughes SE, Dunster H, Manna E, Denniston AK, et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *Lancet Digit Health*. (2023) 5:e168–73. doi: 10.1016/S2589-7500(22)00252-7

32. Voigtlaender S, Pawelczyk J, Geiger M, Vaios EJ, Karschnia P, Cudkowicz M, et al. Artificial intelligence in neurology: opportunities, challenges, and policy implications. *J Neurol*. (2024) 271:2258–73. doi: 10.1007/s00415-024-12220-8

33. Krepper D, Cesari M, Hubel NJ, Zelger P, Sztankay MJ. Machine learning models including patient-reported outcome data in oncology: a systematic literature review and analysis of their reporting quality. *J Patient-Rep Outcomes*. (2024) 8:126. doi: 10.1186/s41687-024-00808-7

34. Heudel PE, Crochet H, Blay JY. Impact of artificial intelligence in transforming the doctor–cancer patient relationship. *ESMO Real World Data Digit Oncol*. (2024) 3:100026. doi: 10.1016/j.esmorw.2024.100026

35. Praet J, Anderhalten L, Comi G, Horakova D, Ziemssen T, Vermersch P, et al. A future of AI-driven personalized care for people with multiple sclerosis. *Front Immunol*. (2024) 15:1446748/full. doi: 10.3389/fimmu.2024.1446748/full

36. Inojosa H, Voigt I, Wenk J, Ferber D, Wiest I, Antweiler D, et al. Integrating large language models in care, research, and education in multiple sclerosis management. *Mult Scler J*. (2024) 30(11-12):1392–401. doi: 10.1177/13524585241277376

37. ISO. ISO 27269:2021. Available online at: https://www.iso.org/standard/79491.html (Accessed February 27, 2024).

38. Erickson CM, Wexler A, Largent EA. Digital biomarkers for neurodegenerative disease. *JAMA Neurol*. (2025) 82:5–6. doi: 10.1001/jamaneurol.2024.3533

39. Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *eBioMedicine*. (2021) 67:103358. doi: 10.1016/j.ebiom.2021.103358

40. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic AI. *arXiv*. (2024). doi: 10.48550/arXiv.2401.05654

41. Maida E, Moccia M, Palladino R, Borriello G, Affinito G, Clerico M, et al. ChatGPT vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. *J Neurol*. (2024) 271:4057–66. doi: 10.1007/s00415-024-12328-x

42. McCradden M, Hui K, Buchman DZ. Evidence, ethics and the promise of artificial intelligence in psychiatry. *J Med Ethics*. (2023) 49:573–9. doi: 10.1136/jme-2022-108447

43. Pearce FJ, Rivera SC, Liu X, Manna E, Denniston AK, Calvert MJ. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997–2022). *Lancet Digit Health*. (2023) 5:e160–7. doi: 10.1016/S2589-7500(22)00249-7

44. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SBM, Savoldelli M, et al. Artificial intelligence in health care: laying the Foundation for Responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Glob Health*. (2020) 16(1):52. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315549/.

45. Guzmán CAF. Global health in the age of AI: Safeguarding humanity through collaboration and action. *PloS Glob Public Health*. (2024) 4:e0002778. doi: 10.1371/journal.pgph.0002778

46. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health*. (2021) 3:e260–5. doi: 10.1016/S2589-7500(20)30317-4

47. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc JAMIA*. (2018) 25:1080–8. doi: 10.1093/jamia/ocy052

48. World Health Organization. *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models*. UK: World Health Organization (2024).

frontiers | Frontiers in Immunology

TYPE Review
PUBLISHED 28 February 2025
DOI 10.3389/fimmu.2025.1514813

OPEN ACCESS

EDITED BY
Marcello Moccia,
University of Naples Federico II, Italy

REVIEWED BY
Rocco Haase,
University Hospital Carl Gustav Carus, Germany
Isabel Voigt,
TUD Dresden University of Technology, Germany

*CORRESPONDENCE
Letizia Leocani
leocani.letizia@hsr.it

RECEIVED 21 October 2024
ACCEPTED 14 February 2025
PUBLISHED 28 February 2025

CITATION
Dini M, Comi G and Leocani L (2025) Digital remote monitoring of people with multiple sclerosis. *Front. Immunol.* 16:1514813. doi: 10.3389/fimmu.2025.1514813

Digital remote monitoring of people with multiple sclerosis

Michelangelo Dini[1,2], Giancarlo Comi[3] and Letizia Leocani[2,3,4]*

[1]Faculty of Psychology, Vita-Salute San Raffaele University, Milan, Italy, [2]Faculty of Medicine, Experimental Neurophysiology Unit, Institute of Experimental Neurology (INSPE), IRCCS-Scientific Institute San Raffaele, Milan, Italy, [3]Department of Neurorehabilitation Sciences, Casa di Cura Igea, Milan, Italy, [4]Faculty of Medicine, Vita-Salute San Raffaele University, Milan, Italy

Introduction: Multiple sclerosis (MS) is a chronic neurodegenerative disease that affects over 2.8 million people globally, leading to significant motor and non-motor symptoms. Effective disease monitoring is critical for improving patient outcomes but is often hindered by the limitations of infrequent clinical assessments. Digital remote monitoring tools leveraging big data and AI offer new opportunities to track symptoms in real time and detect disease progression.

Methods: This narrative review explores recent advancements in digital remote monitoring of motor and non-motor symptoms in MS. We conducted a PubMed search to collect original studies aimed at evaluating the use of AI and/or big data for digital remote monitoring of pwMS. We focus on tools and techniques applied to data from wearable sensors, smartphones, and other connected devices, as well as AI-based methods for the analysis of big data.

Results: Wearable sensors and machine learning algorithms show significant promise in monitoring motor symptoms, such as fall risk and gait disturbances. Many studies have demonstrated their reliability not only in clinical settings and for independent execution of motor assessments by patients, but also for passive monitoring during everyday life. Cognitive monitoring, although less developed, has seen progress with AI-driven tools that automate the scoring of neuropsychological tests and analyse passive keystroke dynamics. However, passive cognitive monitoring is still underdeveloped, compared to monitoring of motor symptoms. Some preliminary evidence suggests that application of AI and big data to other understudied aspects of MS (namely sleep and circadian autonomic patterns) may provide novel insights.

Conclusion: Advances in AI and big data offer exciting possibilities for improving disease management and patient outcomes in MS. Digital remote monitoring has the potential to revolutionize MS care by providing continuous, long-term granular data on both motor and non-motor symptoms. While promising results have been demonstrated, larger-scale studies and more robust validation are needed to fully integrate these tools into clinical practice and generalise their results to the wider MS population.

KEYWORDS

multiple sclerosis, big data, artificial intelligence, monitoring, review

Introduction

Multiple Sclerosis (MS) is a chronic, inflammatory, and neurodegenerative disease that affects the central nervous system (CNS). It is estimated that MS impacts ~2.8 million people globally, with a higher prevalence among women (1). MS can cause a wide range of symptoms, depending on the location of lesions across the CNS. Primarily, MS affects sensorimotor functioning, causing vision loss, sensory alterations, walking difficulties, muscle weakness, spasticity, and problems with coordination and balance (2). Additionally, cognitive impairment can be observed in 30-70% of pwMS (3).

The unpredictable nature of the disease, typically characterised by a relapsing-remitting course and by progressive accrual of disability, profoundly affects the quality of life (QoL) of people with MS (pwMS). Furthermore, recent evidence has shown that many pwMS can experience an insidious disease progression even in the absence of relapses (4). Thus, MS poses significant physical, emotional, and socio-economic burdens on individuals and their families (5). Accurate disease monitoring is crucial to put in place the best possible treatment plans and reduce the negative impact of the disease on patients' QoL. Due to organisational and economical limitations of healthcare systems, however, conventional clinical follow-up assessments are generally performed every 6-12 months, or at the time of a relapse. Thus, clinicians are often unable to detect subtle disease progression and/or to capture all relapses, since they need to rely on patients' recall and infrequent assessments.

The rising adoption of digital health technology in the last decade has sparked an interest in the development, study, and validation of new digital tools for the purpose of monitoring disease progression. Indeed, digital remote monitoring may have the potential to enable longitudinal monitoring of the disease course with a granularity that would otherwise be unobtainable with more costly and less accessible clinical follow-ups (6). A recent European survey found that the vast majority (78%) of patients use commercially-available digital tools (smartphone apps, wearables) to increase awareness of their health, and that 62% of healthcare providers believe that the data obtained from these tools impacts their communication with patients, their understanding of patients' health state, and their decision-making progress (7). Increasing the adoption of validated digital remote monitoring tools into everyday clinical practice would enable clinicians to access a much larger dataset of quantitative measures which could help them to better understand intra-individual disease trajectories and therefore improve the standard of care for pwMS. Digital remote monitoring can cover a wide range of domains (i.e., motor, cognitive and autonomic functions, psychological wellbeing, disease activity, sleep, diet, etc.), and can be carried out using both active and/or passive monitoring techniques. Active monitoring requires patients to consciously provide information, either via patient-reported questionnaires (e.g., asking patients to rate self-perceived fatigue on a scale 1-10), or by performing objective assessments (e.g., by performing a digitalised cognitive test on their smartphone). Passive monitoring leverages data from smart devices and sensors to enable remote monitoring while patients go about their daily life (e.g., daily steps data from accelerometers in a wearable device, or data from a blood glucose monitor placed on the arm). Active and passive methods can be paired to enhance the quality of digital remote monitoring data (e.g., collecting daily steps data from a participant's smartphone, which is also used to administered weekly standardized walking tests designed to be performed while carrying the smartphone in the pocket, to measure the distance walked and other data obtained from the smartphone accelerometer and gyroscope).

The definition of 'big data' keeps evolving, as continuing technological advancement and increasing adoption of devices able to capture more and more data push the boundaries of "big data". However, core properties like high volume (i.e., large quantities of data), velocity (i.e., data which are acquired in real-time) and variety (i.e., data which can be either structured or unstructured) are shared across most definitions (8). Other properties like exhaustivity (i.e., the ability to capture an entire system), high resolution (i.e., the ability to collect many datapoints at short intervals), relationality (i.e., the ability to merge different datasets), scalability (i.e., the ability to expand rapidly in size) have also been proposed (8). In general, data which cannot be easily viewed, processed and analysed using traditional statistical methods and which requires *ad-hoc* processing pipelines to produce meaningful insights could be labelled as big data. A consensus definition for big data in health research was proposed by the Health Directorate-General for Research and Innovation of the EU Commission, stating: "*Big Data in health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points*" (9).

In the context of digital remote monitoring of patients, big data can include structured and/or unstructured data from smart devices, wearables, self-monitoring devices, or electronic health records (EHRs) (10). Data from wearables or data recorded passively from smart devices can easily satisfy the "high volume" and "high velocity" criteria of big data. Indeed, using a single tri-axial accelerometer to monitor motor activity of a single individual over 10 hours, with a sampling frequency of 1 Hz, would yield over ~130,000 raw data points, which would need to be processed and aggregated using custom algorithms to derive basic interpretable metrics (e.g., steps/minute), and then further processed to derive more advanced metrics (e.g., time spent performing moderate vs. intense activity). Data from smart devices used to administer active tests is characterized by significantly lower volume and velocity but can become big data in the context of long-term monitoring, especially as digital remote monitoring allows to administer repeated assessments with higher frequency, longer follow-up times, and to larger cohorts, addressing the "scalability" property of big data. In the context of a simple digital cognitive test for which participants need to respond to 50 stimuli, a typical dataset would contain information on response times, actual responses, correctness of each response, metadata (e.g., date, time, type of device, location, device orientation, stimulus order), resulting in >200 datapoints for each testing session. These raw data would also

need to be processed and aggregated to derive informative metrics (e.g., mean reaction times). Monitoring 20 patients for 12 months through weekly testing with this simple test would result in the collection of ~50,000 datapoints, with longer and more complex assessments increasing the volume of data acquired exponentially. Data from EHRs typically reaches big data status only when large quantities of clinical data are collected for a large number of patients, either longitudinally in a single centre or cross-sectionally through multicentre collaborations. EHRs data also fits the "exhaustivity" property of big data, as they include a wide range of information for each patient (e.g., sociodemographic, clinical, imaging, pharmacological). Another way that EHRs data can fit the criteria for big data is linked to recent developments in Artificial Intelligence (AI) applied to processing and aggregating of unstructured text data, which could enable to start analysing large quantities of unstructured data present in EHRs (e.g., medical notes) in an automated (or semi-automated) quantitative way, thus greatly expanding the dimensionality of EHR datasets.

AI is a term dating back to the 1950s, when it was coined to represent machines exhibiting features akin to human intelligence (e.g., reasoning, learning, vision) (11). In recent years, this term has transitioned more and more from theory to practice, and many subdivisions of AI have been defined, according to their respective properties and use cases (12). Machine Learning (ML) refers broadly to the use of computational algorithms to learn data patterns to make predictions, and then compare the predictions with the actual outcomes, in order to learn iteratively, thus improving the quality of the predictions based on available data at each iteration. Deep Learning (DL) is an evolution of conventional ML, since it follows the same iterative learning approach to improve predictions. However, it differs from ML in that DL models are built from different consecutive hidden layers of 'neurons' (i.e., interconnected processing nodes) which are used to process raw inputs and can be adapted to perform optimally across different specific tasks (i.e., speech recognition, image processing, genomics) (13). One such example are Convolutional Neural Networks (CNN), i.e., DL algorithms built using specific types of connected layers to improve the neural network's ability to perform image recognition tasks, and have thus found large use in radiology, by allowing automated or semi-automated scoring of CT or MRI scans (14). The ever-increasing worldwide dissemination of computing technology means that more and more data is being collected every day, and the increased computational power available today has made it possible to deploy AI in an increasing number of applications (Figure 1).

The aim of this narrative review is to present and discuss recent advancements in the field of digital remote monitoring in MS, with a focus on AI tools and algorithms applied to the analysis of big data from sensors, wearables, smartphones, and other smart devices, as well as data from active digital assessments designed to be performed independently and remotely by patients. Specifically, we aim to discuss how leveraging big data and AI could allow to improve the standard of routine disease monitoring of pwMS across different settings and in different fields, how it could allow researchers to obtain novel insights into specific factors driving disease progression, and what future developments are needed to further advance the state of digital remote monitoring in the future.

Methods

For this narrative review, we focused our literature search on studies of digital remote monitoring of pwMS using AI and/or big data. This includes studies aimed at validating digital monitoring tools designed to enable active or passive digital remote monitoring of MS symptoms and disease progression. To this aim, we conducted a PubMed search for papers containing the following terms in the title and/or abstract: "multiple sclerosis[Title/Abstract]

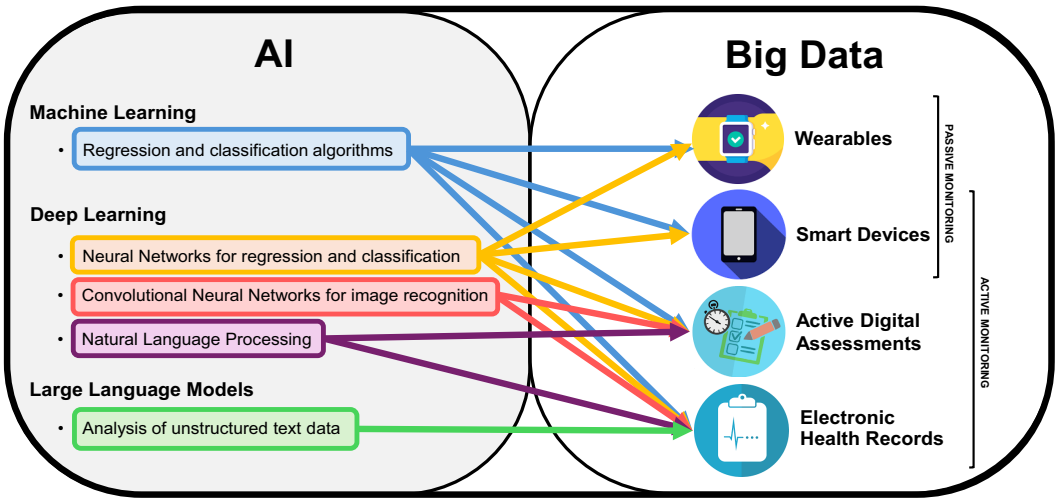


FIGURE 1
AI and Big Data for digital remote monitoring of MSFigure representing the two sets of Artificial Intelligence (AI) and big data, with specific subfields relevant for the field of digital remote monitoring of people with Multiple Sclerosis. The arrows indicate what type of AI-based analysis is best applicable to different types of big data obtainable from different methods of digital remote monitoring.

AND ((‘digital monitor*’[Title/Abstract] OR ‘remote monitor*’[Title/Abstract] OR wearable*[Title/Abstract]) OR (‘artificial intelligence’[Title/Abstract] OR ‘machine learning’[Title/Abstract] OR ‘deep learning’[Title/Abstract]))”. We filtered the search results to only select those published in the last 10 years, i.e., from 1st January 2014 to 1st August 2024.

We then excluded all reviews, meta-analyses, study protocols, opinion papers, editorials. We also excluded all studies where AI or big data where not specifically applied to data from digital remote monitoring or designed to enable it. Therefore, we excluded studies on AI-based processing and analysis of big data from structural (e.g., magnetic resonance) or functional (e.g., positron emission tomography) imaging, robotics-assisted physical rehabilitation, AI-assisted cognitive rehabilitation, AI-based psychological counselling, AI-based analysis of genomics, and those using AI and/or big data to estimate the risk of developing MS or to increase diagnostic accuracy. We also excluded studies of digital remote monitoring in which neither AI nor big data were applicable definitions (i.e., studies aimed at validating the administration of an established clinical test through videoconferencing or other telemedicine approaches, without collection of big data from sensors and/or other electronic devices).

The resulting candidate publications were screened manually by reading the abstracts, to select those who focused on developing and validating data processing and analysis pipelines (including AI) applied or applicable to digital remote data from sensors and/or active remote assessments, as well as those focusing on AI algorithms applied or applicable to the analysis of big data from other sources (e.g., EHRs) to improve the monitoring of disease progression in pwMS.

Results

Our literature search revealed that the majority of studies on digital remote monitoring of pwMS using AI and big data has focused on the use of wearable sensors to assess and monitor motor symptoms. This is not surprising, as motor deficits are one of the most prevalent and invalidating symptoms of MS (2). Therefore, our review begins by providing a report on studies focused on the motor domain, to evaluate the feasibility and validity of digital remote monitoring of motor functions in real-world clinical applications and highlight issues which still require further development. More recently, other studies have also focused on the need to monitor cognitive symptoms, since they are frequently reported as of the main factors which negatively impact the autonomy and QoL of pwMS (15). We present these studies and discuss the potential added benefits of digital remote monitoring of cognition using AI, compared to the current standard of care, as well as the potential to deploy “big data” to enable passive cognitive monitoring. The use of big data and AI for the digital remote monitoring of other symptoms or domains (e.g., sleep, autonomic functions) or to leverage unstructured big data from EHRs to monitor disease progression are still underrepresented in the MS literature. However, the few studies available to date suggest that their further exploration may yield novel insights which would otherwise be unobtainable by using conventional data acquisition, processing and analysis methods. Therefore, we conclude by presenting the studies available to date, to highlight the potential benefits of these different applications of big data and AI to enhance the remote monitoring of pwMS.

Motor domain

Many studies in the last 5-10 years have applied big data analysis and AI to the study of motor symptoms, aiming either to enable continuous passive monitoring, validate remote active motor tests to be used for frequent remote active monitoring, or leverage sensor data to detect digital biomarkers associated with higher odds of disease worsening or relapsing. The three main areas of interest appear to be falls (including both automatic fall detection using sensor data and identification of risk factors), gait (including both passive monitoring and active instrumented tests which can be performed remotely and independently by pwMS), and activity monitoring during everyday life as a digital biomarker of disability progression. The characteristics of all reviewed studies are summarised in Table 1.

Risk of falls

Falls are a major health concern for pwMS, as over 50% of them are estimated to experience at least one fall in a 6-month period, of which half result in injury (16). Continuous remote monitoring of pwMS in real-life environments and automatic falls detection has the potential to increase the detection rate of falls in everyday life, allowing a more precise monitoring of clinical progression. Moreover, it could help identify specific risk factors and consequently develop prevention strategies.

Tulipani et al. (17) investigated the ability to predict fall risk in 37 pwMS wearing a chest and a thigh sensor during sit-stand transitions of daily life and during a standardised sit-stand task in the clinic. Using reported falls in the previous 6 months to dychotomize participants in “fallers” or “non fallers”, they evaluated the ability of sensor data to correctly classify patients in either class. Sit-to-stand transitions in daily life were detected using a DL (long Short-Term Memory) algorithm tuned to detect activity states, which allowed them to select only sensor data from periods of transition from the “sitting” to the “standing” state. Using Receiver Operating Characteristics (ROC) analysis, the best predictor of high fall risk in their study was a chest acceleration metric recorded during execution of the sit-stand task in the clinic (Area Under the Curve [AUC]= 0.89). The best performing sensor metric during the real-life task execution, i.e., average sit-stand time, had slightly lower predictive power (AUC= 0.81). Their results suggest that conventional sensor metrics (e.g., acceleration, total time of execution) may provide useful insights into the fall risk of pwMS, although with reduced accuracy, compared to instrumented functional assessments performed in the clinic. The same research group recently published a longitudinal study (18), with the aim of extending the analysis of sit-stand performance to longitudinal

TABLE 1 Summary of studies on motor domain.

Study	Year	Sample	Study type	Sensor array	Algorithms used	Aim	Monitored activities
Özdemir et al. (23)	2014	14 HCs	Cross-sectional	6 accelerometers - 1 on head - 1 on chest - 1 on waist - 1 on right wrist - 1 on right thigh - 1 on right ankle	5 ML algorithms to distinguish between normal activity and falls (kNN, LSM, SVM, BDM, DTW) + 1 DL (ANN)	Automated falls detection	Simulated falls in a controlled setting
Casilari et al. (22)	2015	4 HCs	Cross-sectional	1 smartphone with embedded inertial sensors (accelerometer + gyroscope) 1 smartwatch with embedded inertial sensors (accelerometer + gyroscope)	Mix of custom and published threshold-based algorithms for automated fall detection	Automated falls detection	Simulated falls in a controlled setting
Chitnis et al. (27)	2019	23 pwMS	Longitudinal (three visits over 24 weeks, 8 weeks of remote monitoring)	3 multi-sensing devices (acceleration, motion, heart rate, skin impedance, body temperature, light exposure, air pressure) - 1 on chest (day only) - 1 on right wrist (day and night) - 1 on right ankle (day and night)	DL for automated detection of activity type and quantification of time spent during each activity phase	Validation of remote gait analysis	Instrumented structured assessments in a controlled setting (baseline, week 16, week 24) Passive real-life remote monitoring (8 weeks)
Bourke et al. (30)	2020	76 pwMS 25 HCs	Longitudinal (24 weeks of remote instrumented testing)	1 smartphone with embedded inertial sensors (accelerometer + gyroscope)	Custom algorithms for automated extraction of gait parameters	Validation of remote gait analysis	Instrumented structured assessments performed remotely and autonomously (1/week for 24 weeks)
Atrsaei et al. (28)	2021	35 pwMS	Cross-sectional	3 inertial sensors (accelerometer + gyroscope) - 1 on lower back - 2 on feet (right and left, used only as validation reference)	Custom threshold-based algorithm based on gait speed from multi-sensor data	Validation of remote gait analysis	Instrumented structured assessments in a controlled setting Instrumented structured assessments performed remotely and autonomously (50% of participants)
					ML (naïve Bayes classifier)	Automated walking bouts detection	Passive real-life remote monitoring for at least 6 hours (50% of participants)
Delahaye et al. (33)	2021	18 pwMS	Cross-sectional	1 wearable GPS receiver, placed on the right shoulder	Custom processing and aggregation pipeline for GPS and altitude data; threshold-based algorithm for walking bout detection	Validation of remote gait analysis	Instrumented structured tests in a controlled setting
Meyer et al. (19)	2021	37 pwMS	Retrospective	2 accelerometers - 1 below clavicle - 1 on right thigh 5 inertial sensors (accelerometer + gyroscope)	Fully automated algorithm for data processing; DL for automatic activity detection (Bidirectional Long ShortTerm Memory)	Fall risk estimation	One-minute walking trial at home

(Continued)

10.3389/fimmu.2025.1514813

TABLE 1 Continued

Study	Year	Sample	Study type	Sensor array	Algorithms used	Aim	Monitored activities
				- 1 on lower sternum - 1 on lower back - 1 on belt line - 2 on shanks (right and left)			
Mosquera-Lopez et al. (20)	2021	25 pwMS	Longitudinal (8 weeks of continuous monitoring)	Wireless time-of-flight home beacons, paired with a wearable smart tag (worn either on the trunk or in the pocket) with embedded accelerometer	Fully automated algorithm for data processing; DL for automatic falls detection (neural network auto-encoder + hyper-ensemble of RFs)	Automated falls detection	Passive real-life remote monitoring
Block et al. (35)	2022	94 pwMS	Longitudinal (12 months of continuous monitoring)	Commercial smart band (Fitbit Flex), including inertial sensors and sleep tracking capabilities	ML (3-compartment GMM)	Automated detection of activity states and behaviour patterns	Passive real-life remote monitoring
Creagh et al. (37)	2022	52 pwMS 24 HCs	Longitudinal (24 weeks of remote instrumented testing)	1 smartphone with embedded inertial sensors (accelerometer + gyroscope) 1 smartwatch with embedded inertial sensors (accelerometer + gyroscope)	CNN applied to raw accelerometer data	Disease severity estimation	Instrumented structured assessments performed remotely and autonomously (1/week for 24 weeks)
Salomon et al. (36)	2022	132 pwMS 90 HCs	Longitudinal (1 week of continuous monitoring)	1 wearable accelerometer, placed on the lower back	Custom algorithms for automated detection of activity fragmentation, circadian and fractal patterns	Automated detection of activity states and behaviour patterns	Passive real-life remote monitoring
Sun et al. (31)	2022	337 pwMS	Longitudinal (10 months of continuous monitoring)	1 commercial smart band (Fitbit Charge 2 or Fitbit Charge 3), including inertial sensors, heart rate and sleep tracking capabilities	Fully automated algorithm for data processing; ML regression (RF, GBT, EN) to predict 6MWT performance	Validation of remote gait analysis	Passive real-life remote monitoring
Tulipani et al. (17)	2022	37pwMS	Cross-sectional	2 accelerometers - 1 on chest - 1 on thigh	Fully automated algorithm for data processing; DL for automatic activity detection (Long Short-Term Memory)	Fall risk estimation	Sit-to-stand transitions in everyday life Standardised sit-to-stand task in the lab
Granja Domínguez et al. (32)	2023	205 pwMS	Cross-sectional	2 insoles with inertial sensors (accelerometer + gyroscope) and pressure sensors	Custom algorithm for automated calculation of gait parameters	Validation of remote gait analysis	Instrumented structured tests in a controlled setting
Kushner et al. (21)	2023	25 pwMS	Longitudinal (8 weeks of continuous monitoring)	Wireless time-of-flight home beacons, paired with a wearable smart tag (worn either on the trunk or in the pocket) with embedded accelerometer	DL for automatic fall detection (Long Short-Term Memory); ML for room detection (kNN)	Automated falls detection	Passive real-life remote monitoring

(Continued)

10.3389/fimmu.2025.1514813

TABLE 1 Continued

Study	Year	Sample	Study type	Sensor array	Algorithms used	Aim	Monitored activities
Salis et al. (26)	2023	128 participants, both HCs and people with mobility issues (20 pwMS)	Cross-sectional	3 inertial sensors (accelerometer + gyroscope) - 2 on feet (right and left) - 1 on lower back 2 time-of-flight infrared sensors placed on ankles (right and left) 2 pressure insoles	Custom threshold-based algorithms for automated walking bout detection	Automated walking bouts detection	Passive real-life remote monitoring (2.5 hours)
					Custom algorithm for automated gait analysis based on integration of multi-sensor data	Validation of remote gait analysis	Structured tests in a controlled setting Simulated activities of daily living in a controlled setting
Stavropoulos et al. (34)	2023	2 pwMS	Proof-of-concept	Commercial smart band (Fitbit Charge 3), including inertial sensors, heart rate and sleep tracking capabilities	Knowledge graphs	Automated detection of activity states and behaviour patterns	Passive real-life remote monitoring
Vandyk et al. (18)	2023	23 pwMS	Longitudinal (6 weeks of continuous monitoring)	3 accelerometers and surface biopotential readers - 1 on left upper chest - 2 on thighs (right and left)	Fully automated algorithm for data processing; DL for automatic activity detection (Long Short-Term Memory)	Fall risk estimation	Passive real-life remote monitoring Sit-to-stand transitions in everyday life (detected automatically)
Kirk et al. (29)	2024	97 participants, both HCs and people with mobility issues (13 pwMS)	Cross-sectional	1 experimental inertial sensor (accelerometer + gyroscope) placed on the lower back 3 reference inertial sensors (accelerometer + gyroscope) - 2 on feet (right and left) - 1 on lower back 2 time-of-flight infrared sensors placed on ankles (right and left) 2 pressure insoles	Mix of custom and ML-based algorithms for automatic gait detection and calculation of gait speed	Validation of remote gait analysis	Instrumented structured tests in a controlled setting Simulated activities of daily living in a controlled setting Passive real-life remote monitoring (2.5 hours)

Articles are listed based on year of publication (in ascending order). 6MWT, 6-Minutes Walking Test; ANN, Artificial Neural Network; BDM, Bayesian Decision making; CNN, Convolutional Neural Network; DL, Deep Learning; DTW, Dynamic Time Warping; EN, Elastic Net; GBT, Gradient Boosted Trees; HCs, Healthy Controls; kNN, k-Nearest Neighbours; LSM, Least Squares Method; ML, Machine Learning; pwMS, people with Multiple Sclerosis; RF, Random Forest; SVM, Support Vector Machine.

10.3389/fimmu.2025.1514813

remote monitoring. They recruited 23 pwMS and monitored them for six weeks, using three wearable sensors worn for all hours of the day (one on the left upper chest, two on the thighs) to record acceleration and surface biopotentials. Furthermore, they applied DL analysis to detect periods of sit-standing transitions. The algorithm identified different fatigue and instability phenotypes which were predictive of fall risk. They also observed that stability tended to decline over the course of the day, providing interesting quantitative insights into daily fluctuations of motor performance. Taken together, these results suggest that DL algorithms may enable to reliably identify activity states remotely and during everyday life, thus allowing to contextualise motor features obtained by the analysis of big data collected continuously from sensors. This is particularly interesting, since novel insights could be obtained by investigating some motor features (e.g., stability) during specific activity states of interest (e.g., sit-to-stand transitions), rather than across the entire range of daily activity states, which would be unfeasible if activity states had to be observed by an examiner or reported by the patient.

DL algorithms were also implemented retrospectively, to detect patients who had a positive recent history of falls (in the previous six months), by leveraging accelerometer data from sensors placed on the sternum, lower back, thigh, and shanks during a one-minute walking task in the clinic (19). This study found that a bidirectional long short-term memory neural network could be used to automatically identify and analyse sensor data from 1-minute walking tests performed remotely and autonomously by pwMS, and identified pwMS who had previously fallen with high accuracy (ROC AUC= 0.88). Notably, this DL algorithm trained on raw sensor data significantly outperformed the classification accuracy of neurologist-administered measures and patient-reported outcome measures, as well as conventional statistical analyses and other traditional ML models (logistic regression, k-nearest neighbours, support vector machine, decision tree) based on conventional aggregate spatiotemporal gait parameters (e.g., average speed). This suggests that AI can leverage big data to capture nonlinear relationships and motor phenotypes associated with an increased risk of falls which are not detected through conventional clinical exams or basic aggregate statistics.

Another key application of big data is the automatic detection of real-world falls in freely moving patients through sensors from wearables and/or smartphones. Mosquera-Lopez et al. (20) developed an algorithm which detects possible falls by combining acceleration and movement features recorded by wearable sensors connected to wireless beacons placed throughout the home. As fall detection was performed in a fully unsupervised way, accuracy of the detection pipeline was tested using 10-fold cross-validation (CV). This system proved highly accurate in detecting falls (sensitivity= 92%, specificity= 98%), producing 0.65 false alarms per day, which translates to roughly 2-3 false alarms per week. However, due to the small sample size and relatively short monitoring time, their dataset was highly imbalanced, with only 270 seconds of fall data compared to over 2,000,000 seconds of total data. In a more recent study (21), the same researchers conducted a secondary analysis of the same dataset, to investigate the relationship between mobility measures (including both movement metrics and location data) and risk of falls in pwMS. They found that half of falls occurred while walking, and that participants were sedentary for most of the time spent at home (>95%). Interestingly, they were able to observe that almost one third (28%) of falls occurred within one second of gait initiation, thus providing quantitative data to highlight the critical role of gait initiation in determining fall risk during everyday life. These results are promising, although the feasibility of this tracking method is obviously lower than that of monitoring devices which do not require altering/adapting the home environment of patients, which could hinder its applicability for real-life long-term monitoring of pwMS. Moreover, such systems cannot be used to assess motor performance in everyday life settings other than patients' homes (e.g., the workplace), limiting the generalizability of their findings. Further studies with much larger samples and longer monitoring durations are required to assess the true feasibility of this monitoring approach, as well as its validity and reliability for real-life clinical applications.

Increasing the range of possible applications of digital remote monitoring is key, to enable monitoring of motor functioning in a more ecological way, which would also allow extend this possibility to a wider range of pwMS. Therefore, more and more studies have tried to leverage commercially available smart devices for remote data collection, as their widespread availability could greatly extend the reach of digital remote monitoring, compared to more experimental and multi-device approaches. A pilot study (22) investigated the ability of a commercially available smartphone and smartwatch to automatically detect falls in an experimental environment, in which healthy controls (HCs) performed a set of simulated falls. Using an experimental setting in which participants performed simulated falls, they were able to directly observe the number of false positives and false negatives produced by the fall detection algorithm, from which they calculated sensitivity and specificity. They found that the joint use of smartphone and smartwatch improved the specificity of all analysed algorithms by a range of 5-15%, compared to smartphone- or smartwatch-only detection, although the issue of false positives alarms remained, as denoted by several false alarms raised during 24h of continuous monitoring. Moreover, the extremely small sample size (N = 4) significantly limits the generalisability of their results. Another study (23) investigated automatic fall detection through a system of tri-axial sensors fitted to six different body parts (head, chest, waist, right wrist, right thigh, right ankle) of HCs performing a standardised set of voluntary falls in an experimental setting. Through ML analyses they were able to reach values >99% for accuracy, sensitivity, and specificity. However, it must be stressed that this result was again observed in a small sample of HCs, performing standardised falls in a controlled setting. Perhaps even more importantly, such a complex sensor array would likely be unfeasible for everyday real-life monitoring of pwMS. It should be noted that studies wishing to evaluate automatic falls detection accuracy through direct observation (i.e., through simulated falls experimental paradigms) are inherently limited, since having pwMS or people with other chronic health conditions performing

simulated falls would pose evident ethical and safety issues. Crucially, this questions the ecological validity of fall detection algorithms validated on young healthy participants. Further research is needed to determine the feasibility and validity of automated fall detection through smartphone and/or wearable data in pwMS and in real-life scenarios.

Gait analysis

Gait disturbances are common in MS, they can present in the early disease stages, and significantly affect QoL by reducing autonomy and impacting negatively on socio-economic status (24). Instrumented assessments of gait are well documented (25) but, until recently, have largely relied on sophisticated lab-based assessments which are costly, cumbersome, and can fail to capture the true walking performance of pwMS in real-life environments. Consequently, most research to date has focused on validating wearable data recorded during laboratory experiments in which participants perform a mix of structured tests and simulated real-life activities. Only recently, researchers have begun leveraging big data gathered from wearables during everyday life to estimate gait parameters of pwMS, or to validate such monitoring devices with a mixed study procedure including both lab-based and remote-based data collection.

Salis et al. (26) validated a multi-sensor system designed to allow real-world monitoring (three inertial sensors, two plantar pressure insoles, and two distance sensors) in 128 participants with different pathologies (including 20 pwMS) who performed a mix of structured tests (e.g., Timed-Up and Go) and simulated activities (e.g., setting the table for dinner). They compared data from the wearable sensors with data from a stereophotogrammetry system, which served as reference. They used intraclass correlation coefficients (ICC) to assess reliability, which can be considered excellent when ICC > 0.90, good when 0.75 < ICC < 0.90, moderate when 0-5 < ICC < 0.75, and poor when ICC < 0.50. The reliability of the wearable system was excellent for structured tests, with ICC values >0.95, while it decreased slightly for simulated activities of everyday life (ICCs between 0.69-0.98). They also evaluated the feasibility of this wearable system for real-life use by recording 2.5 hours of unsupervised activity and reported that the system was well accepted, without major technical or usability issues. However, it must be noted that the real-world part of this study included only 20 healthy young adults and lasted a short time. Further real-life feasibility and acceptability studies with much longer monitoring periods are therefore definitely needed to derive any meaningful conclusions on real world long-term feasibility.

Chitnis et al. (27) collected data remotely from 23 pwMS wearing three sensors (placed on wrist, ankle, and sternum) for eight weeks during real-world daily activities. They designed a workflow for the classification of unstructured raw sensor data, using a DL classifier to distinguish activity periods (i.e., idle, walking, running). Then, they selected only the activity segments classified as "walking" to derive mobility features. Several features extracted from real-world walking bouts (i.e., stance time, swing time, mobility activity time, turning velocity) correlated with gold-standard clinical scales like the Expanded Disability Status Scale (EDSS) and the Multiple Sclerosis Functional Composite (MSFC) and standardised walking tests performed in the clinic (Timed 25-Foot Walk [T25FW]).

While multiple wearable sensors undoubtedly afford a higher degree of precision and provide more data to extract spatiotemporal gait parameters, compared to a single wearable sensor, one must also consider the feasibility of such approaches for longitudinal remote monitoring. Indeed, using multiple sensors imposes higher costs and is more burdensome for patients and researchers alike. This issue grows exponentially with longer follow-up times, limiting the ability to study long-term trends and patterns of motor function in pwMS. More specifically, compared to wearing sensors on multiple body parts, using a single sensor facilitates monitoring in a wider variety of daily life situations (e.g., in public), enhancing the ecological validity of data thus collected. Therefore, some researchers have begun to evaluate the validity of data obtained from a single wearable sensor, which could prove more economical and easier to use, therefore allowing larger studies with longer follow-ups.

Atrsaei et al. (28) developed and validated a ML-based gait estimation approach to predict gait speed and detect waling bouts using a single sensor on the lower back. They recruited 35 pwMS, who performed walking tests in the clinic and at home. and found that reference values obtained from sensors on both feet correlated strongly with gait speed estimated from the sensor on the lower back during a walking test in the clinic (r = 0.96) and at home (r = 0.95); gait speed during daily activities at home were also strongly correlated with reference values recorded in the clinic (r = 0.89). These results show that not only using a single sensor on the back approximates reference values extremely well for walking tests performed in the clinic, but is also able to provide accurate estimation based on real everyday activities. They also tested a ML-based algorithm (naïve Bayes classifier) for automated walking bouts detection and used leave-one-out CV to evaluate its accuracy, using only digital remote data collected during unsupervised daily life activities. The ML-based walking bout detection had high accuracy (96.4%) in detecting walking bouts remotely, during everyday life. Although the authors reported analysing >300 hours of daily activity measurements, the small sample size significantly limits the generalizability of these promising results obtained using a single sensor.

A similar approach (single sensor worn on the lower back; in the clinic and during 2.5 hours of real-world activities) was adopted by a European multicentric study (MOBILISE-D) on N=97 participants with different medical conditions (including 13 pwMS) (29). Reliability was considered good-to-excellent in the clinic (ICC range= 0.79-0.91) and moderate-to-good (ICC range= 0.57-0.88) in real-world activities, compared to a multisensory reference system which included pressure insoles. Although the reliability of the system was lower in the real-world scenario, it was still deemed to remain within a usable range. Predictably, this study found that walking bout duration affected the accuracy of gait speed estimation, with shorter bouts yielding less accurate estimates. It is therefore possible that further studies with more data at the intra-individual level may yield higher accuracy.

Aiming to further explore the use of devices which could be accessible to larger proportions of the population, Bourke et al. (30)

analysed gait parameters recorded by a waist-worn smartphone with built-in accelerometer during a two-minute walking test performed remotely and independently by 76 pwMS and 25 HCs over 24 weeks. The test-retest reliability across consecutive pairs of testing sessions was either excellent or good-to-excellent for 58/92 gait parameters in pwMS, and 29/92 in HCs, indicating higher variability in healthy persons across consecutive test sessions. These results suggest that remote sensor data recorded during active walking tests, using only a waist-worn smartphone, has comparable reliability to sensor data from clinical assessments. This encourages further research, as it could enable a much wider diffusion of instrumented remote walking assessments tanks to the ever-increasing availability of smartphones and wearables, thus expanding the reach of gait monitoring to those with reduced access to clinical services. However, this study involved mainly people with relapsing-remitting MS (RRMS), and only data from 62 participants (51 pwMS, 11 HCs) was used for the analyses (the authors did not explicitly state the reason for excluding almost 40% of the initial sample size). Therefore, further studies with larger sample sizes and more rigorous reporting are needed to establish the feasibility and validity of using smartphone-based sensor data as an endpoint in clinical trials and for real-life clinical monitoring.

All the studies examined so far have been conducted on small samples, and their results cannot therefore be generalised to the wider population of pwMS. The large volume of data obtained through wearable sensors and the high costs associated with specialised sensors has greatly limited the ability of researchers to conduct studies on large samples and with adequately long follow-ups, as evidenced by the studies discussed so far. Multicentric studies on larger samples of pwMS, however, are needed to derive more reliable insights on the validity, reliability, and feasibility of digital remote monitoring tools. As part of the RADAR-CNS initiative, Sun et al. (31), monitored an European cohort (from Italy, Spain, and Denmark) of 337 pwMS over an average duration of 10 months using a commercial wearable (Fitbit). They analysed real-world steps data and applied correlation-based feature selection to select the most relevant features and tested the ability of different ML regression algorithms (random forest, gradient boosted trees, and elastic net) to estimate 6 Minutes Walking Test (6MWT) performance in the clinic by using digital remote monitoring data collected during everyday life. Results show that minute-level features were more predictive than day-level features. Interestingly, they also noted that upper bound statistics (e.g., 90^{th} percentile of minute-level step count) were more strongly related to clinical test scores, indicating that the average performance in clinical gait tests may reflect the upper portions of the distribution of real-life gait abilities. This insight is particularly valuable, as it could mean that the impression of motor functioning that a clinician gets from a patient performing a walking test in the clinic may be an overestimation of their actual day-to-day average motor performance. The accuracy of 6MWT score estimation was quite low, reinforcing the idea that walking performance of pwMS could differ significantly between real life and clinical testing. These findings demonstrate that, in addition to allowing digital remote monitoring, leveraging data from wearables collected during everyday life can provide insights that would not be obtainable through conventional study paradigms, thus improving our understanding of the true validity of gold-standard and widely used clinical tests.

Another study with a large sample size (32) (N = 205 pwMS) focused on validating gait parameters (velocity, ambulation time, cadence, stride length) estimated trough sensor data from connected insoles with pressure and motion sensors, compared to a classic lab-based reference system based on pressure plates. They showed strong concordance between the two systems for gait velocity (ICCs > 0.83), ambulation time (ICC = 0.93), and cadence (ICCs > 0.90), whereas stride length showed poor concordance (ICC = 0.30). Sensorised foot insoles allow continuous data collection in everyday life without requiring visible devices, which could cause stigma or discomfort to some patients. Therefore, this large study provides valuable data on the validity of this gait monitoring device, which may prove particularly useful for patients which are unwilling and/or unable to wear visible devices such as smartwatches or body-mounted sensors. However, one key limitation is the compatibility of insoles with different shoe types, and the need to switch the insoles when changing shoes and when recharging, which could prove burdensome for patients in the long term, and could lead to missing data for extended periods of time or in some specific settings (e.g., while wearing slippers at home).

Whereas most of the literature to date has focused on obtaining gait parameters from accelerometers, Delahaye et al. (33) investigated gait parameters derived from a wearable sensor with integrated Global Positioning System (GPS). Validating GPS-derived walking speed and distance metrics may potentially enable to implement remote monitoring via commercially available and non-wearable devices (e.g., smartphones), thus removing the need for specially designed wearable sensors which may be perceived as cumbersome or that patients may be embarrassed to wear in public. The authors recruited a small convenience sample (N = 18) of pwMS who performed the 6MWT and an outdoor walking session at usual pace (up to 60 minutes). By integrating GPS and altitude data, they were able to measure gait parameters and associate them with variations in the terrain conformation, which could not only allow to better understand variability in motor activity observed through digital remote monitoring, but may also be used to standardize future studies on outdoor walking performance across different centres and countries, They found that walking speed during an outdoor walking session was significantly correlated with 6MWT performance measured in the clinic, whereas maximum walked distance was not. They also noted that 40% of participants did not reach their maximum walking distance during the first walking bout, but on subsequent ones. This suggests that the first stint of a walking task (as is the case for clinical walking tests) may not necessarily yield the best performance. Once again, one can appreciate how real-world motor data collected remotely and digitally was able to provide novel insights which enhance our understanding of the validity of testing procedures performed routinely in clinical or research settings. However, only 12

participants had valid GPS data, which means that GPS data could not be analysed for one third of participants. Therefore, more studies are needed to validate GPS-derived measures, and several technical limitations must be addressed, such as the accuracy of GPS-calculated walked distance for shorter walking bouts, or its accuracy in different environments and settings.

Activity monitoring

Data from wearable sensors may be used to characterise patients not only in terms of their raw quantitative performance metrics (e.g. daily step count), but to infer activity states and behavioural patterns which may be associated with clinical features and/or impact disease progression. This may be done either using knowledge-based frameworks or with a data-driven approach, providing both researchers and clinicians with more readily interpretable outcome measures. Moreover, characterising activity states may enhance the informative value of raw quantitative measures (e.g., by differentiating between steps counted during a light walk or during an intense run).

An example of the knowledge-based approach has been proposed by Stavropoulos et al. (34), who showcased a framework using *a priori* semantic rules to model "problem labels" which could be quickly and easily understood by clinicians and provide added value to raw quantitative data. As an example, "Steps < 500 & Heart Rate < 100 for duration > 800" was a rule used to determine an instance of "Lack of Movement". They then reported the example of a patient for which "Lack of Movement" instances appeared sporadically in the first months of remote monitoring and intensified in time, ultimately occurring almost every day in the last months. This provides a simple and effective way for clinicians to monitor potential risk factors and/or indices of disease worsening without necessarily having to analyse raw data, which may be cumbersome or outright unfeasible depending on the resources of different healthcare centres. However, frameworks based on *a priori* rules strongly rely on the goodness of their assumptions, and the validity of their output must be carefully assessed with *ad-hoc* studies implementing baseline and follow-up clinical assessments to provide quantitative measures of disease progression.

Block et al. (35) adopted a data-driven approach to characterize walking activity, based on minute-to-minute steps data from 94 pwMS who wore a Fitbit continuously for 1 year. They applied an unsupervised ML clustering algorithm (3-compartment Gaussian Mixture Model) to detect the proportion of three levels of activity (low, moderate, high) based on individual participants' steps data, and then evaluated associations with clinical parameters (walking tests, EDSS scores) and patient-reported outcomes. The detected activity levels correlated more strongly with clinical and patient-reported outcomes, compared to raw step count, and the combination of raw steps data and activity levels outperformed both individual metrics. This suggests that the qualitative aspect of steps data plays a pivotal role in predicting key clinical outcomes such as EDSS score. While we can expect patients with lower disability to be more active overall, leveraging AI algorithms to continuously and automatically evaluate the proportion of time spent in low- or high-intensity walking may enable to differentiate two patients which would appear identical if one were to look only at basic aggregate statistics like step count. Indeed, 1000 steps could be performed while doing house chores over 1 hour, or during a short but intense 5-minute walk, two different activities which cannot be accurately distinguished by examining step count alone.

Salomon et al. (36) collected data from 132 pwMS and 90 HCs wearing an accelerometer placed on the lower back for seven days, aiming to uncover daily-living rest-activity fragmentation patterns, circadian rhythms, and fractal regulation parameters. Results showed that pwMS had a more fragmented activity behaviour (likely indicating a greater need for pauses when carrying out prolonged physical activity) and lower amplitude in circadian changes of daily activity (i.e., the difference in activity levels between the five most and least active hours of the day) than HCs. Moreover, both circadian and fragmentation measures were associated with disability severity, as measured by EDSS score. Although a simple general metric like total physical activity remained the strongest discriminator between pwMS and HCs, this study found that incorporating more sophisticated metrics like fragmentation patterns and circadian rhythms detection improved the ability to differentiate between patients and HCs, and between patients with low vs. high disability. This was a cross-sectional study, and therefore could not provide any info on the predictive value of these features on disability progression or relapse risk. However, it is possible that circadian rhythms and fragmentation patterns could also provide novel insights on disease progression (e.g., a patient maintaining the same overall level of activity, but with increased fragmentation due to requiring more frequent rest). Further studies are needed to establish the utility of more advanced activity measured for real life monitoring of pwMS, with specific emphasis on their ability to predict relapse and/or disease progression.

Creagh et al. (37) also adopted a data-driven approach, analysing raw sensor data (smartphone + smartwatch) of 97 participants (24 HCs, 52 pwMS with mild disease severity, 21 pwMS with moderate disease severity) who performed a daily two-minutes walking test remotely for 24 weeks. Raw sensor data were analysed with a deep CNN pre-trained on an open-source human activity recognition dataset, to calculate a continuous quantitative measure of disease severity at each timepoint. Average disease severity across all timepoints correlated strongly with EDSS score. More interestingly, longitudinal disease severity trends were found to be associated with self-reported relapses. These preliminary results suggest that a continuous quantitative measure of disease severity may be more sensitive to change than the EDSS, and that it could also allow to detect trend changes in quasi-real time, which could potentially enable researchers and clinicians to detect relapses and shifts to progressive MS more efficiently. However, significant limitations such as adherence to frequent active testing and reliability of remote tests must be addressed, before such measures can be effectively implemented in everyday clinical practice. Indeed, the authors report that adherence was highly variable across participants, as participants with mild MS showed higher adherence than those with moderate MS and HCs. Moreover, adherence decreased linearly for all subgroups at later timepoints and, in some cases, in concomitance with the onset of reported relapses, as patients stopped performing the walking tests once they began

experiencing a significant worsening of motor function happening. These preliminary findings suggest the need to evaluate adherence to digital remote monitoring via active testing not only as a function of time, but also by uncovering potential associations with sociodemographic data (e.g., economic status, age), clinical features (e.g., cognitive impairment, depressive symptoms), or disease progression (e.g., patients becoming wheelchair-bound).

Cognitive domain

The use of AI and big data for monitoring cognitive function in pwMS has seen significantly less development, compared to the monitoring of motor function. This is likely because evaluating cognitive processes relies much more explicitly on active testing, and it is therefore more laborious to obtain large amounts of data. Indeed, a wearable sensor can detect thousands of datapoints for many motor features passively, just by being worn during everyday activities. The same approach cannot be easily applied to cognitive processes like memory or information processing speed, which are latent variables which need to be evaluated through specifically designed tasks. This significantly limits the ability of researchers to deploy big data to study cognition in MS. Nevertheless, some recent efforts have been made to integrate AI and big data in this field, and their results point to some interesting avenues for future research. The characteristics of all reviewed studies are summarised in Table 2.

Active monitoring

Most efforts have been focused on developing digital versions of established neuropsychological tests, with the aim of enabling automated administration and scoring, thus enabling remote administration and freeing up time for clinicians. In such cases, AI can provide novel ways to automate test administration and scoring, whereas big data has been mainly viewed in the context of granular digital test metrics which would be unfeasible to record manually, but which could enhance the information obtained from the execution of a test, compared to conventional scores.

Birchmeier et al. (38) aimed to digitize the Brief Visuospatial Memory Test – Revised (BVMT-R), a visuospatial learning test which is considered one of the gold-standard cognitive tests in MS (39). Scoring this test is a time-consuming semi-quantitative procedure which requires trained healthcare professionals to evaluate the shape and position of 18 drawings, assigning a score ranging 0-2 to each drawing, and then calculating the final total test score. The authors tested the ability of a CNN to automatically score patients' drawings, and compared its accuracy to clinician ratings, using a validation sample of 135 patients (for a total of 624 drawings). The CNN achieved a good accuracy for perfect or completely wrong drawings (i.e., those scored either 0 or 2 by human raters), while the accuracy for partially wrong drawings (i.e., those scored as 1 by human raters) was unsatisfactory (57%). This suggests that CNNs may not yet substitute clinicians and enable fully automated scoring, especially for drawings which present only slight inaccuracies and are therefore trickier to score, as they require

TABLE 2 Summary of studies on cognitive domain.

Study	Year	Sample	Study type	Cognitive domain	Algorithms used	Aim	Type of monitoring
Birchmeier et al. (38)	2019	135 pwMS	Cross-sectional	Visuospatial learning	CNN for image classification task	Validation of automated test scoring	Active testing
Birchmeier et al. (40)	2020	294 pwMS	Cross-sectional	Visuospatial learning	CNN for image classification task	Validation of automated test scoring	Active testing
Petilli et al. (41)	2021	35 HCs	Cross-sectional	Visuo-constructional ability and visuospatial memory	Custom algorithm for image preprocessing, segmentation and scoring of spatial, procedural and kinematic features	Enhancing the informative value of conventional tests	Active testing
Khaligh-Razavi et al. (42)	2020	91 pwMS 83 HCs	Cross-sectional	Information processing speed	ML multinomial logistic regression	Validation of digital test for autonomous and remote use	Active testing
Lam et al. (45)	2021	102 pwMS 24 HCs	Cross-sectional	–	Custom algorithm for processing and feature extraction from single-keystroke level datapoints	Validation of keystroke dynamic for monitoring of cognition	Passive monitoring
Lam et al. (46)	2022	102 pwMS	Longitudinal (12 months of continuous monitoring and clinical follow-ups every 3 months)	–	Clustering and PCA of features extracted from keystroke data; LMM to evaluate associations with cognitive outcomes	Validation of keystroke dynamic for monitoring of cognition	Passive monitoring

Articles are listed based on year of publication (in ascending order). CNN, Convolutional neural Network; HCs, Healthy Controls; LMM, Linear Mixed Models; ML, Machine Learning; PCA, Principal Component Analysis; pwMS, people with Multiple Sclerosis.

higher-level decision making than what AI can provide as of today. However, AI-based predictions may be implemented to provide preliminary recommendations, thus enabling faster scoring by human raters and reducing organisational burdens. In a subsequent study (40) with a larger validation sample size (1525 drawings), the authors observed that automated ratings matched with 72% of ratings from one neuropsychologist, and with 79% of ratings from another neuropsychologist. Interestingly, when comparing the ratings given by the two neuropsychologists, they observed an agreement in 82% of cases, highlighting the inherent unreliability of such semi-quantitative scoring protocols. This highlights the need to carefully consider the outcome metrics of AI validation studies, especially for semi-quantitative ratings, not only for cognitive tests, but also for other applications (e.g., MRI lesions counting). Indeed, aiming for 100% accuracy, especially while using a small number of human raters as reference may not be the ideal method. In such cases, reaching 100% accuracy could either be impossible, or lead to overfitting (i.e., training the AI algorithm to become an essential copy of that particular group of raters, which lead to poor generalizability and reliability). Conversely, an AI-based support-decision system may allow to increase inter-rater reliability, as AI-based criteria should hypothetically be more consistent that human raters, although *ad-hoc* studies are needed to support this hypothesis.

Another study focused on automated scoring of visuospatial tests (41), with the aim of providing more varied and detailed performance metrics, compared to the conventional scoring procedure, which only yields a single score indicating overall accuracy. They developed a tablet-based version of the Rey Complex Figure copy task, a visuo-constructive and visuospatial memory task which relies on semi-quantitative scoring, similarly to what has been described above. They administered it to 35 HCs and extracted performance indices capturing three different aspects of drawing abilities (spatial, procedural, and kinematic), for which a composite score was also calculated. They showed that automated scoring via CNNs could provide a much richer performance profile, by aggregating large quantities of data which could not be feasibly recorded manually by clinicians administering a test in a clinical setting (e.g., pressure strength, velocity, procedural drawing timeline). This may be very useful for research purposes and may ultimately lead to better classifications of cognitive profiles in MS (i.e., by disentangling the effect of motor, procedural, and visuospatial deficits). Therefore, the potential benefit of automated scoring may not be limited only to reducing test administration and scoring times. Indeed, automated AI-based scoring based on constructional and/or procedural drawing features recorded digitally may ultimately yield higher consistency than current scoring methods based on semi-quantitative ratings made by humans. However, such procedures require a high degree of standardisation; in this study, all participants used the same hardware, and drawings had to be manually screened before AI-based scoring.

Khaligh-Razavi et al. (42) developed a custom computerized image classification task to assess processing speed, and validated it in a sample of 91 pwMS and 83 HCs. The novelty of their approach consists in the embedding of AI (in the form of a ML multinomial logistic regressor) in the testing pipeline, so that their test does not yield a quantitative score, but rather a multi-level prediction on the cognitive status of the examinee, along with its associated predicted probability. This approach aims to predict cognitive status by automatically integrating a multi-dimensional feature set comprised of basic test scores (e.g., classification accuracy),more sophisticated metrics (e.g., intra-trial accuracy over time), and demographic data (e.g., age and education) to produce predictions on cognitive status on a test-by-test basis. By comparing the predictions made by the ML algorithm with cognitive impairment labels based on published cutoff values for gold-standard neuropsychological tests administered in the clinic, they demonstrated excellent discriminant validity for cognitive impairment in MS (AUC = 0.95, sensitivity = 82.9%, specificity = 96.1%). This approach to cognitive testing merits further research, as it may present many significant advantages. For clinical practice, it could reduce time allotted to test administration and scoring, as the test procedure is automated and seamlessly provides a prediction on cognitive status, thus enabling clinicians to dedicate more time to interact with patients and caregivers. For research purposes, an integrated AI data analysis pipeline allows to automatically leverage a larger amount of test performance metrics to derive more detailed insights into the cognitive profile of pwMS. Finally, automated ML-based scoring can leverage consecutively acquired data to continuously upgrade its predictions, likely making it ever more accurate as time progresses and more data is acquired, without the need for repeated validation studies which can be costly and time consuming.

Passive monitoring

Passive monitoring of cognitive functions represents an exciting frontier, as it could potentially enable granular long-term monitoring through big data analysis, without the need for patients to allocate time and energy to actively performing cognitive tests. This could increase the feasibility of continuous monitoring over the years, something which is very hard to achieve through active monitoring, where attrition naturally increases as time progresses (43, 44). However, there is still little evidence on what methods could enable valid and reliable passive monitoring of cognitive functioning.

Lam et al. (45) developed a keyboard app for smartphones, which allows to passively track timing-related keystroke features (e.g., latency between successive key presses, hold time, flight time) and correction-based features (e.g., correction duration, pre-correction slowing). They recruited 102 pwMS and 24 HCs, who were monitored passively as they used the keyboard app for 14 days. Results showed weak-to-moderate correlations with clinical disability, cognitive functioning, and upper limbs dexterity, as measured by the gold-standard clinical tests. Moreover, they observed that most timing-related features were significantly different between HCs and pwMS. In a follow-up longitudinal study (46), they monitored 102 pwMS for 12 months, using the keyboard app for passive monitoring and via clinical follow-ups every three months with clinical tests for upper limb dexterity and cognition. To evaluate associations between passive monitoring features and clinical features, they aggregated keystroke data into a cognition score cluster and a fine motor score cluster. They found

that the cognition score cluster was significantly associated with cognitive functioning at the group level, but not at the individual level, whereas the fine motor score cluster was significantly associated with upper limb dexterity at both the group and individual level.

In conclusion, the evidence available so far indicates that keystroke dynamics may be used to passively monitor longitudinal upper limb dexterity changes at the intra-individual level, whereas the same cannot be yet said for cognitive changes, suggesting that practice effects of repeated testing may have been a confounding factor. Moreover, the concurrent validity of keystroke dynamics is significantly lower than that of digitalized active cognitive tests (47). This is to be expected, as everyday activities such as typing leverage various sensory, motor, and cognitive processes and are not typically performed as rigorously and precisely as cognitive tasks, therefore introducing more noise. Thus, further research is needed, before keystroke dynamics can be considered an effective and reliable passive monitoring tool for cognition in MS. However, the potential to obtain data on cognitive functioning without requiring conscious effort by patients remains an enticing prospect, since it would allow to eliminate the aforementioned issue of loss to follow-up common to active longitudinal testing, and could provide novel, undiscovered insights on the cognitive functioning of pwMS by truly leveraging big data. One key aspect that should be addressed in the future regards the ethics of collecting keystroke data, as it could theoretically allow to uncover patients' sensitive information (passwords, bank details) and warrants a stronger enforcing of data privacy policies.

Other applications

AI and big data can play a significant role in enhancing monitoring capabilities in aspects of MS care/research other than motor and cognitive functioning. These range from passive monitoring of sleep and heart rate variability to the analysis of big data from real-world clinical records. We have grouped these different topics in a single encompassing section, given the small number of publications available thus far, to discuss their potential contribution towards further advancing the standard of care for pwMS, as well as their limitations.

Woelfle et al. (48) recruited 31 pwMS and 31 HCs, with the aim of studying whether remote monitoring of heart rate and sleep parameters could complement step count data in explaining MS severity. Participants wore a commercially available smartwatch (Fitbit Versa 2) for six weeks, during which parameters were extracted for sleep (e.g., sleep efficiency, light/deep/REM sleep duration), heart rate, and activity(e.g., proportion of sedentary/lightly active/fairly active/very active). While activity measures were predictably those most strongly correlated with clinical scales of disability and gait tests, median heart rate and deep sleep proportion also showed moderate correlations. Moreover, incorporating sleep and heart rate measures increased the ability to predict disability (measured by EDSS score), compared to using either baseline sociodemographic data and/or smartwatch-derived motor parameters. This pilot study with a small sample size suggests that sleep and heart rate data may indeed complement activity measures in explaining disease severity. These results are encouraging, especially for the promised ability to track objective sleep parameters remotely and through minimally invasive and economical devices, as compared to portable EEGs or polysomnography performed in the lab, greatly enhancing the feasibility of longitudinal studies of sleep. However, the small sample size warrants further larger studies, to increase the generalizability of results, especially since smartwatch data was lost for 7/62 participants due to synchronization issues, highlighting the need for more reliable data storage and synchronization technologies before such tools can be deemed reliable for larger clinical trials.

Hilty et al. (49) used a previously validated and CE-certified wearable for heart rate detection, with the aim of studying the autonomic nervous system in 56 pwMS and 26 HCs, by analysing circadian trends recorded continuously over a period of two weeks. They applied signal processing algorithms and polynomial regression algorithms to reconstruct circadian trends from big data acquired continuously at 1Hz by the sensor. They observed that circadian trends could distinguish not only pwMS from HCs, but also between pwMS with/without evidence of inflammatory activity (defined either by radiological activity or by a clinical relapse in the prior 12 months), between those with/without evidence of disease progression (defined by neurological deterioration without a relapse event), and between those with low/moderate-to-high disability (defined using an EDSS score cutoff = 3). Their results suggest that continuous heart rate monitoring could enable to uncover specific circadian patterns which distinguish pwMS across inflammatory states (associated with overactive sympathetic activity at night and overall reduced circadian variability) and disease progression (associated with overall reduced heart rate variability and reduced circadian adaptation of the autonomic nervous system). Therefore, autonomic nervous system monitoring with wearable sensors could provide new digital biomarkers and serve as an endpoint in clinical trials for both immunoregulation and symptomatic treatment. Notably, at least seven days of continuous wearing were required to establish robust circadian trends due to high variability of wearable-based heart rate at both the intra-individual and inter-individual level. More studies on larger and more heterogeneous cohorts are needed to confirm these results and increase the generalizability of these results, as >80% of this sample was made up of people with RRMS.

Seccia et al. (50) focused on the application of AI to analyse real-world clinical records of 1624 pwMS (totalling over 18,000 records between 1978 and 2018). They tried to predict the probability of shifting from the relapsing-remitting to the progressive phase at different timepoints (180, 360, 720 days from last visit). They tested predictions based on data from the last available visit using different ML models (visit-oriented approach), or based on the entire clinical history (history-oriented approach) using a specifically designed recurrent neural network (RNN). They found that the visit-oriented approach was better at predicting shifts to progressive MS at 180 days, largely thanks to the inclusion of imaging and liquor

history, suggesting that these two methods are informative on the risk of conversion to progressive MS in the short term. Conversely, the history-oriented approach performed better for predictions of shifting to progressive MS at longer intervals (360 and 720 days), owing largely to its better precision (reflecting less false positives). Crucially, the history-oriented approach was more reliant on clinical features, as both MRI and liquor data was unavailable for the majority of participants at all time points. Taken together, these results indicate that AI can effectively leverage real-world clinical big data to predict the risk of conversion to progressive MS. One key limitation is the intrinsic nature of real-world clinical data, which often contains missing data, as seen for liquor and MRI data in this study. It is crucial that clinical expertise is applied during the planning of analysis and data preprocessing, to determine if missing data are meaningful or not, and how they should be dealt with (e.g., missing liquor data can be expected, as lumbar punctures are not performed at each clinical visit, whereas EDSS score should ideally be available at all timepoints). This once again underlines the importance of data collection and maintenance. A well-structured and well-described feature set allows for much easier collaborations and sharing of data, thus promoting the fusion of different expertise (namely clinical and data science), which could further increase our understanding of MS. Accurate data maintenance could also allow to perform future analyses on data with longer follow-up durations, increasing our understanding of longitudinal disease patterns in MS.

Conclusion

The growing adoption of digital remote monitoring tools has great potential to improve both research and clinical aspects of MS, thanks to remote tracking of motor and non-motor symptoms. This review highlights that connected devices like smartphones and, especially, wearables can effectively monitor motor impairments, such as fall risk and gait disturbances, through continuous, granular data collection during real-world activities. Remote monitoring of physical activity is gaining significant traction in clinical research application. This is demonstrated by the inclusion of remote activity monitoring data as an exploratory endpoint in a recent drug trial (51), albeit through a basic daily step count metric. Further improvements may derive from AI algorithms which can recognize activity states, enriching the quantitative sensor data.

The evidence available on cognitive monitoring still favours the adaptation of active cognitive tests in digital form, to allow remote longitudinal monitoring, which may increase the standard of care for those with reduced mobility and/or access to specialized MS care. Recent advances in AI-driven cognitive tests and keystroke big data provide potential pathways to enable passive cognitive monitoring, but further research is needed to confirm their reliability and clinical utility.

Some studies have explored less-studied domains like sleep and circadian autonomic patterns, with interesting results which suggest that remote monitoring of these domains is feasible and could provide novel insights, compared to traditional research methods. Finally, preliminary exploratory studies have leveraged big data from clinical health records, with promising results, highlighting the need for careful recording, structuring, and maintenance of real-world clinical datasets. Increased awareness of the importance of big data in MS has led to the rising prominence of collaborative databases, both on a national (52–54) and international scale (55, 56), as well as multicentric studies on digital outcomes (57).

However, despite these advancements, challenges remain, including the small sample sizes observed in many studies, which limit the generalizability of their results to different MS populations, namely those with progressive MS, higher disease severity, and reduced access to specialized MS centres. Inclusiveness is a key area which should be addressed more carefully by future studies. Indeed, when assessing the real-world feasibility of digital monitoring for the entire MS population, researchers should be mindful of potential sampling bias, as patients willing/able to undergo such protocols may present distinct features (e.g., younger patients, with lower disability, higher educational attainment, and without cognitive impairment). For the use of AI and ML algorithms, researchers should never forget that an algorithm with many input variables may be very accurate but unusable by non-specialized MS centres which cannot obtain all the clinical/instrumental/sensor data on which the algorithm was trained on. Another significant limitation is the heterogeneity of monitoring methods and study protocols, which negates the possibility to compare feasibility, reliability, and validity data across different studies and devices. Future studies should strive to address these outstanding issues, since feasible, reliable and valid digital monitoring tools represent an invaluable resource for both research and clinical practice.

Finally, the recent rise and diffusion of conversational AI agents (e.g., ChatGPT) has led to some researchers exploring their usefulness in the setting of MS care (58, 59). When applied to disease monitoring, conversational AI could be integrated in eHealth apps as a chatbot, similar to examples from other fields (see for example (60)). This could allow patients to report their symptoms in a conversational manner, instead of having to answer omni-comprehensive and pre-defined structured lists of questions or questionnaires, which could feel alienating and repetitive, leading to low adherence. This may not only be perceived as a more natural and interpersonal approach by patients, but may also reduce their burden, by eliminating the need to answer questions which are not relevant for them at that moment in time. Moreover, an AI-driven closed loop system may also guide the administration of validated patient-reported questionnaires through eHealth apps, by selecting only the questionnaires that are most relevant for each individual patient, based on their reported symptoms at that specific timepoint. We hypothesize that this approach would reduce the time and energy demand on patients, while also providing a more interpersonal, responsive and adaptive monitoring framework, which could then lead to higher adoption and adherence to digital long-term monitoring. However, systematic studies are required to substantiate these hypotheses. Firstly, studies should evaluate the technical feasibility of applying conversational AI to longitudinal symptoms monitoring in MS, focusing particularly on the safety, validity and reliability of the information provided by AI. Secondly, they should investigate the expectations and needs of

patients, caregivers and clinicians toward digital monitoring, to determine if and how AI can be applied to address them.

Author contributions

MD: Writing – original draft, Writing – review & editing. GC: Conceptualization, Supervision, Writing – review & editing. LL: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

GC received consulting and speaking fees from Biogen, Merck, Novartis, Roche, Sanofi -Genzyme, Almirall, Teva, Actelion, Cellgene, BMS, Janssen-Cilag. None related to the present study. LL received consultancy fees from Merck, KGaA and Roche. None related to the present study.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler.* (2020) 26:1816–21. doi: 10.1177/1352458520970841

2. Filippi M, Bar-Or A, Piehl F, Preziosa P, Solari A, Vukusic S, et al. Multiple sclerosis. *Nat Rev Dis Prim.* (2018) 4:1–27. doi: 10.1038/s41572-018-0041-4

3. Benedict RHB, Amato MP, DeLuca J, Geurts JJG. Cognitive impairment in multiple sclerosis: clinical management, MRI, and therapeutic avenues. *Lancet Neurol.* (2020) 19:860–71. doi: 10.1016/S1474-4422(20)30277-5

4. Kappos L, Wolinsky JS, Giovannoni G, Arnold DL, Wang Q, Bernasconi C, et al. Contribution of relapse-independent progression vs relapse-associated worsening to overall confirmed disability accumulation in typical relapsing multiple sclerosis in a pooled analysis of 2 randomized clinical trials. *JAMA Neurol.* (2020) 77:1132. doi: 10.1001/jamaneurol.2020.1568

5. Feigin VL, Nichols E, Alam T, Bannick MS, Beghi E, Blake N, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* (2019) 18:459–80. doi: 10.1016/S1474-4422(18)30499-X

6. Dillenseger A, Weidemann ML, Trentzsch K, Inojosa H, Haase R, Schriefer D, et al. Digital biomarkers in multiple sclerosis. *Brain Sci.* (2021) 11:1–26. doi: 10.3390/brainsci11111519

7. Andrews JA, Craven MP, Lang AR, Guo B, Morriss R, Hollis C. The impact of data from remote measurement technology on the clinical practice of healthcare professionals in depression, epilepsy and multiple sclerosis: survey. *BMC Med Inform Decis Mak.* (2021) 21:1–17. doi: 10.1186/s12911-021-01640-5

8. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc.* (2016) 3:1–10. doi: 10.1177/2053951716631130

9. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: Towards an EU action plan. *Genome Med.* (2016) 8:71. doi: 10.1186/s13073-016-0323-y

10. Piovani D, Bonovas S. Real world-big data analytics in healthcare. *Int J Environ Res Public Health.* (2022) 19:11677. doi: 10.3390/ijerph191811677

11. Turing AM. Computing machinery and intelligence. In: Epstein R., Roberts G., Beber G. (eds) Parsing the Turing Test. 23:24 *Springer.* (2009). doi: 10.1007/978-1-4020-6710-5_3

12. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med.* (2020) 13:69–76. doi: 10.1007/s12178-020-09600-8

13. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

14. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging.* (2018) 9:611–29. doi: 10.1007/s13244-018-0639-9

15. Meca-Lallana V, Gascón-Giménez F, Ginestal-López RC, Higueras Y, Téllez-Lara N, Carreres-Polo J, et al. Cognitive impairment in multiple sclerosis: diagnosis and monitoring. *Neurol Sci.* (2021) 42:5183–93. doi: 10.1007/s10072-021-05165-7

16. Cameron MH, Nilsagard Y. Balance, gait, and falls in multiple sclerosis. In: Elsevier BV, editor. *Handbook of clinical neurology* (2018). Elsevier BV, p. 237–50. doi: 10.1016/B978-0-444-63916-5.00015-X

17. Tulipani LJ, Meyer B, Fox S, Solomon AJ, McGinnis RS. The sit-to-stand transition as a biomarker for impairment: comparison of instrumented 30-second chair stand test and daily life transitions in multiple sclerosis. *IEEE Trans Neural Syst Rehabil Eng.* (2022) 30:1213–22. doi: 10.1109/TNSRE.2022.3169962

18. Vandyk T, Meyer B, Depetrillo P, Donahue N, O'leary A, Fox S, et al. Digital phenotypes of instability and fatigue derived from daily standing transitions in persons with multiple sclerosis. *IEEE Trans Neural Syst Rehabil Eng.* (2023) 31:2279–86. doi: 10.1109/TNSRE.2023.3271601

19. Meyer BM, Tulipani LJ, Gurchiek RD, Allen DA, Adamowicz L, Larie D, et al. Wearables and deep learning classify fall risk from gait in multiple sclerosis. *IEEE J BioMed Heal Inf.* (2021) 25:1824–31. doi: 10.1109/JBHI.2020.3025049

20. Mosquera-Lopez C, Wan E, Shastry M, Folsom J, Leitschuh J, Condon J, et al. Automated detection of real-world falls: modeled from people with multiple sclerosis. *IEEE J BioMed Heal Inf.* (2021) 25:1975–84. doi: 10.1109/JBHI.2020.3041035

21. Kushner T, Mosquera-Lopez C, Hildebrand A, Cameron MH, Jacobs PG. Risky movement: Assessing fall risk in people with multiple sclerosis with wearable sensors and beacon-based smart-home monitoring. *Mult Scler Relat Disord.* (2023) 79:105019. doi: 10.1016/j.msard.2023.105019

22. Casilari E, Oviedo-Jiménez MA. Automatic fall detection system based on the combined use of a smartphone and a smartwatch. *PloS One.* (2015) 10:1–11. doi: 10.1371/journal.pone.0140929

23. Özdemir AT, Barshan B. Detecting falls with wearable sensors using machine learning techniques. *Sensors (Switzerland).* (2014) 14:10691–708. doi: 10.3390/s140610691

24. LaRocca NG. Impact of walking impairment in multiple sclerosis. *Patient Patient-Centered Outcomes Res.* (2011) 4:189–201. doi: 10.2165/11591150-000000000-00000

25. Shanahan CJ, Boonstra FMC, Cofré Lizama LE, Strik M, Moffat BA, Khan F, et al. Technologies for advanced gait and balance assessments in people with multiple sclerosis. *Front Neurol*. (2018) 8:708. doi: 10.3389/fneur.2017.00708

26. Salis F, Bertuletti S, Bonci T, Caruso M, Scott K, Alcock L, et al. A multi-sensor wearable system for the assessment of diseased gait in real-world conditions. *Front Bioeng Biotechnol*. (2023) 11:1143248. doi: 10.3389/fbioe.2023.1143248

27. Chitnis T, Glanz BI, Gonzalez C, Healy BC, Saraceno TJ, Sattarnezhad N, et al. Quantifying neurologic disease using biosensor measurements in-clinic and in free-living settings in multiple sclerosis. *NPJ Digit Med*. (2019) 2:1–8. doi: 10.1038/s41746-019-0197-7

28. Atrsaei A, Dadashi F, Mariani B, Gonzenbach R, Aminian K. Toward a remote assessment of walking bout and speed: application in patients with multiple sclerosis. *IEEE J BioMed Heal Inf*. (2021) 25:4217–28. doi: 10.1109/JBHI.2021.3076707

29. Kirk C, Küderle A, Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, Ullrich M, et al. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. *Sci Rep*. (2024) 14:1–23. doi: 10.1038/s41598-024-51766-5

30. Bourke AK, Scotland A, Lipsmeier F, Gossens C, Lindemann M. Gait characteristics harvested during a smartphone-based self-administered 2-minute walk test in people with multiple sclerosis: Test-retest reliability and minimum detectabl change. *Sensors (Switzerland)*. (2020) 20:1–16. doi: 10.3390/s20205906

31. Sun S, Folarin AA, Zhang Y, Cummins N, Liu S, Stewart C, et al. The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions. *Comput Methods Programs BioMed*. (2022) 227:107204. doi: 10.1016/j.cmpb.2022.107204

32. Granja Domínguez A, Romero Sevilla R, Alemán A, Durán C, Hochsprung A, Navarro G, et al. Study for the validation of the FeetMe® integrated sensor insole system compared to GAITRite® system to assess gait characteristics in patients with multiple sclerosis. *PloS One*. (2023) 18:e0272596. doi: 10.1371/journal.pone.0272596

33. Delahaye C, Chaves D, Congnard F, Noury-Desvaux B, de Müllenheim PY. Measuring outdoor walking capacities using global positioning system in people with multiple sclerosis: Clinical and methodological insights from an exploratory study. *Sensors*. (2021) 21:3189. doi: 10.3390/s21093189

34. Stavropoulos TG, Meditskos G, Lazarou I, Mpaltadoros L, Papagiannopoulos S, Tsolaki M, et al. Detection of health-related events and behaviours from wearable sensor lifestyle data using symbolic intelligence: a proof-of-concept application in the care of multiple sclerosis. *Sensors*. (2021) 21:6230. doi: 10.3390/s21186230

35. Block VJ, Waliman M, Xie Z, Akula A, Bove R, Pletcher MJ, et al. Making every step count: minute-by-minute characterization of step counts augments remote activity monitoring in people with multiple sclerosis. *Front Neurol*. (2022) 13:860008. doi: 10.3389/fneur.2022.860008

36. Salomon A, Galperin I, Buzaglo D, Mirelman A, Regev K, Karni A, et al. Fragmentation, circadian amplitude, and fractal pattern of daily-living physical activity in people with multiple sclerosis: Is there relevant information beyond the total amount of physical activity? *Mult Scler Relat Disord*. (2022) 68:104108. doi: 10.1016/j.msard.2022.104108

37. Creagh AP, Dondelinger F, Lipsmeier F, Lindemann M, De Vos M. Longitudinal trend monitoring of multiple sclerosis ambulation using smartphones. *IEEE Open J Eng Med Biol*. (2022) 3:202–10. doi: 10.1109/OJEMB.2022.3221306

38. Birchmeier ME, Studer T. Automated rating of multiple sclerosis test results using a convolutional neural network. *Stud Health Technol Inform*. (2019) 259:105–8. doi: 10.3233/978-1-61499-961-4-105

39. Langdon DW, Amato MP, Boringa J, Brochet B, Foley F, Fredrikson S, et al. Recommendations for a brief international cognitive assessment for multiple sclerosis (BICAMS). *Mult Scler J*. (2012) 18:891–8. doi: 10.1177/1352458511431076

40. Birchmeier ME, Studer T, Lutterotti A, Penner I-K, Bignens S. Digitalisation of the brief visuospatial memory test-revised and evaluation with a machine learning algorithm. *Stud Health Technol Inform*. (2020) 270:168–72. doi: 10.3233/SHTI200144

41. Petilli MA, Daini R, Saibene FL, Rabuffetti M. Automated scoring for a Tablt-based Rey Figure copy task differentiates constructional, organisational, and motor abilities. *Sci Rep*. (2021) 11:1–19. doi: 10.1038/s41598-021-94247-9

42. Khaligh-Razavi SM, Sadeghi M, Khanbagi M, Kalafatis C, Nabavi SM. A self-administered, artificial intelligence (AI) platform for cognitive assessment in multiple sclerosis (MS). *BMC Neurol*. (2020) 20:1–13. doi: 10.1186/s12883-020-01736-x

43. Midaglia L, Mulero P, Montalban X, Graves J, Hauser SL, Julian L, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. *J Med Internet Res*. (2019) 21:e14863. doi: 10.2196/14863

44. Pless S, Woelfle T, Naegelin Y, Lorscheider J, Wiencierz A, Reyes Ó, et al. Assessment of cognitive performance in multiple sclerosis using smartphone-based training games: a feasibility study. *J Neurol*. (2023) 270:3451–63. doi: 10.1007/s00415-023-11671-9

45. Lam KH, Meijer KA, Loonstra FC, Coerver EME, Twose J, Redeman E, et al. Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. *Mult Scler J*. (2021) 27:1421–31. doi: 10.1177/1352458520968797

46. Lam KH, Twose J, Lissenberg-Witte B, Licitra G, Meijer K, Uitdehaag B, et al. The use of smartphone keystroke dynamics to passively monitor upper limb and cognitive function in multiple sclerosis: longitudinal analysis. *J Med Internet Res*. (2022) 24:1–12. doi: 10.2196/37614

47. Foong YC, Bridge F, Merlo D, Gresle M, Zhu C, Buzzard K, et al. Smartphone monitoring of cognition in people with multiple sclerosis: A systematic review. *Mult Scler Relat Disord*. (2023) 73:104674. doi: 10.1016/j.msard.2023.104674

48. Woelfle T, Pless S, Reyes Ó, Wiencierz A, Kappos L, Granziera C, et al. Smartwatch-derived sleep and heart rate measures complement step counts in explaining established metrics of MS severity. *Mult Scler Relat Disord*. (2023) 80:105104. doi: 10.1016/j.msard.2023.105104

49. Hilty M, Oldrati P, Barrios L, Müller T, Blumer C, Foege M, et al. Continuous monitoring with wearables in multiple sclerosis reveals an association of cardiac autonomic dysfunction with disease severity. *Mult Scler J - Exp Transl Clin*. (2022) 8:205521732211034. doi: 10.1177/20552173221103436

50. Seccia R, Gammelli D, Dominici F, Romano S, Landi AC, Salvetti M, et al. Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. *PloS One*. (2020) 15:1–18. doi: 10.1371/journal.pone.0230219

51. Cree BAC, Cutter G, Wolinsky JS, Freedman MS, Comi G, Giovannoni G, et al. Safety and efficacy of MD1003 (high-dose biotin) in patients with progressive multiple sclerosis (SPI2): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Neurol*. (2020) 19:988–97. doi: 10.1016/S1474-4422(20)30347-1

52. Magyari M, Joensen H, Laursen B, Koch-Henriksen N. The danish multiple sclerosis registry. *Brain Behav*. (2021) 11:e01921. doi: 10.1002/brb3.1921

53. Trojano M, Bergamaschi R, Amato MP, Comi G, Ghezzi A, Lepore V, et al. The Italian multiple sclerosis register. *Neurol Sci*. (2019) 40:155–65. doi: 10.1007/s10072-018-3610-0

54. Hillert J, Stawiarz L. The Swedish MS registry – clinical support tool and scientific resource. *Acta Neurol Scand*. (2015) 132:11–9. doi: 10.1111/ane.12425

55. Glaser A, Butzkueven H, van der Walt A, Gray O, Spelman T, Zhu C, et al. Big Multiple Sclerosis Data network: an international registry research network. *J Neurol*. (2024) 271:3616–24. doi: 10.1007/s00415-024-12303-6

56. Geys L, Parciak T, Pirmani A, McBurney R, Schmidt H, Malbaša T, et al. The multiple sclerosis data alliance catalogue. *Int J MS Care*. (2021) 23:261–8. doi: 10.7224/1537-2073.2021-006

57. Mikolaizak AS, Rochester L, Maetzler W, Sharrack B, Demeyer H, Mazzà C, et al. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement–the Mobilise-D study protocol. *PloS One*. (2022) 17:e0269615. doi: 10.1371/journal.pone.0269615

58. Inojosa H, Gilbert S, Kather JN, Proschmann U, Akgün K, Ziemssen T. Can ChatGPT explain it? Use of artificial intelligence in multiple sclerosis communication. *Neurol Res Pract*. (2023) 5:48. doi: 10.1186/s42466-023-00270-8

59. Patel MA, Villalobos F, Shan K, Tardo LM, Horton LA, Sguigna PV, et al. Generative artificial intelligence versus clinicians: Who diagnoses multiple sclerosis faster and with greater accuracy? *Mult Scler Relat Disord*. (2024) 90:105791. doi: 10.1016/j.msard.2024.105791

60. Haque MDR, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth uHealth*. (2023) 11::e44838. doi: 10.2196/44838

frontiers | Frontiers in Digital Health

TYPE Perspective
PUBLISHED 24 March 2025
DOI 10.3389/fdgth.2025.1507159

OPEN ACCESS

EDITED BY
Axel Faes,
University of Hasselt, Belgium

REVIEWED BY
Aonghus Lawlor,
University College Dublin, Ireland

*CORRESPONDENCE
Lorin Werthen-Brabants
✉ lorin.werthenbrabants@ugent.be

RECEIVED 08 October 2024
ACCEPTED 12 March 2025
PUBLISHED 24 March 2025

CITATION
Werthen-Brabants L, Dhaene T and Deschrijver D (2025) The role of trustworthy and reliable AI for multiple sclerosis. Front. Digit. Health 7:1507159. doi: 10.3389/fdgth.2025.1507159

The role of trustworthy and reliable AI for multiple sclerosis

Lorin Werthen-Brabants*, Tom Dhaene and Dirk Deschrijver

SUMO Lab, IDLab, INTEC, Ghent University – imec, Ghent, Belgium

This paper investigates the importance of Trustworthy Machine Learning (ML) in the context of Multiple Sclerosis (MS) research and care. Due to the complex and individual nature of MS, the need for reliable and trustworthy ML models is essential. In this paper, key aspects of trustworthy ML, such as out-of-distribution generalization, explainability, uncertainty quantification and calibration are explored, highlighting their significance for healthcare applications. Challenges in integrating these ML tools into clinical workflows are addressed, discussing the difficulties in interpreting AI outputs, data diversity, and the need for comprehensive, quality data. It calls for collaborative efforts among researchers, clinicians, and policymakers to develop ML solutions that are technically sound, clinically relevant, and patient-centric.

KEYWORDS

artificial intelligence, multiple sclerosis, trustworthy AI, deep learning, uncertainty quantification

1 Introduction

Machine Learning (ML) is increasingly applied to healthcare applications (1). While traditional statistical methods can help with biomarker discovery and recognizing trends and correlations, modern ML techniques such as Deep Learning (DL), are able to uncover complex correlations and provide better results than traditional, simpler techniques (2) due to their universal nature (3). Conversely, as these techniques become more complex, the need for *reliable* and *trustworthy* models increases (4, 5), especially within healthcare. However, building trust does not have a one-size-fits-all solution, resulting in many techniques to be developed to aid decision making.

For an end-user, be it a clinician or a patient, a model that is trustworthy is one that can provide certain guarantees on its predictions, explain its predictions, and provide a notion of uncertainty. For a complex disease such as Multiple Sclerosis (MS), the need for trustworthy models is especially pertinent, as its progression is non-trivially defined, and the decisions made to hinder its progression are important ones. A machine learning system that does not provide adequate reliability metrics, or trustworthy insights, will be less appealing to the end-user when there are high-stakes consequences. In recent years, the need for Trustworthy ML (TML) has also reached mainstream attention with the use of generative AI becoming more prevalent. For example, though Large Language Models have shown impressive results, they may still provide incorrect results, without any notion of uncertainty or trustworthiness (6). This is also known as the "hallucination" effect (7). Complex data and relationships warrant the use of trustworthiness techniques.

In the following Sections, we provide a summary of techniques present in Trustworthy ML (TML) (Section 2), why TML is necessary for MS (Section 3.1), and the associated challenges (Section 3.2).

2 Trustworthy machine learning

2.1 Out-of-distribution generalization

The many ways in which MS progression can occur (different limbs, locations of lesion growth, etc.), makes the disease variable and patient specific. Therefore, training data will rarely contain enough data to cover the full extent of the ways progression can be observed. Furthermore, due to protocols changing regularly and equipment variability, *concept* or *model drift* (8) may pose a real issue when ML models are deployed in the real world. Model drift occurs when new data do not correspond to the data on which a model was trained. As a result, models must be continually adapted so changes in data distributions are captured.

These issues can be tackled by making use of techniques such as domain adaptation (9, 10), a specific case of transfer learning (11), and synthetic data sampling such as SMOTE (12, 13).

The concept of Out-of-Distribution Generalization can be elucidated by considering a concrete example within the MS context. Imagine an ML model trained on data from North American patients. When this model is applied to patients from different geographical regions with distinct genetic and environmental factors, its predictions may falter due to differences in disease manifestation. Domain adaptation techniques can help here by adjusting the model to account for these regional variations. Similarly, synthetic data sampling, like the aforementioned SMOTE technique, can artificially—not necessarily in a representative way—augment the dataset to include underrepresented samples in a given dataset, improving the model's robustness against a wide range of clinical scenarios. However, it must be stressed that data quality is key, and an underrepresented dataset can not fully capture the underlying factors to guarantee good out-of-distribution generalization.

2.2 Explainability and interpretability

A perfectly interpretable AI provides insights into the inner workings and decision process of an AI system. When it comes to the types of ML systems, they can broadly be divided into two categories: white-box models and black-box models.

2.2.1 White-box models

Models that are inherently explainable and interpretable. These are often simpler methods such as linear or logistic regression, the latter of which can be represented as a nomogram (14), a graphical representation of such models that visually convey the weight of different input variables. These models can be fully dissected, so there may be many ways of representing or explaining them.

2.2.2 Black-box models

Models that can not be interpreted easily, and are regarded as a "black box" out of which little or no knowledge can be derived. However, there are techniques that can provide explainability when working with black-box models, such as making use of Shapley values (15, 16) or making use of Deep Learning specific techniques (17) such as Layer-wise Relevance Propagation (18, 19). These are often post-hoc. In practice, these techniques will show a number of features and their importances expressed as a number. This could also be in the form of a heatmap. These feature importances may not always be as readily interpretable and may need training and education to comprehend adequately. Additionally, they do not necessarily *explain* why those features are important.

A classifier that may perform well in its evaluation metrics (sensitivity, specificity, ROC AUC, etc.) may still benefit from explainability methods. In particular, if models were to take into account many multimodal variables, the primary drivers of a given prediction may offer important insight for the user of the machine learning system.

Related to interpretable AI is explainable AI. Rather than being able to fully comprehend the inner workings of a model, an explainable AI model is able to be queried so that a reasonable explanation to the prediction is provided. Explainable AI can be viewed on different levels as well: Global, cohort, and local explainability. Global explainability provides information about the entire population or dataset. Due to the complex nature of the MS disease, valuable insights on a population level are scarce. Cohort explainability gives insight on subsets of the data, which can be more interesting when taking into account certain covariates. In this way, different groups of patients can be identified and correlations within these groups may offer more helpful insights than looking only at a global level. Lastly, local explainability provides insight on the model's output for a single input example. Every patient has a different profile, and therefore local explainability may help acquire insight into the prediction of the model for that specific patient or observation.

2.3 Uncertainty quantification and calibration

2.3.1 Uncertainty quantification

In machine learning models, uncertainty plays a critical, yet understated role in understanding and interpreting predictions. Healthcare specifically can greatly benefit from uncertainty quantification, as it can add a layer of trust between the user and the model (20–22). Two major sources of uncertainty are aleatoric and epistemic uncertainty (23).

2.3.1.1 Aleatoric uncertainty

This type of irreducible uncertainty is inherent in the data itself. It cannot be reduced by adding more data and manifests as the noise within the data. An example of this uncertainty arises when using very few features. For example, a patient's blood pressure is a crucial health metric, but it exhibits natural variability within an individual due to various factors like stress, activity level, time of day, and even the way it is measured.

This uncertainty can be either homoscedastic, when it remains constant for all values (e.g., base noise of a sensor), or

heteroscedastic, when it varies depending on the value of the sample.

2.3.1.2 Epistemic uncertainty.

Epistemic uncertainty arises from the model's limited knowledge. This reducible uncertainty is high when the model has insufficient data to characterize or capture the target variable. Increasing the size of the data set can help reduce epistemic uncertainty. An intuitive example can be demonstrated as follows: Say there are multiple experts for a single disease such as MS. These experts may disagree on a given prognosis, despite all of them being equally trained for such a task. Analogously, in a machine learning model predicting patient outcomes for MS, the model might exhibit high epistemic uncertainty if it has been trained on a limited or non-representative dataset. Just as the disagreement among experts might stem from variations in their individual experiences and interpretations, the model's uncertainty arises from its limited exposure to the diverse manifestations of the disease. By providing the model with more comprehensive data that captures a wider range of patient histories, symptoms, and outcomes, the epistemic uncertainty can be reduced, leading to more consistent and reliable predictions.

Applying uncertainty quantification in MS involves recognizing and managing the inherent unpredictability in patient responses and disease progression. For instance, a model expressing aleatoric uncertainty might show the variability in a patient's symptoms over time, acknowledging that certain aspects of MS progression cannot be predicted with complete precision. Epistemic uncertainty can be illustrated by a model's varying predictions based on different patient subgroups, reflecting limited knowledge about specific MS manifestations. To quantify and capture these uncertainties, techniques like Monte Carlo Dropout (MCD) (24) can be employed, providing a probabilistic understanding of a model's predictions and helping clinicians make informed decisions under uncertainty.

Uncertainty quantification has been applied to lesion detection in MRI images (25–27), often making use of MCD or other methods of obtaining a model that can express uncertainty (28).

2.3.2 Calibration

A well-calibrated machine learning model is one in which the model's predicted probabilities closely match the probabilities observed in the actual data (29). Mathematically, this is represented as $P(y|\hat{p}(y) = \alpha) = \alpha$. This equation signifies that the probability of an event y occurring, given that the model predicts it with probability α, should ideally be α itself. As a practical example: a model that predicts the probability of 40% disease progression for a patient will ideally be correct 40% of the time of all patients who receive a similar prognosis. For methods such as neural networks, this is not often the case by default, and calibration needs to be improved. Additionally, calibration can also be applied to regressors that output a distribution, rather than a single value. In this case, the confidence interval (such as a 95% confidence interval, for example) can be calibrated to ensure that it matches the observations.

The need for calibration is evident in the lack of information an uncalibrated classifier or regressor provides. Often, as is the case with neural networks, a neural network classifier will collapse to output probabilities close to 100% or 0% consistently, rather than providing accurate probability estimates (29). As a result, a user of such a system needs to blindly trust the classifier rather than being able to take the confidence of the classifier into account.

3 Discussion

3.1 Why trustworthy ML is necessary for MS research

With the current knowledge of MS and performance of state-of-the-art machine learning models in the field, it stands to reason that there may not be a one-size-fits-all solution to detecting disease progression. Although other types of model (such as image classifiers) may perform very well and can reliably be used in most, if not all, cases, this may not be the case for MS. ML models for this purpose will likely be a tool to aid decision making, rather than a decision maker by itself. To that end, an ML model that just states "yes" or "no" is not sufficient. Rather, more information should be supplied to the user. A trustworthy version of this model will highlight parts of the input that contribute greatly to the prediction, show which global and cohort features are important, and also provide a notion of (un)certainty with the prediction. In this way, the user can:

- Select which predictions to trust and keep, both by using aleatoric and epistemic uncertainty as guides
- Analyze the subgroup in which the prediction fits
- Analyze the specific prediction and the features leading to the prediction

For MS research, the use and adoption of ML will be guided by advances in *trustworthy* ML. MS is a disease marked by its heterogeneity in symptoms, progression, and response to treatment, making reliable analysis of significant importance.

The ability of ML models to process and analyze different types of data—from clinical observations to MRI images—can lead to earlier detection and more precise monitoring of the disease's progression. However, the value of these insights depends on their explainability. Clinicians and patients must be able to understand and trust the model's predictions, necessitating a focus on explainable AI. For example, an ML model might identify subtle changes in brain lesions over time, but this information becomes clinically actionable only when it is presented in an understandable manner. Explainable models can elucidate the factors driving a prediction, thereby enhancing the clinician's ability to make informed treatment decisions.

Moreover, the integration of uncertainty quantification in ML models is particularly relevant for MS. Given the variability in how the disease presents and progresses, models that can express their confidence in predictions are invaluable. They provide clinicians with a more nuanced understanding of each prediction,

facilitating more informed risk-benefit analyses when deciding on treatment plans. A model that indicates a high level of uncertainty in its prediction might prompt further testing or closer monitoring, whereas a prediction made with high confidence could lead to more decisive action.

The importance of trustworthy ML in MS research also extends to patient empowerment. Access to understandable and reliable ML-driven insights can foster better patient-clinician dialogues. When patients understand the basis for predictions about their condition, they are better positioned to make informed decisions about their treatment and lifestyle choices.

3.2 Challenges of trustworthy ML for MS

3.2.1 Integration of ML tools to aid clinical decisions

Integrating ML tools into existing clinical workflows presents another layer of complexity. For these tools to be adopted, they must fit into the highly regulated environment of healthcare. This integration involves designing user interfaces and metrics that are intuitive for clinicians, ensuring that ML predictions are presented in a way that complements decision-making processes rather than complicating them (30). Furthermore, imperfect data pose a problem during the training and prediction stages of an ML model. Data collection can be a laborious task, and in some cases the data cannot be accurately represented due to individual differences in disease expression. This rings especially true in the case of MS.

3.2.2 Usability of uncertainty quantification and explainability techniques

As highlighted previously, UQ and explainability techniques have their merit, as they can highlight potential issues when making use of ML assisted decision systems. However, the end-user may not find much use in the way UQ results are represented in literature. Even explainabilty results have varying degrees of success concerning their usability (31). These techniques could benefit from user studies, as their usability hinges on the representation and, in turn, interpretation by the end-user. For example, rather than providing the clinician and/or patient with a numerical value signifying a "trustworthiness" score or certainty otherwise, larger trust could be gained by comparing the patient with other patients that have similar disease trajectories. This opacity can hinder trust and acceptance, especially in a high-stakes field like healthcare where understanding the "why" behind a diagnosis or prognosis is as crucial as the outcome itself (31).

3.2.3 Out-of-distribution data, diverse data, available data

Data diversity and availability are critical factors that significantly influence the development and performance of ML models in MS research. MS is a disease with a highly variable clinical course and a wide range of symptoms that differ from patient to patient. This heterogeneity necessitates a rich and diverse dataset that captures the broad spectrum of the disease. After all, deep learning techniques are prone to overfitting, and may have performance below acceptable levels as a result (21, 32). Initiatives such as MSBase (33, 34) attempt to address the issue of out-of-distribution performance by providing multi-center data. The amount of data by itself may give the end-user a reason to trust a model, given enough diversity. Data quality is another concern, with issues such as missing values, inconsistent data entry, and the need for standardization across different data sources complicating the development of reliable ML models. Introducing diversity by including measurements that stray away from purely medical imaging or clinical data may also provide a new avenue of research, potentially discovering novel biomarkers. Future work should focus on developing models that can adapt to individual patient variations and incorporating emerging data types such as Motor Evoked Potentials (35, 36) into ML models.

4 Conclusion

This paper underscores the importance of trustworthiness in Machine Learning (ML) applications for Multiple Sclerosis (MS). Key aspects such as explainability, uncertainty quantification and calibration, and out-of-distribution generalization have been explored. Additionally, the challenges in integrating ML into clinical workflows and the hurdles posed by data diversity and availability have been discussed.

The authors urge the research community and healthcare providers to prioritize the development and implementation of trustworthy ML solutions for MS (and healthcare in general). There is an urgent need to foster partnerships between computer scientists, neurologists, and patients. This collaboration will ensure the development of ML solutions that are not only technically sound but also clinically relevant and patient-centric. Making comprehensive, high-quality data sets accessible while respecting privacy concerns is crucial. Initiatives should focus on standardizing data collection and sharing practices to aid in the development of more effective ML models. ML tools must be integrated into clinical workflows in a way that is intuitive and enhances decision-making processes. This involves designing user-friendly interfaces and ensuring that clinicians are adequately trained to use these tools effectively.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

LW-B: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. TD: Funding acquisition,

Supervision, Writing – review & editing. DD: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work has been supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* (2017) 2:230–43. doi: 10.1136/svn-2017-000101

2. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with alphafold. *Nature.* (2021) 596:583–9. doi: 10.1038/s41586-021-03819-2

3. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst.* (1989) 2:303–14. doi: 10.1007/BF02551274

4. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis.* (2018) 66:149–53. doi: 10.1093/cid/cix731

5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* (2023) 29:1930–40. doi: 10.1038/s41591-023-02448-8

6. Sejnowski TJ. Large language models and the reverse turing test. *Neural Comput.* (2023) 35:309–42. doi: 10.1162/neco_a_01563

7. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. Data from: Chatgpt and other large language models are double-edged swords (2023)

8. Tsymbal A. *The Problem of Concept Drift: Definitions and Related Work.* Technical report (TCD-CS-2004-15). Dublin: Computer Science Department, Trinity College Dublin (2004). Vol. 106. p. 58.

9. Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. In: *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020* (2021). p. 877–94.

10. Valverde S, Salem M, Cabezas M, Pareto D, Vilanova JC, Ramió-Torrentà L, et al. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage Clin.* (2019) 21:101638. doi: 10.1016/j.nicl.2018.101638

11. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data.* (2016) 3:1–40. doi: 10.1186/s40537-016-0043-6

12. Branco D, Martino B, Esposito A, Tedeschi G, Bonavita S, Lavorgna L. Machine learning techniques for prediction of multiple sclerosis progression. *Soft Comput.* (2022) 26:12041–55. doi: 10.1007/s00500-022-07503-z

13. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res.* (2002) 16:321–57. doi: 10.1613/jair.953

14. Kattan MW, Marasco J. What is a real nomogram? In: *Seminars in Oncology.* Elsevier (2010). Vol. 37. p. 23–6.

15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inform Process Syst.* (2017) 30:4768–77.

16. Basu S, Munafo A, Ben-Amor AF, Roy S, Girard P, Terranova N. Predicting disease activity in patients with multiple sclerosis: an explainable machine-learning approach in the mavenclad trials. *CPT Pharmacom Syst Pharmacol.* (2022) 11:843–53. doi: 10.1002/psp4.12796

17. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: *2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI).* IEEE (2017). p. 1–6.

18. Creagh AP, Lipsmeier F, Lindemann M, Vos MD. Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. *Sci Rep.* (2021) 11:14301. doi: 10.1038/s41598-021-92776-x

19. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. *Explainable AI.* (2019) 11700:193–209. doi: 10.1007/978-3-030-28954-6_10

20. Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013–2023). *Comput Biol Med.* (2023) 165:107441. doi: 10.1016/j.compbiomed.2023.107441

21. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell.* (2019) 1:20–3. doi: 10.1038/s42256-018-0004-1

22. Lambert B, Forbes F, Doyle S, Dehaene H, Dojat M. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artif Intell Med.* (2024) 150:102830. doi: 10.1016/j.artmed.2024.102830

23. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn.* (2021) 110:457–506. doi: 10.1007/s10994-021-05946-3

24. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *International Conference on Machine Learning.* PMLR (2016). p. 1050–9.

25. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med Image Anal.* (2020) 59:101557. doi: 10.1016/j.media.2019.101557

26. Molchanova N, Raina V, Malinin A, La Rosa F, Muller H, Gales M, et al. Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis. In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI).* IEEE (2023). p. 1–5.

27. Tousignant A, Lemaître P, Precup D, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. In: *International Conference on Medical Imaging with Deep Learning.* PMLR (2019). p. 483–92.

28. Lambert B, Forbes F, Doyle S, Tucholka A, Dojat M. Fast uncertainty quantification for deep learning-based MR brain segmentation. In: *EGC 2022-Conference Francophone Pour l'Extraction et la Gestion des Connaissances.* (2022). p. 1–12.

29. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *International Conference on Machine Learning.* PMLR (2017). p. 1321–30.

30. Dabbs ADV, Myers BA, Mc Curry KR, Dunbar-Jacob J, Hawkins RP, Begey A, et al. User-centered design and interactive health technologies for patients. *Comput Inform Nurs.* (2009) 27:175. doi: 10.1097/NCN.0b013e31819f7c7c

31. Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput Biol Med*. (2022) 149:106043. doi: 10.1016/j.compbiomed.2022.106043

32. Mårtensson G, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med Image Anal*. (2020) 66:101714. doi: 10.1016/j.media.2020.101714

33. Butzkueven H, Chapman J, Cristiano E, Grand'Maison F, Hoffmann M, Izquierdo G, et al. Msbase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Mult Scler J*. (2006) 12:769–74. doi: 10.1177/1352458506070775

34. De Brouwer E, Becker T, Werthen-Brabants L, Dewulf P, Iliadis D, Dekeyser C, et al. Machine-learning-based prediction of disability progression in multiple sclerosis: an observational, international, multi-center study. *PLoS Digit Health*. (2024) 3: e0000533. doi: 10.1371/journal.pdig.0000533

35. Rossini PM, Rossi S. Clinical applications of motor evoked potentials. *Electroencephalogr Clin Neurophysiol*. (1998) 106:180–94. doi: 10.1016/S0013-4694(97)00097-7

36. Yperman J, Popescu V, Van Wijmeersch B, Becker T, Peeters LM. Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset. *Sci Data*. (2022) 9:207. doi: 10.1038/s41597-022-01335-0

frontiers | Frontiers in Artificial Intelligence

TYPE Mini Review
PUBLISHED 08 April 2025
DOI 10.3389/frai.2025.1478068

OPEN ACCESS

EDITED BY
Liesbet M. Peeters,
University of Hasselt, Belgium

REVIEWED BY
Diana L. Giraldo,
University of Antwerp, Belgium
Jingpeng Li,
Harvard Medical School, United States

*CORRESPONDENCE
Jean-Pierre R. Falet
✉ jean-pierre.falet@mcgill.ca

†These authors have contributed equally to this work

RECEIVED 27 September 2024
ACCEPTED 19 March 2025
PUBLISHED 08 April 2025

CITATION
Falet J-P, Nobile S, Szpindel A, Barile B, Kumar A, Durso-Finley J, Arbel T and Arnold DL (2025) The role of AI for MRI-analysis in multiple sclerosis—A brief overview. *Front. Artif. Intell.* 8:1478068. doi: 10.3389/frai.2025.1478068

The role of AI for MRI-analysis in multiple sclerosis—A brief overview

Jean-Pierre R. Falet[1,2,3*†], Steven Nobile[1†], Aliya Szpindel[1], Berardino Barile[2,3], Amar Kumar[2,3], Joshua Durso-Finley[2,3], Tal Arbel[2,3] and Douglas L. Arnold[1]

[1]Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [2]Mila - Quebec AI Institute, Montreal, QC, Canada, [3]Department of Electrical and Computer Engineering, Centre for Intelligent Machines, McGill University, Montreal, QC, Canada

Magnetic resonance imaging (MRI) has played a crucial role in the diagnosis, monitoring and treatment optimization of multiple sclerosis (MS). It is an essential component of current diagnostic criteria for its ability to non-invasively visualize both lesional and non-lesional pathology. Nevertheless, modern day usage of MRI in the clinic is limited by lengthy protocols, error-prone procedures for identifying disease markers (e.g., lesions), and the limited predictive value of existing imaging biomarkers for key disability outcomes. Recent advances in artificial intelligence (AI) have underscored the potential for AI to not only improve, but also transform how MRI is being used in MS. In this short review, we explore the role of AI in MS applications that span the entire life-cycle of an MRI image, from data collection, to lesion segmentation, detection, and volumetry, and finally to downstream clinical and scientific tasks. We conclude with a discussion on promising future directions.

KEYWORDS

artificial intelligence, machine learning, magnetic resonance imaging, multiple sclerosis, precision medicine

1 Introduction

Multiple Sclerosis (MS) is a neuro-inflammatory disease of the central nervous system characterized by a wide spectrum of inflammatory and neurodegenerative changes (Compston and Coles, 2008), with clinical manifestations that vary greatly between individuals. Since the 1980s, magnetic resonance imaging (MRI) has been a cornerstone of MS diagnosis and management due to the ability to visualize demyelinating changes and axonal loss resulting from focal inflammation, using a combination of T2 and T1-weighted sequences (Hemond and Bakshi, 2018). The temporal evolution of lesions, which may initially enhance (Filippi et al., 2019), and subsequently expand, remain static, or decrease in size (Koopmans et al., 1989), can also be captured by MRI. A number of MRI biomarkers of MS diagnosis, prognosis, and treatment response, have also been described. These include T2-hyperintense white matter lesions, gadolinium-enhancing lesions, slowly enlarging lesions, paramagnetic rim lesions, cortical/deep gray matter lesions, and leptomeningeal enhancement (Filippi and Agosta, 2010; Filippi et al., 2020). Some of these biomarkers have been found to correlate strongly with key clinical outcomes. One example is the association between new/enlarging T2 lesions and clinical relapses (Rudick et al., 2006; Sormani et al., 2009; Sormani and Bruzzi, 2013).

Despite these advances, MRI-analysis continues to face problems that limit its potential (Maggi and Absinta, 2024). The longer acquisition times and higher field strengths required

to obtain measurements of many recently studied imaging biomarkers introduces new headaches for resource-limited settings. At many clinical sites, the evaluation of MRI continues to be done manually, which is a lengthy, error-prone, and highly variable procedure (Bozsik et al., 2022; Altay et al., 2013). A strongly predictive imaging biomarker of disability progression, especially progression which is independent of relapse activity (Müller et al., 2023), has yet to be found (Filippi et al., 2020). At the therapeutic level, the influx of disease modifying therapies has significantly improved the ability to suppress lesion formation and relapse risk (Amin and Hersh, 2023), but targeting disability progression remains a major challenge. The use of MRI in predicting disease course and facilitating treatment selection is still a work in progress.

The rapid pace of progress in artificial intelligence (AI) has led to new opportunities for MRI-analysis in MS. In contrast to classical statistical methods which focus on acquiring knowledge about a population given data sampled from the same distribution, the field of AI has developed machine learning (ML) methods that focus on learning predictive patterns from a dataset with the aim of making predictions (generalizing) on new data (Bzdok, 2017; Bzdok et al., 2018). Some of this work provides a different perspective on—and a new set of solutions to—the current limitations of MRI-analysis.

When using the MRI modality as part of an AI system, practitioners often prefer to use a set of hand-crafted, image-derived features, which are based on well established image markers (e.g., T2 lesion counts, brain volume). These are typically scalars derived from the voxel-level data, either manually, or through a semi-/fully-automated process. The values for these hand-crafted features, which are easy to interpret, can be stored in tabular form, and used to train a model for a specific task using a variety of ML methods. Alternatively, the raw voxel-level data can be provided directly as an input to ML models. Some types of ML, in particular deep learning (DL), which uses deep artificial neural networks (LeCun et al., 2015), can make use of the high information content in voxel-level data to *learn* (automatically, without explicit guidance from a human expert) abstract, lower-dimensional features of the image that might not be captured by traditional hand-crafted, image-derived features (e.g., the texture of the white matter in a certain brain region). A specific type of deep neural network called the convolutional neural network (CNN) (LeCun et al., 1989; Li et al., 2022) has significantly advanced digital image processing by automatically learning features from images, sometimes leading to superior performance in tasks like image classification and object detection. The theoretical benefits resulting from ML on raw images come at the cost of greater computational and dataset requirements (Berisha et al., 2021), and generally require more expertise in model training. Traditional, hand-crafted features therefore remain valuable, especially in scenarios with limited data or specific constraints (Lin et al., 2020; Zare et al., 2018; O'Mahony et al., 2019).

This review aims to introduce the reader to key areas in which AI is transforming MRI-analysis in MS (see Figure 1 for an overview). Given the vastness of the literature on this topic, this review is meant to provide a high-level overview of selected areas that are of interest to the MS community, showcasing published work on MS-specific applications. As such, this does not represent a comprehensive review of the literature. Where possible, we refer the reader to more in depth, dedicated reviews, in specific sections. First, we will explore how AI can be used for data collection (Section 2), before discussing the traditional tasks of lesion segmentation, detection, and volumetry (Section 3). Finally, we will discuss downstream scientific and clinical tasks (Sections 4, 5, and 6). We end with a discussion on promising future directions (Section 7).

2 Acquisition, pre-processing, and harmonization

MRI has become essential for diagnosing MS and for monitoring it's evolution, primarily because of its higher sensitivity compared to clinical outcome measures of disease activity (McDonald et al., 1994). To reap the benefits of routine monitoring with MRI while minimizing the inconvenience for patients, caregivers, and resource utilization, many have turned to AI to improve the efficiency of MRI data collection. In this section, we will discuss three tasks pertaining to MRI collection: (1) acquiring the MRI images (acquisition), (2) processing the acquired images to improve their signal-to-noise ratio (pre-processing), and (3) transforming the pre-processed images from different scanners/sites to enable direct comparisons (harmonization).

Shortening the MRI acquisition time can be achieved by decreasing the number of sequences in the acquisition protocol, using generative models to synthesize the missing sequences. For example, Wei et al. (2019) showed that it is indeed possible to use a CNN to predict the FLAIR sequence from T1-weighted, T2-weighted, proton density, T1 spin-echo, and double inversion recovery (DIR) sequences. Others provided evidence to suggest that Generative Adversarial Networks [GANs, Goodfellow et al. (2014)] can synthesize DIR from the combination of T1 and T2/FLAIR (Finck et al., 2020, 2022), and T1 from T2-weighted FLAIR (Valencia et al., 2022). Although synthesis of gadolinium-enhanced T1-weighted sequences from low or non-contrast images is under-explored in MS, related work by Narayana et al. (2020) found that the presence of gadolinium-enhancing lesions can be predicted with moderate accuracy from non-contrast MRI.

Another strategy to speed data collection is to acquire lower resolution images, or images with a higher signal-to-noise ratio, and then use ML models in the post-processing phase to reconstruct higher-quality images. Various DL frameworks based on GANs and CNNs have been shown to produce higher-quality reconstructions that can improve lesion visualization and segmentation (Shaul et al., 2020; Zhao et al., 2019; Iwamura et al., 2023; Mani et al., 2021; Falvo et al., 2019). DL has also been used to optimize the more complex processing pipelines used for diffusion weighted imaging sequences (Golkov et al., 2016).

Finally, ML-based harmonization strategies can be used to address a frequently encountered problem in biomedical imaging research: small dataset sizes. Aggregating data from different data collection sites is complicated by the fact that each site may use different scanners and acquisition protocols, resulting in images that do not look alike. This is known to cause variability in tasks such as volume estimation (Clark et al., 2023; Bakshi et al.,

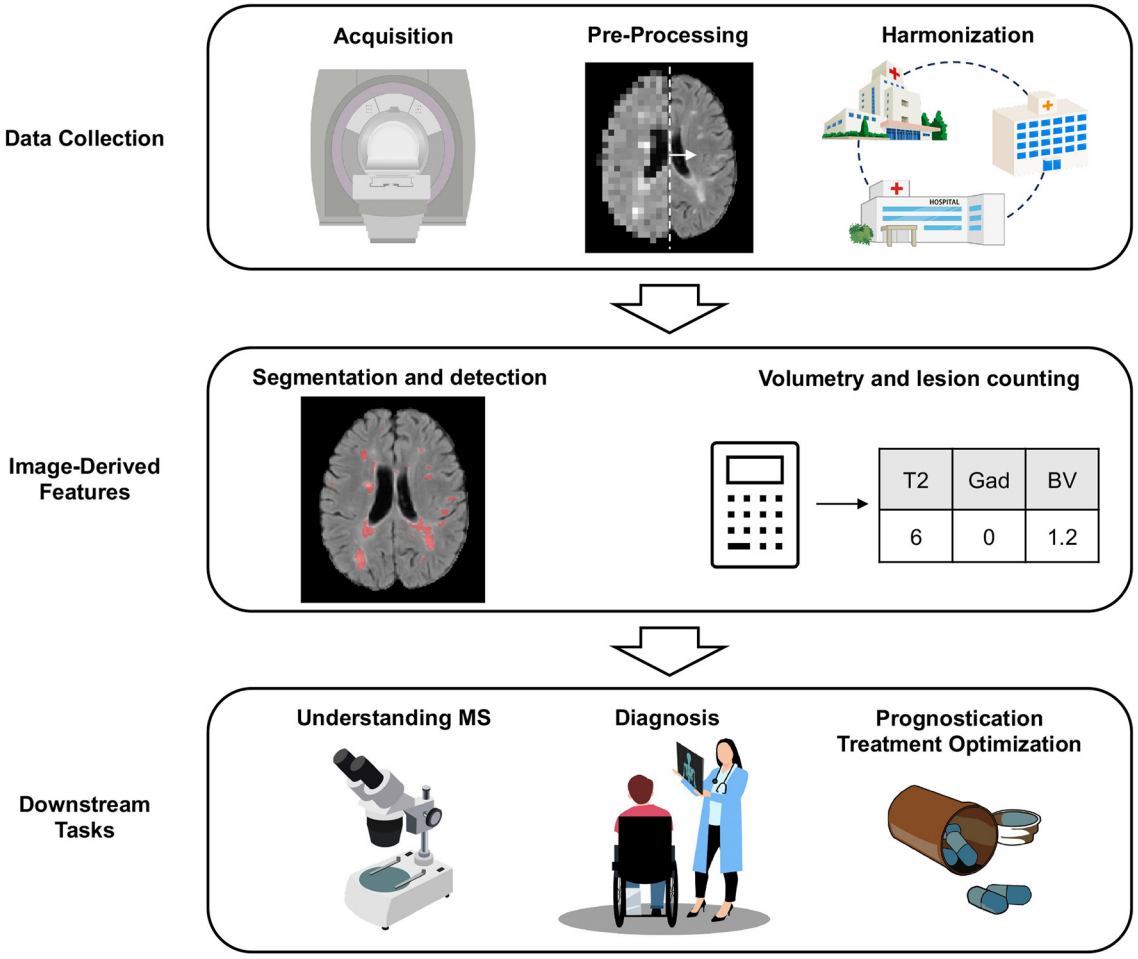


FIGURE 1
An overview of key areas in which AI is being used for MRI-analysis in MS.

2017). "Harmonization" is a solution to this problem that involves transforming the images so they all appear to come from the same distribution. Dewey et al. (2019) found benefits in the downstream task of brain volume estimation when images were first harmonized using DL. If direct visualization or comparisons between images from different datasets is not strictly necessary, one can also bypass the problem of harmonization by training models that are agnostic to the specific combination of sequences that is available for a particular patient (Havaei et al., 2016), or by searching for a set of hyperparameters that lead to comparable performance across a range of datasets (Gentile et al., 2023). It is worth noting that fake images can also be synthesized using DL to augment existing datasets. This is an open research problem and the magnitude of benefit probably depends on the context (Van Tulder and de Bruijne, 2015). Relatively little published research explores MRI generation specifically for MS datasets, but some authors have observed performance gains from augmentation with lesion-containing MRI images that are synthetically generated from the MRI images of healthy subjects (Salem et al., 2019; Basaran et al., 2022).

In summary, AI has shown promise in reducing the time taken to acquire and preprocess the MRI of MS patients, without significantly compromising the quality and utility of the MRI images. AI can also increase the ease with which data from different sources can be pooled together for further analysis, or for increasing the size of datasets which ML models use for training. Many of the methods that were reviewed in this section are at an early stage of development, and these tasks remain an active area of research.

3 Segmentation, lesion detection, and volumetry

Once a patient's MRI has been acquired and pre-processed, it is then ready to be used for clinical management and scientific research. Although the raw, voxel-level data can be fed directly as input to a ML model that is specifically trained for one of the downstream tasks described in Sections 4, 5, 6, there is often added value to taking an intermediate step consisting of identifying and quantifying established radiologic features in the images. These tasks include segmenting radiologic markers of MS, lesion detection, and the volumetric assessment of a variety of brain structures.

Current cross-sectional disease burden assessment typically consists of some variant on lesion volume, lesion count, and brain volume estimation. Monitoring of disease activity over time additionally calls for comparing volume estimates between time-points, and the detection of new or enlarging lesions. In most settings where radiologists and neurologists are responsible for performing these tasks, volume estimation is done qualitatively with high-level descriptors, while lesion detection is done using manual review of 2D slices. The process is lengthy, error-prone, and subject to significant inter- and intra-rater variability (Bozsik et al., 2022; Altay et al., 2013). For these reasons, there has been a growing appetite for at least partially automating these tasks using AI.

The segmentation of T2 lesions is one of the most well studied applications of ML in MS. The literature on automated MS lesion segmentation methods is vast, and methods range from classical ML to DL. We therefore refer the interested reader to several dedicated reviews for more details (García-Lorenzo et al., 2013; Danelakis et al., 2018; Spagnolo et al., 2023; Zeng et al., 2020; Doyle et al., 2018). There has been relatively less work on new (and/or enlarging) T2 lesion segmentation, but more emphasis has been placed on this task during recent challenges (Commowick et al., 2021). Beyond T2 hyper-intense lesions, DL has also been used to segment and detect imaging markers which are not currently integrated in most clinical settings. These include paramagnetic rim lesions (Barquero et al., 2020; Lou et al., 2021; Zhang et al., 2022), central vein sign on susceptibility-weighted images (Maggi et al., 2020), cortical lesions on 7T images (Rosa et al., 2022; La Rosa et al., 2020), gadolinium-enhancing lesions (Gaj et al., 2021; Karimaghaloo et al., 2010; Durso-Finley et al., 2020), and spinal cord lesions (Gros et al., 2019). The task of detecting lesions (including the detection of new lesions on follow-up images) has for the most part been studied in tandem with segmentation (Kamraoui et al., 2022; Salem et al., 2020; McKinley et al., 2020).

Although brain (parenchymal) volumetry has received less attention, DL has been used to segment the thalami of MS patients for the purpose of estimating its volume (Dwyer et al., 2021). DL methods have also been shown to perform well when compared to traditional methods for brain atrophy estimation (Zhan et al., 2023). Moreover, DL-based lesion-filling (or inpainting) has been shown to improve the performance of volumetric estimation methods that are usually sensitive to the presence of lesional tissue (Zhang et al., 2020; Clèrigues et al., 2023). Unfortunately, the large minimal detectable change in volume between clinically relevant intervals and the high inter-scanner variability still limit the utility of brain volume estimation in the clinic (Van Nederpelt et al., 2023). It is worth noting that a number of software packages for automated volumetric analysis and segmentation are available, and some already include DL methods (Billot et al., 2023).

Several challenges have been organized, in which groups compete for best performance on the same lesion segmentation task (either T2 lesion or new T2 lesion segmentation). These were hosted at the IEEE ISBI conference (Carass et al., 2017) and at MICCAI conferences (Styner et al., 2008; Commowick et al., 2018, 2021). In all cases, no model was found to be perfect, when evaluated on the basis of voxel-level segmentation metrics (under or over-segmentation) and lesion detection metrics (e.g., false positive rate), in comparison to the ground-truth segmentation obtained by human expert raters. Rather than indicative of a failure of ML for automatic segmentation, we argue that this finding should lead the community to rethink the way models are evaluated. In all challenges, performance was measured against the segmentation masks obtained from very few human experts, and on relatively small datasets of at most one hundred participants. Despite these challenge's best attempts to address the intra and inter-rater variability associated with the ground-truth lesion masks obtained from human experts (Bozsik et al., 2022; Altay et al., 2013), there remains no accepted consensus on what should constitute "ground truth". Where should one draw the lesion border, given that lesional tissue manifests as a continuous spectrum of intensity on MRI? How do we differentiate an enlarging lesion from confluent new lesions? How do we know if hyperintensities smaller than 3 mm [which are typically disregarded by expert raters (Filippi et al., 2019) to avoid false positive detections], are pathologically significant or not? Without answers to all these questions, finding that DL methods disagree with human experts is arguably insufficient to determine if they are truly inferior. To address this issue, some have proposed explicitly modeling the "label-style" that might be associated with a certain dataset or group of expert-raters (Nichyporuk et al., 2022). Others have avoided the use of ground-truth lesion masks altogether by framing lesion segmentation as an unsupervised anomaly detection task (Behrendt et al., 2023; Castellano et al., 2022; Luo et al., 2023; Pinaya et al., 2022). Training on soft-labels (as opposed to binary labels) (Gros et al., 2021; Lemay et al., 2022) and probabilistic lesion counting (Schroeter et al., 2022) are yet other possible solutions. In recognition of the importance of the problem of model evaluation in the case of image analysis, a large international consortium has recently published recommendations for model evaluation (Maier-Hein et al., 2024; Reinke et al., 2024). Still, more work has to be done to obtain answers to the problems specific to MS lesion segmentation.

To conclude, segmentation, lesion detection, and volumetry, are some of the oldest and most studied ML application in MS. In many cases, they reach performances that are acceptable for many clinical and research settings. More work is needed to determine how best to evaluate automated segmentation frameworks.

4 Improving our understanding of MS

With an increasing number of datasets containing MRI images of MS patients, and the plethora of open questions in MS research, one may ask: could AI help us uncover novel markers of MS diagnosis, evolution, and treatment response? For years, patients with MS have been categorized into a binary classification system consisting of relapsing-remitting and progressive clinical phenotypes (Lublin and Reingold, 1996). It was later found that significant overlap exists in disease evolution across these subtypes, prompting the introduction of subtype-agnostic evolution-focused terminology such as "relapse-associated worsening (RAW)" and "progression independent of relapse-activity (PIRA)" (Lublin et al., 2022). The current most accepted perspective is that individual differences in disease course can be traced back to different combinations of inflammatory, neurodegenerative, and

compensatory processes that lie along a continuous spectrum (Lassmann, 2019; Pitt et al., 2022; Vollmer et al., 2021).

This paradigm-shift, coupled with the fact that none of the existing MRI biomarkers have been particularly predictive of the key clinical outcome of disability progression (Filippi et al., 2020), has led researchers to search for alternative MRI-markers that could better explain the observed heterogeneity in disease evolution and treatment response. Notably, Eshaghi et al. (2021) and Pontillo et al. (2022) used an unsupervised ML algorithm called SuStaIn (Young, 2018) to identify disease subtypes characterized by distinct temporal progression patterns on MRI. Both groups found subtypes characterized by early cortical or deep gray matter atrophy, early signal changes in normal appearing white matter, and early T2 lesion accumulation. More work is needed to externally validate these subtypes and better understand their clinical correlates.

ML has also been used more directly to assist scientists in uncovering novel MRI markers. One strategy involves taking a pre-trained classifier (e.g., a model trained to predict MS diagnosis, or future disease activity) and producing "saliency-maps". These allow researchers to visualize the features that are thought to be "important" according to the classifier; for example, features associated with a diagnosis of MS, poorer prognosis, or specific phenotypes. By using heatmaps generated using layer-wise relevance propagation, Eitel et al. (2019) found that a CNN classifier pre-trained to predict MS diagnosis focused on T2-lesions and their location, along with non-lesional or gray matter areas that included the thalamus. Storelli et al. (2022) produced heatmaps from a CNN that was trained to predict EDSS-worsening, and identified differences in periventricular regions, white matter lesions and the corpus callosum, for EDSS-worsened patients. Zhang et al. (2021) interrogated different heatmap-generating techniques to better understand crucial brain regions that could help distinguish MS phenotypes, finding that the abnormalities associated with SPMS were more extensive compared to RRMS, the latter involving primarily the occipital region and, to a lesser extent, the frontal region. Finally, Kumar et al. (2022) proposed to identify candidate biomarkers of future new/enlarging T2 lesions in an RRMS population through a process called counterfactual image synthesis; specifically, by predicting how a patient's MRI would look like if they had a different future outcome (a counterfactual), and by taking the difference between the real (factual) and counterfactual images, markers that are predictive of future outcomes (in this case, lesion activity) can be revealed.

AI can therefore be useful to better understand disease evolution and heterogeneity. While exciting, this work remains largely at the level of methodological development, and more translational research will be needed.

5 Diagnosis

It is imperative that an MS diagnosis be confirmed rapidly, and accurately, to ensure that patients receive the best possible care. MS is currently diagnosed according to the 2017 McDonald criteria, which combines historical, MRI, and laboratory data (Thompson et al., 2018). While significant efforts have been made to accelerate MS diagnosis, the heterogeneity of the disease and broad differential diagnosis still continues to put the clinician at risk of misdiagnoses, which can delay the initiation of an adequate treatment (Solomon et al., 2019; Brownlee and Solomon, 2021). Recent diagnostic criteria might provide increased sensitivity for the diagnosis, but at the cost of reduced specificity (Mescheriakova et al., 2018; Habek et al., 2018). In this section, we will discuss the use of AI for improving the accuracy and reliability of MS diagnosis. Note that there is some overlap with Section 3, since the detection of MS lesions on MRI is an important component of the diagnostic criteria (but not the only one). In the current section, the focus will be on the classification task of MS diagnosis, with the understanding that automated lesion segmentation and detection methods could be used upstream to provide image-derived features to an MS classifier.

Both classical ML and DL methods have been applied to the task of MS diagnosis, with MRI being the most common input modality for the classifier [we refer the reader to dedicated reviews on this topic for more details (Nabizadeh et al., 2022; Aslam et al., 2022; Shoeibi et al., 2021)]. Reported diagnostic sensitivity, and especially specificity, can be quite high [pooled sensitivity 92% (95%CI: 90%, 95%) and specificity 93% (95%CI: 90%, 96%), respectively, according to a recent meta-analysis (Nabizadeh et al., 2023)]. Even simple image-derived scalars such as the average of T1, T2*, and the total/myelin bound water content, have been found to be highly predictive (when used as input to train a supervised ML classifier) of an MS diagnosis (Neeb et al., 2019).

Differentiating MS from other diseases that can mimic it's presentation is also an important task in the clinic. Rocca et al. (2021) used a basic 3D-CNN with MRI as input to differentiate MS from neuromyelitis optica spectrum disorder (NMOSD), central nervous system vasculitis, and migraine, and found that the diagnostic accuracy exceeded that of human experts. Similarly, Kim et al. (2020) showed that MS could be differentiated from NMOSD using a 3D-CNN based on the ResNet architecture (He et al., 2016), as accurately as two neurologists. Huang et al. (2022) found that a transformer-based image classifier (Xu et al., 2021) could differentiate MS from NMOSD and myelin oligodendrocyte glycoprotein antibody disease as accurately as two neuroradiologists. MS could also be differentiated from hereditary diffuse leukodystrophy with spheroids using linear discriminant analysis (Mangeat et al., 2020), and from low grade tumors using MR-spectroscopy-derived features as input to a variety of ML models (Ekşi et al., 2021; Preul et al., 1996).

Overall, there is a growing amount of evidence supporting the use of AI in MS diagnosis.

6 Prognostication and treatment optimization

One of the main challenges for the clinician evaluating a patient with a new diagnosis of MS is to predict long-term prognosis (the evolution of the disease over time). The related task of treatment optimization (predicting which treatment will have the most beneficial effect) often depends on having an accurate prognosis. This begs the question: can AI do any better? Many early research efforts were focused on predicting the occurrence or timing of clinically-defined MS subtype transitions, using these as surrogate

markers of poor prognosis. However, as discussed in Section 4, there has been a tendency to de-emphasize these subtypes in the diagnosis and management of MS. Prognostication tasks that we will focus on in this section therefore involve the prediction of the evolution of specific manifestations of the disease, which include radiologic activity (new/enlarging T2 lesions), relapses, disability accumulation, and patient-reported outcomes.

Prognostication with respect to disability outcomes turns out to be a very challenging task, even for AI (Seccia et al., 2021). When predicting disability progression from hand-crafted, image-derived tabular features, Pellegrini et al. (2020) found that a variety of classical ML models could achieve only modest predictive performance (C-index $\leq$ 0.65). Nonetheless, predictive performance can vary greatly depending on what features are used as input, on the model, and on the optimization procedure. With regards to the input, Zhao et al. (2017) found that classical ML methods performed better when adding image-derived features from a 1-year follow-up MRI visit to the set of inputs, which otherwise consisted of data recorded at a baseline visit. The benefit of longitudinal follow-up was also highlighted in work that used SuStaIn (Young, 2018) for unsupervised temporal modeling of imaging trajectories. Specifically, Pontillo et al. (2022) were able to identify a "deep-gray-matter-first" subtype that was associated with long-term cognitive impairment, and Eshaghi et al. (2021) could identify a "lesion-led" subtype that was associated with both confirmed disability progression and relapse rate. Using long term clinical (non-imaging) follow-up data has also been shown to lead to a considerable performance boost when predicting progression (De Brouwer et al., 2021). All this evidence suggests that ML on longer-term MRI data represents a promising, though challenging, research direction.

With regards to the model type, Zhao et al. (2020) found that ensembles of gradient-boosted trees such as XGBoost and LightGBM performed better than alternative ML methods when predicting 5-year EDSS worsening from logitudinal data collected over 2 years, with an area under the curve (AUC) ranging from 0.79 to 0.83. Interestingly, their feature importance analysis [and that of others (Law et al., 2019)] suggests that clinical disability metrics (which includes the EDSS) might be more predictive than tabular image-derived features for this particular task.

It is possible that voxel-level MRI data, which has been understudied for the task of predicting clinical prognosis, could harbor more predictive features of prognosis than traditional image-derived features. In support of this hypothesis, Storelli et al. (2022) were able to train a CNN to predict 2-year EDSS and SDMT worsening with 75.0% sensitivity, and 87.5% specificity. It is also possible that non-trivial implementation details, such as the inclusion of a T2-lesion mask along with the raw MRI as input, could further boost performance (Tousignant et al., 2019). These studies hint at DL's potential to improve upon tabular, hand-crafted, image-derived features (e.g., T2 lesion volume). In an attempt to elucidate the relative contribution of voxel-level data to predicting disability progression Zhang et al. (2023) studied a dataset of 300 MS patients, with a very large feature set spanning numerous MRI sequences, laboratory data, demographic information, disability scores, and unstructured clinical notes. Imaging, tabular data, and notes were encoded and fused using various neural network architectures, and used for predicting EDSS milestones 3-years later. While their best performing model made use of all three modalities (AUC 0.8380), a model trained without the MRI modality was only marginally worse (AUC 0.8078). Their study is limited by a small dataset size, with a comparatively large feature set, which could result in poor model optimization. More research is therefore needed to explore this important question, but this will require larger datasets, and additional methodological advances.

DL has also been used on radiologic markers of disease activity, which in certain cases are more sensitive to disease evolution than clinical measurements. A few studies have shown promising preliminary results in predicting the future appearance of new/enlarging T2 lesions from baseline MRI (Prabhakar et al., 2023; Durso-Finley et al., 2023, 2022). Tabular, hand-crafted image-derived features have also been used to classify a lesion as active or inactive (Peng et al., 2021). Similar to the task of predicting clinical prognosis (which focuses on predicting future disability-related outcomes), there remains the possibility that non-trivial methodological contributions may yield significant performance gains.

AI tools that aid in prognostication can be used for treatment optimization (for example, by favoring a more potent drug for a patient predicted to have highly active disease); however, it is also useful to consider the related task of estimating the "treatment effect" of a medication on the disease course. The most common treatment effect estimand that clinicians consider as part of treatment-related decisions is the *average* treatment effect, which typically is estimated using randomized clinical trials, and represents the average effect of a treatment on a population (compared to placebo or to a baseline drug). Some of the ML research cited in previous sections have presented results pertaining to treatment effect estimation. For example, the "lesion-led" subtype discovered by Eshaghi et al. (2021) appears to be specify a sub-group of individuals that experience a larger average treatment effect. Another line of work in causal ML aims to personalize treatment recommendations by predicting the treatment effect for a particular *individual* given their unique characteristics (Curth et al., 2024). For example, Durso-Finley et al. (2022) proposed a multi-headed CNN to predict the individual treatment effect of several treatments on new/enlarging T2-lesions, which used a person's MRI as input. Beyond treatment optimization, individual treatment effect estimation could also play a role in improving the statistical power of clinical trials by preferentially randomizing individuals who are predicted to benefit from an experimental therapy (Falet et al., 2022; Kanber et al., 2019).

In conclusion, although prognostication and treatment optimization remain challenging tasks, MRI-based ML research continues to improve upon previous baselines through diverse methodological innovations. Some models appear to identify subgroups of individuals that are more responsive to certain disease modifying therapies. These results are therefore paving the path toward precision medicine.

7 Discussion

In this review, we have presented several tasks where AI systems might already reliably outperform human experts in MS-specific applications. Indeed, a recent validation study by

Barnett et al. (2023) provided evidence supporting the use of AI tools for lesion detection and volumetric analyses, in both clinical settings and research studies. We also discussed tasks which are hardly feasible without recent advances in DL, such as MRI sequence synthesis and automated biomarker discovery.

As the performance of AI tools continues to improve, we will arguably see increasing interest in trustworthiness, because these AI systems are expected to take part in high-risk human decision-making. Trust in AI systems is built in numerous ways, one of which is by giving them the ability to explain the rationale behind a model's predictions, resulting in "explainable" AI systems (Došilović et al., 2018). Additionally, users should be aware of the level of confidence that a model has in a particular prediction, and how much this reflects the actual errors that a model might make. This line of work, often referred to as "uncertainty" estimation (and the related problem of calibration), allows users to know when to trust a model's predictions (Gawlikowski et al., 2023). In addition, to trust that a model will behave well in practice, there should be a good understanding of how it will generalize to new data, and whether or not it will be robust to distribution shifts (for example, if there is a change in acquisition protocol). The field of causal machine learning (Sanchez et al., 2022), which models the data generative process using causal models, promises improved out-of-distribution generalization, and represents an active field of research. MS researchers have begun to address all three of these topics, specifically explainable methods (see examples in Section 4), probabilistic modeling for uncertainty estimation (Nair et al., 2020; Durso-Finley et al., 2023), and structural causal models of MRI image generation (Reinhold et al., 2021), but more work is needed to truly enable trustworthy AI-assisted MRI analysis in MS.

Looking forward, it seems clear that highly capable AI systems based on large foundation models (Brown et al., 2020; Devlin et al., 2018; Touvron et al., 2023; Ramesh et al., 2021) will have a major impact on biomedical imaging research, including in MS. Certain chat-bots based on large language models (LLMs) can now arguably pass the Turing test (Jannai et al., 2023), and score higher than the average human on medical exams (Achiam, 2023). LLMs are increasingly being used in medical applications (Agbavor and Liang, 2022; Patel and Lam, 2023; Singhal et al., 2023; Jiang et al., 2023), and multi-modal inputs (which includes biomedical imaging) are becoming more common (Moor et al., 2023). Although foundation models remain understudied in MS applications, interesting future directions include using foundation models to improve generalization from small MS-specific datasets, through in-context learning (Dong et al., 2024), or fine-tuning. That said, in order to reap all the benefits of foundation models for MS-specific applications, several open problems need to be solved. These include sub-par reasoning capabilities (Rae et al., 2021; McKenzie et al., 2023; Arkoudas, 2023) which could be dangerous in high-stakes environments such as healthcare (Richens et al., 2020; Fraser et al., 2018), broader concerns regarding AI safety (Bommasani et al., 2021; Anderljung et al., 2023; Urbina et al., 2022), and predictions that may be unacceptably skewed to the detriment of a particular group of people (Mehrabi et al., 2021). As more solutions to these problems are found, we can expect an increasing focus on large foundation models in the coming years, to help solve some of the most challenging tasks in MS MRI-analysis.

Author contributions

J-PF: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. SN: Writing – original draft, Writing – review & editing. AS: Writing – original draft, Writing – review & editing, Visualization. BB: Writing – original draft, Writing – review & editing. AK: Writing – original draft, Writing – review & editing. JD-F: Writing – original draft, Writing – review & editing. TA: Supervision, Writing – original draft, Writing – review & editing. DA: Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. TA acknowledges support from the Canada Institute for Advanced Research (CIFAR) AI Chairs program and the Natural Sciences and Engineering Research Council of Canada. J-PF acknowledges support from the Fonds de recherche du Québec–Santé/Ministère de la Santé et des Services sociaux and was a recipient of the Vanier Canada Graduate Scholarships Doctoral Award (CGV-192746).

Conflict of interest

DA reports consulting fees from Biogen, Celgene, Frequency Therapeutics, Genentech, Merck, Novartis, Race to Erase MS, Roche, Sanofi–Aventis, Shionogi, and Xfacto Communications, grants from Immunotec and Novartis and an equity interest in NeuroRx.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv preprint* arXiv:2303.08774.

Agbavor, F., and Liang, H. (2022). Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health* 1:e0000168. doi: 10.1371/journal.pdig.0000168

Altay, E. E., Fisher, E., Jones, S. E., Hara-Cleaver, C., Lee, J. C., and Rudick, R. A. (2013). Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol.* 70, 338–344. doi: 10.1001/2013.jamaneurol.211

Amin, M., and Hersh, C. M. (2023). Updates and advances in multiple sclerosis neurotherapeutics. *Neurodegener. Dis. Manag.* 13, 47–70. doi: 10.2217/nmt-2021-0058

Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., et al. (2023). Frontier ai regulation: Managing emerging risks to public safety. *arXiv* [preprint] arXiv:2307.03718. doi: 10.48550/arXiv.2307.03718

Arkoudas, K. (2023). GPT-4 can't reason. *arXiv* [preprint] arXiv:2308.03762. doi: 10.48550/arXiv.2308.03762

Aslam, N., Khan, I. U., Bashamakh, A., Alghool, F. A., Aboulnour, M., Alsuwayan, N. M., et al. (2022). Multiple sclerosis diagnosis using machine learning and deep learning: challenges and opportunities. *Sensors* 22:7856. doi: 10.3390/s22207856

Bakshi, R., Roy, S., Stern, W., Tummala, S., Yousuf, F., Zhu, A., et al. (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *Am. J. Neuroradiol.* 38:1501–1509. doi: 10.3174/ajnr.A5254

Barnett, M., Wang, D., Beadnall, H., Bischof, A., Brunacci, D., Butzkueven, H., et al. (2023). A real-world clinical validation for ai-based MRI monitoring in multiple sclerosis. *NPJ Digital Med.* 6:196. doi: 10.1038/s41746-023-00940-6

Barquero, G., La Rosa, F., Kebiri, H., Lu, P.-J., Rahmanzadeh, R., Weigel, M., et al. (2020). Rimnet: A deep 3D multimodal MRI architecture for paramagnetic RIM lesion assessment in multiple sclerosis. *NeuroImage: Clinical* 28:102412. doi: 10.1016/j.nicl.2020.102412

Basaran, B. D., Qiao, M., Matthews, P. M., and Bai, W. (2022). "Subject-specific lesion generation and pseudo-healthy synthesis for multiple sclerosis brain images," in *International Workshop on Simulation and Synthesis in Medical Imaging* (Cham: Springer), 1–11.

Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., and Schlaefer, A. (2023). "Patched diffusion models for unsupervised anomaly detection in brain MRI," in *Proceedings of Machine Learning Research-Preprint* (Athens: IEEE), 1–14.

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., et al. (2021). Digital medicine and the curse of dimensionality. *NPJ Digital Med.* 4:153. doi: 10.1038/s41746-021-00521-5

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., et al. (2023). Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* 86:102789. doi: 10.1016/j.media.2023.102789

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv* [preprint] arXiv:2108.07258. doi: 10.48550/arXiv.2108.07258

Bozsik, B., Tóth, E., Polyák, I., Kerekes, F., Szabó, N., Bencsik, K., et al. (2022). Reproducibility of lesion count in various subregions on MRI scans in multiple sclerosis. *Front. Neurol.* 13:843377. doi: 10.3389/fneur.2022.843377

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901. doi: 10.48550/arXiv.2005.14165

Brownlee, W. J., and Solomon, A. J. (2021). Misdiagnosis of multiple sclerosis: time for action. *Multiple Sclerosis J.* 27:805–806. doi: 10.1177/13524585211005367

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Front. Neurosci.* 11:543. doi: 10.3389/fnins.2017.00543

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nat. Methods* 15:233. doi: 10.1038/nmeth.4642

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *Neuroimage* 148, 77–102. doi: 10.1016/j.dib.2017.04.004

Castellano, G., Placidi, G., Polsinelli, M., Tulipani, G., and Vessio, G. (2022). "Unsupervised brain MRI anomaly detection for multiple sclerosis classification," in *International Conference on Pattern Recognition* (Cham: Springer), 644–652.

Clark, K. A., O'Donnell, C. M., Elliott, M. A., Tauhid, S., Dewey, B. E., Chu, R., et al. (2023). Intersite brain MRI volumetric biases persist even in a harmonized multisubject study of multiple sclerosis. *J. Neuroimag.* 33, 941–952. doi: 10.1111/jon.13147

Clèrigues, A., Valverde, S., Salvi, J., Oliver, A., and Lladó, X. (2023). Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry. *Comp. Med. Imag. Graph.* 103:102157. doi: 10.1016/j.compmedimag.2022.102157

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention* (Strasbourg: Hal Open Science), 126.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8:13650. doi: 10.1038/s41598-018-31911-7

Compston, A., and Coles, A. (2008). Multiple sclerosis. *Lancet*. 372:1502–1517. doi: 10.1016/S0140-6736(08)61620-7

Curth, A., Peck, R. W., McKinney, E., Weatherall, J., and van der Schaar, M. (2024). Using machine learning to individualize treatment effect estimation: challenges and opportunities. *Clini. Pharmacol. Therapeut.* 115, 710–719. doi: 10.1002/cpt.3159

Danelakis, A., Theoharis, T., and Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comp. Med. Imag. Graphics* 70, 83–100. doi: 10.1016/j.compmedimag.2018.10.002

De Brouwer, E., Becker, T., Moreau, Y., Havrdova, E. K., Trojano, M., Eichau, S., et al. (2021). Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. *Comput. Methods Programs Biomed.* 208:106180. doi: 10.1016/j.cmpb.2021.106180

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint] arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., et al. (2019). Deepharmony: a deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* 64, 160–170. doi: 10.1016/j.mri.2019.05.041

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., et al. (2024). A survey on in-context learning. *arXiv* [preprint] arXiv:2301.00234. doi: 10.18653/v1/2024.emnlp-main.64

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). "Explainable artificial intelligence: a survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (Opatija: IEEE), 0210–0215.

Doyle, A., Elliott, C., Karimaghaloo, Z., Subbanna, N., Arnold, D. L., and Arbel, T. (2018). "Lesion detection, segmentation and prediction in multiple sclerosis clinical trials," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017* (Quebec City: Springer), 15–28.

Durso-Finley, J., Arnold, D. L., and Arbel, T. (2020). "Saliency based deep neural network for automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019* (Shenzhen: Springer), 108–118.

Durso-Finley, J., Falet, J.-P., Mehta, R., Arnold, D. L., Pawlowski, N., and Arbel, T. (2023). "Improving image-based precision medicine with uncertainty-aware causal models," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023* (Vancouver: MICCAI 2023),14224.

Durso-Finley, J., Falet, J.-P. R., Nichyporuk, B., Arnold, D. L., and Arbel, T. (2022). "Personalized prediction of future lesion activity and treatment effect in multiple sclerosis from baseline MRI," in *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning* (New York: PMLR), 172.

Dwyer, M., Lyman, C., Ferrari, H., Bergsland, N., Fuchs, T. A., Jakimovski, D., et al. (2021). DeepGRAI (Deep Gray Rating via Artificial Intelligence): fast, feasible, and clinically relevant thalamic atrophy measurement on clinical quality T2-FLAIR MRI in multiple sclerosis. *NeuroImage Clin.* 30, 102652. doi: 10.1016/j.nicl.2021.102652

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A. U., Ruprecht, K., Giess, R. M., et al. (2019). Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical* 24:102003. doi: 10.1016/j.nicl.2019.102003

Ekşi, Z., Özcan, E., Çakıroğlu, M., Öz, C., and Aralaşmak, A. (2021). Differentiation of multiple sclerosis lesions and low-grade brain tumors on MRS data: machine learning approaches. *Neurol Sci* 42, 389–3395. doi: 10.1007/s10072-020-04950-0

Eshaghi, A., Young, A. L., Wijeratne, P. A., Prados, F., Arnold, D. L., Narayanan, S., et al. (2021). Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat. Commun.* 12:2078. doi: 10.1038/s41467-021-22265-2

Falet, J.-P. R., Durso-Finley, J., Nichyporuk, B., Schroeter, J., Bovis, F., Sormani, M.-P., et al. (2022). Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nat. Commun.* 26:1. doi: 10.1038/s41467-022-33269-x

Falvo, A., Comminiello, D., Scardapane, S., Scarpiniti, M., and Uncini, A. (2019). "A multimodal dense U-Net for accelerating multiple sclerosis MRI," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (Pittsburgh, PA: IEEE), 1–6.

Filippi, M., and Agosta, F. (2010). Imaging biomarkers in multiple sclerosis. *J. Magnetic Res. Imag.* 31, 770–788. doi: 10.1002/jmri.22102

Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., et al. (2019). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain* 142, 1858–1875. doi: 10.1093/brain/awz144

Filippi, M., Preziosa, P., Langdon, D., Lassmann, H., Paul, F., Àlex Rovira, S.choonheim, M. M., et al. (2020). Identifying progression in multiple sclerosis: new perspectives. *Ann. Neurol.* 88, 438–452. doi: 10.1002/ana.25808

Finck, T., Li, H., Grundl, L., Eichinger, P., Bussas, M., Mühlau, M., et al. (2020). Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection. *Invest. Radiol.* 55, 318–323. doi: 10.1097/RLI.0000000000000640

Finck, T., Li, H., Schlaeger, S., Grundl, L., Sollmann, N., Bender, B., et al. (2022). Uncertainty-aware and lesion-specific image synthesis in multiple sclerosis magnetic resonance imaging: a multicentric validation study. *Front. Neurosci.* 16:889808. doi: 10.3389/fnins.2022.889808

Fraser, H., Coiera, E., and Wong, D. (2018). Safety of patient-facing digital symptom checkers. *Lancet* 392, 2263–2264. doi: 10.1016/S0140-6736(18)32819-8

Gaj, S., Ontaneda, D., and Nakamura, K. (2021). Automatic segmentation of gadolinium-enhancing lesions in multiple sclerosis using deep learning from clinical MRI. *PLoS ONE* 16:e0255939. doi: 10.1371/journal.pone.0255939

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.* 17, 1–18. doi: 10.1016/j.media.2012.09.004

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al. (2023). A survey of uncertainty in deep neural networks. *Artif. Intellig. Rev.* 56, 1513–1589. doi: 10.1007/s10462-023-10562-9

Gentile, G., Jenkinson, M., Griffanti, L., Luchetti, L., Leoncini, M., Inderyas, M., et al. (2023). Bianca-ms: An optimized tool for automated multiple sclerosis lesion segmentation. *Hum. Brain Mapp.* 44, 4893–4913. doi: 10.1002/hbm.26424

Golkov, V., Dosovitskiy, A., Sperl, J. I., Menzel, M. I., Czisch, M., Sämann, P., et al. (2016). Q-space deep learning: twelve-fold shorter and model-free diffusion MRI scans. *IEEE Trans. Med. Imaging* 35, 1344–1351. doi: 10.1109/TMI.2016.2551324

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *arXiv* [preprint] arXiv.1406.2661. doi: 10.48550/arXiv.1406.2661

Gros, C., De Leener, B., Badji, A., Maranzano, J., Eden, D., Dupont, S. M., et al. (2019). Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *Neuroimage* 184, 901–915. doi: 10.1016/j.neuroimage.2018.09.081

Gros, C., Lemay, A., and Cohen-Adad, J. (2021). Softseg: Advantages of soft versus binary training for image segmentation. *Med. Image Anal.* 71:102038. doi: 10.1016/j.media.2021.102038

Habek, M., Pavičić, T., Ruška, B., Pavlović, I., Gabelić, T., Barun, B., et al. (2018). Establishing the diagnosis of multiple sclerosis in croatian patients with clinically isolated syndrome: 2010 versus 2017 mcdonald criteria. *Mult. Scler. Relat. Disord.* 25, 99–103. doi: 10.1016/j.msard.2018.07.035

Havaei, M., Guizard, N., Chapados, N., and Bengio, Y. (2016). "Hemis: Hetero-modal image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference* (Athens: Springer), 469–477.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778.

Hemond, C. C., and Bakshi, R. (2018). Magnetic resonance imaging in multiple sclerosis. *Cold Spring Harb. Perspect. Med.* 8:5. doi: 10.1101/cshperspect.a028969

Huang, C., Chen, W., Liu, B., Yu, R., Chen, X., Tang, F., et al. (2022). Transformer-based deep-learning algorithm for discriminating demyelinating diseases of the central nervous system with neuroimaging. *Front. Immunol.* 13:897959. doi: 10.3389/fimmu.2022.897959

Iwamura, M., Ide, S., Sato, K., Kakuta, A., Tatsuo, S., Nozaki, A., et al. (2023). Thin-slice two-dimensional T2-weighted imaging with deep learning-based reconstruction: improved lesion detection in the brain of patients with multiple sclerosis. *Magnetic Res. Med. Sci.* 2022:0112. doi: 10.2463/mrms.mp.2022-0112

Jannai, D., Meron, A., Lenz, B., Levine, Y., and Shoham, Y. (2023). Human or not? A gamified approach to the turing test. *arXiv* [preprint] arXiv:2305.20010. doi: 10.48550/arXiv.2305.20010

Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., et al. (2023). Health system-scale language models are all-purpose prediction engines. *Nature* 2023, 1–6. doi: 10.1038/s41586-023-06160-y

Kamraoui, R. A., Mansencal, B., Manjon, J. V., and Coupé, P. (2022). Longitudinal detection of new MS lesions using deep learning. *Front. Neuroimag.* 1:948235. doi: 10.3389/fnimg.2022.948235

Kanber, B., Nachev, P., Barkhof, F., Calvi, A., Cardoso, J., Cortese, R., et al. (2019). High-dimensional detection of imaging response to treatment in multiple sclerosis. *NPJ Digital Med.* 2:49. doi: 10.1038/s41746-019-0127-8

Karimaghaloo, Z., Shah, M., Francis, S. J., Arnold, D. L., Collins, D. L., and Arbel, T. (2010). "Detection of gad-enhancing lesions in multiple sclerosis using conditional random fields," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2010: 13th International Conference* (Beijing: Springer), 41–48.

Kim, H., Lee, Y., Kim, Y.-H., Lim, Y.-M., Lee, J. S., Woo, J., et al. (2020). Deep learning-based method to differentiate neuromyelitis optica spectrum disorder from multiple sclerosis. *Front. Neurol.* 11:599042. doi: 10.3389/fneur.2020.599042

Koopmans, R. A., Li, D. K., Oger, J. J., Mayo, J., and Paty, D. W. (1989). The lesion of multiple sclerosis: Imaging of acute and chronic stages. *Neurology* 39, 959–963. doi: 10.1212/WNL.39.7.959

Kumar, A., Hu, A., Nichyporuk, B., Falet, J.-P. R., Arnold, D. L., Tsaftaris, S., et al. (2022). "Counterfactual image synthesis for discovery of personalized predictive image markers," in *MICCAI Workshop on Medical Image Assisted Biomarkers'* Discovery (Springer), 113–124.

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P.-J., Galbusera, R., et al. (2020). Multiple sclerosis cortical and wm lesion segmentation at 3t MRI: a deep learning method based on flair and mp2rage. *NeuroImage: Clinical* 27:102335. doi: 10.1016/j.nicl.2020.102335

Lassmann, H. (2019). Pathogenic mechanisms associated with different clinical courses of multiple sclerosis. *Front. Immunol.* 9:3116. doi: 10.3389/fimmu.2018.03116

Law, M. T., Traboulsee, A. L., Li, D. K., Carruthers, R. L., Freedman, M. S., Kolind, S. H., et al. (2019). Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression. *Multiple Sclerosis J.-Exp. Transl. Clini.* 5:2055217319885983. doi: 10.1177/2055217319885983

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lemay, A., Gros, C., Naga Karthik, E., and Cohen-Adad, J. (2022). Label fusion and training methods for reliable representation of inter-rater uncertainty. *Mach. Learn. Biomed. Imag.* 1:1–27. doi: 10.59275/j.melba.2022-db5c

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2022). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 6999–7019. doi: 10.1109/TNNLS.2021.3084827

Lin, W., Hasenstab, K., Cunha, G. M., and Schwartzman, A. (2020). Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. *Sci. Rep.* 10:77264. doi: 10.1038/s41598-020-77264-y

Lou, C., Sati, P., Absinta, M., Clark, K., Dworkin, J. D., Valcarcel, A. M., et al. (2021). Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3t susceptibility-based MR imaging. *NeuroImage: Clinical* 32:102796. doi: 10.1016/j.nicl.2021.102796

Lublin, F. D., Häring, D. A., Ganjgahi, H., Ocampo, A., Hatami, F., Čuklina, J., et al. (2022). How patients with multiple sclerosis acquire disability. *Brain* 145, 3147–3161. doi: 10.1093/brain/awac016

Lublin, F. D., and Reingold, S. C. (1996). Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on clinical trials of new agents in multiple sclerosis. *Neurology* 46:907–911. doi: 10.1212/WNL.46.4.907

Luo, G., Xie, W., Gao, R., Zheng, T., Chen, L., and Sun, H. (2023). Unsupervised anomaly detection in brain MRI: Learning abstract distribution from massive healthy brains. *Comput. Biol. Med.* 154:106610. doi: 10.1016/j.compbiomed.2023.106610

Maggi, P., and Absinta, M. (2024). Emerging MRI biomarkers for the diagnosis of multiple sclerosis. *Multiple Sclerosis J.* 30, 1704–1713. doi: 10.1177/13524585241293579

Maggi, P., Fartaria, M. J., Jorge, J., Rosa, F. L., Absinta, M., Sati, P., et al. (2020). CVSNet: a machine learning approach for automated central vein sign assessment in multiple sclerosis. *NMR Biomed.* 33:4283. doi: 10.1002/nbm.4283

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., et al. (2024). Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* 21, 195–212. doi: 10.1038/s41592-023-02151-z

Mangeat, G., Ouellette, R., Wabartha, M., Leener, B. D., Plattén, M., Karrenbauer, V. D., et al. (2020). Machine learning and multiparametric brain MRI to differentiate hereditary diffuse leukodystrophy with spheroids from multiple sclerosis. *J. Neuroimag.* 30, 674–682. doi: 10.1111/jon.12725

Mani, A., Santini, T., Puppala, R., Dahl, M., Venkatesh, S., Walker, E., et al. (2021). Applying deep learning to accelerated clinical brain magnetic resonance imaging for multiple sclerosis. *Front. Neurol.* 12:685276. doi: 10.3389/fneur.2021.685276

McDonald, W. I., Miller, D. H., and Thompson, A. J. (1994). Are magnetic resonance findings predictive of clinical outcome in therapeutic trials in multiple sclerosis? the dilemma of interferon-beta. *Ann. Neurol.* 36, 14–18. doi: 10.1002/ana.410360106

McKenzie, I. R., Lyzhov, A., Parrish, A., Prabhu, A., Mueller, A., Kim, N., et al. (2023). Inverse scaling: when bigger isn't better. *Transact. Mach. Learn. Res*. Available online at: https://openreview.net/forum?id=DwgRm72GQF

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., et al. (2020). Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage: Clinical* 25:102104. doi: 10.1016/j.nicl.2019.102104

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv*. 54:6. doi: 10.1145/3457607

Mescheriakova, J. Y., Wong, Y. Y. M., Runia, T. F., Jafari, N., Samijn, J. P., de Beukelaar, J. W., et al. (2018). Application of the 2017 revised mcdonald criteria for multiple sclerosis to patients with a typical clinically isolated syndrome. *JAMA Neurol*. 75, 1392–1398. doi: 10.1001/jamaneurol.2018.2160

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., et al. (2023). Foundation models for generalist medical artificial intelligence. *Nature* 616, 259–265. doi: 10.1038/s41586-023-05881-4

Müller, J., Cagol, A., Lorscheider, J., Tsagkas, C., Benkert, P., Yaldizli, Ö., et al. (2023). Harmonizing definitions for progression independent of relapse activity in multiple sclerosis: A systematic review. *JAMA Neurol*. 80, 1232–1245. doi: 10.1001/jamaneurol.2023.3331

Nabizadeh, F., Masrouri, S., Ramezannezhad, E., Ghaderi, A., Sharafi, A. M., Soraneh, S., et al. (2022). Artificial intelligence in the diagnosis of multiple sclerosis: a systematic review. *Mult. Scler. Relat. Disord*. 59:103673. doi: 10.1016/j.msard.2022.103673

Nabizadeh, F., Ramezannezhad, E., Kargar, A., Sharafi, A. M., and Ghaderi, A. (2023). Diagnostic performance of artificial intelligence in multiple sclerosis: a systematic review and meta-analysis. *Neurol. Sci*. 44, 499–517. doi: 10.1007/s10072-022-06460-7

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med. Image Anal*. 59:101557. doi: 10.1016/j.media.2019.101557

Narayana, P. A., Coronado, I., Sujit, S. J., Wolinsky, J. S., Lublin, F. D., and Gabr, R. E. (2020). Deep learning for predicting enhancing lesions in multiple sclerosis from noncontrast MRI. *Radiology* 294, 398–404. doi: 10.1148/radiol2019191061

Neeb, H., Schenk, J., Neeb, H., and Schenk, J. (2019). Multivariate prediction of multiple sclerosis using robust quantitative MR-based image metrics. *Z. Med. Phys*. 29, 262–271. doi: 10.1016/j.zemedi.2018.10.004

Nichyporuk, B., Cardinell, J., Szeto, J., Mehta, R., Falet, J.-P., Arnold, D. L., et al. (2022). Rethinking generalization: The impact of annotation style on medical image segmentation. *Machine Learn. Biomed. Imag*. 1, 1–37. doi: 10.59275/j.melba.2022-2d93

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., et al. (2019). Deep learning vs. traditional computer vision. *Adv. Intellig. Syst. Comput*. 943, 128–144. doi: 10.1007/978-3-030-17795-9_10

Patel, S. B., and Lam, K. (2023). ChatGPT: the future of discharge summaries? *Lancet Digital Health* 5, e107-e108. doi: 10.1016/S2589-7500(23)00021-3

Pellegrini, F., Copetti, M., Sormani, M. P., Bovis, F., de Moor, C., Debray, T. P., et al. (2020). Predicting disability progression in multiple sclerosis: Insights from advanced statistical modeling. *Multiple Sclerosis Journal* 26:1828–1836. doi: 10.1177/1352458519887343

Peng, Y., Zheng, Y., Tan, Z., Liu, J., Xiang, Y., Liu, H., et al. (2021). Prediction of unenhanced lesion evolution in multiple sclerosis using radiomics-based models: a machine learning approach. *Mult. Scler. Relat. Disord*. 53:102989. doi: 10.1016/j.msard.2021.102989

Pinaya, W. H., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., et al. (2022). Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Med. Image Anal*. 79:102475. doi: 10.1016/j.media.2022.102475

Pitt, D., Lo, C. H., Gauthier, S. A., Hickman, R. A., Longbrake, E., Airas, L. M., et al. (2022). Toward precision phenotyping of multiple sclerosis. *Neurol.-Neuroimmunol. Neuroinflammat*. 9:200025. doi: 10.1212/NXI.0000000000200025

Pontillo, G., Penna, S., Cocozza, S., Quarantelli, M., Gravina, M., Lanzillo, R., et al. (2022). Stratification of multiple sclerosis patients using unsupervised machine learning: a single-visit MRI-driven approach. *Eur. Radiol*. 32, 5382–5391. doi: 10.1007/s00330-022-08610-z

Prabhakar, C., Li, H. B., Paetzold, J. C., Loehr, T., Niu, C., Mühlau, M., et al. (2023). "Self-pruning graph neural network for predicting inflammatory disease activity in multiple sclerosis from brain MR images," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2023*, eds. H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, et al. (Cham: Springer Nature Switzerland), 226–236.

Preul, M. C., Caramanos, Z., Collins, D. L., Villemure, J. G., Leblanc, R., Olivier, A., et al. (1996). Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy. *Nature Medicine 1996 2:3* 2:323–325. doi: 10.1038/nm0396-323

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: methods, analysis & insights from training gopher. *arXiv* [preprint] arXiv:2112.11446. doi: 10.48550/arXiv.2112.11446

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., et al. (2021). "Zero-shot text-to-image generation," in *International Conference on Machine Learning* (New York: PMLR), 8821–8831.

Reinhold, J. C., Carass, A., and Prince, J. L. (2021). "A structural causal model for MR images of multiple sclerosis," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference* (Strasbourg: Springer), 782–792.

Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., et al. (2024). Understanding metric-related pitfalls in image analysis validation. *Nat. Methods* 21, 182–194. doi: 10.1038/s41592-023-02150-0

Richens, J. G., Lee, C. M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun*. 11:3923. doi: 10.1038/s41467-020-17419-7

Rocca, M. A., Anzalone, N., Storelli, L., Poggio, A. D., Cacciaguerra, L., Manfredi, A. A., et al. (2021). Deep learning on conventional magnetic resonance imaging improves the diagnosis of multiple sclerosis mimics. *Invest. Radiol*. 56, 252–260. doi: 10.1097/RLI.0000000000000735

Rosa, F. L., Beck, E. S., Maranzano, J., Todea, R. A., van Gelderen, P., de Zwart, J. A., et al. (2022). Multiple sclerosis cortical lesion detection with deep learning at ultra-high-field MRI. *NMR Biomed*. 35:e4730. doi: 10.1002/nbm.4730

Rudick, R. A., Lee, J.-C., Simon, J., and Fisher, E. (2006). Significance of T2 lesions in multiple sclerosis: a 13-year longitudinal study. *Ann. Neurol*. 60, 236–242. doi: 10.1002/ana.20883

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2019). Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-Net. *IEEE Access* 7, 25171–25184. doi: 10.1109/ACCESS.2019.2900198

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical* 25:102149. doi: 10.1016/j.nicl.2019.102149

Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., ONeil, A. Q., and Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *R. Soc. Open Sci*. 9:220638. doi: 10.1098/rsos.220638

Schroeter, J., Myers-Colet, C., Arnold, D. L., and Arbel, T. (2022). "Segmentation-consistent probabilistic lesion counting," in *International Conference on Medical Imaging with Deep Learning* (New York: PMLR), 1034–1056.

Seccia, R., Romano, S., Salvetti, M., Crisanti, A., Palagi, L., and Grassi, F. (2021). Machine learning use for prognostic purposes in multiple sclerosis. *Life* 11, 1–18. doi: 10.3390/life11020122

Shaul, R., David, I., Shitrit, O., and Raviv, T. R. (2020). Subsampled brain MRI reconstruction by generative adversarial neural networks. *Med. Image Anal*. 65:101747. doi: 10.1016/j.media.2020.101747

Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., et al. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput. Biol. Med*. 136:104697. doi: 10.1016/j.compbiomed.2021.104697

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180. doi: 10.1038/s41586-023-06291-2

Solomon, A. J., Naismith, R. T., and Cross, A. H. (2019). Misdiagnosis of multiple sclerosis: Impact of the 2017 McDonald criteria on clinical practice. *Neurology* 92, 26–33. doi: 10.1212/WNL.0000000000006583

Sormani, M. P., Bonzano, L., Roccatagliata, L., Cutter, G. R., Mancardi, G. L., and Bruzzi, P. (2009). Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytic approach. *Ann. Neurol*. 65, 268–275. doi: 10.1002/ana.21606

Sormani, M. P., and Bruzzi, P. (2013). MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. *Lancet Neurol*. 12, 669–676. doi: 10.1016/S1474-4422(13)70103-0

Spagnolo, F., Depeursinge, A., Schädelin, S., Akbulut, A., Müller, H., Barakovic, M., et al. (2023). How far MS lesion detection and segmentation are integrated into the clinical workflow? a systematic review. *NeuroImage: Clinical* 39:103491. doi: 10.1016/j.nicl.2023.103491

Storelli, L., Azzimonti, M., Gueye, M., Vizzino, C., Preziosa, P., Tedeschi, G., et al. (2022). A deep learning approach to predicting disease progression in multiple sclerosis using magnetic resonance imaging. *Invest. Radiol*. 57, 423–432. doi: 10.1097/RLI.0000000000000854

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., et al. (2008). 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. *Midas J*. 2008, 1–6. doi: 10.54294/lmkqvm

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *Lancet Neurol*. 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Tousignant, A., Lemaître, P., Precup, D., Arnold, D. L., and Arbel, T. (2019). "Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, eds. M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, et al. (New York: PMLR), 483–492.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv* [preprint] arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intellig.* 4, 189–191. doi: 10.1038/s42256-022-00465-9

Valencia, L., Clèrigues, A., Valverde, S., Salem, M., Oliver, A., Rovira, A., et al. (2022). Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis. *Front. Neurosci.* 16:954662. doi: 10.3389/fnins.2022.954662

Van Nederpelt, D. R., Amiri, H., Brouwer, I., Noteboom, S., Mokkink, L. B., Barkhof, F., et al. (2023). Reliability of brain atrophy measurements in multiple sclerosis using MRI: an assessment of six freely available software packages for cross-sectional analyses. *Neuroradiology* 65, 1459–1472. doi: 10.1007/s00234-023-03189-8

Van Tulder, G., and de Bruijne, M. (2015). "Why does synthesized data improve multi-sequence classification?," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference* (Munich: Springer), 531–538.

Vollmer, T. L., Nair, K. V., Williams, I. M., and Alvarez, E. (2021). Multiple sclerosis phenotypes as a continuum the role of neurologic reserve. *Neurology* 11, 342–351. doi: 10.1212/CPJ.0000000000001045

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Colliot, O., Stankoff, B., et al. (2019). Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis. *J. Med. Imag.* 6, 014005–014005. doi: 10.1117/1.JMI.6.1.014005

Xu, W., Xu, Y., Chang, T., and Tu, Z. (2021). "Co-scale conv-attentional image transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Los Alamitos, CA: IEEE Computer Society), 9961–9970.

Young, A. L. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature Communications* 9, 1–16. doi: 10.1038/s41467-018-05892-0

Zare, M. R., Alebiosu, D. O., and Lee, S. L. (2018). "Comparison of handcrafted features and deep learning in classification of medical x-ray images," in *Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)* (Kota Kinabalu: IEEE), 1–5.

Zeng, C., Gu, L., Liu, Z., and Zhao, S. (2020). Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front. Neuroinform.* 14:610967. doi: 10.3389/fninf.2020.610967

Zhan, G., Wang, D., Cabezas, M., Bai, L., Kyle, K., Ouyang, W., et al. (2023). Learning from pseudo-labels: deep networks improve consistency in longitudinal brain volume estimation. *Front. Neurosci.* 17:1196087. doi: 10.3389/fnins.2023.1196087

Zhang, H., Bakshi, R., Bagnato, F., and Oguz, I. (2020). "Robust multiple sclerosis lesion inpainting with edge prior," in *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020* (Lima: Springer), 120–129.

Zhang, H., Nguyen, T. D., Zhang, J., Marcille, M., Spincemaille, P., Wang, Y., et al. (2022). QSMRim-Net: imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps. *NeuroImage: Clinical* 34:102979. doi: 10.1016/j.nicl.2022.102979

Zhang, K., Lincoln, J. A., Jiang, X., Bernstam, E. V., and Shams, S. (2023). Predicting multiple sclerosis severity with multimodal deep neural networks. *BMC Med. Inform. Decis. Mak.* 23, 1–17. doi: 10.1186/s12911-023-02354-6

Zhang, Y., Hong, D., Mcclement, D., Oladosu, O., Pridham, G., and Slaney, G. (2021). Grad-cam helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J. Neurosci. Methods* 353:109098. doi: 10.1016/j.jneumeth.2021.109098

Zhao, C., Shao, M., Carass, A., Li, H., Dewey, B. E., Ellingsen, L. M., et al. (2019). Applications of a deep learning method for anti-aliasing and super-resolution in MRI.*Magn Reson Imaging* 64, 132–141. doi: 10.1016/j.mri.2019.05.038

Zhao, Y., Healy, B. C., Rotstein, D., Guttmann, C. R., Bakshi, R., Weiner, H. L., et al. (2017). Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE* 12:e0174866. doi: 10.1371/journal.pone.0174866

Zhao, Y., Wang, T., Bove, R., Cree, B., Henry, R., Lokhande, H., et al. (2020). Ensemble learning predicts multiple sclerosis disease course in the summit study. *NPJ Digit. Med.* 3:135. doi: 10.1038/s41746-020-00338-8

frontiers | Frontiers in Neurology

TYPE Brief Research Report
PUBLISHED 10 September 2025
DOI 10.3389/fneur.2025.1620469

OPEN ACCESS

EDITED BY
Axel Faes,
University of Hasselt, Belgium

REVIEWED BY
Nada Haj Messaoud,
University of Monastir, Tunisia
Stijn Denissen,
Vrije Universiteit Brussel, Belgium

*CORRESPONDENCE
Sarah Hindawi
✉ sarah.hindawi@roche.com

RECEIVED 27 May 2025
ACCEPTED 15 August 2025
PUBLISHED 10 September 2025

CITATION
Hindawi S, Szubstarski B, Boernert E, Tackenberg B and Wuerfel J (2025) Federated learning for lesion segmentation in multiple sclerosis: a real-world multi-center feasibility study.
Front. Neurol. 16:1620469.
doi: 10.3389/fneur.2025.1620469

Federated learning for lesion segmentation in multiple sclerosis: a real-world multi-center feasibility study

Sarah Hindawi[1]*, Bartlomiej Szubstarski[2], Eric Boernert[3], Björn Tackenberg[3] and Jens Wuerfel[3]

[1]Hoffmann-La Roche Limited, Mississauga, ON, Canada, [2]Roche Polska Sp. z o.o., Warsaw, Poland, [3]F. Hoffmann-La Roche AG, Basel, Switzerland

Multiple sclerosis (MS) is a chronic neuroinflammatory disease driven by immune-mediated central nervous system damage, often leading to progressive disability. Accurate segmentation of MS lesions on MRI is crucial for monitoring disease and treatment efficacy; however, manual segmentation remains time-consuming and prone to variability. While deep learning has advanced automated segmentation, robust performance benefits from large-scale, diverse datasets, yet data pooling is restricted by privacy regulations and clinical performance remains challenged by inter-site heterogeneity. In this proof-of-concept work, we aim to apply and adopt Federated Learning (FL) in a real-world hospital setting. We assessed FL for MS lesion segmentation using the self-configuring nnU-Net model, leveraging 512 MRI cases from three sites without sharing raw patient data. The federated model achieved Dice scores ranging from 0.66 to 0.80 across held-out test sets. While performance varied across sites, reflecting data heterogeneity, the study demonstrates the potential of FL as a scalable and secure paradigm for advancing automated MS analysis in distributed clinical environments. This work supports adopting secure, collaborative AI in neuroimaging, offering utility for privacy-sensitive clinical research and a starting point for medical AI development, bridging the gap between model generalizability and regulatory compliance.

KEYWORDS

federated learning, MRI lesion segmentation, privacy-preserving AI, distributed deep learning, multi-site training

1 Introduction

Multiple sclerosis is a chronic autoimmune disorder of the central nervous system (CNS) and is a leading cause of non-traumatic neurological disability among young adults (1). MS affects more than 2.8 million individuals worldwide (2). The disease is characterized by inflammatory demyelinating CNS lesions (3), which appear as hyperintense areas in white matter on T2-weighted/FLAIR MRI and are crucial for diagnosis and monitoring disease progression. Lesion burden correlates with disability (4), making accurate lesion segmentation vital for evaluating treatment efficacy.

Manual MS lesion segmentation is the clinical gold standard but is labor-intensive and prone to observer variability. Recent convolutional neural network approaches, including 3D U-Net variants, have achieved Dice scores of 0.6–0.8 for automated MS lesion segmentation on benchmark datasets (5). We employed nnU-Net, a self-configuring framework with strong performance across diverse medical segmentation tasks (6). Clinical adoption of automated segmentation methods remains limited due to the heterogeneity of MRI data, including

variations in acquisition protocols, scanner types, and lesion characteristics across patient populations. Models trained on single-center data may generalize poorly to external data due to distribution shifts (7). Privacy regulations limit data sharing across centers, limiting the ability to curate sufficiently large and diverse training datasets. This fragmentation of data impedes the development of generalizable AI models and continues to hinder machine learning (ML) translation into clinical settings (8).

Federated learning (FL) has emerged as a promising solution by enabling collaborative model training without exchanging raw data. In a federated learning paradigm, each institution (client) trains a local copy of the global model on-site. Instead of transferring patient data, only the model's learned parameters (e.g., weight updates) are shared with a central server. The server aggregates the updates from the participating clients to construct a consensus global model, enabling collaborative learning while addressing privacy concerns and utilizing otherwise inaccessible datasets. Despite its promise, FL is still in the early stages of medical deployment (9) Two studies from the same research group have investigated FL for MS lesion segmentation. These studies used simulated FL environments with clinical and public datasets (fewer than 200 subjects across scenarios) and reported moderate Dice scores ranging from 54 to 77% (10, 11). A recent study (12) also investigated FL for MS lesion segmentation as part of a broader benchmark of five neuroimaging tasks, conducted in a simulated FL environment, reporting Dice scores ranging from 63.2 to 70.2% on MSSEG dataset (13).

In contrast, our study deploys a federated learning framework for MS lesion segmentation in a real-world, multi-institutional setting, addressing legal and regulatory constraints that often hinder clinical translation. These challenges, typically underexplored in simulated environments, are addressed through a secure, end-to-end deployment in which each site retains full ownership and control of its data, demonstrating the practical feasibility of integrating FL into clinical practice under strict data governance. We trained and evaluated the model across three clinical institutions on a total of 512 MRI cases, integrating both academic research and routine clinical data. Specifically, we aim to establish a federated architecture for distributed image analysis and assess the feasibility of training a model for segmenting T2-weighted hyperintense MS lesions across sites. By demonstrating FL's application to MS lesion segmentation, we aim to strengthen the groundwork for privacy-preserving, collaborative AI in neuroimaging.

2 Methods

2.1 Federated framework architecture

To enable privacy-preserving, multi-center training for MS lesion segmentation, we extended our federated learning platform with imaging capabilities by integrating it with an established open-source framework for radiology image processing. Specifically, we utilized Kaapana, an open-source platform described in (14, 15), to coordinate local imaging processing and computational workflows. Kaapana is a modular toolkit for medical image analysis that enables decentralized data access, data management, and remote execution of containerized algorithms. It supports private cloud development and integrates seamlessly with local clinical IT infrastructure. The platform employed a client–server FL architecture to train the model across three participating sites. Each client maintained a local copy of the model and trained it on its own dataset of MR images. A central server acted as the coordinating node, aggregating received model parameters using the Federated Averaging (FedAvg) algorithm, which computes a weighted average of the clients' model weights (16). To address operational, security and collaboration network scalability needs in real-world clinical environments, we extended our setup with additional enterprise-grade computational governance capabilities developed by Apheris, enabling institutions to collaborate securely on distributed data within a governed and privacy-preserving framework. This integration allowed all collaborating institutions to retain end-to-end control over algorithm execution. Although open-source solutions offer transparency and adaptability, their integration into clinical workflows can introduce operational overhead, including the need for manual code reviews. To mitigate this challenge and reduce risk, we implemented a centralized algorithm review process with a controlled algorithm pull mechanism from a central container registry, ensuring reproducibility, data and model governance, and streamlined collaboration without exposing sensitive data.

By design, the federated model should be exposed to a wider variety of imaging patterns (patient demographics, scanner types, artifact profiles) than any single-site model, ideally resulting in a more generalizable model. MRI data were preprocessed using a standardized pipeline applied consistently across all sites to ensure uniform orientation and registration. We employed nnU-Net, which automatically configures its architecture, preprocessing, and training pipelines to the given dataset, enabling site-specific adaptation and efficient deployment with minimal computational and implementation overhead (6). The model was trained across sites using a uniform configuration and shared hyperparameters. Each site used locally managed infrastructure, typically comprising GPUs with at least 24 GB of VRAM (NVIDIA Turing or newer) and at least 64 GB of RAM. The training was done in a synchronous federated manner such that all sites participated in each round. By the end of training, the final federated model was evaluated on held-out test sets at each participating site.

Throughout the federated training, no MR images or patient identifiers were ever exchanged. Only data fingerprints, containing image sizes, voxel spacings, and intensity characteristics for model initialization, along with model parameters were shared during FL iterations. Dataset fingerprints were required for the adaptive, rule-based configuration of the segmentation pipeline, including the selection of the patch size, network topology, and batch size, all of which depend on image properties (6). This approach together with a decentralized architecture inherently preserves data privacy, as an adversary cannot directly access the underlying images through the central server. To further secure communications, all network traffic between the server and client nodes was encrypted using state-of-the-art protocols. Each participating site deployed and operated a local platform within its own firewall, allowing the central orchestration server to invoke nnU-Net federated training workflows on local data. We implemented local basic authentication for the nodes and an external identity and access mechanism for the central node. This design enables more autonomous, isolated and efficient deployment at each site. Node authentication within the federated network is based on a centrally generated token that each site receives via independent media during registration. This token includes: 1. an

SSL certificate, 2. an authentication token, 3. connection details of the central instance and 4. a symmetrical encryption key as an additional protection mechanism for data in transit. To simplify deployment and avoid dependencies on potential vulnerabilities in the open-source network stack, we opted not to implement a dedicated virtual network infrastructure for federated nodes. Instead, we introduced a symmetric encryption layer implemented explicitly at the federated client and server applications. Due to its inherent speed, this mechanism was well-suited for encrypting client-generated weights at each round. Additionally, it served as a safeguard to ensure secure communication between clients and the central node, effectively replicating the protection typically provided by a virtual private network.

This setup guarantees the authenticity of the contributing clients and prevents spoofing or tampering within the federated network. Each client application maintained a list of approved datasets and workflows for federated processing, allowing site personnel to contribute to model training without relinquishing control over their data. This approach is compliant with data protection regulations and addresses the ethical concerns of data sharing.

The federated learning architecture (illustrated in Figure 1) supports the following key user workflows:

1. Model publication by ML Engineer - A locally tested model is converted into its federated version and uploaded to a central model repository. Once approved by the site Data Custodian, it becomes available for execution at the corresponding sites.
2. Upload of data assets and data access policies by the Data Custodian - At each site, the Data Custodian defines which models are authorized to access the uploaded data. This enables Gateway agents to accept requests to execute approved ML models.
3. Federated Workflow Execution by Data Scientist - Using the Python SDK, the Data Scientist interacts with the Federated Learning Orchestrator to initiate computation pods at the federated nodes (workers) and the central platform

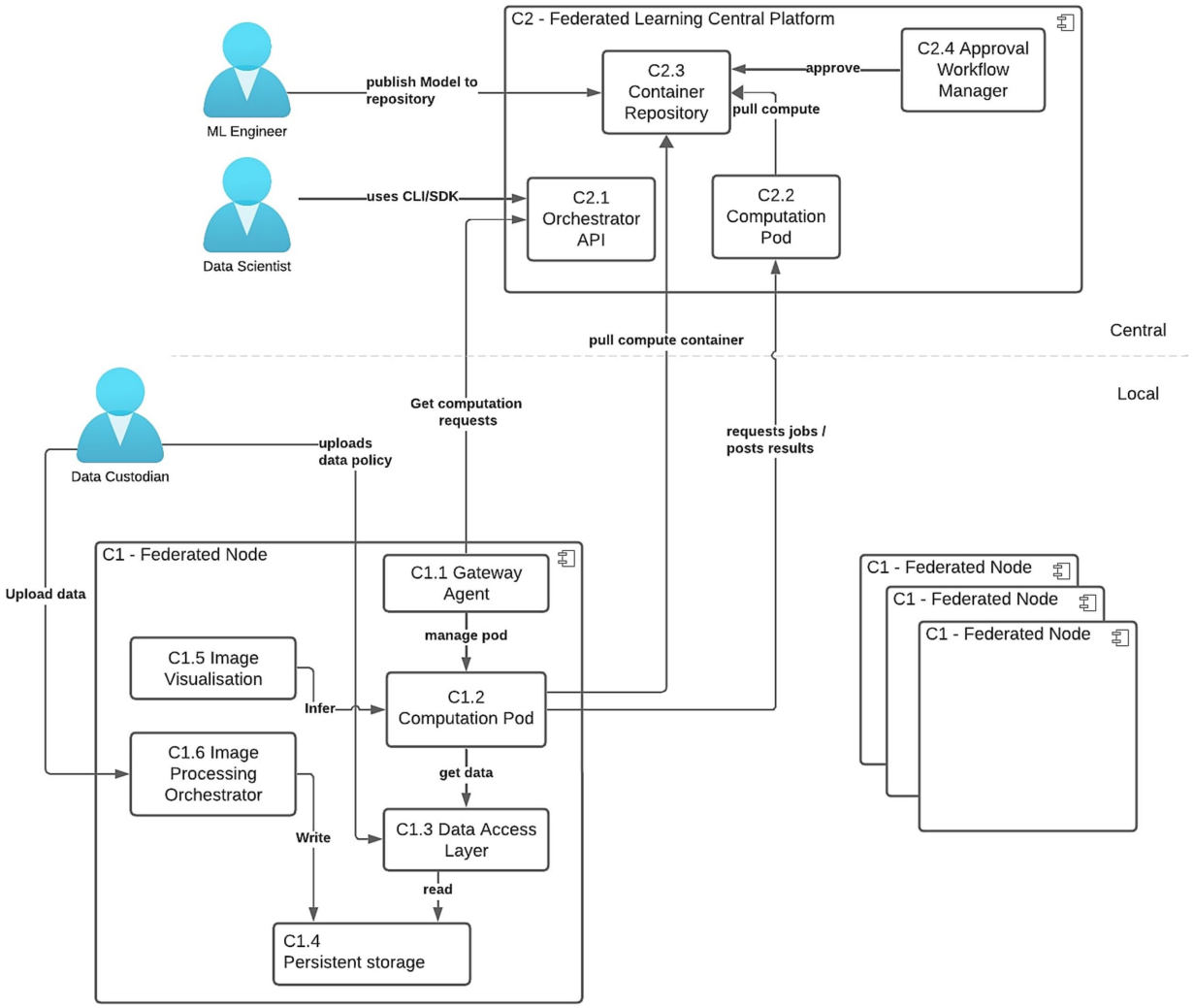


FIGURE 1
High-level solution architecture showing the integration of Kaapana and Apheris into a single solution. This extended architecture allows site personnel to control processing pipelines executed on their data. It supports three primary user workflows and illustrates basic component interactions at a single federated node.

(aggregator), in accordance with the approved data access policies. This setup facilitates the full execution of nnU-Net training across the participating sites.

2.2 Description of datasets

This is a multi-center, multi-country study utilizing anonymized MRI data from patients with MS. The study involved in-house data from a previous Roche-sponsored trial at our site (Site A, 149 cases, each consisting of paired T1w and FLAIR images), as well as anonymized observational data from two academic medical centers: one in Switzerland (Site B, 325 cases) and one in Germany (Site C, 38 cases). A total of 512 expert-annotated MRI cases were used, of which 380 were allocated to training and validation. Patients were uniquely assigned to either the training/validation or test sets to avoid data leakage and ensure unbiased model evaluation. In accordance with data protection principles, all data remained local at each site and were never shared centrally.

The datasets included 1 mm isotropic 3D T2-weighted/FLAIR and T1-weighted sequences (with a tolerance of ±0.1 mm, ranging from 0.9 to 1.1 mm). Scans were acquired on Siemens, Philips, and GE Medical Systems scanners at a field strength of 3 Tesla, and each site followed its own routine clinical MRI protocol, resulting in some heterogeneity in image resolution and contrast. Site A contributed data from a range of scanner models across the three vendors: Siemens (Skyra, Verio, Prisma, Prisma_fit, TrioTim), Philips (Achieva, Achieva dStream, Intera, Ingenia), and GE Medical Systems (Signa HDxt, Discovery MR750, SIGNA Premier). Site B provided data acquired on Siemens Skyra and Skyra Fit scanners, while Site C used the Siemens Skyra Fit. Table 1 summarizes dataset characteristics across sites. The diversity of imaging sources and clinical presentations should reduce site-specific biases and enhance generalizability.

To ensure data consistency, T1-weighted images were registered to their corresponding FLAIR sequences, and automated quality control was applied to identify potential image quality issues. This diverse dataset, representing multiple sites with varying imaging protocols, was used to assess the federated approach under realistic conditions of inter-site heterogeneity.

2.3 Preprocessing for image standardization

A standardized automated preprocessing pipeline was applied to ensure data consistency across all sites. This process included automated quality control procedures assessing key image properties. Signal-to-noise ratios (SNR) were computed in modality-specific anatomical regions to estimate overall image quality. T1w SNR was calculated in the brain parenchyma, while FLAIR SNR was calculated in the cerebrospinal fluid. Artifact presence was estimated using the MAI-Lab sorting and artifacts detection tool (17), and cropping was detected by evaluating brain coverage across anatomical boundaries. Voxel dimensions were validated against the expected isotropic resolution (1.0 ± 0.1 mm), and inter-modality brain mask volume similarity was assessed to detect major discrepancies or modality-specific artifacts. All MRI data were reoriented to a standardized axial orientation to ensure uniform spatial alignment. T1-weighted images were registered to their corresponding FLAIR images, correcting for positional misalignment. These preprocessing steps were performed locally at each site using Kaapana and integrated into the federated learning workflow, ensuring uniformity in the input data across sites for subsequent model training.

2.4 Model selection and training

We selected nnU-Net for its robust performance across diverse medical segmentation tasks, offering automatic adaptation and competitive results without manual customization (6). nnU-Net handles preprocessing, architecture selection, and postprocessing, reducing the need for extensive manual intervention. It is also well-suited for 3D multi-modal input, automatically configuring an appropriate 3D U-Net architecture based on input image dimensions and hardware constraints.

Local training at each site adhered to the standard nnU-Net training configuration and hyperparameters, with configurable values set to a learning rate of 0.01, weight decay of 3×10^{-5}, 250 training batches per epoch, and 33% foreground oversampling. The model used the standard Dice loss combined with cross-entropy, as provided by the default configuration of nnU-Net. We conducted 50 rounds of federated training, with each round corresponding to one local epoch at each site. Limiting local training to a single epoch helped prevent models from overfitting to local data and drifting from the global objective. Training progress was monitored by tracking site-level training and validation losses after each federated round to ensure stability and detect potential divergence. After each round, the server aggregated client weight updates using the FedAvg algorithm to generate a new global model, which was then redistributed to all sites.

The final global model obtained after 50 rounds of federated training was evaluated independently at each site using its respective held-out test set. Model evaluation at the three participating sites included both quantitative metrics such as Dice score, sensitivity, and precision, as well as a qualitative review by a neuroradiologist to assess overall performance, including true positive detection and tendencies to miss lesions across anatomical regions. To benchmark against a non-federated scenario, we trained and tested a baseline nnU-Net

TABLE 1 Summary of site-specific data including number of cases, scanner vendors, and lesion characteristics.

Site	Train/Validation cases	Test cases	Scanner vendors	Median Lesion volume (cc)	Median Lesion count
Site A	105	44	Siemens, Philips, GE	4.54 [2.35–9.36]	44 [25–67]
Site B	247	78	Siemens	4.90 [1.53–13.81]	33 [19–55]
Site C	28	10	Siemens	2.48 [0.58–3.91]	27 [8–65]

model locally using our site's data. This enabled a comparison between the performance of the federated model and the locally trained model, both evaluated on the same test set from our institution.

2.5 Privacy and security considerations

Patient privacy was a core requirement of our FL framework, which inherently avoids sharing raw imaging data. All images were anonymized at their source by removing identifying metadata (e.g., DICOM headers), ensuring that no personally identifiable information was accessible. Federated training was conducted within protected compute environments, with each site's data remaining on secure local infrastructure. Our configuration follows enterprise-grade governance principles, ensuring that each client site retains full control over which algorithms are executed on its data. This level of control allows individual node administrators to prevent the execution of unauthorized or potentially malicious code, thereby strengthening overall system security.

As described in the Architecture section, only sites that received a secret, unique token were allowed to contribute to the central model updates, thus limiting potential poisoning attacks. Since communications between sites and the central server were TLS-encrypted, and additionally encrypted at the sites with a symmetric key shared within the token, the risk of an adversarial attack was minimal. The central cloud-based environment employed AWS Well-Architected Framework mechanisms, with access to the Federated Orchestrator restricted to a predefined IP range. This setup limited the marginal risk of reconstruction or inference attacks and allowed the use of original parameters and weights from individual nodes.

While FL reduces data privacy risks by design, it is not entirely immune to threats such as model inversion or membership inference attacks. To mitigate these risks, we adopted strict security principles integrated directly into the framework. Execution of any machine learning code or federated learning configuration requires explicit approval from each participating site. Comprehensive encryption and tightly controlled access to both site and central nodes further minimize the risk of sensitive data leakage or attacks by unauthorized, potentially malicious actors.

From a regulatory standpoint, this study adhered to data protection laws. Since only model parameters and not raw data were exchanged, each institution maintained full control over its data. Our framework serves as a starting point for multi-center collaborations, promoting secure AI development in medical imaging.

3 Results

3.1 Quantitative analysis

We conducted 50 rounds of federated training, with each round corresponding to one local epoch per site. In our setting, preliminary experiments with additional local epochs per round resulted in abrupt performance degradation, which may reflect FedAvg's sensitivity to data heterogeneity (18). This aligns with observations in the literature where non-IID data or class imbalances can cause gradient misalignment, driving local models away from the global objective (19). Model performance over 50 federated rounds is shown in Figure 2. Most reductions in training and validation losses occurred within the first 10–15 rounds, after which learning progressed more gradually. As the system retains only model weights from the final

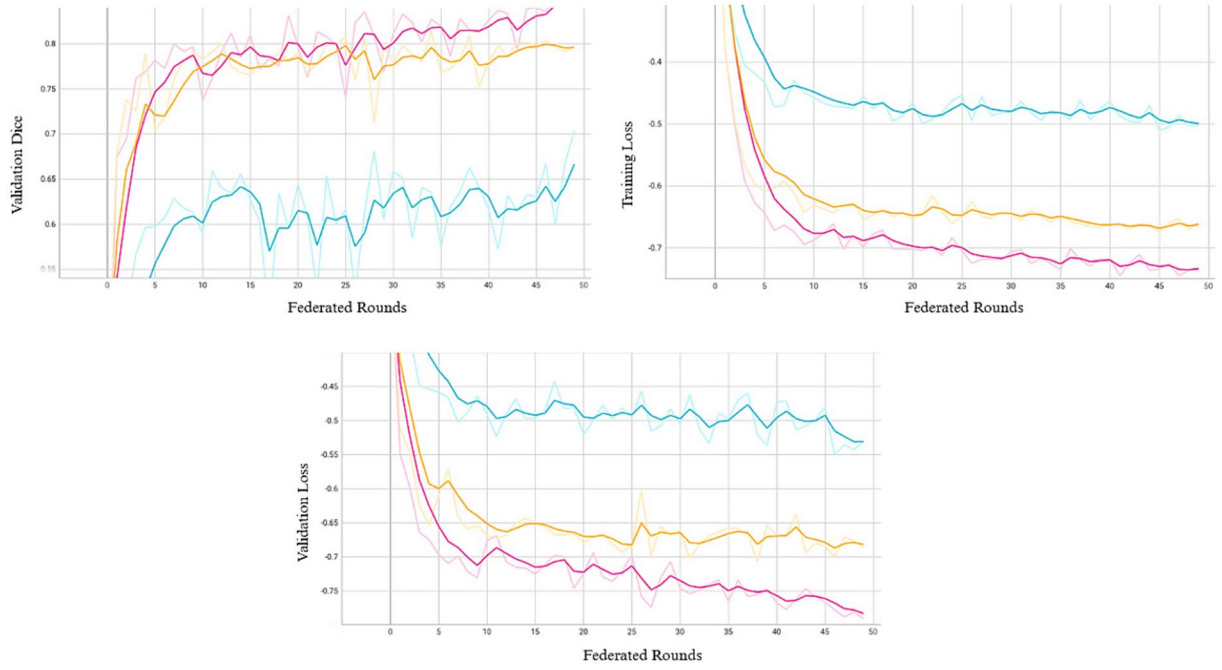


FIGURE 2
Validation Dice (top left), training loss (top right), and validation loss (bottom) across 50 federated rounds. Each curve represents one of the three participating sites (Site A: pink, Site B: orange, Site C: blue).

training round, we adopted a fixed training schedule rather than an adaptive early stopping strategy based on convergence. Extended training revealed that additional rounds improved performance at certain sites, while others experienced a decline, potentially due to model drift or overfitting to dominant patterns. Although 50 rounds may not represent the global optimum, this configuration provided a balanced trade-off across all participating sites.

To assess model performance, we compared a locally trained model to its federated counterpart using a held-out test set of 44 MRI cases from our site. Both models were trained using the same hyperparameters and configuration to ensure a fair comparison. A local nnU-Net model trained for 50 epochs using only our site's training set achieved a mean Dice score of 0.88 ± 0.04, sensitivity of 0.85 ± 0.05, and precision of 0.90 ± 0.05 on our site's test set. In comparison, the federated model, trained for 50 rounds with one local epoch per round across the three sites, achieved a mean Dice score of 0.80 ± 0.07 on the same test set. While the federated model showed a lower Dice score, it demonstrated higher sensitivity (0.89 ± 0.07 vs. 0.85), indicating improved lesion detection, albeit with reduced precision. One-sided Wilcoxon tests indicated that the local model had significantly higher Dice and precision ($p = 5.7 \times 10^{-14}$ for both). In contrast, the federated model showed significantly higher sensitivity based on a one-sided paired t-test ($p = 2.4 \times 10^{-7}$). In a clinical context, higher sensitivity is valuable for minimizing the risk of missed lesions; however, the corresponding decrease in precision reflects a higher rate of false positives, which may result in unwarranted diagnostic procedures, increased clinician workload, and patient distress.

To evaluate the cross-site generalizability of the federated model, we evaluated it on held-out test sets from the other two participating sites, comprising 78 and 10 cases, where it achieved mean Dice scores of 0.71 ± 0.15 and 0.66 ± 0.16, respectively. These results are summarized in Table 2. A Kruskal-Wallis test across all sites showed significant site-dependent variability in Dice scores ($p = 3.05 \times 10^{-5}$). Given the limited test sample size at Site C, we further conducted a two-sided Mann–Whitney U test between Site A and Site B, which also indicated a statistically significant difference in Dice scores between the two sites ($p = 3 \times 10^{-5}$).

Figure 3 presents the distribution of performance metrics for each site, with Sites A and B showing relatively more consistent distributions and Site C exhibiting broader variability, reflecting inter-site differences in model generalization. Although performance varied, likely due to differences in imaging protocols, scanner types, or annotation standards, the model maintained moderate segmentation performance across diverse clinical environments without access to raw patient data. Importantly, federated training does not preclude subsequent site-level adaptation. Fine-tuning the global model on local data can help capture site-specific patterns, offering a balanced approach that preserves the robustness gained from diverse data while recovering the precision of locally optimized models.

TABLE 2 Federated model performance on the test set from each participating site.

Site	Dice Score	Sensitivity	Precision
Site A	0.80 ± 0.07	0.89 ± 0.07	0.74 ± 0.11
Site B	0.71 ± 0.15	0.74 ± 0.11	0.70 ± 0.17
Site C	0.66 ± 0.16	0.64 ± 0.19	0.74 ± 0.23

Metrics are reported as mean ± standard deviation for Dice score, sensitivity, and precision, highlighting inter-site variability in segmentation accuracy.

3.2 Qualitative assessment

To complement the quantitative evaluation, a qualitative radiological assessment was conducted to examine the alignment between visual observations and metric-based performance. A board-certified neuroradiologist and MS expert assessed aspects not fully captured by global quantitative metrics, such as pathological plausibility (e.g., false negatives and false positives), anatomical consistency (e.g., periventricular, subcortical, and other region-specific biases), and morphological correctness (e.g., small versus large lesions). The expert reviewed lesion masks generated by (1) the federated model trained across all sites and (2) a model trained solely on local data from our site. As in the quantitative evaluation, the comparison was performed on outputs generated from the held-out test set at our site, with the models' outputs reviewed side by side to identify clinically meaningful differences in segmentation behavior.

Figure 4 presents a visual comparison on a FLAIR slice from our site's test set, with model predicted segmentation masks overlaid on the image. The results highlight key differences between the models, with the federated model detecting more lesions, reflecting higher sensitivity, but also introducing more false positives. While further validation is warranted, these findings demonstrate the feasibility of federated learning for automated MS lesion segmentation, underscoring its potential for broader clinical application.

4 Discussion

This Proof of Concept study demonstrates the end-to-end technical feasibility of deploying federated learning as a scalable, privacy-preserving framework across clinical institutions, each with distinct privacy constraints, data governance policies, and technical environments. Our work addresses a gap often overlooked in simulated FL research by preserving full data governance at each site while supporting scalable algorithm integration and institutional participation. By integrating Kaapana and Apheris, our framework enables autonomous data curation and enforces consensus-based algorithm approval prior to execution at each site, enhancing both privacy and operational security. This design allows each institution to manage its own imaging workflows while safeguarding against unauthorized computation, making the approach particularly well-suited for sensitive clinical environments. This federated setup is inherently portable and supports scalable, efficient deployment. It can be extended to additional institutions by deploying a platform instance at each site with secure client-to-server communication. This modular architecture emphasizes flexibility, reproducibility, and compatibility with diverse governance policies, enabling broader future adoption.

Building on this infrastructure, we evaluated the federated model on the held-out test set from each participating site. For comparative analysis, we also compared its performance on our site's held-out test set relative to a model trained and tested locally. Although the federated model showed a lower Dice score compared to the locally trained model at our site, it achieved higher recall, which may indicate improved lesion detection. This trade-off reflects a core challenge in

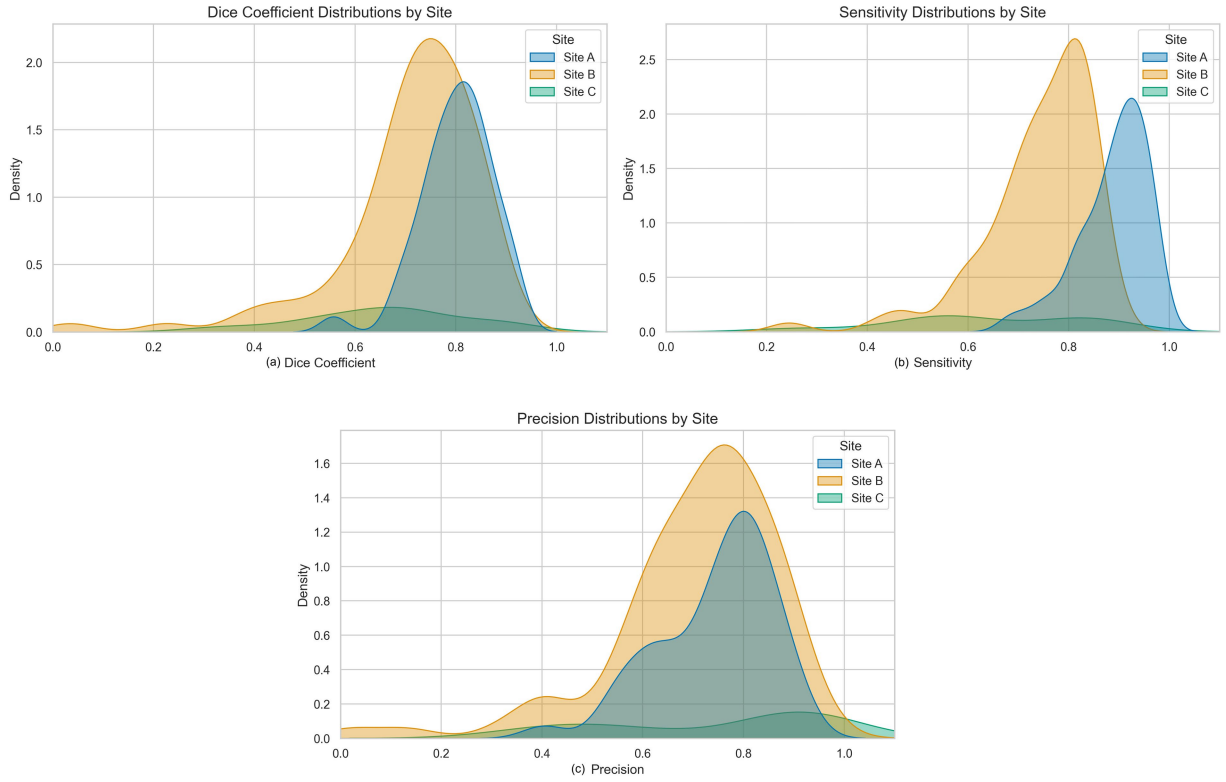


FIGURE 3
Density plots of segmentation metrics across sites. The plots show the distribution of **(a)** Dice score, **(b)** sensitivity, and **(c)** precision for Site A, Site B, and Site C, reflecting inter-site variability in segmentation performance.

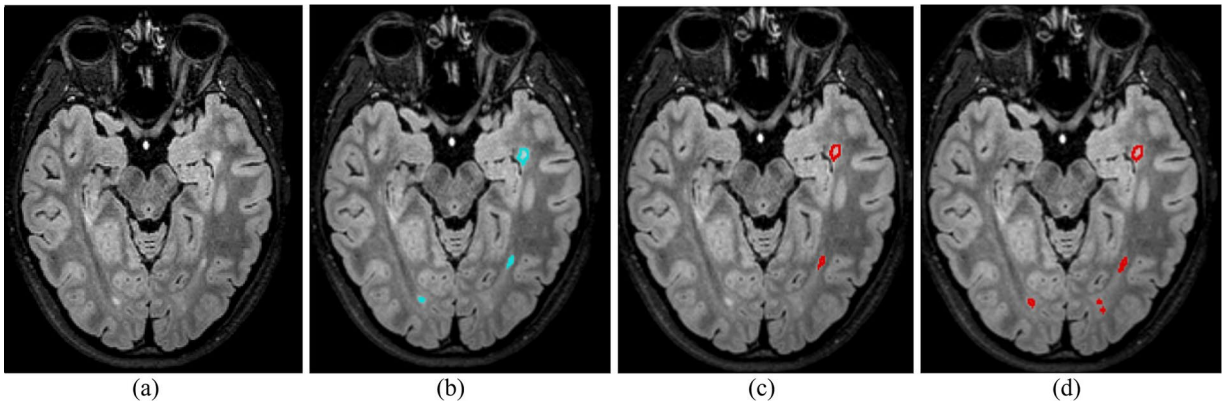


FIGURE 4
Comparative visualization of lesion segmentation between training paradigms: **(a)** FLAIR MRI slice without annotations; **(b)** Ground truth manual segmentation; **(c)** Prediction from the local model (trained solely on our data); **(d)** Prediction from the federated model (trained across three sites). The federated model detects more lesions but also introduces additional false positives, reflecting the trade-off between sensitivity and precision.

FL as it requires balancing global generalization with site-specific optimization. The observed performance gap in Dice score likely stems from the federated model's exposure to heterogeneous, non-IID data across institutions, which encourages learning generalized representations rather than overfitting to any specific site's patterns. Federated models are optimized to perform robustly across diverse data distributions, enhancing sensitivity to subtle or atypical lesions that may be underrepresented in any single site's dataset. However, this improvement in sensitivity was accompanied by reduced precision, as the federated model might not fully adapt to site-specific imaging features and annotation styles. This misalignment may cause the model to over-segment or misclassify challenging regions,

resulting in an increased number of false positives. Additionally, while local training on homogeneous data can converge rapidly, federated learning may require more rounds to achieve comparable performance due to the challenges of learning from fragmented and non-IID data distributions.

Beyond performance trade-offs, our study highlights several practical challenges that are often overlooked in simulated FL settings. First, in synchronous FL workflows, training requires all sites to remain active; resource outages or downtime at any site can halt the entire federated round. Second, training local models for comparison with the federated model requires technical expertise at all participating sites, which may not always be readily available. In contrast, participation in federated training and quantitative evaluation of the federated model in our setup did not require machine learning expertise. Third, centralized baseline models trained on pooled multi-site data, which are commonly used as performance upper bounds for federated models, are often infeasible in real-world clinical settings due to data privacy regulations, as was the case in our study. These constraints underscore the gap between FL in theory and its real-world implementation.

It is also worth noting that the federated model in this Proof of Concept study was not intended to optimize performance, and thus was only trained on a relatively small dataset (380 cases), whereas many deep learning studies rely on datasets exceeding 1,000 cases (20) or even tens of thousands in population-scale initiatives like UK Biobank (e.g., 39,694 subjects (21)). While expanding to larger, more diverse cohorts is expected to improve generalizability, site-specific accuracy gains may require complementary strategies. For instance, fine-tuning the federated model on local data can improve local performance, but risks catastrophic forgetting, where local adaptation distorts generalizable representations learned during federated training, leading to degraded performance on external datasets. To address this, personalized FL strategies such as FedBN (22), which retains local batch normalization statistics to account for domain shifts, and Ditto (23), which optimizes a personalized objective while maintaining alignment with the global model, have shown promise in non-IID settings. Additionally, adaptive aggregation adjusts client contributions to better manage data skew, with methods like FedProx (24) introducing a proximal term to reduce client drift and improve convergence stability.

While this study demonstrates the technical feasibility of FL in real-world settings, future research should explore integrating adaptive aggregation, personalized FL strategies, and expanding datasets to further improve model performance in heterogeneous environments. Overall, these findings establish a starting point for adopting federated learning in clinical practice, with potential for future scaling to multi-modal and longitudinal MS studies.

Software and resources

The federated learning infrastructure was implemented using the open-source Kaapana platform (https://github.com/kaapana/kaapana), with the nnU-Net training pipeline available at https://github.com/kaapana/kaapana/tree/develop/data-processing/processing-pipelines/nnunet. Additional computational governance capabilities were supported by Apheris (https://www.apheris.com), enabling secure collaboration across participating institutions.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the datasets consist of anonymized clinical MRI data from a previous Roche-sponsored trial and observational data from two academic medical centers. All data remained local to each site and were not transferred or shared centrally, in accordance with data protection regulations. The datasets are not publicly available due to patient privacy considerations and institutional data protection agreements, and therefore are not included in the manuscript or supplementary files. Requests for further information regarding the data or methodology may be directed to the corresponding author, Sarah Hindawi (sarah.hindawi@roche.com).

Ethics statement

The studies involving humans were approved by the Ethics Committee of Northwestern and Central Switzerland and the Ethics Committee of Charité - University Medicine Berlin, Germany. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

SH: Writing – original draft, Conceptualization, Writing – review & editing. BS: Writing – review & editing. EB: Writing – review & editing, Conceptualization, Supervision. BT: Supervision, Writing – review & editing. JW: Conceptualization, Writing – review & editing, Supervision.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We would like to acknowledge MIAC, Hycean and Apheris; especially Marco Duering, Vitor Gouveia, Jonas Scherer, Klaus Kades, Johannes Forster, Ian Hales and Christopher Woodward supporting the development and implementation of the federated learning architecture and solution. In addition, we would like to express our sincere gratitude to Prof. Dr. Cristina Granziera and Univ.-Prof. Dr. Friedemann Paul for their invaluable scientific

leadership and significant contributions to the data that made this study possible.

Conflict of interest

SH, BS, BT, JW, and EB are employees and/or stockholders of F. Hoffmann-La Roche Ltd.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that Generative AI was used in the creation of this manuscript. Portions of the manuscript's language and coherence were refined using GPT-4o (OpenAI, 2024), based on author-provided drafts and edits.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1. Dimitrov LG, Turner B. What's new in multiple sclerosis? *Br J Gen Pract*. (2014) 64:612–3. doi: 10.3399/bjgp14X682609

2. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: insights from the atlas of MS, third edition. *Mult Scler*. (2020) 26:1816–21. doi: 10.1177/1352458520970841

3. Barkhof F, Koeller KK. Demyelinating diseases of the CNS (brain and spine). In: J Hodler, RA Kubik-Huch and SchulthessGK von, editors. Diseases of the brain, head and neck, spine 2020–2023: Diagnostic imaging. Cham: Springer; (2020).

4. Nabizadeh F, Zafari R, Mohamadi M, Maleki T, Fallahi MS, Rafiei N. MRI features and disability in multiple sclerosis: a systematic review and meta-analysis. *J Neuroradiol*. (2024) 51:24–37. doi: 10.1016/j.neurad.2023.11.007

5. Zeng C, Gu L, Liu Z, Zhao S. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front Neuroinform*. (2020) 14:610967. doi: 10.3389/fninf.2020.610967

6. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. Nn U-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

7. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell*. (2022) 4:e210064. doi: 10.1148/ryai.210064

8. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med*. (2020) 3:119. doi: 10.1038/s41746-020-00323-1

9. Rehman MHU, Hugo Lopez Pinaya W, Nachev P, Teo JT, Ourselin S, Cardoso MJ. Federated learning for medical imaging radiology. *Br J Radiol*. (2023) 96:20220890. doi: 10.1259/bjr.20220890

10. Liu D, Cabezas M, Wang D, Tang Z, Bai L, Zhan G, et al. Multiple sclerosis lesion segmentation: revisiting weighting mechanisms for federated learning. *Front Neurosci*. (2023) 17:1167612. doi: 10.3389/fnins.2023.1167612

11. Bai L, Wang D, Wang H, Barnett M, Cabezas M, Cai W, et al. Improving multiple sclerosis lesion segmentation across clinical sites: a federated learning approach with noise-resilient training. *Artif Intell Med*. (2024) 152:102872. doi: 10.1016/j.artmed.2024.102872

12. Wagner F, Xu W, Saha P, Liang Z, Whitehouse D, Menon D, et al. (2024); Feasibility of federated learning from client databases with different brain diseases and MRI modalities. Available online at: https://openaccess.thecvf.com/content/WACV2025/papers/Wagner_Feasibility_of_Federated_Learning_from_Client_Databases_with_Different_Brain_WACV_2025_paper.pdf (Accessed July 30, 2025).

13. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci Rep*. (2018) 8:13650. doi: 10.1038/s41598-018-31911-7

14. Kades K, Scherer J, Zenk M, Kempf M, Maier-Hein K. Towards real-world federated learning in medical image analysis using Kaapana In: S Albarqouni, S Bakas, S Bano, MJ Cardoso, B Khanal and B Landmanet al, editors. Distributed, collaborative, and federated learning, and affordable AI and healthcare for resource diverse Global Health. Cham: Springer (2022). 130–40.

15. Scherer J, Nolden M, Kleesiek J, Metzger J, Kades K, Schneider V, et al. Joint imaging platform for federated clinical data analytics. *JCO Clin Cancer Inform*. (2020) 4:1027–38. doi: 10.1200/CCI.20.00045

16. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. Seattle, WA: DBLP (2016).

17. Gao R, Luo G, Ding R, Yang B, Sun H. A lightweight deep learning framework for automatic MRI data sorting and artifacts detection. *J Med Syst*. (2023) 47:124. doi: 10.1007/s10916-023-02017-z

18. Zhu H, Xu J, Liu S, Jin Y. Federated learning on non-IID data: a survey. *Neurocomputing*. (2021) 465:371–90. doi: 10.1016/j.neucom.2021.07.098

19. Wang J, Liu Q, Liang H, Joshi G, Poor H. Tackling the objective inconsistency problem in heterogeneous federated optimization. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems Curran Associates Inc. Red Hook, NY

20. Gabr RE, Coronado I, Robinson M, Sujit SJ, Datta S, Sun X, et al. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: a large-scale study. *Mult Scler*. (2020) 26:1217–26. doi: 10.1177/1352458519856843

21. Alfaro-Almagro F, McCarthy P, Afyouni S, Andersson JLR, Bastiani M, Miller KL, et al. Confound modelling in UK biobank brain imaging. *NeuroImage*. (2021) 224:117002. doi: 10.1016/j.neuroimage.2020.117002

22. Peng Z, Song Y, Wang Q, Xiao X, Tang Z. Fed BN: a communication-efficient federated learning strategy based on blockchain. In 2024 27th international conference on computer supported cooperative work in design (CSCWD). Sydney: IEEE; (2024). p. 754–759.

23. Li T, Hu S, Beirami A, Smith V. (2020) Ditto: fair and robust federated learning through personalization. Available online at: https://proceedings.mlr.press/v139/li21h/li21h.pdf (Accessed July 18, 2025).

24. Li T, Sahu AK, Zaheer M, Sanjabi M, Smith AT (2020) 1.5. Federated optimization in heterogeneous networks. Available online at: https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf (Accessed July 18, 2025).

Frontiers in Immunology

Explores novel approaches and diagnoses to treat immune disorders.

The official journal of the International Union of Immunological Societies (IUIS) and the most cited in its field, leading the way for research across basic, translational and clinical immunology.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact