Supplementary material

1 Treatment effect modeling

We are interested in predicting the individual treatment effect (ITE), τ_i, defined according to the Neyman/Rubin Potential Outcome Framework:[10]

$$\tau_i := Y_i(1) - Y_i(0) \qquad (1)$$

where $Y_i(1)$ and $Y_i(0)$ represent the outcome of participant i when given treatment and control medications, respectively. The Fundamental Problem of Causal Inference[11] is that the ITE is unobservable because only one of the two outcomes is realized in any given patient, dictated by their treatment allocation. $Y_i(1)$ and $Y_i(0)$ are therefore termed potential outcomes or, alternatively, factual (observed) and counterfactual (not observed) outcomes.

Ground-truth can nevertheless be observed at the group level. The average treatment effect (ATE) is defined as the expected difference between both potential outcomes:

$$ATE := \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \qquad (2)$$

The ATE is still defined using causal quantities which are not observable, so additional assumptions are needed (a detailed discussion of the underlying assumptions is beyond the scope of this paper). In specific situations, such as randomized control trials where the outcome is independent of treatment allocation, in this case $T \in \{0,1\}$, the ATE can identified from the observed outcome Y:

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \qquad (3)$$

Broadly speaking, the ATE (sometimes formulated as a ratio instead of a difference) is the quantity estimated in clinical trials, but here we seek to estimate the ATE of a sub-group of patients conditioned on their baseline characteristics, a d-dimensional feature vector $x \in \mathcal{X} \subseteq R^d$. This conditional average treatment effect (CATE) is therefore defined as:

$$CATE := \mathbb{E}[Y(1) - Y(0) \mid X = x] \qquad (4)$$

and can similarly be rewritten in terms of the observed outcome Y in the context of assumptions that hold in randomized controlled trials (where $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$):

$$\tau(x) = \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0] = \mu_1(x) - \mu_0(x) \quad (5)$$

A CATE estimator $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ can be parametrized by a neural network trained on an observational dataset $\mathcal{D} = \{x_i, y_i, t_i\}_{i=1}^{n}$. In this paper we learn a multitask multilayer perceptron in which $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ share parameters in the earlier layers but have distinct parameters in the output heads. The best CATE estimator can be shown to also be the best estimator for the ITE in terms of mean squared error (see Equation 2 in Künzel et al.[5]).

2 Treatment effect correction

Given that our tuning metrics are rank based, they do not guarantee that the model will provide an accurate point estimate for the outcome of interest. We therefore correct for shifts between the ground-truth and predicted treatment effects by computing a shifting scalar, σ. This scalar represents the difference between the empirical mean estimate of the ATE from a sample n composed of n_1 treated and n_0 control participants, and the mean of predicted ITEs:

$$\sigma = \frac{1}{n_1}\sum_{i=1}^{n} y_i t_i - \frac{1}{n_0}\sum_{i=1}^{n} y_i(1 - t_i) - \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)] \quad (6)$$

This learned shifting scalar is then applied at inference time to produce the ITE estimate:

$$\hat{\tau}_i := \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i) + \sigma \quad (7)$$

3 Slope outcome

We assume that progression is slow over the course of the 1-2 year duration of a phase 2 or 3 clinical trial such that it can be modeled using linear regression:

$$f(t) = w^* t + w_0^* \quad (8)$$

where $f(t)$ is the predicted Expanded Disability Status Scale (EDSS) score of a patient at time t, and w^* and w_0^* are the weight and bias, respectively, that were learned through the method of least squares. Each patient i will have a separate slope of progression, w_i^*. This individual slope is used as the outcome of interest for training the model, such that

$$y_i = w_i^* \qquad (9)$$

Note that in order to match the definition of Confirmed Disability Progression (CDP) used in our dataset, where CDP is defined as an increase of 0.5 in EDSS if the baseline EDSS > 5.5 and 1.0 if the baseline EDSS ≤ 5.5, we scale all changes in the EDSS by a factor of 2 for values beyond 5.5. This is done before fitting the linear regression model.

4 Weighted average treatment difference curve

Following Zhao et al.,[22] we define a conditional expectation, *AD(c)*, which reflects the CATE of a sub-group of patients who are predicted to respond more than a certain response threshold, *c*, to the active medication:

$$AD(c) = \mathbb{E}[Y(1) - Y(0) \mid \hat{\tau}_i \geq c] \qquad (10)$$

where $\hat{\tau}_i$ is the predicted ITE from our neural network. The conditional expectation for $Y(1) - Y(0)$ here is estimated using the restricted mean survival time (RMST) for the time-to-CDP24, truncated at 2 years.[33] By defining the conditional expectation in terms of the RMST instead of the slope outcome used as the target for our neural network, the *AD(c)* better reflects how well our model can identify responders using a survival-based metric, which is ultimately what clinical trials will be interested in.

The *AD(c)* behaves as a population selector, whereby patients expected to respond with effect size greater than a desirable *c* can be enrolled in a clinical trial or recommended the medication in a clinical setting.

We derive a metric from the *AD(c)* to evaluate model performance in predicting treatment effect. If patients are ranked accurately according to their predicted responsiveness to the active medication, the resultant *AD(c)* curve should be increasing almost monotonically, with a large area under the curve, AD_{auc}. We compute the AD_{auc} using polygon approximation, with operating points every 10 percentiles from 0 until the 80th percentile. The units are therefore in years. We use Zhao et al.'s definition of the AD_{abc} to subtract the effect size of the entire population to facilitate comparison of different models:

$$AD_{abc} = AD_{auc} - AD(\hat{\tau}_0) \qquad (11)$$

where $\hat{\tau}_0$ represents the minimum predicted ITE. We further weigh the AD_{abc} by multiplying it to a measure of monotonicity to promote a monotonically increasing *AD(c)*. We chose to use the C-index between the AD_{abc} values and the percentile thresholds given that we used the C-index elsewhere in this work, but using other metrics such as Spearman's rank correlation coefficient could accomplish the same purpose.

5 Supplementary Tables and Figures

Supplementary Table 1 Main inclusion criteria for ORATORIO, OLYMPUS and ARPEGGIO

Inclusion criteria[a]	ORATORIO	OLYMPUS	ARPEGGIO
Age	18 - 55 years	18 - 65 years	22 - 55 years
Diagnosis	PPMS according to 2005 revised McDonald criteria	PPMS according to 2001 McDonald criteria	PPMS according to 2010 revised McDonald criteria
EDSS	3.0 - 6.5	2.0 - 6.5	3.0 - 6.5
FSS-Pyramidal	≥ 2	≥ 2	≥ 2
Disease duration[b]	< 15 years if EDSS > 5.0 or < 10 years if EDSS ≤ 5.0	≥ 1 year	
CSF	Elevated IgG index or at least one IgG OCB	Elevated IgG index or at least one IgG OCB	
MRI			Lesions consistent with PPMS at baseline
Disability progression			Documented worsening of clinical disability in the 2 years prior to screening

[a]Please refer to the original trial publications[6-8] for detailed inclusion and exclusion criteria.
[b]Disease duration is measured from the time of symptom onset.
FSS=Functional Systems Score.

Supplementary Table 2 MLP hyperparameter sets for random search.

Hyperparameter	Search Set
Learning rate	{0.01,0.001,0.0001}
L2 regularization coefficient	{0.01,0.001}
Max-norm constraint on the weights	{4, 3, *None*}
Hidden layer width	{8, 64, 264}
Hidden layer dropout probability	{0.2, 0.3, 0.4, 0.5}
Input layer dropout probability	{0.0, 0.1, 0.2}

Supplementary Table 3 Group statistics for predicted responders and non-responders to anti-CD20-Abs at the 75th percentile threshold

	Responders		Non-responders		P-value[a]	
	Single[b]	Crogging[c]	Single	Crogging	Single	Crogging
Trial contribution:						
OLYMPUS	22	86	98	331		
ORATORIO	59	195	157	507		
Demographics:						
Age (years)	43.26 (9.08)	42.20 (8.64)	45.90 (7.13)	45.31 (7.87)	**0.019**	**<0.001**
Sex (% male)	53.09%	58.72%	46.67%	46.78%	0.372	**0.001**
Height (cm)	170.95 (9.78)	171.72 (9.74)	170.95 (9.09)	170.08 (9.23)	0.996	**0.014**
Weight (kg)	69.69 (14.00)	75.00 (17.93)	77.81 (16.92)	74.70 (16.42)	**<0.001**	0.804
Disease duration (years)[d]	6.67 (4.46)	6.57 (3.79)	7.94 (5.96)	7.64 (5.46)	**0.044**	**<0.001**
Disability Scores:						
EDSS	4.91 (1.25)	5.05 (1.19)	4.64 (1.26)	4.60 (1.25)	0.094	**<0.001**
FSS-Bowel and Bladder	1.42 (0.89)	1.35 (0.87)	1.15 (0.91)	1.18 (0.90)	**0.021**	**0.004**
FSS-Brainstem	1.03 (1.02)	1.29 (0.98)	0.74 (0.84)	0.68 (0.82)	**0.025**	**<0.001**
FSS-Cerebellar	2.53 (0.76)	2.55 (0.82)	1.92 (1.07)	1.94 (1.01)	**<0.001**	**<0.001**
FSS-Cerebral	1.22 (0.83)	1.34 (0.84)	1.01 (0.87)	0.93 (0.86)	0.059	**<0.001**
FSS-Pyramidal	2.95 (0.62)	2.93 (0.66)	2.75 (0.78)	2.77 (0.71)	**0.018**	**<0.001**
FSS-Sensory	1.45 (1.01)	1.67 (1.05)	1.59 (1.03)	1.50 (1.04)	0.270	**0.021**
FSS-Visual	1.12 (1.03)	1.19 (1.01)	0.60 (0.82)	0.67 (0.85)	**<0.001**	**<0.001**
Mean T25FW (sec)	18.94 (23.91)	17.75 (21.32)	12.18 (17.04)	11.29 (14.04)	**0.021**	**<0.001**
Mean 9HPT dominant hand (sec)	41.12 (37.60)	40.63 (38.11)	27.72 (19.01)	27.93 (16.20)	**0.003**	**<0.001**
Mean 9HPT non-dominant hand (sec)	51.88 (54.58)	43.98 (45.83)	29.79 (20.62)	31.98 (29.44)	**0.001**	**<0.001**
MRI metrics:						
Gad count	1.90 (8.67)	2.23 (8.43)	0.78 (4.04)	0.67 (2.84)	0.264	**0.002**
T2 lesion volume (mL)	13.39 (16.82)	13.44 (14.30)	10.28 (13.04)	10.28 (13.43)	0.132	**0.001**
Normalized brain volume (L)	1.48 (0.07)	1.47 (0.08)	1.47 (0.09)	1.47 (0.08)	0.280	0.388

[a]P-values for continuous and ordinal variables are calculated using a two-sided Welch's t-test due to unequal variances/sample sizes. P-value for the categorical variable "Sex" is calculated using a two-sided Fisher's exact test due to unequal and relatively small sample sizes.
[b]Single refers to the single anti-CD20-Ab test set (30% of the mixed dataset).
[c]Crogging refers to the nested cross validation aggregation procedure in the outer testing loop (100% of the mixed dataset).
[d]Disease duration is measured from the time of symptom onset.
Standard deviation shown in brackets following each value.
FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test.
P-values < 0.05 are shown in bold.

Supplementary Table 4 Group statistics for predicted responders and non-responders to laquinimod at the 50th percentile threshold

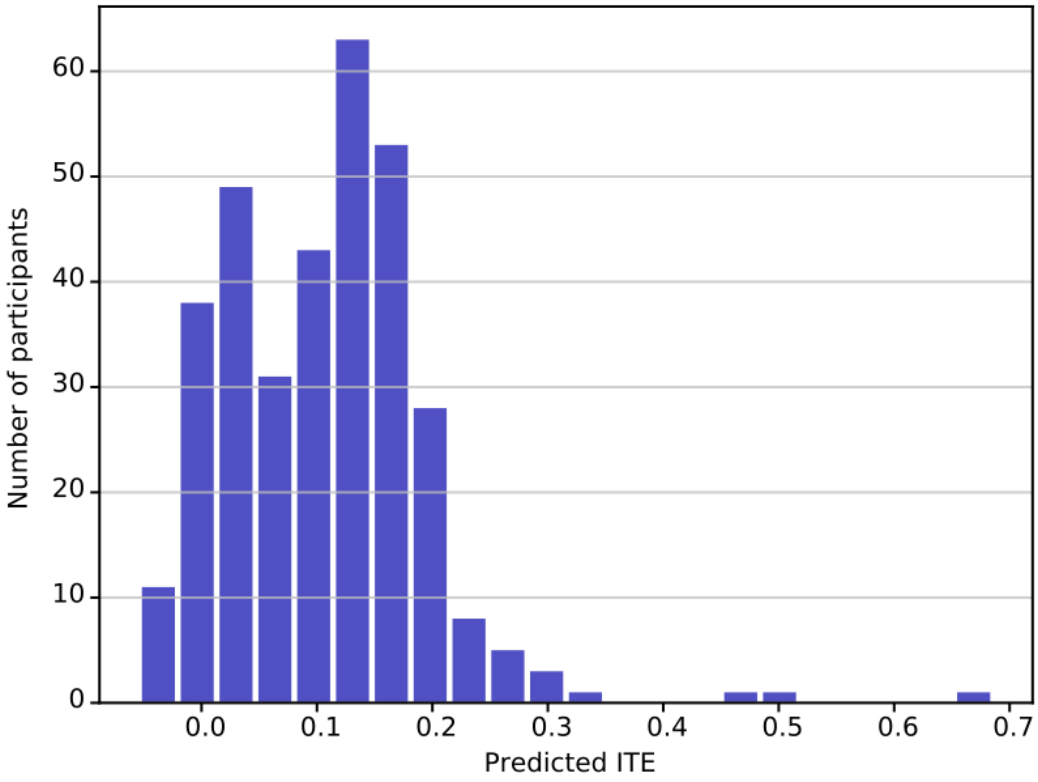
	Responders		Non-responders		P-value[a]	
	Original[b]	Retrained[c]	Original	Retrained	Original	Retrained
Trial contribution:						
ARPEGGIO	158	158	165	165		
Demographics:						
Age (years)	45.83 (7.38)	44.57 (7.41)	47.10 (6.14)	48.31 (5.60)	0.095	**<0.001**
Sex (% male)	60.13%	62.03%	49.09%	47.27%	0.057	**0.010**
Height (cm)	173.28 (9.19)	173.62 (9.20)	170.43 (9.60)	170.11 (9.49)	**0.007**	**0.001**
Weight (kg)	73.61 (15.54)	75.53 (16.14)	75.62 (16.05)	73.78 (15.49)	0.254	0.321
Disease duration (years)[d]	6.51 (4.41)	6.95 (5.09)	9.10 (6.61)	8.68 (6.27)	**<0.001**	**0.007**
Disability Scores:						
EDSS	4.60 (0.93)	4.52 (0.97)	4.36 (0.95)	4.43 (0.92)	**0.024**	0.401
FSS-Bowel and Bladder	1.36 (0.95)	1.28 (0.95)	1.09 (0.86)	1.17 (0.87)	**0.008**	0.287
FSS-Brainstem	1.03 (0.89)	1.22 (0.94)	0.96 (0.96)	0.77 (0.86)	0.512	**<0.001**
FSS-Cerebellar	2.35 (0.77)	2.37 (0.77)	1.88 (0.87)	1.86 (0.85)	0.852	**<0.001**
FSS-Cerebral	0.96 (0.91)	0.82 (0.90)	0.85 (0.89)	0.99 (0.89)	0.285	0.100
FSS-Pyramidal	2.96 (0.56)	2.92 (0.66)	2.82 (0.62)	2.87 (0.53)	**0.038**	0.446
FSS-Sensory	1.61 (1.08)	1.76 (1.07)	1.87 (0.95)	1.72 (0.98)	**0.024**	0.739
FSS-Visual	1.39 (1.49)	1.42 (1.47)	0.39 (0.68)	0.36 (0.67)	**<0.001**	**<0.001**
Mean T25FW (sec)	10.29 (9.78)	9.96 (9.60)	9.00 (6.48)	9.31 (6.78)	0.166	0.487
Mean 9HPT dominant hand (sec)	31.47 (14.88)	30.61 (14.86)	25.34 (7.90)	26.16 (8.46)	**<0.001**	**0.001**
Mean 9HPT non-dominant hand (sec)	34.69 (20.65)	32.11 (15.50)	26.20 (6.84)	28.67 (15.96)	**<0.001**	0.051
MRI metrics:						
Gad count	0.25 (0.56)	0.22 (0.41)	0.18 (0.38)	0.21 (0.54)	0.151	0.954
T2 lesion volume (mL)	3.17 (4.16)	3.06 (3.93)	2.84 (3.72)	2.95 (3.96)	0.451	0.804
Normalized brain volume (L)	1.50 (0.11)	1.51 (0.12)	1.45 (0.12)	1.45 (0.11)	**<0.001**	**<0.001**

[a]P-values for continuous and ordinal variables are calculated using a two-sided Welch's t-test due to unequal variances/sample sizes. P-value for the categorical variable "Sex" is calculated using a two-sided Fisher's exact test due to unequal and relatively small sample sizes.
[b]The original model trained on 70% of the anti-CD20-Ab dataset.
[c]The model trained on 100% of the anti-CD20-Ab dataset.
[d]Disease duration is measured from the time of symptom onset.
Standard deviation shown in brackets following each value.
FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test.
P-values < 0.05 are shown in bold.

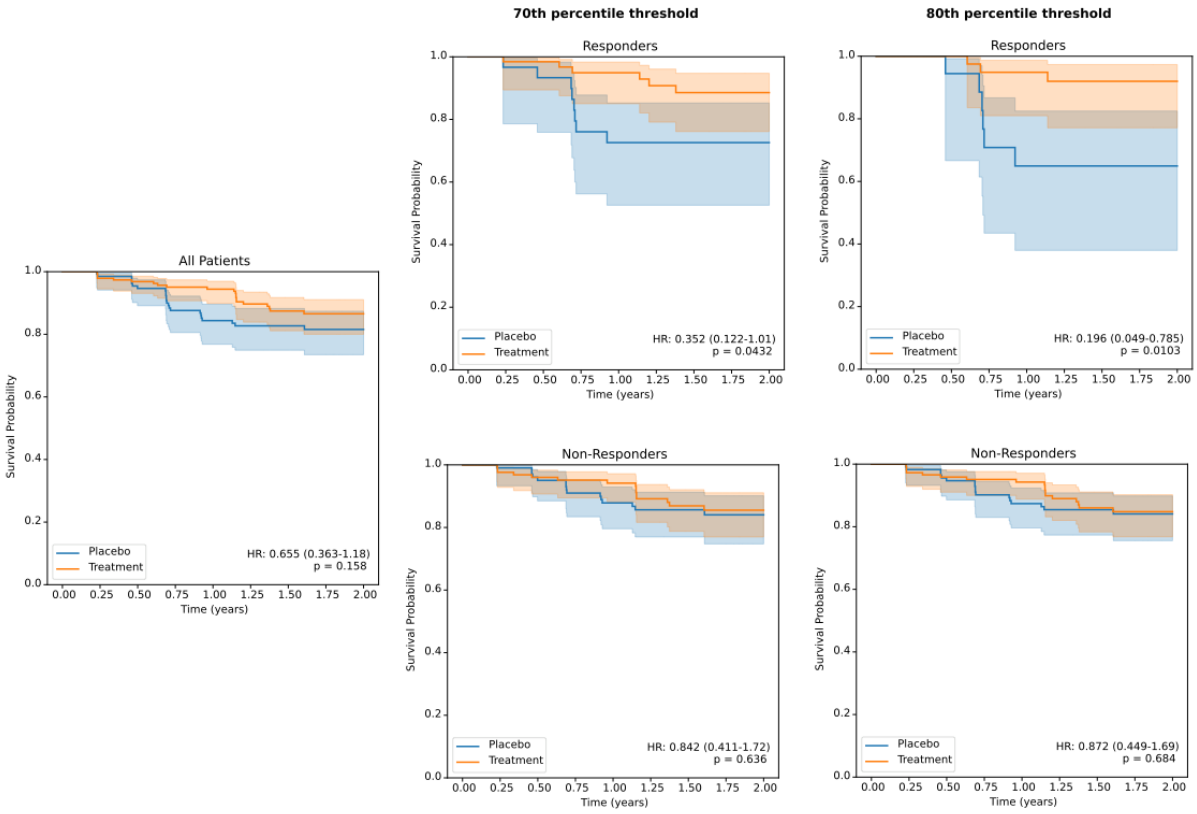
Supplementary Table 5 Group statistics for predicted responders and non-responders to laquinimod at the 75th percentile threshold

	Responders		Non-responders		P-value[a]	
	Original[b]	Retrained[c]	Original	Retrained	Original	Retrained
Trial contribution:						
ARPEGGIO	78	78	245	245		
Demographics:						
Age (years)	45.28 (7.43)	42.18 (7.77)	46.86 (6.55)	47.85 (5.84)	0.097	**<0.001**
Sex (% male)	61.54%	64.10%	52.24%	51.43%	0.192	0.052
Height (cm)	173.03 (9.66)	173.99 (9.31)	171.44 (9.43)	171.13 (9.47)	0.209	**0.021**
Weight (kg)	72.41 (15.46)	74.71 (16.55)	75.35 (15.88)	74.62 (15.60)	0.150	0.966
Disease duration (years)[d]	6.31 (4.77)	6.71 (5.36)	8.32 (6.00)	8.19 (5.87)	**0.003**	**0.040**
Disability Scores:						
EDSS	4.84 (0.89)	4.52 (0.89)	4.36 (0.93)	4.46 (0.96)	**<0.001**	0.639
FSS-Bowel and Bladder	1.54 (0.97)	1.40 (0.98)	1.12 (0.87)	1.17 (0.89)	**0.001**	0.068
FSS-Brainstem	1.14 (0.92)	1.40 (0.97)	0.94 (0.92)	0.86 (0.87)	0.101	**<0.001**
FSS-Cerebellar	2.56 (0.67)	2.51 (0.71)	1.96 (0.85)	1.98 (0.85)	0.855	**<0.001**
FSS-Cerebral	1.12 (0.95)	0.79 (0.88)	0.84 (0.87)	0.94 (0.90)	**0.026**	0.204
FSS-Pyramidal	3.00 (0.60)	2.86 (0.63)	2.86 (0.59)	2.90 (0.58)	0.070	0.598
FSS-Sensory	1.71 (1.09)	1.90 (1.03)	1.75 (1.01)	1.69 (1.02)	0.743	0.125
FSS-Visual	1.87 (1.67)	1.87 (1.64)	0.57 (0.87)	0.57 (0.89)	**<0.001**	**<0.001**
Mean T25FW (sec)	11.93 (12.64)	10.36 (12.22)	8.90 (6.12)	9.40 (6.54)	**0.045**	0.510
Mean 9HPT dominant hand (sec)	35.33 (18.96)	32.38 (17.35)	26.11 (7.88)	27.05 (9.72)	**<0.001**	**0.012**
Mean 9HPT non-dominant hand (sec)	39.88 (26.97)	34.09 (16.70)	27.32 (7.78)	29.17 (15.35)	**<0.001**	**0.023**
MRI metrics:						
Gad count	0.18 (0.38)	0.23 (0.42)	0.22 (0.51)	0.21 (0.50)	0.409	0.695
T2 lesion volume (mL)	2.44 (3.14)	3.22 (3.70)	3.18 (4.16)	2.93 (4.02)	0.099	0.565
Normalized brain volume (L)	1.50 (0.12)	1.53 (0.13)	1.47 (0.12)	1.46 (0.11)	**0.021**	**<0.001**

[a]P-values for continuous and ordinal variables are calculated using a two-sided Welch's t-test due to unequal variances/sample sizes. P-value for the categorical variable "Sex" is calculated using a two-sided Fisher's exact test due to unequal and relatively small sample sizes.
[b]The original model trained on 70% of the anti-CD20-Ab dataset.
[c]The model trained on 100% of the anti-CD20-Ab dataset.
[d]Disease duration is measured from the time of symptom onset.
Standard deviation shown in brackets following each value.
FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test.
P-values < 0.05 are shown in bold.



Supplementary Figure 1 Histogram of predicted ITE for the anti-CD20-Ab test set. Positive predicted ITEs indicate a benefit from anti-CD20-Abs over placebo.



Supplementary Figure 2 Kaplan-Meyer curves for predicted responders to laquinimod at different percentile thresholds for response. Survival probability is measured in terms of time-to-CDP24. Censorship times are clamped at 2 years. P-values are calculated using log-rank tests.