medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Estimating treatment effect for individuals with progressive multiple sclerosis using deep learning

Jean-Pierre R. Falet[1,2], Joshua Durso-Finley[2,3], Brennan Nichyporuk[2,3], Julien Schroeter[2], Francesca Bovis[4], Maria-Pia Sormani[4], Doina Precup[5], Tal Arbel[2,3] and Douglas Lorne Arnold[1,6]

Abstract

Modeling treatment effect could identify a subgroup of individuals who experience greater benefit from disease modifying therapy, allowing for predictive enrichment to increase the power of future clinical trials. We use deep learning to estimate the conditional average treatment effect for individuals taking disease modifying therapies for multiple sclerosis, using their baseline clinical and imaging characteristics. Data were obtained as part of three placebo-controlled randomized clinical trials: ORATORIO, OLYMPUS and ARPEGGIO, investigating the efficacy of ocrelizumab, rituximab and laquinimod, respectively. A shuffled mix of participants having received ocrelizumab or rituximab, anti-CD20-antibodies, was separated into a training (70%) and testing (30%) dataset, but we also performed nested cross-validation to improve the generalization error estimate. Data from ARPEGGIO served as additional external validation. An ensemble of multitask multilayer perceptrons was trained to predict the rate of disability progression on both active treatment and placebo to estimate the conditional average treatment effect. The model was able to separate responders and non-responders across a range of predicted effect sizes. Notably, the average treatment effect for the anti-CD20-antibody test set during nested cross-validation was significantly greater when selecting the model's prediction for the top 50% (HR 0.625, p=0.008) or the top 25% (HR 0.521, p=0.013) most responsive individuals, compared to HR 0.835 (p=0.154) for the entire group. The model trained on the anti-CD20-antibody dataset could also identify responders to laquinimod, finding a significant treatment effect in the top 30% of individuals (HR 0.352, p=0.043). We observed enrichment across a broad range of baseline features in the responder subgroups: younger, more men, shorter disease duration, higher disability scores, and more lesional activity. By simulating a 1-year study where only the 50%

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

predicted to be most responsive are randomized, we could achieve 80% power to detect a significant difference with 6 times less participants than a clinical trial without enrichment. Subgroups of individuals with primary progressive multiple sclerosis who respond favourably to disease modifying therapies can therefore be identified based on their baseline characteristics, even when no significant treatment effect can be found at the whole-group level. The approach allows for predictive enrichment of future clinical trials, as well as personalized treatment selection in the clinic.

Author affiliations:

1 Department of Neurology and Neurosurgery, McGill University

2 Centre for Intelligent Machines, McGill University, Canada

3 Department of Electrical and Computer Engineering, McGill University, Canada

4 Department of Health Sciences (DISSAL), University of Genova, Italy

5 School of Computer Science, McGill University, Canada

6 NeuroRx Research

Correspondence to: Jean-Pierre R. Falet

Full address: 3801 University St. #WB-321, Montreal, Quebec H3A 2B4, Canada

E-mail: jean-pierre.falet@mail.mcgill.ca

Running title: Multiple sclerosis treatment effect

Keywords: multiple sclerosis; progression; deep learning; precision medicine; clinical trials

Abbreviations: CATE = Conditional average treatment effect; CDP = Confirmed Disability Progression; CPH = Cox Proportional Hazards; CV = Cross-validation; EDSS = Expanded Disability Status Scale; Gad = Gadolinium-enhancing lesion; Individual treatment effect = ITE; MLP = Multi-layer perceptron; PPMS = Primary progressive multiple sclerosis; RRMS = Relapsing-remitting multiple sclerosis

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Introduction

Multiple disease modifying therapies have been successfully developed for the treatment of RRMS using the strategy of performing relatively short and small phase 2 trials with an MRI endpoint for establishing proof of concept and finding the optimal dose, before proceeding to longer, more expensive phase 3 trials. The absence of analogous MRI endpoints for progressive multiple sclerosis have hampered progress in developing drugs for this clinical phase of the disease. As proceeding directly to large, phase 3 trials is expensive and risky, most programs having followed this path have failed to adequately demonstrate efficacy.

It is often the case that medications are more effective in some patients than others. Selecting such a subgroup for inclusion in a clinical trial in order to increase its power is a technique called predictive enrichment.[1] A drug proven to be efficacious in an enriched trial can later be tested in a larger, more inclusive population. This sequence prevents efficacious medications from having their effect diluted in early clinical trials due to inclusion of a population that is too heterogeneous. As an example, Bovis et al.[2] used CPH models to predict a more responsive sub-group of RRMS patients to laquinimod, a medication whose average treatment effect in the original phase 3 studies was insufficient for drug approval.

To achieve the goal of enriching clinical trials with more responsive individuals, a machine learning problem can be formulated as the prediction of the ITE, or the difference between a person's rate of progression on treatment and that on placebo. This formulation is grounded in the theory of causal inference, and the related sub-fields of heterogeneous treatment effects and uplift modeling have contributed numerous approaches adapted to machine learning (reviewed elsewhere[3]). Arguably some of the most popular methods have been tree-based approaches (see Radcliffe and Surry[4] for an example) which model treatment effect directly, and meta-learning approaches[5] which use base models trained on the outcome of interest for the downstream task of treatment effect prediction.

We implement a learning framework based on an ensemble of multitask MLPs, a type of deep neural network architecture, to predict ITE using readily available clinical information (demographic characteristics and clinical disability scores) and scalar MRI metrics (lesional and volumetric) obtained at a screening visit. This approach can be used to identify a sub-

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

group of more responsive individuals for the purpose of clinical trial enrichment, using a desirable ITE threshold. In this work, we study the population of patients with PPMS that were exposed to anti-CD20 monoclonal antibodies in two clinical trials, ORATORIO (NCT01194570) and OLYMPUS (NCT00087529). We also test our model on individuals exposed to laquinimod as part of the ARPEGGIO trial (NCT02284568), a medication with a completely different mechanism of action, to assess whether learned predictors of response could be mechanism-agnostic.

Materials and methods

Data

Data were acquired as part of three randomized placebo-controlled clinical trials: ORATORIO,[6] OLYMPUS[7] and ARPEGGIO.[8] Anti-CD20 monoclonal antibodies, henceforth abbreviated anti-CD20-Abs (ocrelizumab in ORATORIO and rituximab in OLYMPUS) are primarily thought to act through B-cell depletion. Laquinimod (in ARPEGGIO) has a different and complex mechanism of action, thought to involve immunomodulatory effects on innate immune cell lineages (dendritic cells and monocytes peripherally, and microglia and astrocytes centrally). All three trials enrolled adults with PPMS and had similar inclusion criteria, shown in Supplementary Table 1. We excluded participants who spent less than 24 weeks in the trial, who had less than two clinical visits, or who were missing one or more input features at the baseline visit.

We used all clinical and MRI metrics (features) that were consistently recorded as part of the data available from the three trials, amounting to 19 features. Values were recorded at the baseline visit (immediately before randomized treatment allocation), and are a combination of binary (e.g. sex), ordinal (e.g. disability scores), and continuous variables (e.g. age, T2 lesion volume). Lesion segmentation and brain volume estimation were done according to the individual study's methodology. Means and standard deviations for each feature distribution are shown in Table 1. The feature distributions are similar for all three treatment groups. However, some clinically meaningful differences are found in the MRI metrics, where the anti-CD20-Ab group has a higher average Gad count and T2 lesion volume compared to the other treatment group. The following right-skewed distributions were log-transformed: normalized brain volume, T2 lesion volume, timed 25-foot walk (T25FW), and 9-hole peg

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

test (9HPT). Gad counts were binned into ten bins as follows: 0, 1, 2, 3, 4, 5-6, 7-9, 10-14, 15-19 and 20+ lesions. Finally, to improve convergence during gradient descent,[9] all non-binary features were standardized by subtracting the mean and dividing by the standard deviation, both calculated from the training dataset.

Treatment effect modeling

We are interested in predicting the 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 in specific situations, such as randomized control trials, because treatment allocation is independent of the outcome. We provide a detailed discussion of two important estimands, the average treatment effect and the CATE in Supplementary Material Section 1. Briefly, the average treatment effect represents the average effect when considering the entire population, while CATE considers a sub-population selected based on patient characteristics. The following discussion focuses on the problem of CATE estimation, which we use to frame the problem of predicting treatment response.

The best estimator for the CATE is conveniently also the best estimator for the ITE in terms of mean squared error (see Equation 2 in Künzel et al.[5]). Several frameworks have been developed to model the CATE, but a simple metalearning approach which decomposes the estimation into sub-tasks that can be solved using any supervised machine learning model provides a flexible starting point.[5] For a broader survey of methods, see the survey on uplift modeling by Gutierrez & Gérardy[3] (the uplift literature has contributed extensively to the field of causal inference, particularly when dealing with randomized experiments from an econometrics perspective).

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

In the present work, a MLP was used as the base model for its high representational power, flexible architecture, and ability to integrate into larger end-to-end-trainable neural networks consisting of different modules (such as convolutional neural networks or attention modules). We used a multitask architecture with two output heads, one for the modeling the potential outcome on treatment, $\hat{\mu}_1(x)$, and the other to model the potential outcome on placebo, $\hat{\mu}_0(x)$. For inference, we compute can compute the CATE estimate $\hat{\tau}(x_i)$ given a feature vector x_i as:

$$\hat{\tau}(x_i) = \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i) \qquad (2)$$

An additional correction to increase robustness to distributional shifts that could occur as a result of our tuning procedures is applied to $\hat{\tau}(x_i)$ at inference time to produce our final ITE estimate, $\hat{\tau}_i$ (see Supplementary Material Section 1 and 2 for details).

This multitask approach can be seen as a variant of the T-Learner described by Künzel et al.,[5] except that the two base models in our case share weights in the first layers. Our network is similar to that conceptualized by Alaa et al,[12] but without the propensity network used to correct for any dependence between the treatment allocation and the outcome conditional on the features, given that our dataset is from a randomized controlled trial.

To decrease the size of the hyperparameter search space, we fixed the number of layers and only tuned the layer width. We used two common hidden layers and one treatment-specific hidden layer (Fig. 1). More common or treatment-specific layers could be used if necessary, but given the low dimensionality of our feature-space and the relatively low number of instances, we opted to keep the network's depth small to avoid over-fitting. The inductive bias behind our choice of using a multitask architecture is that disability progression can have both disease-specific and treatment-specific predictors which can be encoded into the common and treatment-specific hidden layer representations, respectively. Consequently, the common hidden layers can learn from all the available data, irrespective of treatment allocation. Rectified linear unit (ReLU) activation functions were used for non-linearity.

We compared the performance of the multitask MLP on our validation metric, the weighted AD_{abc} (defined below in the "Validation metrics" subsection) to the following linear (or log-linear) models as baselines: ridge regression and CPH. We also compared it to a popular non-linear heterogeneous treatment effect estimator, uplift forest.[4] Ridge regression models were used as base models inside both an S-Learner and T-learner configuration (as defined by Künzel et al.[5]), while the CPH model was used as the base model for a T-learner only (as

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

implemented by Bovis et al.[2]). The uplift forest was used as a standalone heterogeneous treatment effect estimator.

Outcome definition

The primary outcome used in clinical trials assessing the efficacy of therapeutic agents on disease progression is the time to CDP at 12, or 24 weeks. We chose to use CDP at 24 weeks (CDP24) because it is a more robust indication that the disability accrual will be sustained over 5 years.[13] CDP24 is based on progression on the EDSS, a scale from 0 (no disability) to 10 (death), in discrete 0.5 increments. We define CDP24 as an increase in the EDSS of 0.5 for baseline EDSS values > 5.5, or an increase of 1.0 for baseline EDSS values ≤ 5.5, sustained over 24 weeks. This difference in the increment required to confirm disability progression is commonly adopted in clinical trials, and partially accounts for the finding that patients transition through different stages of the EDSS at different rates.[14]

While it is possible to predict time-to-event using traditional machine learning methods if workarounds are used to address right-censored data or using machine learning frameworks specifically developed to model survival data (reviewed elsewhere[15]), we choose not to model time-to-CDP24 because of limitations inherent in this metric. As outlined by Healy et al.,[16] CDP reflects not only the rate of progression but also the current stage of disease, which is problematic because the stage is represented by a discretized EDSS at a single baseline visit. This results in a noisy outcome label which could make it harder for a model to learn a representation that relates to the progressive biology which we are trying to model.

We therefore model the rate of progression directly by fitting a linear regression model onto the EDSS values of each individual participant over multiple visits (see Supplementary Material Section 3 for details) and take its slope to be the outcome label y_i that our model uses for training. One advantage of the slope outcome over time-to-CDP24 is that it can be modeled using any type of regression model. After model training using the slope outcome, we revert to time-to-CDP24 for model evaluation to facilitate comparison with treatment effect survival metrics reported in the original clinical trial publications.

Training

We trained our model using mini-batch gradient descent with a batch size of 128 and the Adam optimizer.[17] To prevent overfitting, we monitored the concordance index (C-index) on

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

the validation set during CV, and early-stopped model training at the epoch with the highest C-index, up to a maximum of 100 epochs. Dropout and L2 regularization were used, along with a max-norm constraint on the weights[18] to further prevent overfitting.

Batches were sampled in a stratified fashion to maintain the proportions of participants receiving active treatment and placebo. The mean squared error was computed from each instance's squared error at the output-head with the available factual (ground-truth) outcome. Furthermore, the squared errors from each output head were weighted by $n_s/(2 * n_t)$, where n_s represents the total number of participants in the training split and n_t represents the number of participants in the treatment arm corresponding to the output head of interest. This compensates for treatment allocation imbalance in the dataset.

Hyperparameter tuning and experimental setup

A random search over 100 different hyperparameter combinations was used to find the combination with the best validation performance (see Supplementary Table 2 for the full search set).

In our first experiment, we randomly shuffled data from ORATORIO and OLYMPUS given that the two active arms test drugs (ocrelizumab and rituximab, respectively) with the same mechanism of action (n = 1,119). We will refer to this dataset as the anti-CD20-Ab dataset. We then split this dataset into a training (70%) and testing (30%) set. We kept the data from ARPEGGIO as a second held-out test set (n = 323). The training set was subjected to 10-fold CV for hyperparameter tuning.

In our second experiment, we used 5x8 nested CV[19] on the anti-CD20-Ab dataset to ensure our model training procedure was robust to the choice of training and test split. This involves adding an outer loop to the usual CV procedure, such that the entire dataset is split into five sets, one of which (20% of the total) is used as an unseen test set for each outer fold. Its corresponding training set (the 80% remaining) is subjected to 8-fold CV (validation sets are therefore 10% of the original dataset) for tuning and model selection in the inner loop. This results in five different models being selected and tested on a different test set, amounting to test predictions for every individual in the dataset.

In our third experiment, we retrained a model on the entire anti-CD20-Ab dataset and tested it on the ARPEGGIO dataset. The training set was subjected to 10-fold CV.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

We used CV aggregation, or crogging,[20] to improve the generalization error estimate using our metrics. Crogging involves aggregating all validation set predictions (rather than the validation metrics) and computing one validation metric for the entire CV procedure. We also used crogging in the outer loop of the nested CV experiment, therefore computing one test metric for the entire dataset.

We aimed to reduce variance by using the early-stopped models obtained from each CV fold as members of an ensemble. This ensemble's prediction is the median of its members' predictions (as opposed to the mean, which is more sensitive to outliers), and is used for inference on the unseen test set.

Hyperparameter tuning for the baseline models (ridge regression and CPH) was done on same folds and with the same metrics as for the MLP. The uplift forest was tuned only for the weighted AD_{abc} (defined below) given that it estimates treatment effect directly.

Validation metrics

The best model was selected during CV on the basis of two validation metrics: the C-index of the factual predictions, and the weighted AD_{abc}.

The first validation metric, the C-index, is defined as the proportion of correctly ordered pairs over all admissible pairs.[21] We used the C-index instead of the mean squared error for tuning because we found empirically that models tuned for C-index performed better during CV in terms of their weighted AD_{abc}. It has the advantage over other non-parametric ranking metrics such as Spearman's rank correlation coefficient of being able to deal with censored data, if the chosen outcome label of interest would be a time-to-event variable, but in the case where a slope outcome is used either would provide an appropriate measure of rank ordering.

The second validation metric, a modified (weighted) version of the AD_{abc} described by Zhao et al.,[22] directly measures the quality of the treatment effect prediction. It is a measure derived from the area under the *AD(c)* curve, which is the average treatment effect for individuals who are predicted to respond (according to $\hat{\tau}_i$) more than a desired response threshold *c*. A model capable of ranking responders appropriately should have an *AD(c)* curve that is almost monotonically increasing with a large area under the curve. AD_{abc} is in unit years, and a larger positive number therefore indicates the ability for the model to separate responders across a range of predicted treatment effect sizes. See Supplementary Material Section 4 for a more detailed discussion including how we weigh the AD_{abc}. For ease of interpretation, we

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

will refer to c in terms of the percentile for c among all $\hat{\tau}_i$ in the test set (e.g. the 80th percentile threshold for c represents the top 20% most responsive individuals according to our model's prediction $\hat{\tau}_i$).

We combine both validation metrics during tuning by choosing the model with the highest weighted AD_{abc} among all models that fall within 1 SD of the best performing model based on the C-index. The SD of the best performing model's C-index is calculated from the C-indices obtained in the individual CV folds.

Statistical analysis

Hazard ratios were calculated using CPH models with associated p-values from log-rank tests. Right censoring times were clamped at 2 years at inference for all experiments except for the effect size calculation in the simulated phase 2 clinical trial in the Results section "Simulating a one-year phase 2 clinical trial enriched with predicted responders", where the right censoring times were clamped at 1 year.

Note that we multiplied all treatment effect values by −1 to simplify interpretation in the Results section, such that a positive effect indicates improvement, while a negative effect indicates worsening on treatment.

Software

All experiments were implemented in Python 3.8.[23] MLP models were implemented using the Pytorch library.[24] Scikit-Learn[25] was used for the implementation of ridge regression, while Lifelines[26] was used for CPH. CausalML[27] was used for implementing the uplift forest. For reproducibility, the same random seed was used for data splitting and model initialization across all experiments.

Data availability

Data used in this work are controlled by pharmaceutical companies and therefore are not publicly available. Access requests should be forwarded to data controllers via the corresponding author.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Results

Predicting response to anti-CD20 monoclonal antibodies

We first trained an ensemble of multitask MLPs on 70% (n = 784) of the mixed dataset consisting of participants from ORATORIO and OLYMPUS. A histogram of test predictions on the 30% (n = 335) test set is shown in Supplementary Fig. 1. Varying the threshold c from the 10th percentile to the 80th percentile of predicted treatment effect results in the *AD(c)* curve shown in Fig. 2 (AD_{abc} = 0.0679). It is appropriately increasing throughout its range with a positive AD_{abc}, demonstrating the ability for the model to rank response to anti-CD20-Abs accurately, in a way that reflects the group-level ground-truth.

Kaplan-Meyer curves of the factual time-to-CDP24 in predicted responders ($\hat{\tau}_i \geq c$) and non-responders ($\hat{\tau}_i < c$) are shown in Fig. 4 at various percentile thresholds for c. We observe a decrease in HR from 0.787 (p=0.292) when including the entire test population to 0.395 (p=0.0279) at the 75th percentile threshold for c (including only the top 25% most responsive individuals). At this 75th percentile threshold, the nonresponder group has a HR of 1.02 (p=0.938), which suggests that a large part of the beneficial effect visible at the whole-group level comes from a very small proportion of patients. The lowest HR achieved is 0.3 (95% CI 0.100-0.901, p=0.0229) at the 85th percentile threshold, but beyond this point the sample size becomes too small and the confidence intervals too large to provide a meaningful estimate.

Due to the relatively small test size (n = 335), to better estimate the generalization error of the training and model selection procedure, we performed 5x8 nested CV,[19] yielding test-time predictions for the entire dataset (n = 1,119). The outer-fold crogging is expectedly consistent with our previous experiment (AD_{abc} = 0.0415), and provides more confidence in the treatment effect estimates due to the increased test set size. The effect is very similar at the 50th percentile threshold (HR 0.625, p=0.008), while it slightly less beyond the 75th percentile. From this estimate, we expect to reach a HR of 0.521 (p=0.013) with a 75th percentile threshold, with a peak HR of 0.338 (p=0.013) at the 92nd percentile threshold.

We verified whether the predictive ability of the model is similar if we consider men and women separately. The model tested on men alone achieved a weighted AD_{abc} of 0.057, while the one tested on women alone achieved 0.093, indicating a greater power in isolating responders in women. Taking the 75th percentile for comparison, the model reaches a HR of 0.293 (p<0.001) with women compared to 0.555 (p=0.090) with men.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Simulating a one-year phase 2 clinical trial enriched with predicted responders

To understand the effect of enriching a future clinical trial studying novel B-cell depleting agents for reduction of disability progression, we simulated a 1-year randomized clinical trial using populations enriched with predicted responders. Using our model to predict sub-groups of responders to anti-CD20-Abs across a variety of thresholds, we can calculate the 1-year CDP24 event rate and 1-year HR for these sub-groups, which can then be used for sample size estimation. Table 2 shows the sample size needed to have 80% power to detect a significant difference with $\alpha = 0.05$ across various degrees of enrichment.

The 50th percentile is the threshold for inclusion of participants that provides the best compromise between needing a small number (n = 497) of participants to show a significant treatment effect, while still randomizing most screened participants. This process involves screening a total of 944 individuals and selecting the top 50% who are predicted to be most responsive for enrollment in the trial. This leads to a 6-fold reduction in the number of patients that need to be randomized, and a 3-fold reduction in the number of patients that need to be screened, compared to the scenario where all participants are randomized into a one-year study (n = 3,068).

Group characteristics of predicted responders and non-responders

We provide group-level characteristics of the predicted responders in comparison with non-responders at the 50th percentile threshold in Table 3. See Supplementary Table 3 for the statistics at the 75th percentile threshold. We observe enrichment across a broad range of input features in the responder sub-group: younger, more men, shorter disease duration, higher disability scores, and more lesional activity. Normalized brain volume was the only feature which did not differ between the two groups at either threshold. A lower weight was only significantly different in the single test set, and not in the nested CV aggregate.

Predicting response to laquinimod

Finally, to determine whether the same model could be predictive of treatment response to medications with different mechanisms of action, and to provide a second validation for the

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

model trained on the single 70% training set in the first anti-CD20-Ab experiment, we tested it on a separate clinical trial that was not used during training: ARPEGGIO. As a second step, we retrained the model on 100% of the anti-CD20-Ab dataset, and again tested this retrained model on the ARPEGGIO dataset. The results are shown in Table 4. We see that the model trained on the single anti-CD20-Ab training set also generalizes to this second unseen test set (AD_{abc} = 0.024). The treatment effect increases almost monotonically and reaches significance ($p < 0.05$) at the 80th percentile threshold. The retrained model performs better (AD_{abc} = 0.0436), obtaining a HR of 0.352 (p=0.043) at the 70th percentile and 0.196 (p=0.010) at the 80th percentile. The Kaplan-Meyer curves for the 70th and 80th percentile thresholds are shown in Supplementary Fig. 2. We show the group characteristics for predicted responders using a 50th percentile and a 75th percentile threshold in Supplementary Tables 4 and 5, respectively. The significant groupwise differences are largely similar to those obtained on the anti-CD20-Ab dataset, with a few notable exceptions. A taller height and a higher normalized brain volume are present in the ARPEGGIO responder group, whereas no difference is observed in weight, Gad count and T2 lesion volume. Altogether, this out of distribution generalization provides strong evidence suggesting that this approach can truly find underlying predictors of treatment response which are at least partly drug mechanism agnostic.

Comparison to baseline models

To determine whether non-linear models (such as an MLP with ReLU activations) provide any performance gains over linear models, we compared the multitask MLP to the commonly used ridge regression trained on the slope outcome and to a CPH model (which is log-linear) trained directly on the time-to-CDP24 outcome. We used an uplift forest as a non-linear baseline.

The AD_{abc} achieved by each model when tested on both the anti-CD20-Ab and the laquinimod datasets is shown in Table 5, along with the average p-value obtained from log-rank tests at each of the 8 percentile thresholds used to compute the *AD(c)*. The multitask MLP outperforms all other linear (and log-linear) baselines, along with the non-linear uplift forest which fails to identify responders (negative weighted AD_{abc}) on our dataset.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Discussion

This work addresses an important limitation in the current treatment of PPMS, whereby no biomarker adequately predictive of treatment response is available to guide the choice of treatment for individuals, either in the clinic or in clinical trials. It is well recognized that the disability trajectory of individuals with multiple sclerosis and their response to specific treatments are highly heterogeneous. We therefore set the stage for treatment effect prediction with the primary goal of increasing the efficiency of early phase clinical trials in PPMS, but also secondarily to help make better treatment decisions in the clinic. We use a multitask MLP architecture for CATE estimation, and through a one-year phase 2 clinical trial simulation demonstrate the benefit of predictive enrichment, which could reduce the number of patients randomized by 6-fold while screening 3 times less patients than would be required for a short (one year) clinical trial in PPMS. We go through numerous steps to validate our approach, in particular by using nested CV and crogging to better estimate the generalization error and by showing generalization to a different unseen test set. The latter demonstrates that this approach can likely be useful on future clinical trials even if they study different medications.

Although our results illustrate a potential benefit of using predictive models in the clinic to assist in selecting patients who are more likely to benefit from anti-CD20-Abs, it remains an open question whether patients for whom our model predicted minimal response over two years could benefit from longer periods of administration. Answering this question would require longer-term observational data.

Interpretability of neural networks is a growing field in the machine learning literature (reviewed elsewhere[28]) and how to best interpret the predictions of neural networks remains an area of active research. We therefore do not attempt to provide in-depth interpretations of the network's predictions, but instead analyse the group-level statistics of the predicted responders at various thresholds of response. The more responsive groups are enriched in numerous baseline features, including a younger age, more men, higher disability scores, and more lesional activity. Similarly, in subgroup analyses from OLYMPUS, an age less than 51 years and presence of gad lesions at baseline was also found to be associated with increased response.[7] In a study by Bovis et al.,[2] a response scoring function obtained via a T-learner composed of CPH models in RRMS found increased age, female sex, more Gad lesions and a higher normalized brain volume to be associated with better response. Conversely, Signori et

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

al.[29] found an opposite relationship between age and response in RRMS, demonstrating that different types of models looking at different combinations of features can find different optimal solutions.

Finding more individuals with greater lesional activity (particularly Gad, and to a lesser extent T2 lesions) in the responder subgroups would support a role for these lesions in progressive biology. However, this role appears to be indirect, as most of the progression in multiple sclerosis occurs independent of relapse activity. The presence of multiple pathophysiological mechanisms to explain progression in multiple sclerosis involving both inflammation and neurodegeneration has been postulated,[30] and smouldering inflammation associated with slowly enlarging lesions[31] could be a potential effector for anti-CD20-Ab's effect on progression. The alternative hypothesis that lesional activity in the responder group could be a marker of more aggressive disease for which CDP is more likely to occur is not supported by our results, since our model does not work solely by identifying a more rapidly progressive sub-group. This is evidenced by the reduction in the frequency of CDP in the treatment group for the more responsive individuals (see Table 2).

Interestingly, despite a balanced dataset with respect to gender, our model was better at identifying responders in women compared to men. This is particularly interesting because most of the responders are men. To deploy ethically fair point-of-care tools, every effort should be made to ensure that easily identifiable groups of individuals (e.g. based on gender, age, and ethnicity) all benefit equally from such tools. Potential ways to address this in future work would be through the use of fine tuning on larger observational datasets, optimizing models for each identifiable sub-group separately, or through more complex loss-weighing schemes.

Limitations of this work include the choice of model. Although our MLP outperformed linear baselines, MLPs are notoriously more difficult to tune and at higher risk of overfitting. We made heavy use of several regularization schemes to prevent this (shallow/narrow network, dropout, weight decay, max-norm constraint and early-stopping). This approach is not the easiest to implement nor the most computationally efficient, but it provided the best results, suggesting inherent non-linearities in the dataset that benefit from ReLU networks and their compositional expressivity. Our hyperparameter tuning procedure is also one of many that can be designed. Optimizing for the weighted AD_{abc} directly could potentially provide better results and is the subject of ongoing work. Finally we used MRI-derived metrics that came

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

from the individual clinical trials and that offered expert corrected lesion counts and volumes. Extracting features from the MRI images themselves through convolutional neural networks is the subject of ongoing work.

In conclusion, we demonstrate the utility of CATE estimation for predictive enrichment of clinical trials aimed at increasing the efficiency of the drug development process in PPMS. We were able to find subgroups of increasingly responsive individuals to anti-CD20 therapies. Our model was able to generalize to a medication with a very different mechanism of action, laquinimod, suggesting that there might be common predictors for treatment effect independent of mechanism, which would facilitate the use of such a model for planning future clinical trials. This flexible training paradigm and multitask model architecture can easily be integrated into larger neural networks to benefit from data-driven imaging feature extraction through convolutional neural networks. The use of this approach is not limited to enrichment of clinical trials and can also be used for precision medicine in the clinic when deciding whether initiation of a therapy is worthwhile, by predicting response of individual patients based on their unique characteristics.

Funding

This investigation was supported (in part) by an award from the International Progressive Multiple Sclerosis Alliance (award reference number PA-1412-02420), by an endMS Personnel Award from the Multiple Sclerosis Society of Canada (Falet, JR), and by a Canada Graduate Scholarship-Masters Award from the Canadian Institutes of Health Research (Falet, JR). Falet, JR is also being supported through the Fonds de recherche Santé / Ministère de la Santé et des Services sociaux training program for specialty medicine residents with an interest in pursuing a research career, Phase 1.

Competing interests

Arnold, DL, reports consulting fees from Albert Charitable Trust, Alexion Pharma, Biogen, Celgene, Frequency Therapeutics, Genentech, Med-Ex Learning, Merck, Novartis, Population Council, Receptos, Roche, and Sanofi-Aventis, grants from Biogen, Immunotec and Novartis, and an equity interest in NeuroRx. Sormani, MP, has received personal compensation for consulting services and for speaking activities from Merck, Teva, Novartis,

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Roche, Sanofi Genzyme, Medday, GeNeuro, and Biogen. Precup, D, works part-time for DeepMind. The remaining authors report no competing interests.

References

1. Temple R. Enrichment of Clinical Study Populations. *Clinical Pharmacology & Therapeutics.* 2010;88:774–8.

2. Bovis F, Carmisciano L, Signori A, et al. Defining responders to therapies by a statistical modeling approach applied to randomized clinical trial data. *BMC Medicine* 2019;17(1):113.

3. Gutierrez P, Gérardy JY. Causal Inference and Uplift Modelling: A Review of the Literature. In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Proceedings of Machine Learning Research. 2017;67:1–13.

4. Radcliffe N and Surry PD. Real-World Uplift Modelling with Significance-Based Uplift Trees. 2012. Accessed October 1 2021. https://stochasticsolutions.com/pdf/sig-based-up-trees.pdf

5. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. In: *Proceedings of the National Academy of Sciences*. 2019;116:4156–65.

6. Montalban X, Hauser SL, Kappos L, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *New England Journal of Medicine*. 2017;376:209-220.

7. Hawker K, O'Connor P, Freedman MS, et al. Rituximab in patients with primary progressive multiple sclerosis: Results of a randomized double-blind placebo-controlled multicenter trial. *Annals of Neurology*. 2009;66:460–71.

8. Giovannoni G, Knappertz V, Steinerman JR, et al. A randomized, placebo-controlled, phase 2 trial of laquinimod in primary progressive multiple sclerosis. *Neurology*. 2020;95(8):e1027-e1040.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

9. LeCun YA, Bottou L, Orr GB, Müller KR. Efficient Backprop. In: Montavon G, Orr GB, Müller KR, eds. *Neural Networks: Tricks of the Trade: Second Edition*. Springer Berlin Heidelberg; 2012:9–48.

10. Imbens GW, Rubin DB, eds. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press; 2015.

11. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association*. 1986;81:945– 60.

12. Alaa AM, Weisz M, Schaar M van der. Deep Counterfactual Networks with Propensity-Dropout. In: *Proceedings of the 34th International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2017:70.

13. Kalincik T, Cutter G, Spelman T, et al. Defining reliable disability outcomes in multiple sclerosis. *Brain*. 2015;138:3287–98.

14. Zurawski J, Glanz BI, Chua A, et al. Time between expanded disability status scale (EDSS) scores. *Multiple Sclerosis and Related Disorders*. 2019;30:98–103.

15. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys.* 2019;51(6):1-36.

16. Healy BC, Glanz BI, Swallow E, et al. Confirmed disability progression provides limited predictive information regarding future disease progression in multiple sclerosis. *Multiple Sclerosis Journal - Experimental, Translational and Clinical*. 2021;7(2).

17. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *Proceedings of the 3rd International Conference for Learning Representations*. 2015.

18. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. 2014;15:1929–58.

19. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? *arXiv*. 2021. [Preprint] arXiv:2104.00673 [stat.ME].

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

20. Barrow DK, Crone SF. Crogging (cross-validation aggregation) for forecasting: A novel algorithm of neural network ensembles on time series subsamples. In: *The 2013 International Joint Conference on Neural Networks*. 2013:1–8.

21. Harrel Jr. FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996;15:361–87.

22. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association*. 2013;108:527–39.

23. Van Rossum G, Drake FL. *Python 3 Reference Manual*. CreateSpace; 2009.

24. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Proceedings of Advances in Neural Information Processing Systems 32*. 2019;32.

25. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–30.

26. Davidson-Pilon C. Lifelines: survival analysis in Python. *Journal of Open Source Software*. 2019;4:1317.

27. Chen H, Harinen T, Lee JY, Yung M, Zhao Z. CausalML: Python Package for Causal Machine Learning. *arXiv*. 2020. [preprint] arXiv:2002.11631 [cs.CY].

28. Zhang Y, Tiňo P, Leonardis A, Tang K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2021:1–17.

29. Signori A, Schiavetti I, Gallo F, Sormani MP. Subgroups of multiple sclerosis patients with larger treatment benefits: a meta-analysis of randomized trials. *European Journal of Neurology*. 2015;22:960–6.

30. Faissner S, Plemel JR, Gold R, Yong VW. Progressive multiple sclerosis: from pathophysiology to therapeutic strategies. Nature Reviews Drug Discovery. 2019;18:905–22.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

31. Elliott C, Wolinsky JS, Hauser SL, et al. Slowly expanding/evolving lesions as a magnetic resonance imaging marker of chronic active multiple sclerosis lesions. *Multiple Sclerosis Journal*. 2019;25(14):1915-1925.

32. Royston P, Parmar MK. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*. 2013;13:152.

Tables and Figures

Table 1 Feature distribution per treatment arm

	Anti-CD20-Ab *n* = 745	**Laquinimod** *n* = 190	**Placebo** *n* = 507
Trial contribution:			
ARPEGGIO	0	190	133
OLYMPUS	276	0	141
ORATORIO	469	0	233
Demographics:			
Age (years)	44.65 (8.08)	46.34 (6.56)	44.91 (8.13)
Sex (% male)	51.28	56.84	47.93
Height (cm)	170.67 (9.44)	172.20 (9.35)	170.44 (9.39)
Weight (kg)	74.91 (17.10)	75.59 (15.52)	74.18 (16.22)
Disease duration (years)[a]	7.53 (5.12)	8.15 (6.14)	7.14 (5.12)
Disability Scores:			
EDSS	4.73 (1.25)	4.49 (0.97)	4.62 (1.18)
FSS-Bowel and Bladder	1.24 (0.88)	1.27 (0.94)	1.18 (0.91)
FSS-Brainstem	0.84 (0.91)	1.00 (0.91)	0.86 (0.91)
FSS-Cerebellar	2.10 (1.02)	2.11 (0.82)	2.09 (0.95)
FSS-Cerebral	1.03 (0.88)	0.95 (0.91)	0.98 (0.87)
FSS-Pyramidal	2.80 (0.71)	2.92 (0.55)	2.83 (0.68)
FSS-Sensory	1.54 (1.03)	1.74 (1.04)	1.60 (1.07)
FSS-Visual	0.81 (0.92)	0.95 (1.34)	0.79 (0.97)
Mean T25FW (sec)	13.34 (17.42)	9.57 (8.78)	11.44 (12.83)
Mean 9HPT dominant hand (sec)	31.79 (27.45)	28.49 (12.28)	29.35 (15.35)
Mean 9HPT non-dominant hand (sec)	34.68 (34.13)	31.33 (17.89)	33.87 (3‘.49)
MRI metrics:			
Gad count	1.07 (4.60)	0.22 (0.53)	0.82 (4.78)
T2 lesion volume (mL)	11.47 (14.09)	3.00 (3.95)	8.38 (11.74)
Normalized brain volume (L)	1.47 (0.08)	1.48 (0.12)	1.47 (0.10)

[a]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.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Table 2 Estimated sample size to detect a significant treatment effect from anti-CD20-Abs by predictive enrichment

Percentile[a]	CDP control[b]	CDP treatment[b]	HR[c]	P-value[d]	Sample size control[e]	Sample size treatment[e]	Number screened[f]
0	0.148	0.109	0.745	0.088	1023	2045	3068
10	0.159	0.115	0.695	0.043	620	1239	2066
20	0.150	0.115	0.739	0.117	931	1862	3491
30	0.166	0.112	0.646	0.021	422	844	1809
40	0.183	0.101	0.516	0.002	180	360	900
50	0.172	0.088	0.478	0.003	166	331	994
60	0.154	0.080	0.485	0.012	183	366	1373
70	0.180	0.089	0.460	0.013	139	278	1390
80	0.239	0.075	0.283	<0.001	49	97	730

[a]Percentile threshold for selecting predicted responders into the simulation. The 0th percentile represents an unenriched population, while the 80th percentile leads to inclusion of only the top 20% most responsive individuals (i.e. a greater percentile represents a more enriched study population).
[b]Proportion of CDP24 events after one year for the responder groups corresponding to each percentile threshold.
[c]HR for time-to-CDP24 after one year for the responder groups corresponding to each percentile threshold.
[d]P-value obtained from a log-rank test.
[e]Sample size estimates reflect the number of participants that need to be randomized into the study and are based on the 1-year CDP24 rate and 1-year HR of responder groups in the anti-CD20-Ab dataset.
[f]Number of participants that need to be screened to reach the corresponding sample size estimate for randomization. This is dictated by the amount of predictive enrichment applied at randomization (see Percentile column).

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Table 3 Group statistics for predicted responders and non-responders to anti-CD20-Abs at the 50th percentile threshold

	Responders		Non-responders		P-value[a]	
	Single[b]	**Crogging[c]**	**Single**	**Crogging**	**Single**	**Crogging**
Trial contribution:						
OLYMPUS	52	199	68	218		
ORATORIO	113	361	103	341		
Demographics:						
Age (years)	44.93 (8.22)	43.58 (8.47)	45.59 (7.21)	45.48 (7.77)	0.436	**<0.001**
Sex (% male)	52.73	53.57	43.86	45.97	0.126	**0.012**
Height (cm)	171.75 (9.65)	170.77 (9.30)	170.19 (8.80)	170.21 (9.46)	0.123	0.313
Weight (kg)	73.93 (16.15)	74.20 (16.78)	77.72 (16.88)	75.35 (16.82)	**0.037**	0.251
Disease duration (years)[d]	6.99 (5.07)	6.97 (4.52)	8.25 (6.11)	7.78 (5.62)	**0.041**	**0.009**
Disability Scores:						
EDSS	4.85 (1.25)	4.79 (1.24)	4.57 (1.27)	4.63 (1.26)	**0.041**	**0.032**
FSS-Bowel and Bladder	1.32 (0.89)	1.28 (0.88)	1.12 (0.93)	1.16 (0.91)	0.051	**0.027**
FSS-Brainstem	0.96 (0.95)	1.04 (0.94)	0.67 (0.82)	0.63 (0.82)	**0.003**	**<0.001**
FSS-Cerebellar	2.33 (0.87)	2.28 (0.96)	1.82 (1.12)	1.90 (1.01)	**<0.001**	**<0.001**
FSS-Cerebral	1.09 (0.86)	1.10 (0.87)	1.03 (0.87)	0.96 (0.88)	0.521	**0.012**
FSS-Pyramidal	2.99 (0.63)	2.88 (0.64)	2.61 (0.80)	2.74 (0.76)	**<0.001**	**0.001**
FSS-Sensory	1.37 (1.00)	1.57 (1.04)	1.73 (1.03)	1.52 (1.05)	**0.001**	0.482
FSS-Visual	1.01 (0.97)	0.94 (0.96)	0.45 (0.73)	0.67 (0.87)	**<0.001**	**<0.001**
Mean T25FW (sec)	16.45 (22.81)	14.50 (18.81)	11.27 (14.30)	11.32 (13.42)	**0.014**	**0.001**
Mean 9HPT dominant hand (sec)	37.46 (34.70)	34.63 (30.86)	24.67 (5.60)	27.61 (14.33)	**<0.001**	**<0.001**
Mean 9HPT non-dominant hand (sec)	44.47 (45.59)	37.42 (36.75)	26.09 (6.98)	32.56 (32.32)	**<0.001**	**0.019**
MRI metrics:						
Gad count	1.22 (6.19)	1.45 (6.16)	0.88 (4.83)	0.67 (3.24)	0.569	**0.008**
T2 lesion volume (mL)	11.89 (15.60)	11.48 (13.20)	10.20 (12.44)	10.67 (14.22)	0.275	0.322
Normalized brain volume (L)	1.48 (0.07)	1.47 (0.08)	1.47 (0.09)	1.47 (0.08)	0.447	0.761

[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.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Table 4 Treatment effect for predicted responders to laquinimod at various response percentile thresholds

	Original[b]		Retrained[c]	
Percentile[a]	**HR**	**P-value**	**HR**	**P-value**
20	0.691	0.272	0.668	0.198
30	0.778	0.492	0.533	0.066
40	0.651	0.261	0.508	0.082
50	0.567	0.187	0.486	0.067
60	0.636	0.333	0.641	0.341
70	0.445	0.119	0.352	**0.043**
80	0.275	**0.028**	0.196	**0.010**

[a]Percentile threshold for selecting predicted responders to laquinimod. The 20th percentile considers the top 80% most responsive individuals to be "responders", while the 80th percentile considers only the top 20% most responsive individuals to be "responders".
[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.
P-values are calculated using a log-rank test.
P-values < 0.05 are shown in bold.

Table 5 Comparison of model performance on the anti-CD2-Ab and laquinimod datasets

	Anti-CD20-Ab[a]		Laquinimod[b]	
	AD_{abc}[c]	**P-value[d]**	**AD_{abc}**	**P-value**
Ridge Regression (S-Learner)	0.009	0.145	0.002	0.255
Ridge Regression (T-Learner)	-0.003	0.279	-0.003	0.107
CPH (T-Learner)	-0.005	0.547	-0.001	0.237
Uplift forest	-0.002	0.267	-0.002	0.380
Multitask MLP	**0.042**	**0.028**	**0.0436**	**0.0887**

[a]Refers to the anti-CD20-Ab dataset subjected to the nested cross validation aggregation procedure in the outer testing loop (test metrics are therefore computed from test predictions on 100% of the dataset).
[b]Refers to the laquinimod test set, with models being trained on 100% of the anti-CD20-Ab dataset.
[c]The weighted AD_{abc} values. A larger positive number indicates better performance at ranking responders.
[d]This p-value is an average of p-values obtained from log-rank tests performed on responder subgroups selected from percentile thresholds ranging from the 0th to the 80th percentile (in increments of 10). This average p-value is not to be interpreted as a standard p-value, but rather an intuitive summary of the significance achieved across a range of response thresholds.
Best performance for each metric is shown in bold.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

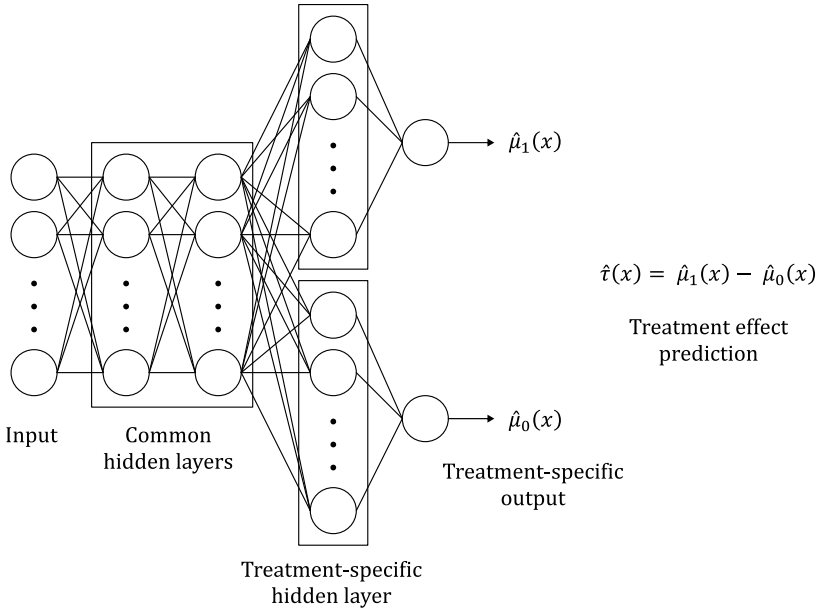

Figure 1 Multitask MLP architecture for CATE estimation. $\hat{\tau}(x)$: CATE estimate given a feature vector x. $\hat{\mu}_0(x)$: predicted potential outcome on control medication. $\hat{\mu}_1(x)$: predicted potential outcome on treatment.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

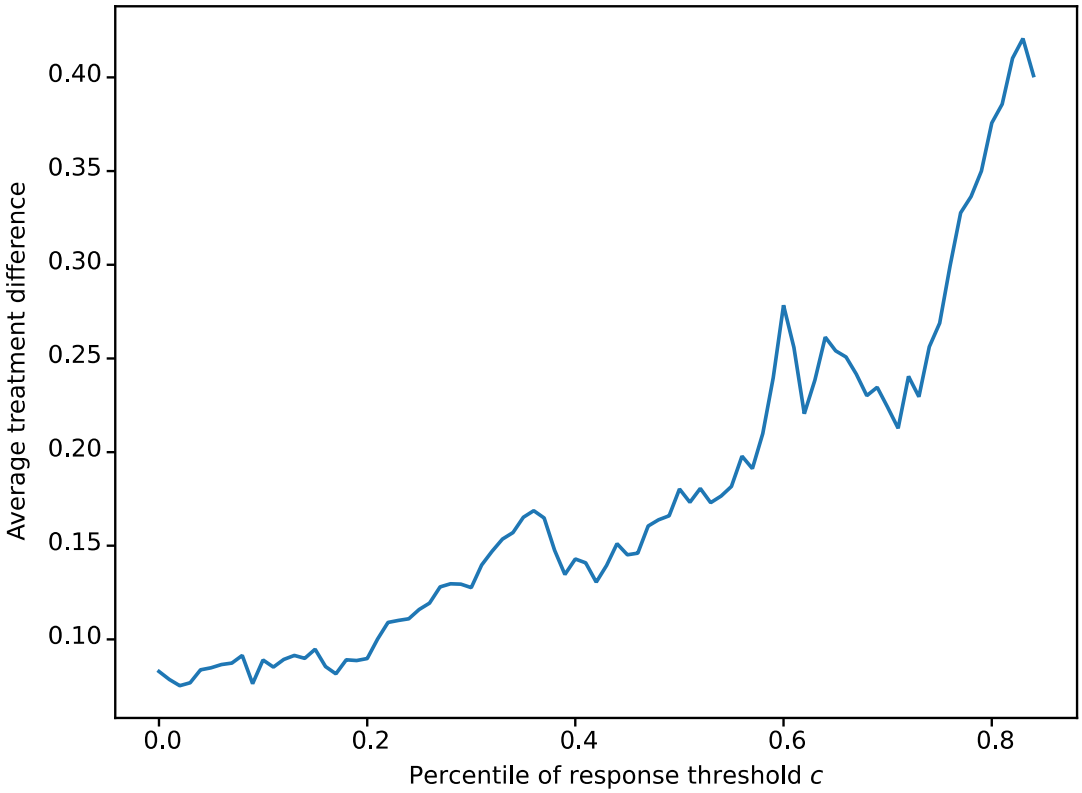

Figure 2 Average treatment difference curve for the anti-CD20-Ab test set. Average treatment difference represents the difference in the restricted mean survival time at 2 years between anti-CD20-Abs and placebo.

medRxiv preprint doi: https://doi.org/10.1101/2021.10.31.21265690; this version posted November 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

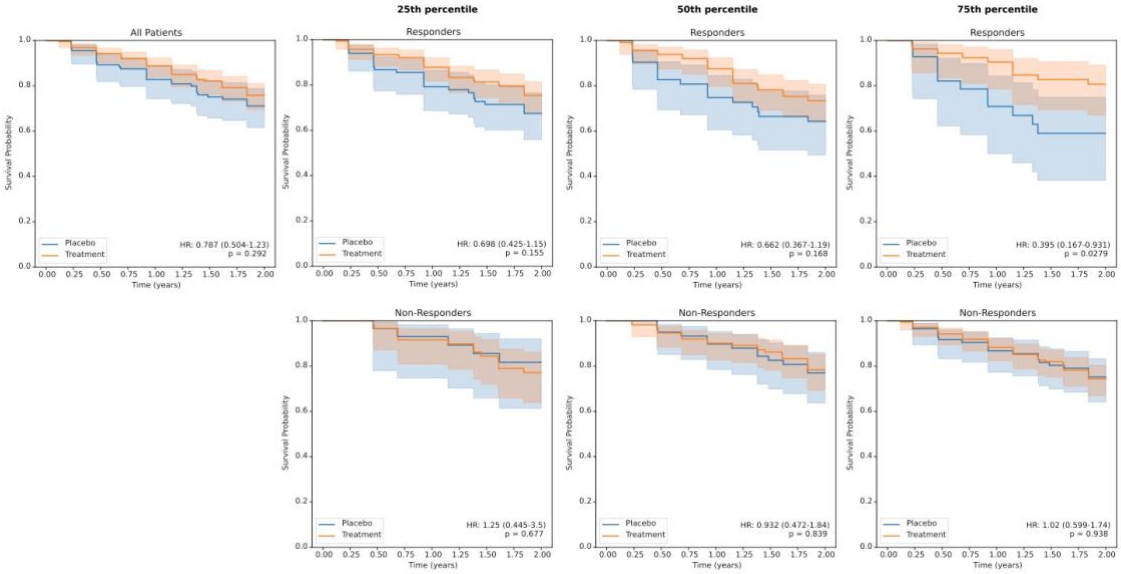

Figure 3 Kaplan-Meyer curves for predicted responders to anti-CD20-Abs at different percentile thresholds for response. Survival probability is measured in terms of time-to-CDP24. Censorship times are clamped at 2 years. P-values are calculated using log-rank tests.