npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01788-8

# Personalized federated learning for predicting disability progression in multiple sclerosis using real-world routine clinical data

Check for updates

Ashkan Pirmani[1,2,3,4], Edward De Brouwer[1], Ádám Arany[1], Martijn Oldenhof[1], Antoine Passemiers[1], Axel Faes[2,3,4], Tomas Kalincik[5], Serkan Ozakbas[6], Riadh Gouider[7], Barbara Willekens[8,9], Dana Horakova[10], Eva Kubala Havrdova[10], Francesco Patti[11], Alexandre Prat[12], Alessandra Lugaresi[13], Valentina Tomassini[14], Pierre Grammond[15], Elisabetta Cartechini[16], Izanne Roos[5], Cavit Boz[17], Raed Alroughani[18], Maria Pia Amato[19,20], Katherine Buzzard[21], Jeannette Lechner-Scott[22], Joana Guimarães[23,24], Claudio Solaro[25], Oliver Gerlach[26], Aysun Soysal[27], Jens Kuhle[28], Jose Luis Sanchez-Menoyo[29], Daniele Spitaleri[30], Tunde Csepany[31], Bart Van Wijmeersch[32], Radek Ampapa[33], Julie Prevost[34], Samia J. Khoury[35], Vincent Van Pesch[36], Nevin John[37], Davide Maimone[38], Bianca Weinstock-Guttman[39], Guy Laureys[40], Pamela McCombe[41], Yolanda Blanco[42], Ayse Altintas[43], Abdullah Al-Asmi[44], Justin Garber[45], Anneke Van der Walt[46], Helmut Butzkueven[46], Koen de Gans[47], Csilla Rozsa[48], Bruce Taylor[49], Talal Al-Harbi[50], Attila Sas[51], Cecilia Rajda[52], Orla Gray[53], Danny Decoo[54], William M. Carroll[55], Allan G. Kermode[56], Marzena Fabis-Pedrini[57], Deborah Mason[58], Angel Perez-Sempere[59], Mihaela Simu[60], Neil Shuey[61], Bhim Singhal[62], Marija Cauchi[63], Todd A. Hardy[64], Sudarshini Ramanathan[65], Patrice Lalive[66], Carmen-Adella Sirbu[67], Stella Hughes[68], Tamara Castillo Trivino[69], Liesbet M. Peeters[2,3,4] & Yves Moreau[1] ✉

Early prediction of disability progression in multiple sclerosis (MS) remains challenging despite its critical importance for therapeutic decision-making. We present the first systematic evaluation of personalized federated learning (PFL) for 2-year MS disability progression prediction, leveraging multi-center real-world data from over 26,000 patients. While conventional federated learning (FL) enables privacy-aware collaborative modeling, it remains vulnerable to institutional data heterogeneity. PFL overcomes this challenge by adapting shared models to local data distributions without compromising privacy. We evaluated two personalization strategies: a novel AdaptiveDualBranchNet architecture with selective parameter sharing, and personalized fine-tuning of global models, benchmarked against centralized and client-specific approaches. Baseline FL underperformed relative to personalized methods, whereas personalization significantly improved performance, with personalized FedProx and FedAVG achieving ROC-AUC scores of 0.8398 ± 0.0019 and 0.8384 ± 0.0014, respectively. These findings establish personalization as critical for scalable, privacy-aware clinical prediction models and highlight its potential to inform earlier intervention strategies in MS and beyond.

Multiple Sclerosis (MS) is a complex neurological disorder affecting millions of people worldwide[1]. In the absence of a cure, current treatment strategies focus on controlling disease progression and preventing relapses[2]. However, the heterogeneity of MS complicates disease management, as each patient experiences unique disease progressions and varying responses to treatment[3]. The primary challenge lies in capturing this heterogeneity to enable personalized, data-driven treatment strategies[4–6].

A promising approach for personalizing care involves leveraging the increasing availability of Real-World Data (RWD) through the application of Machine Learning (ML)[7]. Previous studies have shown that ML can

A full list of affiliations appears at the end of the paper. ✉e-mail: Yves.Moreau@esat.kuleuven.be

significantly improve our understanding of MS progression, uncover new biomarkers, and predict individual treatment responses[8–13].

Despite recent advances, developing advanced ML models for MS remains constrained by limited access to large-scale, high-quality datasets, which often require *data centralization*[14]. Although MS impacts an estimated 2.8 million individuals globally[1], the clinical data needed for precision modeling remain fragmented and siloed across healthcare institutions. Aggregating such data is complicated by legitimate but complex regulatory constraints, data ownership concerns, and inconsistent data quality standards[15–18]. Consequently, these factors present significant obstacles to conventional centralized model training, motivating the need for alternative approaches.

However, this centralization challenge is not unique to MS and has been observed in other fields as well, motivating the development of Federated Learning (FL)[19,20]. FL is a decentralized learning paradigm that enables training ML models while preserving data localization[21,22]. This decentralized approach is strongly aligned with data privacy and protection standards, offering a solution to the dilemmas inherent in data centralization[23–26].

Within healthcare, FL has shown success in a broad spectrum of applications, ranging from predicting hospitalization for cardiac events[21,27], to enhancing whole-brain magnetic resonance imaging segmentation[28], and even advancing drug discovery[29]. The potential of FL in MS is evident, although only a few studies have investigated this synergy, primarily focusing on imaging data[30–32].

Nevertheless, conventional FL methods rely on a single global model shared across all clients, which often performs poorly when local data distributions differ significantly. This challenge is particularly pronounced in MS, where clinical presentation, disease progression, and treatment response vary markedly across patients and institutions. In such heterogeneous settings, conflicting client updates can hinder convergence during training, while the absence of client-specific adaptation limits the model's relevance to local contexts. These shortcomings not only impair overall performance but may also reduce the incentive for participation among underrepresented clients[33]. Personalized Federated Learning (PFL) has emerged to address these gaps[33], enabling models to incorporate local data characteristics and thereby improving both predictive accuracy and robustness in diverse clinical environments such as MS.

Building on this motivation, our study evaluates the practical applicability of FL and PFL for analyzing routine clinical RWD in MS. We assessed multiple strategies for predicting disability progression and examined their feasibility in real-world healthcare environments. In doing so, we aimed to provide both empirical evidence and actionable insights to guide the effective deployment of FL-based solutions in clinical practice.

Specifically, we investigated five main data analysis paradigms: (1) centralized modeling, where all data are pooled into a single dataset; (2) baseline FL, which trains a joint model on decentralized data; (3) and (4) two PFL approaches that enable personalization; and (5) local modeling, where each client trains its model independently.

The first PFL approach introduces a novel ML architecture specifically designed for FL, called *AdaptiveDualBranchNet*. This method modifies the learning process by maintaining individual models with varying depths for each client and federating only partial model parameters. The second approach involves fine-tuning, where each client personalizes its own model after the FL setup using local data. The frameworks for baseline FL, as well as the adaptive and fine-tuned PFL approaches, are illustrated in Fig. 1, highlighting their respective architectures and workflows.

Our experiments leveraged the MSBase registry, the largest global database of MS patients, and simulated a realistic data partitioning scenario to reflect the heterogeneity observed in real-world clinical settings[34,35]. This comprehensive experimental design enabled us to identify the conditions under which FL can be effectively applied to MS research.

Contributions: (1) Systematic evaluation of FL and PFL in MS research: we present the first comprehensive assessment of federated and PFL approaches for modeling disability progression in MS using routine clinical RWD. (2) Identification of conditions for effective FL deployment: Through extensive benchmarking, we identify key factors that influence the success of FL in MS, offering actionable guidance for researchers and healthcare practitioners. (3) Development of AdaptiveDualBranchNet: we introduce AdaptiveDualBranchNet, a novel FL architecture that enables partial model sharing and demonstrates improved performance compared to existing FL baselines.

## Results

The federated experiments incorporated different strategies, including the FedAVG[36], FedProx[37], and FedOpt (FedYogi, FedAdam, FedAdagrad) algorithms[38]. For the binary classification task of this study, we carried out an extensive hyperparameter tuning using grid search for each FL strategy[39] to find the best set of parameters. To maintain consistency and strengthen the reliability of our results, we repeated each experiment 10 times to confirm the robustness of our findings. The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for these results are shown in Fig. 2, with detailed metrics provided in Table 1. The results of the experiment runtime as another metric also presented in Table 2.

### Centralized superiority in overall performance, bridged by personalized federated learning

Among these results, the centralized model consistently outperformed all federated models, achieving the highest ROC–AUC (0.8092 ± 0.0012) and AUC–PR (0.4605 ± 0.0043) scores. This highlights the benefits of having access to centralized data, which enables more effective model training.

Within the federated models, FedAdam and FedYogi demonstrated the best performance, with ROC–AUC and AUC–PR scores of 0.7920 ± 0.0031 and 0.4488 ± 0.0061 for FedAdam, and 0.7910 ± 0.0028 and 0.4420 ± 0.0078 for FedYogi, respectively. However, these gains came at the cost of higher computational demands, as FedAdam required the longest training time at 236 min, about 23% more than FedProx.

FedAVG provided a balanced alternative, with a ROC–AUC of 0.7840 ± 0.0019 and an AUC–PR of 0.4030 ± 0.0059, completing training in ~193 min. FedProx, while offering similar performance to FedAVG, reduced training time by around 11%, completing in 172 min, making it a practical option for faster execution.

FedAdagrad, although slightly lower in performance (ROC–AUC: 0.7762 ± 0.0021, AUC–PR: 0.3913 ± 0.0061), showed a comparable training duration of 190 min, balancing efficiency and accuracy.

Nevertheless, baseline FL remained limited in overall performance. As shown in Fig. 4, challenges such as non-IID data distributions, varying dataset sizes, and class imbalance hindered the model's ability to generalize. To overcome these limitations, we evaluated two personalization strategies aimed at improving overall performance and compared their results in the following analysis.

To begin with, *AdaptiveDualBranchNet* (referred to as "Adaptive") demonstrated clear improvements in model performance compared to the baseline FL paradigm. As shown in Table 1, the adaptive method consistently increased both ROC–AUC and AUC–PR across all FL strategies. However, as Table 2 indicates, these gains came at the cost of increased computational time.

FedProx, which is developed as a federated strategy to tackle system heterogeneity, variations in client data distributions and computational resources, demonstrated a 7.2% improvement in ROC–AUC and a 31% increase in AUC–PR over the baseline federated model. This suggests that FedProx's adaptive personalization enhances its ability to capture diverse data patterns effectively. However, this increased flexibility came with a 27% longer training time, highlighting a trade-off between improved accuracy and computational efficiency. Using FedAVG, a simpler strategy that averages neural network parameters across clients, showed a similar improvement in AUC–PR (28%) while requiring only a 7% increase in computational time. This indicates that Adaptive FedAVG benefits from personalization with minimal added computational burden, making it a strong choice in resource-constrained scenarios where personalization is desired. The adaptive approach also boosted the performance of FedAdagrad, FedYogi, and FedAdam, all of which utilize adaptive optimization

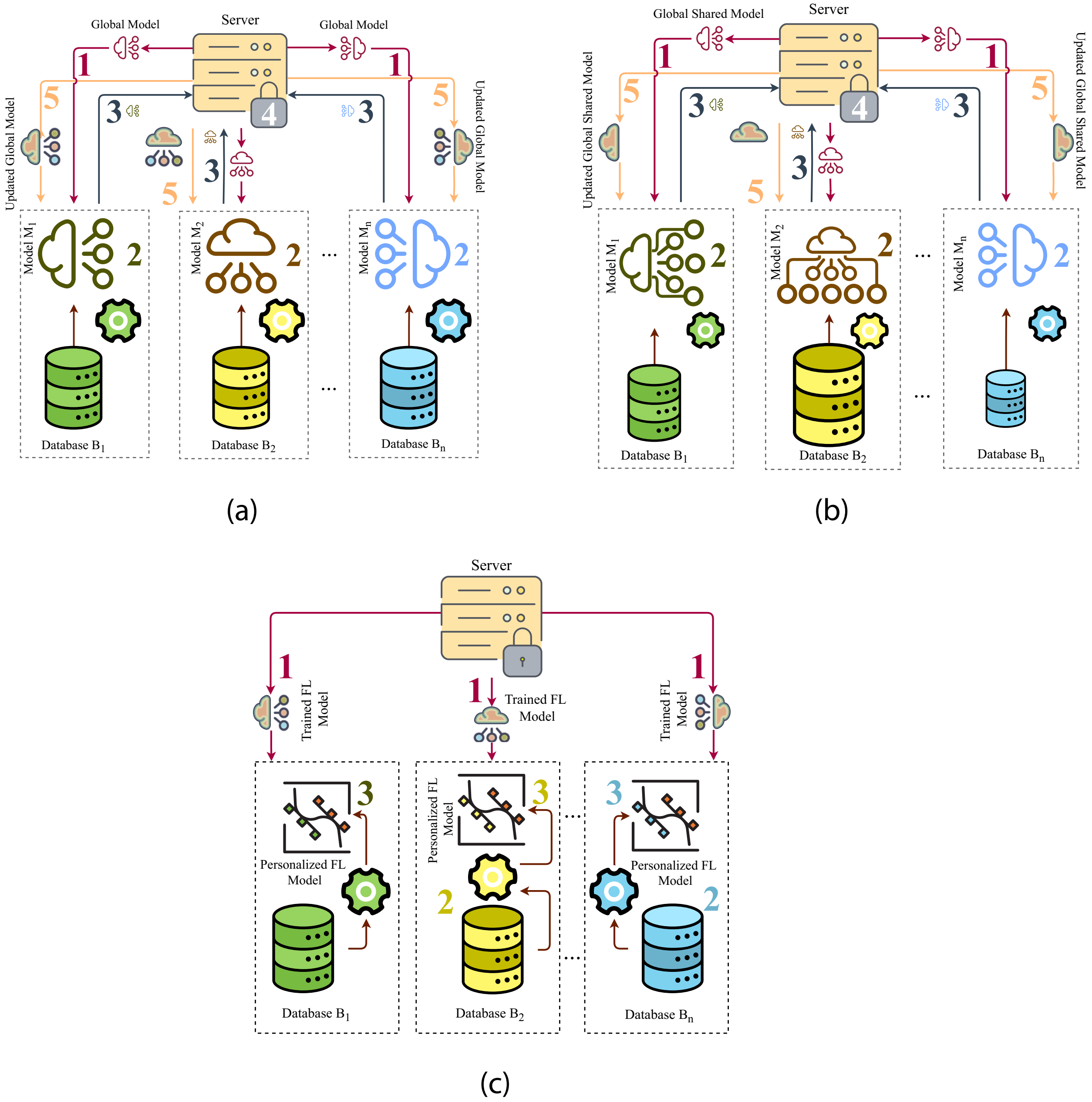


**Fig. 1 | Analysis paradigms for predicting disability progression in multiple sclerosis using real world data. a** Baseline Federated Learning (FL): Depicts the classic iterative process in which multiple clients (e.g., clinical sites) collaboratively train a single global model. Each client gets the current global model (Step 1) and trains it locally on its private dataset (Step 2). The locally updated parameters are then uploaded to a central server (Steps 3) and aggregated (Step 4), refining the global model without exchanging any raw patient data. (Step 5) Dissemination of the updated global model to all clients for continued local training based on aggregated knowledge. **b** Personalized Federated Learning (PFL) with Adaptive Partial Parameter Exchanges (AdaptiveDualBranchNet): Illustrates a dual branch architecture where each client's model is split into a shared core (federated across all clients) and local extension layers (trained solely with private data). During each federated round (Steps 1, 3, 5), only the shared core parameters are exchanged and aggregated at the central server, preserving common knowledge. The local extension layers remain entirely onsite (Step 2), allowing each client to further personalize its model based on unique data distributions or sample sizes. **c** PFL via Fine Tuning: Shows how a pre-trained global FL model (Step 1) is shared with each client. Each client fine tunes this model on its local dataset (Step 2), creating a personalized version (Step 3) that reflects client specific characteristics. This approach retains the benefits of cross site collaboration while allowing for tailored predictions. Collectively, these paradigms form part of a broader analysis that also includes centralized (pooled data) and local (client specific) training baselines. This holistic evaluation framework helps elucidate the strengths and trade offs of each approach in leveraging real world data for predicting disability progression in multiple sclerosis.

techniques. FedAdagrad demonstrated a 7.7% improvement in ROC–AUC and a 31% increase in AUC–PR, though it came with a 21% increase in training time. FedYogi improved AUC–PR by 6.4%, but required a 28% increase in time. On the other hand, FedAdam achieved a 16% increase in AUC–PR with only a 4% increase in training time, indicating its efficiency in handling personalization through optimized learning rate and momentum adjustments, although its initial training time was relatively high.

Following adaptive personalization, fine-tuning also improved model performance by adapting the global model to each client's local data distribution, allowing it to better capture individual patient patterns. Fine-tuned models consistently demonstrated improvements in both ROC–AUC and AUC–PR metrics, as summarized in Table 1. For instance, both FedProx and FedAVG achieved the highest performance following fine-tuning. The FedProx model reached a ROC–AUC of

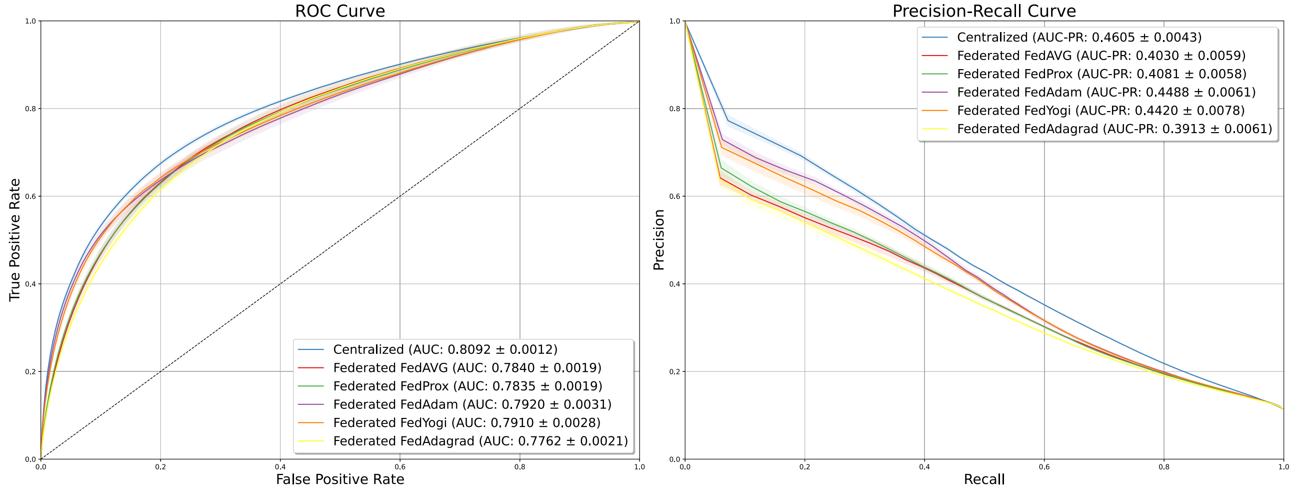


**Fig. 2 | Comparative analysis of federated learning model performance.** Evaluating ROC–AUC and AUC–PR Metrics Across Different Strategies. (Left) Receiver Operating Characteristic: The centralized model achieves the highest performance with an ROC–AUC = 0.8092 ± 0.0012, demonstrating the advantage of having a pooled dataset. Among FL strategies, FedAdam and FedYogi perform best, with ROC–AUC values of 0.7920 ± 0.0031 and 0.7910 ± 0.0028, respectively. The other FL methods, including FedAvg and FedProx, show slightly lower performance, underscoring the challenges of a global federated model in heterogeneous data settings. (Right) Precision-Recall Curve: Again, the centralized model outperforms with an AUC–PR = 0.4605 ± 0.0043. Among FL methods, FedAdam achieves the highest AUC–PR of 0.4488 ± 0.0061, while FedYogi and FedProx follow closely with values of 0.4420 ± 0.0078 and 0.4081 ± 0.0058, respectively. The drop in performance compared to the centralized approach reflects the difficulty of capturing minority class predictions in federated settings. These results emphasize the performance gap between centralized and federated learning strategies, particularly in heterogeneous and imbalanced data scenarios.

**Table 1 | Performance metrics (ROC–AUC and AUC–PR) for personalized federated learning models compared to federated learning baseline across various strategies**

| | ROC–AUC ↑ | | | AUC–PR ↑ | | |
|---|---|---|---|---|---|---|
| | Personalized FL | | FL | Personalized FL | | FL |
| Experiments | Fine-tuned | Adaptive | Baseline | Fine-tuned | Adaptive | Baseline |
| FedAVG | **0.8370 ± 0.0016** | **0.8384 ± 0.0014** | 0.7840 ± 0.0019 | 0.5156 ± 0.0046 | **0.5290 ± 0.0062** | 0.4030 ± 0.0059 |
| FedProx | **0.8375 ± 0.0019** | **0.8398 ± 0.0019** | 0.7834 ± 0.0019 | 0.5221 ± 0.0044 | **0.5346 ± 0.0029** | 0.4081 ± 0.0058 |
| FedAdagrad | 0.8340 ± 0.0012 | 0.8361 ± 0.0021 | 0.7762 ± 0.0021 | 0.5043 ± 0.0043 | 0.5131 ± 0.0062 | 0.3913 ± 0.0061 |
| FedYogi | 0.8369 ± 0.0027 | 0.8178 ± 0.0026 | **0.7910 ± 0.0028** | **0.5379 ± 0.0072** | 0.4702 ± 0.0059 | **0.4420 ± 0.0078** |
| FedAdam | 0.8339 ± 0.0015 | 0.8324 ± 0.0032 | **0.7920 ± 0.0031** | **0.5383 ± 0.0050** | 0.5197 ± 0.0073 | **0.4488 ± 0.0061** |
| Centralized | | 0.8092 ± 0.0012 | | | 0.4605 ± 0.0043 | |

"Centralized" results are included for comparison purposes and do not fall under the FL or PFL categories. For brevity, the term "AdaptiveDualBranchNet" will be referred to simply as "Adaptive" throughout this manuscript. Additionally, the non-personalized FL model is commonly referred to as the baseline FL paradigm. The value after "±" denotes the standard deviation of the measurements. The bold values indicate the best-performing results within each row, where higher values are better, as shown by the arrows in the column headers.

**Table 2 | Experiment timing comparison: personalized vs. baseline federated learning (in min)**

| Method | Experiment Time (min) ↓ | |
|---|---|---|
| | Adaptive | Baseline |
| FedAVG | **206.81 ± 3.670** | 192.66 ± 5.370 |
| FedProx | 217.51 ± 13.18 | **171.77 ± 9.440** |
| FedAdagrad | 230.91 ± 18.27 | 190.02 ± 2.240 |
| FedYogi | 246.74 ± 5.420 | 193.21 ± 41.89 |
| FedAdam | 225.69 ± 7.910 | 236.08 ± 13.42 |

Where lower timing values are preferred, bold is used to highlight the best timing for each model type (i.e., the better-performing federated strategy between Adaptive and Baseline for that specific method). In cases where two values are very close and not meaningfully different, both are bolded to reflect comparable performance.

0.8375, reflecting a 6.91% improvement over its baseline of 0.7834, while FedAVG closely followed with a ROC–AUC of 0.8370. For AUC–PR, FedProx saw an increase from 0.4081 to 0.5221, representing an ~28% gain.

Among the strategies tested, FedAdagrad benefited the most from fine-tuning, with a 7.5% increase in ROC–AUC and a 28.9% rise in AUC–PR. This suggests that FedAdagrad was particularly responsive to fine-tuning, allowing it to better adapt to the data distributions specific to individual clients.

Figure 3 also confirms that fine-tuned models generally outperform federated models. This is evident from the consistently positive differences in both ROC-AUC and AUC-PR scores across all methods. Moreover, the fact that these differences rarely cross zero indicates that the advantages of Adaptive models are statistically meaningful. When it comes to comparing adaptive and fine-tuned models, the differences are minor and hover around zero, suggesting that these two approaches are quite similar overall. While adaptive models hold a slight edge in methods like FedAdam and FedYogi. Note that the *x*-axis scale varies across the plots.

### Federated models struggle against centralized model at the client level

To perform a detailed client-level comparison across countries, we evaluated five key paradigms: federated, fine-tuned, adaptive, centralized, and local. For consistency, FedProx was selected as the representative federated model

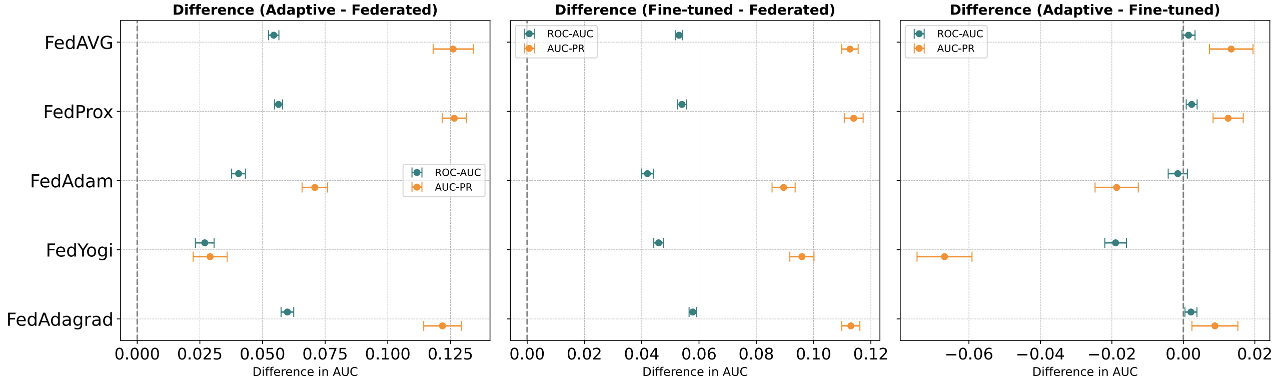


**Fig. 3 | Comparison of ROC–AUC and AUC–PR differences among adaptive, fine-tuned, and federated models.** This figure shows average ROC–AUC and AUC–PR score differences across five strategies (FedAVG, FedProx, FedAdam, FedYogi, and FedAdagrad) for each pairwise comparison of model types. Error bars represent 95% confidence intervals from multiple runs, with a dashed vertical line at zero indicating no difference between models.

**Table 3 | Performance scores for different learning paradigms across countries**

| | | ROC–AUC ↑ | | | | | AUC–PR ↑ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **Dataset Size** | **Federated** | **Adaptive** | **Fine-tuned** | **Centralized** | **Local** | **Federated** | **Adaptive** | **Fine-tuned** | **Centralized** | **Local** |
| CZ | 55435 | 0.8909 | **0.9211** | 0.9193 | 0.8781 | 0.8669 | 0.5875 | **0.7214** | 0.6990 | 0.5568 | 0.4600 |
| IT | 54354 | 0.7946 | **0.8168** | 0.8108 | 0.7769 | 0.7603 | 0.4590 | **0.5099** | 0.4924 | 0.4312 | 0.3491 |
| TR | 37853 | 0.8365 | **0.8887** | 0.8854 | 0.8467 | 0.8498 | 0.4427 | **0.5953** | 0.5716 | 0.483 | 0.4161 |
| ES | 33396 | 0.7680 | **0.8027** | 0.7969 | 0.7713 | 0.7609 | 0.3819 | **0.4564** | 0.4393 | 0.4055 | 0.3381 |
| CA | 27131 | 0.7370 | **0.7619** | 0.7615 | 0.7488 | 0.7332 | 0.3383 | **0.3965** | 0.3949 | 0.3893 | 0.3407 |
| AU | 23906 | 0.7300 | **0.7679** | 0.7657 | 0.7490 | 0.7287 | 0.3098 | **0.3962** | 0.3805 | 0.3882 | 0.3129 |
| PT | 6884 | 0.7449 | 0.8475 | **0.8525** | 0.8252 | 0.7972 | 0.2627 | 0.4524 | **0.4774** | 0.449 | 0.3562 |
| BE | 6534 | 0.6495 | 0.8115 | **0.8156** | 0.7963 | 0.7987 | 0.1559 | 0.3980 | **0.4050** | 0.3553 | 0.2987 |
| KW | 5725 | 0.7445 | 0.9137 | 0.9128 | 0.8761 | **0.9164** | 0.1661 | 0.5104 | **0.5111** | 0.4558 | 0.4850 |
| HU | 4892 | 0.7128 | 0.9608 | **0.9632** | 0.9495 | 0.9549 | 0.3099 | **0.7810** | 0.779 | 0.7311 | 0.5924 |
| NL | 4869 | 0.5614 | 0.6595 | 0.6782 | 0.6873 | **0.7107** | 0.1797 | 0.2539 | 0.2756 | **0.2826** | 0.2650 |
| TN | 4780 | 0.7857 | **0.9535** | 0.9535 | 0.9319 | 0.9312 | 0.503 | 0.8639 | **0.8664** | 0.8178 | 0.8040 |
| CH | 3836 | 0.6212 | 0.7650 | 0.7700 | **0.7925** | 0.7274 | 0.1232 | 0.3042 | **0.3196** | 0.3084 | 0.1952 |
| IR | 2980 | 0.6396 | 0.8269 | **0.8330** | 0.8158 | 0.7471 | 0.2514 | 0.5570 | **0.5702** | 0.5345 | 0.3682 |
| AR | 2440 | 0.6714 | **0.8856** | 0.8719 | 0.8274 | 0.8784 | 0.2625 | 0.6287 | **0.6331** | 0.5801 | 0.5946 |
| LB | 1937 | 0.5955 | **0.7589** | 0.7398 | 0.7314 | 0.6553 | 0.1171 | **0.3343** | 0.3187 | 0.2756 | 0.2255 |
| US | 1344 | 0.5627 | 0.7303 | 0.7368 | **0.7437** | 0.7044 | 0.1383 | 0.2493 | 0.2797 | **0.3128** | 0.2433 |
| IL | 1140 | 0.6937 | 0.8750 | **0.8782** | 0.8537 | 0.8503 | 0.2604 | 0.5986 | **0.6310** | 0.5198 | 0.5181 |
| OM | 969 | 0.5339 | 0.8472 | **0.8731** | 0.8093 | 0.7981 | 0.0962 | 0.5421 | **0.5897** | 0.4763 | 0.2995 |
| CU | 782 | 0.5625 | 0.7971 | 0.8050 | 0.8062 | **0.8266** | 0.1864 | 0.4190 | 0.4744 | 0.4775 | **0.5260** |
| BR | 578 | 0.5768 | **0.8063** | 0.7680 | 0.7307 | 0.7070 | 0.1434 | 0.4308 | 0.4200 | **0.4677** | 0.4605 |
| SA | 256 | 0.6749 | 0.8915 | 0.8677 | **0.9374** | 0.8827 | 0.2466 | 0.659 | 0.5619 | 0.6851 | **0.7674** |
| GB | 221 | 0.6520 | 0.6060 | 0.6880 | **0.8510** | 0.5333 | 0.2576 | 0.3250 | 0.3007 | **0.4774** | 0.2529 |
| NZ | 110 | 0.3286 | **0.6095** | 0.4857 | 0.4190 | 0.5873 | 0.0738 | **0.1247** | 0.1094 | 0.0810 | 0.1057 |
| GR | 99 | 0.7240 | 0.8714 | 0.8703 | **0.9292** | 0.7998 | 0.6495 | 0.8794 | 0.8777 | **0.9082** | 0.8048 |
| WA | | 0.7835 | **0.8398** | 0.8375 | 0.8092 | 0.7983 | 0.4081 | **0.5346** | 0.5221 | 0.4605 | 0.3874 |

The table reports ROC–AUC and AUC–PR metrics for federated learning, adaptive, fine-tuned, centralized, and local approaches. The bold values indicate the best-performing results within each row, where higher values are better, as shown by the arrows in the column headers.

because of its balance of performance and computational efficiency. While FedAdam showed marginally better performance, FedProx offered a more favorable trade-off between accuracy and resource usage. Nonetheless, comprehensive results for all federated strategies are provided in Supplementary Tables 1–4.

The results presented in Table 3 indicated that the centralized model outperformed all baseline federated approaches, achieving a mean weighted average ROC–AUC of 0.8092 and an AUC–PR of 0.4605. This represents an improvement of ~3.3% in ROC–AUC and 12.8% in AUC–PR over the federated model. These results highlight the advantage of centralized training, as it fully leverages the entire dataset, leading to better performance compared to federated models.

### Local models capture specific client insights but lack generalization

To provide a more comprehensive perspective, we also calculated the weighted averages for the local models. The local model achieved a ROC–AUC of 0.7983 and an AUC–PR of 0.3874, placing its ROC–AUC performance between that of the centralized and federated models, though its AUC–PR lagged behind the federated models. This suggests that while local models benefit from training on specific client data, they may lack the broader insights captured by centralized models and the generalized patterns learned by federated models.

### Personalization enhances client-level performance in federated learning

Bringing together the results from PFL, baseline FL, centralized, and local models, we analyzed the performance scores presented in Table 3. Notably, we observed that while the federated model initially lagged behind the centralized model, the PFL approach allowed it to close the performance gap and eventually surpass the centralized model. This suggests that personalization can significantly boost the effectiveness of FL by adapting to client-specific data distribution.

From a broader perspective, the adaptive and fine-tuned paradigms are the top performers, with adaptive models leading in 11 countries and fine-tuned models in 6. This demonstrates the clear advantage of PFL approaches, which consistently achieve the highest ROC–AUC scores across different countries. The centralized paradigm ranks next, leading in five countries, showing its occasional competitiveness. The local model outperformed others in three countries, while federated paradigms did not achieve the highest performance in any country. This suggests that, without the personalized adjustments or aggregation benefits seen in centralized models, federated approaches struggle to match the predictive accuracy of the other paradigms. A similar pattern was observed for AUC–PR.

### Federated vs. local models: impact of dataset size

To assess whether countries experienced greater benefits from federated or local models (*excluding the centralized and PFL paradigms*), our analysis of the ROC–AUC metric revealed that 19 cases favored local models, while six cases showed better results with the federated approach. Focusing on dataset size, particularly for intermediate-sized countries (ranging from BE to AR), we observed significant performance differences in favor of local models, with an average advantage of 15.98% in ROC–AUC.

In contrast, these differences were much less pronounced in countries with larger datasets, such as those from CZ to AU, where federated models outperformed local models in five of the top six cases.

Further analysis using Spearman correlation confirmed these observations, revealing a moderate negative correlation ($\rho = -0.503$, $p = 0.010$) between dataset size and the performance gap (ROC–AUC difference) between federated and local models. This pattern suggests that as dataset size increases, federated models can generalize more effectively, achieving performance comparable to or even surpassing that of local models. Countries with smaller datasets, the federated model typically underperforms in comparison to local models, with notable differences in metrics scores. It appears that local models, when trained on smaller, more specific datasets, are able to capture unique dataset characteristics that the federated model—due to its aggregated, generalized approach—may fail to recognize. Similarly, the performance gap between federated and centralized models showed a strong negative correlation ($\rho = -0.761$, $p < 0.001$), underscoring the ability of federated models to approach centralized model performance when trained on larger datasets.

### Performance trends in the largest data-contributing countries

Focusing on analyzing the top six countries, CZ, IT, TR, ES, CA, AU, which hold 82% of the data as highlighted in Fig. 4d, Our goal is to identify the most effective alternative paradigms in comparison to one another. The findings from this analysis summarized in Table 4 illustrates a notable trend: across the six countries, the adaptive paradigm outperforms other approaches, closely followed by the fine-tuned model. These two paradigms frequently demonstrate superior ROC–AUC scores, surpassing local, centralized and federated models. A detailed breakdown highlighting the best-performing models for each country is provided in Supplementary Table 5.

## Discussion

Advancing ML models for complex conditions such as MS requires access to large and diverse datasets. However, centralizing sensitive patient data from multiple institutions presents significant regulatory, logistical, and ethical challenges[14–18]. FL offers a privacy-aware alternative by enabling collaborative model training across decentralized datasets[19,20]. Yet, baseline FL methods, such as FedAVG, which learn a single global model, often underperform in the presence of substantial statistical heterogeneity (non-IID data) common in multi-institutional clinical datasets[23,36]. These limitations motivate the development of PFL approaches, which aim to adapt models to the unique characteristics of each client while preserving the collaborative benefits of FL[33].

To systematically investigate this challenge, we compared multiple FL paradigms against centralized and local baselines for predicting MS disability progression. The results revealed a clear performance hierarchy (Tables 1, 3): PFL strategies, including Adaptive and fine-tuning methods, achieved the highest discrimination performance, surpassing both centralized and local models. Although centralized models performed well and ranked third overall, their feasibility is often constrained by privacy regulations and logistical barriers that complicate large-scale data aggregation. In many practical healthcare scenarios, assembling centralized datasets remains challenging. These findings underscore the relevance of federated approaches, which enable collaborative model development without compromising data sovereignty. Importantly, our results show that with appropriate personalization strategies, FL can become a practical and privacy-respecting alternative to centralized training.

Beyond discrimination performance, clinical applicability also requires strong calibration and reliable risk estimation. While top-performing PFL models, such as Adaptive FedProx (ROC-AUC 0.8398), achieved discrimination comparable to published benchmarks for similar MS prediction tasks[10,13], discrimination alone does not guarantee clinical utility. For meaningful clinical use, particularly at the individual patient level, predictions must accurately reflect the true risk. Current model performance may support applications such as cohort-level monitoring or population health analyses, but further improvements in predictive certainty and calibration are needed before safe deployment in high-stakes clinical decision-making. In particular, comprehensive reporting of calibration metrics, including Brier scores, expected calibration error, and calibration diagrams, would provide a more complete assessment of model reliability[40]. Ensuring that predictive outputs are trustworthy is essential to avoid patient harm and misinformed management[41].

Ensuring clinical applicability not only requires discrimination and calibration but also demands flexible modeling approaches that can adapt to local data characteristics without sacrificing global knowledge transfer. Addressing this need, our proposed AdaptiveDualBranchNet architecture contributes to architectural personalization strategies by enhancing model flexibility. It introduces dynamic depth in the personalized branch, allowing complexity to scale with local data volume. Unlike prior approaches such as FedPer[42], which rely on fixed partitions between shared and private layers, AdaptiveDualBranchNet retains the full global model and integrates client-specific layers in parallel. This design preserves the expressive capacity and transferability of the shared global representation while enabling client-specific adaptation. Benchmarking against representative PFL baselines, detailed in Supplementary Note Section 1.4, further supports the effectiveness of this approach.

Building upon these improvements, future work could explore dynamic or learning-based adaptations to enhance personalization flexibility. Rather than relying on heuristically defined thresholds for scaling

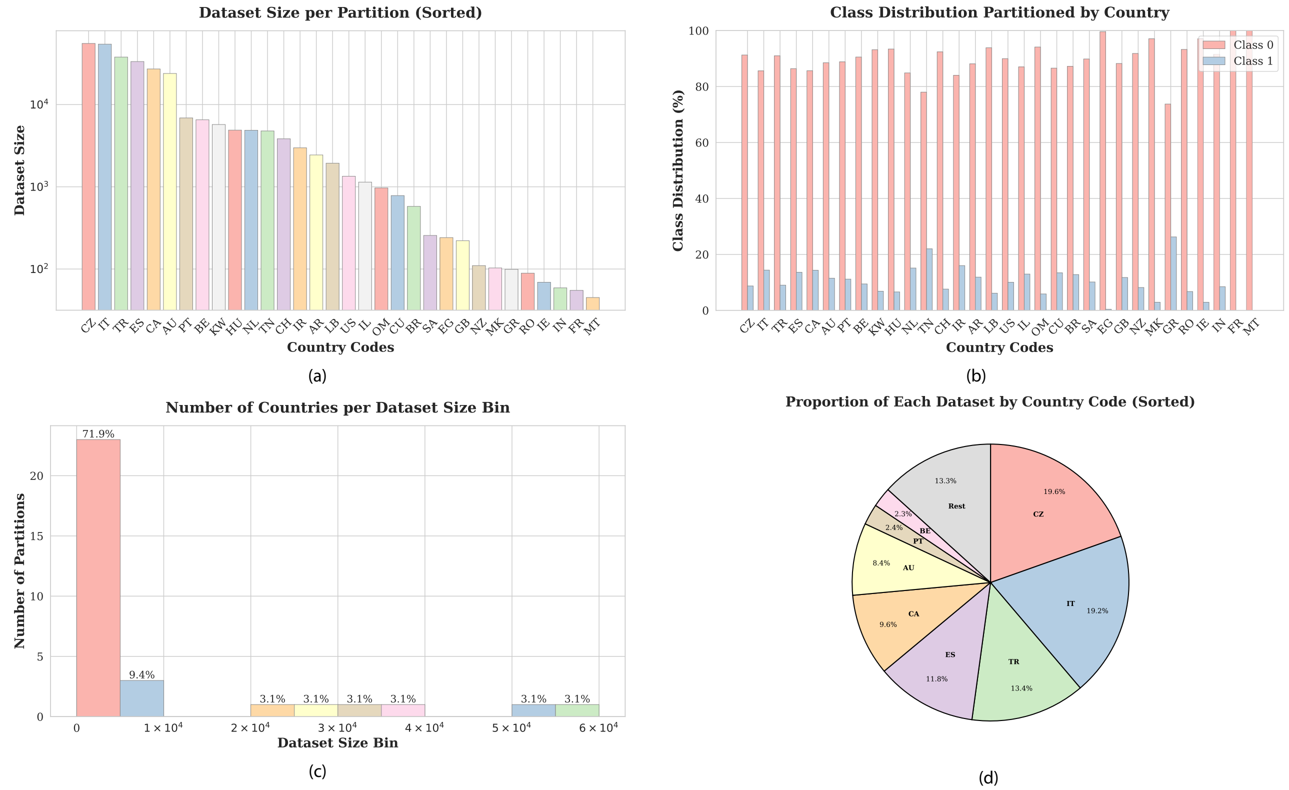


**Fig. 4 | Heterogeneity of country-specific data partitions for federated learning.** **a** Log-scaled distribution of country-specific dataset sizes $D_{C_i}$, sorted in descending order, highlighting disparities in data contributions. **b** Class imbalance across countries, showing underrepresentation of Class 1 (MS worsening confirmed) relative to Class 0. **c** Histogram of dataset sizes using 5K bin intervals, emphasizing skewed availability across participating centers. **d** Pie chart illustrating the proportional contribution of each country to the overall dataset. Together, these analyses demonstrate the significant variability in both data quantity and label distributions across clients, underscoring the challenges faced by federated learning models operating in real-world clinical settings.

**Table 4 | Performance comparison of top six countries by dataset size across different paradigms**

| Country | Best Paradigm | Second Best Paradigm | Third Best Paradigm | Worst Paradigm |
|---|---|---|---|---|
| CZ | Fine-tuned FedYogi 0.9222 | Adaptive FedProx 0.9211 | Federated FedYogi 0.901 | Local 0.8669 |
| | Best-Worst: 0.0553 | Second Best-Worst: 0.0542 | Third Best-Worst: 0.0341 | |
| IT | Adaptive FedProx 0.8168 | Fine-tuned FedYogi 0.8151 | Federated FedAVG 0.7960 | Local 0.7603 |
| | Best-Worst: 0.0565 | Second Best-Worst: 0.0548 | Third Best-Worst: 0.0357 | |
| TR | Adaptive FedAVG 0.8891 | Fine-tuned FedYogi 0.8881 | Local 0.8498 | Federated FedAdagrad 0.8339 |
| | Best-Worst: 0.0552 | Second Best-Worst: 0.0542 | Third Best-Worst: 0.0159 | |
| ES | Adaptive FedAVG 0.8028 | Fine-tuned FedYogi 0.8020 | Federated FedYogi 0.7793 | Local 0.7609 |
| | Best-Worst: 0.0419 | Second Best-Worst: 0.0411 | Third Best-Worst: 0.0121 | |
| CA | Adaptive FedProx 0.7619 | Fine-tuned FedAdagrad 0.7617 | Centralized 0.7488 | Federated FedAdagrad 0.7327 |
| | Best-Worst: 0.0292 | Second Best-Worst: 0.0290 | Third Best-Worst: 0.0161 | |
| AU | Adaptive FedAVG 0.7692 | Fine-tuned FedAVG 0.7678 | Centralized 0.749 | Federated FedAdagrad 0.7187 |
| | Best-Worst: 0.0505 | Second Best-Worst: 0.0491 | Third Best-Worst: 0.0303 | |

The second line of each multi-row represents the difference between the given paradigm's performance and the worst performance.

model depth, allowing models to autonomously adjust their complexity in response to richer client-specific data signals may further improve generalization and robustness. Beyond architectural adaptivity, optimization dynamics also represent an important axis for personalization. In this study, a uniform learning rate was applied across all parameters of the models, with dynamic adjustment governed by a ReduceLROnPlateau scheduler[43]. Preliminary investigations into more granular strategies, such as distinct learning rates for global and personalized branches, differential regularization, and gradient clipping, suggested additional opportunities for performance gains. Although not included in the final experiments to maintain a controlled baseline, these approaches highlight promising future directions. Moreover, client-specific hyperparameter adaptation, modulating local learning rates, batch sizes, or regularization strengths based on client characteristics[44–46], may further improve fairness, stability, and adaptability across heterogeneous federated settings. Together, dynamic architectural scaling and personalized optimization strategies could enable more resilient and equitable predictive modeling in decentralized healthcare applications.

Complementary to adaptive architectures, post-hoc fine-tuning provided an additional strategy for personalization. Fine-tuning global models on local data distributions improved model performance but showed dependency on data sufficiency. Clients with sparse data faced greater risks of overfitting and unstable evaluation due to small or imbalanced test sets. Stratified analysis (Supplementary Fig. 2) revealed that fine-tuning yielded the most consistent improvements for clients with intermediate data volumes, particularly those initially underserved by the global model. Detailed stratified results are provided in Supplementary Note Section 1.5. These findings emphasize the need for personalization methods that remain effective even in low-data settings, such as adapter-based fine-tuning[47,48] or selective layer updating combined with lightweight regularization[49].

Several limitations of this study should be acknowledged. First, the study relied on simulation due to data governance constraints. Although the simulation framework was designed to accurately model algorithmic behavior and data heterogeneity, it cannot fully capture real-world factors such as network variability, system heterogeneity, or participant dropout[50,51]. Consequently, certain absolute metrics, such as the ~15% increase in training time observed for the Adaptive model, may not generalize directly. Although computational overhead was manageable in simulation, real-world deployments will be necessary to properly assess practical resource demands and communication efficiency. Nevertheless, the relative performance hierarchies observed across paradigms offer valuable hypotheses for future validation under real-world constraints.

Further limitations include the use of retrospective RWD[7,52], which may introduce missingness or bias, and the country-based data partitioning schema. While pragmatic, country-level partitioning may not fully capture natural clinical or institutional boundaries. Exploring alternative partitioning strategies, such as clinic-level, regionally grouped, or quantity-skewed clients, could provide further insights into model robustness under varied federated topologies. In addition, the use of weighted averaging during evaluation may introduce bias favoring larger clients, as discussed in more detail in Supplementary Note 1.1. Future work could explore complementary evaluation strategies, such as server-side global testing, to provide a more balanced assessment of model generalizability.

Translating FL models into clinical practice will require rigorous external validation using independent cohorts across diverse real-world settings, patient populations, and infrastructures. Such efforts could be facilitated by initiatives like the European Health Data Space[53,54] and should align with established reporting guidelines such as TRIPOD[55] to ensure methodological transparency and reproducibility.

Sustainable and scalable deployment of FL in healthcare will also require supportive ecosystem development, including trusted intermediaries, standardized governance protocols, and mechanisms for equitable benefit-sharing. Initiatives such as the Global Data Sharing Initiative in MS[14,56] and projects like MELLODDY[57] illustrate promising models for federated collaboration across institutions and industries.

Future technical enhancements should explore integrating multimodal data sources (e.g., MRI, Evoked Potentials[58]), adopting advanced privacy-preserving techniques (e.g., differential privacy, secure multi-party computation[59,60]) while balancing trade-offs, and developing refined personalization strategies, particularly for low-resource clients.

Ultimately, unlocking the clinical potential of FL will depend not only on technical advances but also on embedding FL within a broader healthcare ecosystem that supports data harmonization, clinician engagement, regulatory alignment on fairness, interpretability[61], and privacy. Demonstrating real-world clinical utility through prospective impact studies will be essential to validate technical performance and build trust in FL-enabled decision support as a safe, fair, and effective tool for patient care[62].

## Methods

### Cohort definition and episode extraction

Data of individuals diagnosed with MS were systematically collected and combined from 146 distinct centers, as documented in the MSBase registry up to September 2020[34,35]. The data was collected during routine clinical care at tertiary MS centers. The preliminary extraction of data from MSBase was governed by certain inclusion criteria: a minimum follow-up period of 12 months, a minimum age of 18 years, and a diagnosis of either relapsing remitting MS, secondary progressive MS, primary progressive MS, or clinically-isolated syndrome. The resulting dataset encompassed 44,886 patients. To uphold the integrity of the data, several quality assurance measures were employed. These entailed the elimination of duplicate or inconsistent visits recorded on the same day, removal of visits dated before 1970, and exclusion of patients exhibiting clinically isolated syndrome at their last documented visit.

Each patient's clinical trajectory was segmented into multiple, potentially overlapping, *episodes*, using the exact methodology for definition and extraction established and validated in prior work by De Brouwer et al.[13]. For clarity and completeness within this manuscript, we specify the definition used: Each episode represents a distinct instance for predicting disability progression and comprises three core components:

1. A **Baseline EDSS Measurement**: A single Expanded Disability Status Scale (EDSS) score recorded at time $t = 0$.
2. An **Observation Window**: This includes the complete available clinical history for the patient prior to the baseline measurement ($t \leq 0$), encompassing all recorded EDSS scores, Kurtzke Functional Systems (KFS) scores, relapse information, treatment history, and other relevant covariates from the MSBase registry. The duration of this observation window is therefore variable, depending on the length of the patient's recorded history up to $t = 0$.
3. A **Disability Progression Label**: A binary outcome indicating whether confirmed disability progression occurred within the two-year period following the baseline EDSS measurement ($0 < t \leq 2$ years). Confirmed disability progression required demonstrating a sustained increase in EDSS, based on thresholds defined by Kalincik et al.[63], confirmed over a period of at least six months, and excluding any EDSS measurements taken within one month of a recorded relapse.

An episode was considered valid for inclusion only if: (i) the observation window contained at least three EDSS measurements within the 3.25 years immediately preceding the baseline ($t = 0$), ensuring sufficient recent data density; and (ii) adequate follow-up data existed after $t = 0$ to ascertain the confirmed disability progression status within the 2-year prediction window. Critically, although episodes from the same patient may share common historical data, each valid episode, defined by its unique baseline time point and subsequent 2-year outcome period, was treated as an independent instance for model training and evaluation.

This curation and episode extraction process yielded a final dataset $D$ comprising $|D| = 283{,}115$ valid episodes derived from 26,246 unique patients. This dataset forms the basis for the binary classification task aimed at predicting disease disability progression within a two-year horizon. For a more comprehensive description of data variables, their definitions, and their preprocessing, we refer to the publication by De Brouwer et al[13].

As the objective of this study is to evaluate the effectiveness of FL, it was essential to partition the centralized global dataset $D$ to set up FL experiments. The global dataset (preprocessed dataset from ref. 13) $D$ included a key feature indicating the geographical origin of the data. Using this feature, the dataset $D$ was divided into 32 disjoint subsets $D_{C_i}$, each corresponding to a different country, as defined by: $D = \bigcup_{i=0}^{31} D_{C_i}, \quad |D_{C_i}| \geq 5, \quad \forall i \in \{0, 1, \ldots, 31\}$. Within each subset, the data was split into 60% training, 20% validation, and 20% test. Normalization was performed independently for each partition using statistics (mean and standard deviation) derived from its respective training set.

Upon detailed examination of this partitioning scheme, significant variations were evident in the dataset sizes across the created countries, as depicted in Fig. 4a. Figure 4c highlights that, ~75% of countries included fewer than 5000 samples. This partitioning was particularly revealing when considered alongside the pie chart analysis in Fig. 4d, which showed that six countries (CZ, IT, TR, ES, CA, AU) accounted for 82% of the total cohort size. This comparison indicated that while the majority of the dataset was concentrated in a few countries, many countries did not hold a significant share of the total data.

Further analysis revealed substantial class imbalance across countries, captured in Fig. 4b and Table 5. Both illustrate pronounced variability in the proportions of Class 0 ("MS worsening not confirmed") and Class 1 ("MS worsening confirmed"), with several countries exhibiting complete absence

**Table 5 | Country-level dataset sizes and class distributions, sorted by dataset size and organized column-wise**

| Country | Dataset Size | Class | | Country | Dataset Size | Class | | Country | Dataset Size | Class | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Class 0 (%) | Class 1 (%) | | | Class 0 (%) | Class 1 (%) | | | Class 0 (%) | Class 1 (%) |
| **CZ** | 55435 | 91.26 | 8.74 | **TN** | 4780 | 77.99 | 22.01 | **EG** | 241 | 99.59 | 0.41 |
| **IT** | 54354 | 85.60 | 14.40 | **CH** | 3836 | 92.39 | 7.61 | **GB** | 221 | 88.24 | 11.76 |
| **TR** | 37853 | 90.98 | 9.02 | **IR** | 2980 | 83.99 | 16.01 | **NZ** | 110 | 91.82 | 8.18 |
| **ES** | 33396 | 86.39 | 13.61 | **AR** | 2440 | 88.11 | 11.89 | **MK** | 103 | 97.09 | 2.91 |
| **CA** | 27131 | 85.65 | 14.35 | **LB** | 1937 | 93.86 | 6.14 | **GR** | 99 | 73.74 | 26.26 |
| **AU** | 23906 | 88.50 | 11.50 | **US** | 1344 | 89.96 | 10.04 | **RO** | 89 | 93.26 | 6.74 |
| **PT** | 6884 | 88.83 | 11.17 | **IL** | 1140 | 87.02 | 12.98 | **IE** | 69 | 97.10 | 2.90 |
| **BE** | 6534 | 90.54 | 9.46 | **OM** | 969 | 94.12 | 5.88 | **IN** | 59 | 91.53 | 8.47 |
| **KW** | 5725 | 93.15 | 6.85 | **CU** | 782 | 86.57 | 13.43 | **FR** | 55 | 100.00 | 0.00 |
| **HU** | 4892 | 93.40 | 6.60 | **BR** | 578 | 87.20 | 12.80 | **MT** | 45 | 100.00 | 0.00 |
| **NL** | 4869 | 84.86 | 15.14 | **SA** | 256 | 89.84 | 10.16 | | | | |

The table summarizes the dataset size and the proportion of Class 0 (*MS worsening not confirmed*) and Class 1 (*MS worsening confirmed*) for each participating country. Countries are sorted by dataset size in descending order. The data highlight substantial heterogeneity across countries, both in the number of available samples and in class balance. While Class 0 generally dominates, several countries exhibit severe class imbalance or complete absence of one class, underscoring the challenges of federated learning across non-identically distributed clinical datasets.

of one class. Although Class 0 predominated overall, both the magnitude of label imbalance and the variation in dataset sizes differed markedly across clients. This compounded heterogeneity in outcome distributions and sample availability reflects a fundamental deviation from the classical assumption of identically and independently distributed (IID) data, presenting additional challenges for federated model development in real-world clinical settings.

## Predicting disability progression

This analysis sets the stage for addressing a key clinical question in MS research: *the progression of disability*. This dimension of MS research is critical, as underscored in the literature[9], due to its substantial impact on PwMS. The precise prediction and thorough monitoring of disability progression are instrumental for clinicians in formulating effective treatment strategies, personalizing patient care, and ultimately, enhancing patient outcomes[64,65]. Our study contributes to this by investigating methodologies that not only aim to augment patient care but also seek to expand the medical community's comprehension of MS. This is achieved by harnessing insights from RWD, stepping towards the conversion of these insights into tangible real world evidence[66].

Building on the foundational work of De Brouwer et al.[13], this study adopted the FL approach for predicting confirmed disability progression over a two-year period with a 6-month confirmation window, utilizing RWD. This research leveraged the decentralized and privacy-preserving attributes of FL, marking a significant shift from conventional centralized data analyses. The investigation stood at the convergence of clinical need and technological innovation, with the potential to optimize the utilization of RWD in MS research. In the following section, we outline the experimental setup used in this study, which includes federated, adaptive, fine-tuned federated, local, and centralized models.

## Federated model

After partitioning the dataset by countries, we trained the FL models using each country's dataset. The experiment simulated a server-client architecture, with the server coordinating the learning process and the clients participating in distributed training. The server initiated the training process by setting up the model and distributing this initial model parameters to all available clients. Following this, each client starts local training on their respective dataset.

During each training cycle, or federation round, clients train their local models, these being the global model received from the server, for $E$ epochs. After training these $E$ epochs, the clients send their updated models back to the server, along with relevant metrics and the sizes of their test set. The server then executes the *federated strategy* to update the global model.

This process is iterative, with the server distributing the updated global model to the clients in each subsequent federation round. The cycle continues until a predetermined number of federation rounds $F$ were completed.

In our experiments, we selected a Multi-Layer Perceptron (MLP) with 42 input features as the baseline model to facilitate a comprehensive analysis. While De Brouwer et al.[13] explored various architectural frameworks in a centralized setting, we chose the MLP for its reliable performance and lack of significant differences compared to other models in our analyses.

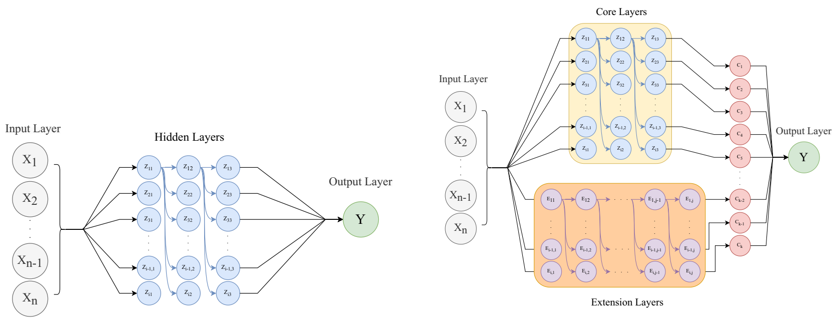
The training parameters were set with a batch size of $B = 512$, a local client learning rate $\eta_k$ = 1e-4, and a maximum number of epochs $E$ set to either 10 or 20, depending on the specific experiment. Weight decay was applied with $\lambda$ = 5e-5, and early stopping was employed with a patience parameter $P = 5$, indicating that training would halt if validation loss did not improve after five epochs. Regarding the model parameters, both the baseline and AdaptiveDualBranchNet (Core layers) models had $h = 512$ hidden units, a dropout rate of $\delta = 0.1$, and $l = 5$ layers. The AdaptiveDualBranchNet model was further enhanced with the ability to dynamically add up to $l_k^{\text{ext}} = 5$ extra layers, each comprising $h_{\text{ext}} = 64$ hidden units. For the FL setup, we conducted $F = 350$ federation rounds across $K = 32$ clients, with all clients participating in both training and evaluation processes. The entire experimental process was repeated $R = 10$ times for robustness.

In terms of the specific federated optimization strategies, FedProx was configured with a proximal term of $\mu$ = 1e-3. For the FedYogi optimization, parameters included $\eta$ = 1e-2, $\eta_k$ = 9.5e-2, $\tau$ = 1e-8, $\beta_1 = 0.6$, and $\beta_2 = 0.999$. The FedAdam strategy shared the server learning rate of $\eta$ = 1e-2 and local learning rate $\eta_k$ = 9.5e-2, with a regularization value $\tau$ = 1e-8, and momentum parameters $\beta_1 = 0.6$ and $\beta_2 = 0.999$. Lastly, the FedAdagrad model used $\eta$ = 1e-2, $\eta_k$ = 1e-2, and $\tau$ = 1e-8.

For training, we used Python 3.9.19 with PyTorch 1.13.1, Flower 1.8.0, Scikit-learn 1.5.0, and Pandas 2.2.2. The complete source code of this study is openly accessible at https://github.com/ashkan-pirmani/FL-MS-RWD. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation, Flanders (FWO) and the Flemish Government utilizing Intel Xeon Platinum 8468 CPUs (Sapphire Rapids) cluster.

## Personalized federated learning

In FL, applying a uniform model architecture across heterogeneous clients poses significant challenges[33]. In our setting, clients varied substantially in dataset sizes and class distributions, deviating from the classical assumption of IID data. Although it is common practice to deploy a fixed architecture, for example an MLP with static depth and width, we observed that such one-size-fits-all designs introduce important inefficiencies: larger models often overfit on



(a) Baseline

(b) AdaptiveDualBranchNet

**Fig. 5 | The diagram depicts the structure of the Baseline and AdaptiveDualBranchNet models. a** The Baseline network features a standard feedforward architecture. It begins with an Input Layer, which feeds into a series of Hidden Layers. Each hidden layer comprises neurons arranged in a fully connected structure, with arrows indicating the flow of information from one layer to the next. The connections show that each neuron in one layer is connected to every neuron in the subsequent layer, enabling the network to capture complex relationships between inputs. The network's final layer is the Output Layer, which aggregates the learned features from the hidden layers to produce the output *Y*. The straightforward structure of this network is designed for general-purpose learning tasks without additional branching or specialized layers. **b** The AdaptiveDualBranchNet architecture extends the Baseline by introducing a dual-branch structure comprising Core Layers and Extension Layers. The Core Layers, highlighted in yellow, retain the fully connected structure of the Baseline's Hidden Layers and are shared across all clients, being trained in a FL setup to capture fundamental and generalizable features from the data. In contrast, the Extension Layers, shown in orange, are client-specific and designed to learn personalized representations. These layers receive input from the same Input Layer as the Core Layers but follow a distinct structural design tailored to capture additional, domain- or client-specific variations in the data. Unlike the Core Layers, which are updated through FL aggregation, the Extension Layers remain locally trained, enabling each client to adapt the model to its unique distribution while benefiting from the shared knowledge encoded in the Core Layers. At the final stage, both branches feed into a set of processing nodes (depicted as *c* -units in red), which consolidate the learned representations before reaching the Output Layer *Y*. This separation between federated (global) and local (personalized) training allows the AdaptiveDualBranchNet to balance generalization and personalization, making it particularly effective in heterogeneous data environments where both shared knowledge and client-specific adaptations are necessary.

data-sparse partitions, whereas smaller models fail to fully leverage data-rich clients. This discrepancy underscores the need for adaptive modeling strategies that can adjust dynamically to local data characteristics.

Allowing each client to have a distinct architecture would conceptually address this issue but would render aggregation across clients infeasible due to mismatched model structures. To overcome this, we propose *AdaptiveDualBranchNet*, a model that dynamically modulates its complexity while maintaining architectural compatibility for aggregation. The network features a dual-branch design: a Core branch, comprising five fixed hidden layers with 512 neurons each, and a flexible Extension branch whose depth is determined by the local data volume. The Extension branch can add up to five additional hidden layers, each with 64 neurons, following a logarithmic scaling heuristic. Clients with larger datasets (e.g., more than 25000 samples) utilize all extension layers, while clients with smaller datasets (e.g., fewer than 2000 samples) omit the Extension branch entirely to reduce overfitting risk. Intermediate clients proportionally incorporate one to four extension layers based on their dataset size. Outputs from the Core and Extension branches are merged before the final prediction layer, ensuring a shared representational space across all clients. During federated training, only non-extension layers parameters are communicated and aggregated globally, preserving consistency while enabling localized adaptation. The design parameters and scaling thresholds were empirically optimized using development data to balance predictive accuracy, computational efficiency, and generalization across heterogeneous client populations. Pseudocode for the AdaptiveDualBranchNet algorithm is provided in Algorithm 1, and a schematic comparison with a standard MLP is shown in Fig. 5.

**Algorithm 1**. AdaptiveDualBranchNet: A Step-by-step Pseudocode Representation

```
Initialization:
for each client $k$ in $K$ do:
    Obtain the dataset size $n_k$
    Compute the number of extension layers $l_k^{\text{ext}}$ using the function $l_k^{\text{ext}} = \text{calculate\_extension}\,(\mathrm{n_k})$
    Initialize the local model parameters $\Theta_{k,f} = \{\Theta_{k,f}^{\text{input}}, \Theta_{k,f}^{\text{core}}, \Theta_{k,f}^{\text{ext}}, \Theta_{k,f}^{\text{combine}}, \Theta_{k,f}^{\text{output}}\}$
    where:
        $\Theta_{k,f}^{\text{input}}$ are the parameters of the input layer,
        $\Theta_{k,f}^{\text{core}}$ are the parameters of core layers,
        $\Theta_{k,f}^{\text{ext}}$ are the parameters of the $l_k^{\text{ext}}$ extension layers,
        $\Theta_{k,f}^{\text{combine}}$ are the parameters of the combining layer, and
        $\Theta_{k,f}^{\text{output}}$ are the parameters of the output layer.
end for
for federation round $f = 1$ to $F$do:
    Local Training (Round$f$):
    for each client $k$ in $K$do:
        Train the local model $\Theta_{k,f}$on local data for $E$ epochs or until an early stopping criterion is met
        Update local model parameters to $\Theta'_{k,f} = \{\Theta'^{\text{input}}_{k,f}, \Theta'^{\text{core}}_{k,f}, \Theta'^{\text{ext}}_{k,f}, \Theta'^{\text{combine}}_{k,f}, \Theta'^{\text{output}}_{k,f}\}$
    end for
    Local Model Upload (Round$f$):
    for each client $k$do:
        Send parameters of all non-extension layers to the central server: $\{\Theta'^{\text{input}}_{k,f}, \Theta'^{\text{core}}_{k,f}, \Theta'^{\text{combine}}_{k,f}, \Theta'^{\text{output}}_{k,f}\}$
    end for
    Aggregation (Round$f$):
    Aggregate uploaded non-extension layers' parameters using the chosen Federated Strategy:
    $\Phi_f = \text{FedStrategy}\,(\{\Theta'^{\text{input}}_{1,\text{f}}, \Theta'^{\text{core}}_{1,\text{f}}, \Theta'^{\text{combine}}_{1,\text{f}}, \Theta'^{\text{output}}_{1,\text{f}}, \ldots, \Theta'^{\text{input}}_{K,f}, \Theta'^{\text{core}}_{K,f}, \Theta'^{\text{combine}}_{K,f}, \Theta'^{\text{output}}_{K,f}\})$
        Example of FedAVG Strategy:
    $\Phi_f = \sum_{k=1}^{K} \frac{n_k}{n} \left(\Theta'^{\text{input}}_{k,f} + \Theta'^{\text{core}}_{k,f} + \Theta'^{\text{combine}}_{k,f} + \Theta'^{\text{output}}_{k,f}\right)$
    Global Model Distribution (Round$f$):
    Send updated global non-extension model $\Phi_f$ to clients
    for each client $k$ do:
        Update local model to $\Theta_{k,f+1} = \{\Phi_f + \Theta'^{\text{ext}}_{k,f}\}$
    end for
end for
Repeat:
Repeat steps 3–9 for $F$ federation rounds
```

Our experimental design aimed to go beyond only training FL models. To achieve this, we introduced an additional fine-tuning phase. This phase was crucial for evaluating the impact on model performance, as it aimed to further optimize the models following the initial FL process. To make the process clear, initially, a global model was trained using a federated approach. Upon completion of this global training, the model was disseminated back to each client for further refinement. Subsequently, each model locally underwent retraining exclusively with data from its corresponding client. To be more specific, each client's model is individually optimized using client-specific data. The strategic importance of fine-tuning lies in its ability to enhance the models' sensitivity to the unique attributes of their respective clients. This process synergizes the extensive, general learning acquired during the global federated training with the detailed, localized understanding extracted from each client's data. The objective was to strike a balance between the global model's generalization capabilities and the local datasets' specificity.

For fine-tuning, the local client learning rate was reduced to $\eta_k = 1\text{e-}4$ and dynamically adjusted using a scheduler with a patience threshold of five epochs. To allow the model to better explore the data, the batch size was also reduced to $B = 128$ across all clients, with training extended up to $E = 50$ epochs. This setup enabled the model to process the data more thoroughly during the fine-tuning phase.

### Local model

In the local model setup, each client independently trained a model using only their own partition, without any data federation or pooling. This approach, lacking centralized coordination or parameter sharing, served as a baseline for comparing the efficacy of FL methods from another viewpoint.

### Centralized model

In this setting, the global dataset $D$, was employed to train a centralized model. This served as another benchmark, where all data were aggregated and utilized in a conventional, non-federated manner for model training and evaluation.

### Evaluation method

In centralized learning, performance is typically gauged using a unified global test set. However, FL introduces the flexibility of performing evaluations either on the server-side or directly on the client-side[67]. Server-side evaluation necessitates the existence of one global test set located on the server. This approach encounters significant obstacles due to the distributed nature of sensitive RWD across multiple stakeholders and the rigorous demands of data privacy and regulatory standards. These challenges severely limit the feasibility of consolidating a singular global test set on the server-side.

Another challenge with server-side evaluation is ensuring the representativeness of the test set. Particularly with heterogeneous, non-IID settings, there is an increased risk of not accurately capturing the full diversity of the distributed datasets. Such biases could inadvertently skew the analysis, leading to findings that are less reliable or generalizable, thus compromising the study's validity.

Considering the need to reflect a real-world scenario, our study chose a federated (client-side) evaluation approach. To guarantee a fair and representative assessment and to avoid reliance on a potentially biased global test set, we implemented a consistent test set across all experiments (including centralized, federated and fine-tuned). This dataset, selected based on data partitioned by each country. The local model was tested on the unseen test sets of each country, and the performance metrics from these tests were aggregated using a weighted average based on the test set size of each country.

To explain the evaluation process on the client side clear, let $K$ be the total number of clients (in our experiment $K = 32$). The size of the dataset for the $i$-th client is $n_i$, where $i$ ranges from 1 to $K$. The evaluation metric achieved by the $i$-th client is represented as $M_i$. The total size of all clients' datasets combined is $N$, calculated as $N = \sum_{i=1}^{K} n_i$. The overall evaluation metric $E$ for the global FL model is given by the formula: $E = \frac{1}{N}\sum_{i=1}^{K}(n_i E_i)$, which reflects the model's performance across all clients. This method gives weight to the individual characteristics of each client, thus offering a detailed understanding of the model's performance in different environments. However, this can cause issues, especially if some clients have much larger datasets than others, leading to a biased evaluation metric. This potential bias makes it difficult to directly compare the performance of the FL model with that of a centralized model gauged on independent test set.

### Metrics

During all experimental setups, the evaluation metrics used included the ROC–AUC, AUC–PR, and the total experiment time, measured from the beginning of training to the end of the last federation round. These metrics served as robust indicators for assessing both the performance and computational efficiency of the models under investigation.

### Ethics declarations

This study was submitted to KU Leuven's Privacy and Ethics platform (PRET). The project application was scrutinized in view of principles and obligations laid down in the General Data Protection Regulation of the European Parliament and of the Council of 27 April 2016. Based on the information presented and given the researcher's explicit declaration that the project carried out accordingly, KU Leuven issued a favorable advice and confirmed that the project may be implemented as such. All relevant information concerning the processing of personal data in the framework of this project has been registered in KU Leuven's records of processing activities. The project application was also reviewed by the Social and Societal Ethics Committee (SMEC) of KU Leuven. The Committee confirmed that the project application meets the expected ethical standards regarding the voluntary involvement of human participants in scientific research. SMEC's decision regarding this protocol is favorable (PRET approval number: G-2023-6771).

## Data availability

The dataset used in this study can be accessed by requesting permission from the MSBase principal investigators involved. MSBase acts as the central contact point, facilitating data-sharing agreements with individual data custodians to ensure compliance with ownership requirements. Requests should be directed to info@msbase.org. Access is controlled to protect patient data and adhere to data ownership policies.

## Code availability

In accordance with the FAIR principles in scientific research, all code utilized in this study has been made publicly available. The preprocessing scripts can be accessed at https://gitlab.com/edebrouwer/ms_benchmark, while the training pipeline is available at https://github.com/ashkan-pirmani/FL-MS-RWD. The ML pipeline was developed using PyTorch[68], and metrics were computed using Scikit-learn. Figures were generated with Matplotlib, and the FL process was facilitated using Flower[69]. A complete list of dependencies is provided in the environment file located within the training repository.

Received: 13 February 2025; Accepted: 10 June 2025;
Published online: 24 July 2025

## References

1. Walton, C. et al. Rising prevalence of multiple sclerosis worldwide: insights from the atlas of MS, third edition. *Mult. Scler. J.* **26**, 1816–1821 (2020).
2. McGinley, M. P., Goldschmidt, C. H. & Rae-Grant, A. D. Diagnosis and treatment of multiple sclerosis: a review. *JAMA* **325**, 765–779 (2021).

3. Reich, D. S., Lucchinetti, C. F. & Calabresi, P. A. Multiple sclerosis. *N. Engl. J. Med.* **378**, 169–180 (2018).
4. Degenhardt, A., Ramagopalan, S. V., Scalfari, A. & Ebers, G. C. Clinical prognostic factors in multiple sclerosis: a natural history review. *Nat. Rev. Neurol.* **5**, 672–682 (2009).
5. Pellegrini, F. et al. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. *Mult. Scler. J.* **26**, 1828–1836 (2020).
6. Seker, B. I. O. et al. Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. *Cochrane Database Syst. Rev.* **2020**, CD013606 (2020).
7. Sherman, R. E. et al. Real-world evidence—what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
8. Brown, F. S. et al. Systematic review of prediction models in relapsing remitting multiple sclerosis. *PLoS ONE* **15**, 1–13 (2020).
9. Havas, J. et al. Predictive medicine in multiple sclerosis: a systematic review. *Mult. Scler. Relat. Disord.* **40**, 101928 (2020).
10. Seccia, R. et al. Machine learning use for prognostic purposes in multiple sclerosis. *Life* **11**, 122 (2021).
11. Hartmann, M., Fenton, N. & Dobson, R. Current review and next steps for artificial intelligence in multiple sclerosis risk research. *Comput. Biol. Med.* **132**, 104337 (2021).
12. Brouwer, E. D. et al. Longitudinal modeling of MS patient trajectories improves predictions of disability progression. *Comput. Methods. Prog. Biomed.* **208**, 106180 (2020).
13. De Brouwer, E. et al. Machine-learning-based prediction of disability progression in multiple sclerosis: an observational, international, multi-center study. *PLOS Digit. Health* **3**, 1–25 (2024).
14. Pirmani, A. et al. The journey of data within a global data sharing initiative: a federated 3-layer data analysis pipeline to scale up multiple sclerosis research. *JMIR Med. Inform.* **11**, e48030 (2023).
15. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
16. Wu, J., Roy, J. & Stewart, W. F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med. Care* **48**, S106–S113 (2010).
17. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* **20**, 144–151 (2013).
18. Wilkinson, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
19. Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A. & Eskofier, B. Federated learning for healthcare: systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol.* **13**, 1–23 (2022).
20. Xu, J. et al. Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **5**, 1–19 (2021).
21. Li, S. et al. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. *J. Am. Med. Inform. Assoc.* **30**, 2041–2049 (2023).
22. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
23. Brisimi, T. S. et al. Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* **112**, 59–67 (2018).
24. Yin, X., Zhu, Y. & Hu, J. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions. *ACM Comput. Surv.* **54**, 1–36 (2021).
25. Wang, W. et al. A privacy preserving framework for federated learning in smart healthcare systems. *Inf. Process. Manag.* **60**, 103167 (2023).
26. Truex, S. et al. A hybrid approach to privacy-preserving federated learning. *Inform. Spektrum* **42**, 356–357 (2019).
27. Donkada, S. et al. Uncovering promises and challenges of federated learning to detect cardiovascular diseases: a scoping literature review (2023).
28. Yi, L. et al. Su-net: an efficient encoder-decoder model of federated learning for brain tumor segmentation. In *Artificial Neural Networks and Machine Learning - ICANN 2020: 29th International Conference on Artificial Neural Networks,* Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I, 761–773. https://doi.org/10.1007/978-3-030-61609-0_60 (Springer-Verlag, Berlin, Heidelberg, 2020).
29. Oldenhof, M. et al. Industry-scale orchestrated federated learning for drug discovery. *Proc. AAAI Conf. Artif. Intell.* **37**, 15576–15584 (2024).
30. Liu, D. et al. Multiple sclerosis lesion segmentation: revisiting weighting mechanisms for federated learning. *Front. Neurosci.* **17**, 1167612 (2023).
31. Denissen, S. et al. Towards multimodal machine learning prediction of individual cognitive evolution in multiple sclerosis. *J. Pers. Med.* **11**, 1349 (2021).
32. Denissen, S. et al. Transfer learning on structural brain age models to decode cognition in MS: a federated learning approach. *medRxiv.* https://www.medrxiv.org/content/early/2023/04/26/2023.04.22.23288741 (2023).
33. Tan, A. Z., Yu, H., Cui, L. & Yang, Q. Towards personalized federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 9587–9603 (2023).
34. Butzkueven, H. et al. Msbase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis. *Mult. Scler. J.* **12**, 769–774 (2006).
35. Kalincik, T. & Butzkueven, H. The MSBase registry: informing clinical practice. *Mult. Scler. J.* **25**, 1828–1834 (2019).
36. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data https://arxiv.org/abs/1602.05629 (2023).
37. Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A. & Smith, V. On the convergence of federated optimization in heterogeneous networks (2018).
38. Reddi, S. J. et al. Adaptive federated optimization. In *Proc. International Conference on Learning Representations* https://openreview.net/forum?id=LkFG3lB13U5 (2021).
39. Biewald, L. Experiment tracking with weights and biases https://www.wandb.com/. Software available from wandb.com (2020).
40. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In: *Proc. 34th International Conference on Machine Learning - Volume 70*, ICML'17, 1321–1330 (JMLR, 2017).
41. Davis, S. E., Greevy, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **112**, 103611 (2020).
42. Arivazhagan, M. G., Aggarwal, V., Singh, A. K. & Choudhary, S. Federated learning with personalization layers https://arxiv.org/abs/1912.00818 (2019).
43. PyTorch. ReduceLROnPlateau—PyTorch 2.6 documentation. https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html. Accessed 28 Mar 2025 (2019).
44. Khodak, M. et al. Federated hyperparameter tuning: challenges, baselines, and connections to weight-sharing https://arxiv.org/abs/2106.04502 (2021).
45. Zawad, S. & Yan, F. Hyperparameter tuning for federated learning—systems and practices. In Nguyen, L. M., Hoang, T. N. & Chen, P.-Y. (eds.) *Federated Learning*, 219–235. https://www.sciencedirect.com/science/article/pii/B9780443190377000211 (Academic Press, 2024).
46. Zhang, H. et al. Federated learning hyperparameter tuning from a system perspective. *IEEE Internet Things J.* **10**, 14102–14113 (2023).
47. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In: *Proc. 28th International Conference on Neural Information Processing Systems—Volume 2*, NIPS'14, 3320–3328 (MIT Press, 2014).
48. Houlsby, N. et al. Parameter-efficient transfer learning for NLP. *ArXiv*abs/1902.00751. https://doi.org/10.48550/arXiv.1902.00751 (2019).

49. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. In *Fine-tuning*, 328–339 (2018).
50. Kairouz, P. et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14**, 1–210 (2021).
51. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**, 50–60 (2020).
52. Pirmani, A., Moreau, Y. & Peeters, L. M. Unlocking the power of real-world data: a framework for sustainable healthcare. *Stud. Health Technol. Inform.* **316**, 1582–1583 (2024).
53. EU. European Health Data Space regulation (EHDS). https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en. Accessed: 26 Mar 2025.
54. Auffray, C. et al. Making sense of big data in health research: Towards an EU action plan. *Genome Med.* **8**, 71 (2016).
55. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC Med.* **13**, 1 (2015).
56. Peeters, L. M. et al. Covid-19 in people with multiple sclerosis: a global data sharing initiative. *Mult. Scler. J.* **26**, 1157–1162 (2020).
57. Heyndrickx, W. et al. Melloddy: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *J. Chem. Inf. Model.* **64**, 2331–2344 (2024).
58. Andorra, M. et al. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. *J. Neurol.* **271**, 1133–1149 (2024).
59. Dwork, C. Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V. & Wegener, I. (eds.) *Automata, Languages and Programming*, 1–12 (Springer Berlin Heidelberg, 2006).
60. Goldreich, O. Secure multi-party computation. *Manuscript. Preliminary version* **78** (1998).
61. López-Blanco, R., Alonso, R. S., González-Arrieta, A., Chamoso, P. & Prieto, J. Federated learning of explainable artificial intelligence (FED-XAI): a review. In Ossowski, S. et al. (eds.) *Distributed Computing and Artificial Intelligence, 20th International Conference*, 318–326 (Springer Nature, 2023).
62. Choi, A. et al. A novel deep learning algorithm for real-time prediction of clinical deterioration in the emergency department for a multimodal clinical decision support system. *Sci. Rep.* **14**, 30116 (2024).
63. Kalincik, T. et al. Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. *Brain* **140**, 2426–2443 (2017).
64. Leray, E. et al. Evidence for a two-stage disability progression in multiple sclerosis. *Brain* **133**, 1900–1913 (2010).
65. Andersson, P. B., Waubant, E., Gee, L. & Goodkin, D. E. Multiple sclerosis that is progressive from the time of onset: clinical characteristics and progression of disability. *Arch. Neurol.* **56**, 1138–1142 (1999).
66. Liu, M., Qi, Y., Wang, W. & Sun, X. Toward a better understanding about real-world evidence. *Eur. J. Hosp. Pharm.* **29**, 8–11 (2022).
67. Pirmani, A., Oldenhof, M., Peeters, L. M., De Brouwer, E. & Moreau, Y. Accessible ecosystem for clinical research (federated learning for everyone): development and usability study. *JMIR Form. Res.* **8**, e55496 (2024).
68. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. In: *Proc. 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2019).
69. Beutel, D. J. et al. Flower: a friendly federated learning research framework https://doi.org/10.48550/arXiv.2007.14390 (2020).

## Acknowledgments

A.Pirmani., Y.M. and M.O. are funded by (1) VLAIO PM: Augmenting Therapeutic Effectiveness through Novel Analytics (HBC.2019.2528); Research Council KU Leuven: Symbiosis 4 (C14/22/125), Symbiosis3 (C14/18/092; CELSA—Active Learning (CELSA/21/019). Y.M., M.O., and A.Pirmani. are affiliated to Leuven.AI and received funding from the Flemish Government (AI Research Program, FWO SBO (S003422N), and ELIXIR Belgium (I002819N)). E.D.B. and A.Passiemers. was funded by a FWO-SB grant. R.G. received honoraria as consultant on scientific advisory boards or as speaker from Biogen, Celgene-BMS, Janssen, Merck, Novartis, Roche, Sanofi-Genzyme, Sandoz, Teva and Viatris. B.W. received consultancy and advisory board fees from Roche, Sanofi-Genzyme, Biogen, Merck-Serono, Bayer-Schering, Novartis and Allergan; received congress support from Biogen, Merck-Serono, Teva and Roche. She has also received research support from Novartis, Biogen, Roche, FWO (Research Foundation Flanders) and Fonds D.V. (Ligue Nationale Belge de la Sclerose en Plaques, Fondation Roi Baudouin). D.H .was supported by the Charles University: Cooperatio Program in Neuroscience, by the project National Institute for Neurological Research (Program EXCELES, ID Project No. LX22NPO5107) —Funded by the European Union—Next Generation EU, and by General University Hospital in Prague project MH CZ-DRO-VFN64165.She also 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. E.K.H. received travel and/or consultancy compensation from Sanofi-Genzyme, Roche, Teva, Merck, Novartis, Celgene, Biogen. F.P. received personal compensation for serving on advisory board by Almirall, Alexion, Biogen, Bristol, Janssen, Merck, Novartis and Roche. He further received research grant by Alexion, Almirall, Biogen, Bristol, Merck, Novartis and Roche and by FISM, Reload Association (Onlus), Italian Health Minister, and University of Catania.AL received speakers fees and travel grants from Novartis, Biogen, T'évalua, Sanofi. V.T. served on scientific advisory boards for Roche, Janssen, Sanofi-Genzyme, Novartis and Merck, received conference fee and travel support from Novartis, Biogen, Sanofi-Genzyme, Teva, Abbvie and Merck and received educational event support from Novartis. E.C. received honoraria/research support from Biogen, Merck Serono, Novartis, Roche, and Teva; has been member of advisory boards for Actelion, Biogen, Celgene, Merck Serono, Novars, and Sanofi Genzyme; received honoraria/research support from Biogen, Merck Serono, Novartiss, Roche, and Teva; has been member of advisory boards for Actelion, Biogen, Celgene, Merck Serono, Novartis, and Sanofi Genzyme; and has been supported by the Czech Ministry of Education—project Cooperatio LF1, research area Neuroscience, and the project National Institute for Neurological Research (Program EXCELES, ID project No LX22NPO5107) —funded by the European Union-Next Generation EU. I.R. received speaking/consulting fees and/or travel funding from Almirall, Biogen, Bristol Myers Squibb, Janssen, Merck, Novartis, Roche, Sanofi-Genzyme and Teva. C.B. received speaking honoraria from Biogen, Novartis, Sanofi, Merck, Roche, Almirall and Teva. R.A. has received speaker honoraria and consultant fees from Biogen Idec, Novartis, Merck, Janssen, Bristol-Meyers, Bayer, Sanofi Genzyme, Roche and Teva. K.B. has no disclosures to declare. C.S. received personal compensation for consulting, serving on a scientific advisory board, speaking or other activities from Alexion, Amgen/Horizon, Biogen, Bristol Myers Squibb, Janssen/Johnson\&Johnson, Merck Serono, Novartis, Roche and Sanofi/Genzyme, and her institutions received research grants from Novartis, Roche and Sanofi/GenzymeOG has received consultation and speaker fees, travel grants and research support from: Biogen, Sanofi Genzyme, Merck, Novartis, Roche, Alexion, Viatris, Janssen, Bristol Myers Squibb, Almirall, Lundbeck. A.S. has served in advisory boards for Novartis, EMD Serono, Roche, Biogen idec, Sanofi Genzyme, Pendopharm and has received grant support from Genzyme and Roche, has received research grants for his institution from Biogen idec, Sanofi Genzyme, EMD Serono. J.K. has no disclosures. J.L.S.M. served on scientific advisory boards or as a consultant for MS International Federation and World Health Organisation, Therapeutic Goods Administration, BMS, Roche, Janssen, Genzyme, Novartis, Merck and Biogen, received conference travel support and/or speaker honoraria from WebMD Global, Merck, Sandoz, Novartis, Biogen, Roche, Eisai, Genzyme, Teva and BioCSL and received research or educational event support from Biogen,

Novartis, Genzyme, Roche, Celgene and Merck. D.S. received speaking honoraria from Biogen, Novartis, Merck and Teva. T.C. served on scientific advisory boards, received conference travel support and/or speaker honoraria from Roche, Novartis, Merck and Biogen. I.R. is supported by MS Australia and the Trish MS Research Foundation. BVM received conference travel support from Biogen, Novartis, Bayer-Schering, Merck and Teva; has participated in clinical trials by Sanofi Aventis, Roche and Novartis. R.A. received honoraria as a speaker and for serving on scientific advisory boards from Bayer, Biogen, GSK, Merck, Novartis, Roche and Sanofi-Genzyme.VVPreceived honoraria as consultant on scientific advisory boards by Biogen, Bayer-Schering, Merck, Teva and Sanofi-Aventis; has received research grants by Biogen, Bayer-Schering, Merck, Teva and Novartis. N.J. received speaker honoraria and/or education support from Biogen, Teva, Novartis, Genzyme-Sanofi, Roche, Merck and Alexion; has been a member of advisory boards for Merck and Biogen. D.M. received honoraria and consulting fees from Bayer Schering, Novartis, Merck, Biogen and Genzyme. B.W.G. received travel grants from Novartis, Bayer-Schering, Merck and Teva; has participated in clinical trials by Sanofi Aventis, Roche and Novartis. G.L. travel compensation from Novartis, Biogen, Roche and Merck. Her institution receives the honoraria for talks and advisory board commitment as well as research grants from Biogen, Merck, Roche and Novartis. R.G. has received research grant and / or advisory board honoraria from Biogen, Hikma, Merck, Roche and Sanofi. A.A. accepted travel compensation from Novartis, Biogen, Genzyme, Teva, and speaking honoraria from Biogen, Novartis, Genzyme and Teva. A.A.A. has received a MENACTRIMS clinical fellowship grant (2020). A.V.D.W. received honoraria or research funding from Biogen, Genzyme, Novartis, Teva Neurosciences, and ATARA Pharmaceuticals.Helmut Butzkueven served on scientific advisory boards for Merck, Genzyme, Almirall, and Biogen; received honoraria and travel grants from Sanofi Aventis, Novartis, Biogen, Merck, Genzyme and Teva. B.T. received compensation for serving on IDMC for Biogen TAH has received travel grants from Merck Healthcare KGaA (Darmstadt, Germany), Biogen, Sanofi, Bristol Meyer Squibb, Almirall, Roche and Eisai. His institution has received research grants and consultancy fees from Roche, Biogen, Sanofi, Merck Healthcare KGaA (Darmstadt, Germany), Bristol Meyer Squibb, Janssen, Almirall, Novartis Pharma, Alexion, Neuraxpharm and Eisai. C.R. received speaker fees, research support, travel support, and/or served on advisory boards by Swiss MS Society, Swiss National Research Foundation (320030_189140/1), University of Basel, Progressive MS Alliance, Alnylam, Bayer, Biogen, Bristol Myers Squibb, Celgene, Immunic, Merck, Neurogenesis, Novartis, Octave Bioscience, Quanterix, Roche, Sanofi, Stata DX. O.G. accepted travel compensation from Novartis, Merck and Biogen, speaking honoraria from Biogen, Novartis, Sanofi, Merck, Almirall, Bayer and Teva and has participated in clinical trials by Biogen, Merck and Roche. D.D. received honoraria as a consultant on scientific advisory boards by Bayer-Schering, Novartis and Sanofi-Aventis and compensation for travel from Novartis, Biogen, Sanofi-Aventis, Teva and Merck. A.G.K. received research funding, speaker honoraria and compensation for travel from and served as a consultant on advisory board for Bayer-Schering, Teva, Biogen, Merck, Genzyme and Novartis. M.F.P. has no relevant disclosure. M.S. received speaker honoraria/conference travel support from Biogen, Merck, Novartis, Roche, Sanofi-Aventis and Teva. N.S. has received honoraria as a speaker from Biogen, Merck, Novartis and Sanofi-Genzyme and for serving on scientific advisory boards from Novartis. B.S. Richard Macdonell or his institution have received remuneration for his speaking engagements, advisory board memberships, research and travel from Biogen, Merck, Genzyme, Bayer, Roche, Teva, Novartis, CSL, BMS, MedDay and NHMRC. M.C. NAJ is a PI on commercial MS studies sponsored by Novartis, Roche and Sanofi. He has received speaker's honoraria and consultancy fees from Merck. He has had conference travel and registration reimbursement; and consultancy fees from Novartis. PM received speaker honoraria for Advisory Board and travel grants from Alexion, Almirall, Bayer, Biogen, Bristol Myers Squibb, Merck, Novartis, Roche, Sanofi-Genzyme, and Teva. P.L. received conference travel support from Novartis, Teva, Biogen, Bayer and Merck and has participated in a clinical trials by Biogen, Novartis, Teva and Actelion. C.A.S. served as a consultant for Biogen, EMD Serono, Novartis, Genentech, Celgene/Bristol Meyers Squibb, Sanofi Genzyme, Bayer, Janssen, Labcorp, Horizon and SANA. Dr. Weinstock-Guttman also has received grant/research support from Novartis, Biogen, Horizon/Amgen. She serves in the editorial board for Children, CNS Drugs, MS International, Journal of Neurology Frontiers Epidemiology. S.H. has received research funding from the National Health and Medical Research Council (NHMRC, Australia), the Petre Foundation, the Brain Foundation, the Royal Australasian College of Physicians, and the University of Sydney. She is supported by an NHMRC Investigator Grant (GNT2008339). She serves as a consultant on the International Steering Committee for a clinical trial led by UCB (NCT05063162). She is on the advisory board for educational activities led by Limbic Neurology. She has been an invited speaker for educational/research sessions coordinated by Biogen, Alexion, Novartis, Excemed and Limbic Neurology. She is on the medical advisory board (non-remunerated positions) of The MOG Project and the Sumaira Foundation.

## Author contributions

A.Pirmani. conceived the study, performed the experiments, analyzed the data, and led the manuscript writing. E.D.B., L.M.P., and Y.M. coordinated and supervised the study design, contributed to result interpretation, and critically reviewed the manuscript. M.O., A.A., A.Passiemers., and A.F. participated in data analysis and manuscript revision. Authors from T.K. to T.C.T. (the MSBase Study Group) contributed data. All authors reviewed and approved the final manuscript for submission.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01788-8.

**Correspondence** and requests for materials should be addressed to Yves Moreau.

**Reprints and permissions information** is available at http://www.nature.com/reprints

[1]STADIUS, ESAT, KU Leuven, Leuven, Belgium. [2]Biomedical Research Institute, Hasselt University, Hasselt, Belgium. [3]Data Science Institute, Hasselt University, Hasselt, Belgium. [4]University Multiple Sclerosis Center, Hasselt University, Hasselt, Belgium. [5]Department of Neurology, Neuroimmunology Centre, Royal Melbourne Hospital, Melbourne, VIC, Australia. [6]Medical Point Hospital, Izmir University of Economics, Izmir, Turkey. [7]Department of Neurology, LR 18SP03, Clinical Investigation Centre Neurosciences and Mental Health, Razi University Hospital, Tunis, Tunisia. [8]Department of Neurology, Antwerp University Hospital, Edegem, Belgium. [9]University of Antwerp, Antwerp, Belgium. [10]Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University in Prague and General University Hospital, Prague, Czechia. [11]Department of Medical and Surgical Sciences and Advanced Technologies, GF Ingrassia, Catania, Italy. [12]CHUM and Universite de Montreal, Montreal, QC, Canada. [13]Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, Bologna, Italy. [14]Institute for Advanced Biomedical Technologies (ITAB), Dept Neurosciences, Imaging and Clinical Sciences, University G. d'Annunzio of Chieti-Pescara, Chieti, Italy. [15]CISSS Chaudière-Appalache, Levis, QC, Canada. [16]Neurology Unit, AST Macerata, Macerata, Italy. [17]Department of Neurology, Medical Faculty, Karadeniz Technical University, Trabzon, Turkey. [18]Division of Neurology, Department of Medicine, Amiri Hospital, Sharq, Kuwait. [19]Department NEUROFARBA, University of Florence, Florence, Italy. [20]IRCCS Fondazione Don Carlo Gnocchi, Florence, Italy. [21]Department of Neurosciences, Box Hill Hospital, Box Hill, VIC, Australia. [22]Hunter Medical Research Institute, University Newcastle, Newcastle, NSW, Australia. [23]Department of Neurology, Unidade Local de Saúde de São João, Porto, Portugal. [24]Department of Clinical Neurosciences and Mental Health, Faculty of Medicine of University of Porto, Porto, Portugal. [25]Neurology Unit, Galliera Hospital, Genova, Italy. [26]Academic MS Center Zuyd, Department of Neurology, Zuyderland Medical Center, Sittard-Geleen, The Netherlands. [27]Bakirkoy Education and Research Hospital for Psychiatric and Neurological Diseases, Istanbul, Turkey. [28]Department of Neurology, University Hospital and University of Basel, Basel, Switzerland. [29]Department of Neurology, Galdakao-Usansolo University Hospital, Osakidetza-Basque Health Service, Galdakao, Spain. [30]Azienda Ospedaliera di Rilievo Nazionale San Giuseppe Moscati Avellino, Avellino, Italy. [31]Faculty of Medicine, University of Debrecen, Debrecen, Hungary. [32]Noorderhart Hospitals, Rehabilitation & MS University MS Centre, Hasselt-Pelt, Belgium. [33]Nemocnice Jihlava, Jihlava, Czechia. [34]CSSS Saint-Jérôme, Saint-Jerome, QC, Canada. [35]Nehme and Therese Tohme Multiple Sclerosis Center, American University of Beirut Medical Center, Beirut, Lebanon. [36]Department of Neurology, Cliniques Universitaires Saint-Luc, Brussels, Belgium. [37]Department of Medicine, School of Clinical Sciences, Monash University, Clayton, VIC, Australia. [38]Centro Sclerosi Multipla, UOC Neurologia, Azienda Opsedaliera per l'Emergenza Cannizzaro, Catania, Italy. [39]Department of Neurology, Jacobs MS Center for Treatment and Research, New York, NY, USA. [40]Department of Neurology, Universitary Hospital Ghent, Ghent, Belgium. [41]Department of Neurology, Royal Brisbane Hospital, Brisbane, QLD, Australia. [42]Service of Neurology, Center of Neuroimmunology, Hospital Clinic de Barcelona, Barcelona, Spain. [43]Department of Neurology, School of Medicine and Koc University Research Center for Translational Medicine (KUTTAM), Koc University, Istanbul, Turkey. [44]College of Medicine & Health Sciences, Sultan Qaboos University, Al-Khodh, Oman. [45]Department of Neurology, Westmead Hospital, Sydney, NSW, Australia. [46]Department of Neurology, The Alfred Hospital, Melbourne, VIC, Australia. [47]Groene Hart Ziekenhuis, Gouda, The Netherlands. [48]Jahn Ferenc Teaching Hospital, Budapest, Hungary. [49]Royal Hobart Hospital, Hobart, TAS, Australia. [50]Neurology Department, King Fahad Specialist Hospital-Dammam, Dammam, Saudi Arabia. [51]Department of Neurology and Stroke, BAZ County Hospital, Miskolc, Hungary. [52]Department of Neurology, University of Szeged, Szeged, Hungary. [53]South Eastern HSC Trust, Belfast, UK. [54]AZ Alma Ziekenhuis, Damme, Belgium. [55]Perron Institute for Neurological and Translational Science, Sir Charles Gairdner Hospital, The University of Western Australia, Perth, WA, Australia. [56]Perron Institute, QEII Medical Centre, University of Western Australia, Nedlands, WA, Australia. [57]Perron Institute for Neurological and Translational Science, The University of Western Australia, Perth, WA, Australia. [58]Christchurch Hospital, Christchurch, New Zealand. [59]Neurology Unit, Hospital General Universitario de Alicante, Alicante, Spain. [60]University of Medicine and Pharmacy Victor Babes Timisoara, Timisoara, Romania. [61]St Vincents Hospital Fitzroy, Melbourne, VIC, Australia. [62]Bombay Hospital Institute of Medical Sciences, Mumbai, India. [63]Neurosciences Department, Mater Dei Hospital, Birkirkara, Malta. [64]Department of Neurology, Concord Repatriation General Hospital, Sydney, NSW, Australia. [65]Translational Neuroimmunology Group, Kids Neuroscience Centre and Brain and Mind Centre, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia. [66]Department of Clinical Neurosciences, Division of Neurology, Unit of Neuroimmunology, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland. [67]Clinical Neurosciences Department, 'Carol Davila' University of Medicine and Pharmacy, Bucharest, Romania. [68]Royal Victoria Hospital, Belfast, UK. [69]Hospital Universitario Donostia and IIS Biodonostia, San Sebastián, Spain. ✉e-mail: Yves.Moreau@esat.kuleuven.be