Using deep learning for predictive enrichment of clinical trials in progressive multiple sclerosis

Jean-Pierre René Falet

Integrated Program in Neuroscience

McGill University, Montreal

February, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Several treatments for the acute inflammatory manifestations of multiple sclerosis (MS) were identified using the strategy of conducting a phase 2 study with an imaging-based biomarker outcome. In contrast, progress in identifying treatments that slow the progressive manifestations of MS has been hampered by the absence of suitable biomarkers. In this work, we hypothesize that a predictive enrichment strategy, where individuals predicted to be more responsive to a treatment are preferentially randomized into a clinical trial, can circumvent this problem by increasing a trial's statistical power. We propose an artificial neural network for estimating the conditional average treatment effect (CATE) on disability progression, taking as input an individual's pre-treatment clinical and imaging characteristics. We trained and validated the model on a pooled dataset from six randomized clinical trials ($n = 3830$), revealing large increases in statistical power that could render short, proof-of-concept clinical trials feasible. More responsive individuals tended to be younger, with a shorter disease duration, higher disability scores, and more lesion activity on magnetic resonance imaging (MRI) of the brain. Additional experiments showed that a model trained to estimate CATE for one drug can generalize to a drug from a different class, and that our model was superior to several alternative approaches. Altogether, our proposed enrichment strategy could facilitate progress in identifying treatments for disability progression in MS.

Résumé

Plusieurs traitements efficaces contre les manifestations inflammatoires aiguës de la sclérose en plaques (SEP) ont été identifiés en utilisant une stratégie consistant à mener une étude de phase 2 avec une mesure de l'efficacité du traitement basée sur un biomarqueur radiologique. En revanche, les progrès dans l'identification de traitements qui ralentissent les manifestations progressives de la SEP ont été entravés par l'absence de biomarqueurs appropriés. Dans ce travail, nous émettons l'hypothèse qu'une stratégie d'enrichissement prédictif, où les individus prédits comme répondant mieux à un traitement sont préférentiellement randomisés dans un essai clinique, peut contourner ce problème en augmentant la puissance statistique d'un essai. Nous proposons un réseau neuronal artificiel pour estimer l'effet moyen conditionnel du traitement (CATE) sur la progression de l'invalidité, en utilisant les caractéristiques cliniques et d'imagerie d'un individu enregistré avant le début d'un traitement. Nous avons entraîné et validé ce modèle sur des données comportant six essais cliniques randomisés ($n = 3830$), démontrant de fortes augmentations de la puissance statistique qui pourraient rendre réalisables des essais cliniques courts pour des preuves de concept. Les répondants étaient plus jeunes, avec une durée de la maladie plus courte, des scores d'invalidités plus élevés et plus d'activité lésionnelle. Des expériences supplémentaires ont montré qu'un modèle entraîné pour estimer le CATE d'un médicament peut généraliser à un médicament d'une classe différente, et que notre modèle était supérieur à plusieurs approches alternatives. Dans l'ensemble, notre stratégie d'enrichissement pourrait faciliter les progrès dans l'identification des traitements contre la progression de l'invalidité en SEP.

Acknowledgments

I would first like to thank my supervisors, Dr. Douglas Lorne Arnold and Dr. Tal Arbel, for their tremendous guidance and support. Dr. Doina Precup and Dr. Yasser Iturria-Medina, as members of my advisory committee, and Dr. Boris Bernhardt, as my Integrated Program in Neuroscience mentor, have also been instrumental in providing feedback on my work and in navigating the various milestones during graduate studies. In addition, I am grateful for the valuable input of our collaborators Dr. Behrooz Mahasseni, Dr. Maria-Pia Sormani, and Dr. Francesca Bovis.

I thank my colleagues Brennan Nichyporuk and Justin Szeto for assisting me in setting up my computational pipeline and providing much needed advice throughout my studies. I am also greatful for the fruitful collaborations with my colleagues Joshua Durso-Finley, Amar Kumar, and Anjun Hu. I also thank Zografos Caramanos and Istvan Morocz for managing the database used in this work.

I am thankful for the support of the Integrated Program in Neuroscience, in particular in granting me access to invaluable coursework in mathematics and computer science. Moreover, the resources made available for this project at the Montreal Neurological Institute and at the Centre for Intelligent Machines were very much appreciated.

I thank the companies who generously provided the clinical trial data that made this work possible: Biogen, BioMS, MedDay, Novartis, Roche / Genentech, and Teva. My graduate studies were also supported by an endMS Personnel Award from the Multiple Sclerosis Society of Canada, a Canada Graduate Scholarship-Masters Award from the Canadian Institutes of Health Research, and the Fonds de recherche du Québec - 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.

Preface and Contribution of Authors

This thesis is presented in a manuscript-based format. The work described was performed under the co-supervision of Dr. Tal Arbel and Dr. Douglas Lorne Arnold. This thesis contains three chapters. Chapter 1 is an introduction and review of relevant literature. Chapter 2 is a journal article published in *Nature Communications*, with its supplementary information included in Appendix A. Chapter 3 includes a general discussion and concluding remarks.

Contribution of Authors

The contribution of authors for each chapter is as follows:

- Chapter 1: I conducted the literature review and wrote the introduction independently. Tal Arbel and Douglas Lorne Arnold provided feedback and edits.

- Chapter 2 and Appendix A (published in Falet, J.-P. R. *et al.* Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nature Communications* **13,** 5645 (1 2022)): Douglas Lorne Arnold oversaw data collection, and I contributed to quality-checks. I pre-processed the data independently. The problem statement was written by myself and Douglas Lorne Arnold. I conducted the review of literature, and conceived of the hypothesis and objectives. I implemented all the methods, with advice on hyperparamater tuning and architecture design choices provided by Joshua Durso-Finley, Brennan Nichyporuk and Julien Schroeter. Brennan Nichyporuk additionally reviewed my code to ensure reproduceability. I conducted all experiments, with suggestions for

baseline models in Table 2.3 provided by Francesca Bovis, Maria-Pia Sormani, and Doina Precup. Francesca Bovis, Maria-Pia Sormani additionally provided advice on the methodology used for estimation of sample size in Table 2.4. I wrote the entire manuscript and produced all figures and tables. Feedback and edits were provided by all the aforementioned authors, including the co-supervisors of this work, Tal Arbel and Douglas Lorne Arnold.

- Chapter 3: I wrote the discussion and conclusions independently. Tal Arbel and Douglas Lorne Arnold provided feedback and edits. The discussion includes references to four other published papers I co-authored during my master's, but that are not part of this thesis. My contributions for these are as follows:

  – Durso-Finley, J. *et al.* Personalized Prediction of Future Lesion Activity and Treatment Effect in Multiple Sclerosis from Baseline MRI. *Medical Imaging with Deep Learning, PMLR* **172** (2022): I provided advice for clinical data pre-processing, and contributed significantly to the methods (which are partly based on the methods in Chapter 2), to the experimental setup and analysis, and to manuscript writing.

  – Kumar, A. *et al.* Counterfactual Image Synthesis for Discovery of Personalized Predictive Image Markers. *Medical Imaging Computing and Computer Assisted Intervention Society, Workshop on Medical Image Assisted Biomarkers Discovery (MIABID22), LNCS* **13602,** 113–124 (2022): I had minor contributions to the conception of the hypothesis and objectives, to the methods, experimental setup and analysis of results, and significant contributions to manuscript writing.

  – Hu, A. *et al.* Clinically Plausible Pathology-Anatomy Disentanglement in Patient Brain MRI with Structured Variational Priors. *NeurIPS 2022 Machine Learning for Health Workshop* (2022): I had minor contributions to the conception of the hypothesis and objectives, to the experimental setup and analysis

of results, and significant contributions to manuscript writing.

– Nichyporuk, B. *et al.* Rethinking Generalization: The Impact of Annotation Style on Medical Image Segmentation. *Machine Learning for Biomedical Imaging* **1** (December 2022 2022): I had minor contributions to the analysis of results and manuscript writing.

Contents

List of Figures

List of Tables

List of Abbreviations

9HPT 9-hole peg test 32, 33, 50

anti-CD20-Abs anti-CD20 monoclonal antibodies 11, 14–17, 21–23, 25, 28, 30, 47, 50, 53

ATE average treatment effect 3, 34, 56, 58

BBB blood brain barrier 1

CATE conditional average treatment effect 4–6, 9, 10, 16, 23, 24, 28, 34, 35, 47, 49, 52, 53, 56

CDP confirmed disability progression at 24 weeks 17, 33, 34, 57, 58

CDP confirmed disability progression 3, 33, 50, 57, 58

CPH Cox proportional hazards 13, 23, 28, 38, 39

CV cross-validation 36–38

DAWM diffusely abnormal white matter 7, 48

DL deep learning 3, 9–11, 47, 48, 52–54

DMT disease modifying therapy 2, 48

EDSS Expanded Disability Status Scale 15, 16, 29, 32–34, 50, 57, 58

FSS Functional Systems Scores 20, 22, 32

FWML focal white matter lesion 7

Gad gadolinium-enhancing 1, 2, 6, 25, 28, 30, 32–34, 47

GFAP glial fibrillary acidic protein 8

GM grey matter 7

IFNb-1a interferon beta-1a 14

ITE individual treatment effect 4, 5, 34, 35, 52, 55

LCLA low contrast letter acuity 50

ML machine learning 3, 4

MLP multilayer perceptron 14–16, 22–24, 28, 31, 34–36, 38, 49, 53, 56

MRI magnetic resonance imaging 1–3, 6–10, 12–14, 20, 25, 31, 32, 48, 51, 53, 54

MS multiple sclerosis 1, 2, 7–9, 47, 48, 53, 54

MSE mean squared error 35, 36, 38

MSFC multiple sclerosis functional composite 50

NAWM normal appearing white matter 7, 8

NBV normalized brain volume 22, 32, 33

NfL neurofilament light chain 7, 8

OCT optical coherence tomography 8

PASAT Paced Auditory Serial Addition Test 50

PET positron emission tomography 8

PIRA progression independent of relapse activity 1, 8, 9, 48

PMS progressive multiple sclerosis 1–3, 7, 9, 29, 30, 53

PPMS primary progressive multiple sclerosis 1, 14, 15, 20, 29, 30, 32, 36, 48

PRL paramagnetic rim lesion 6, 7

RMST restricted mean survival time 15, 17, 58

RRMS relapsing-remitting multiple sclerosis 1–3, 6–9, 12–15, 23, 24, 28–30, 32, 36, 48

SDMT Symbols Digit Modalities Test 50

SEL slowly expanding lesion 6, 7

SPMS secondary progressive multiple sclerosis 1, 6, 7, 29, 48

T25FW timed 25-foot walk 20, 22, 32, 33, 50

TSPO translocator protein 8

VAE variational autoencoder 49

Chapter 1

Introduction

Multiple sclerosis (MS) is a common inflammatory and neurodegenerative condition affecting the central nervous system. The most common subtype, relapsing-remitting MS (RRMS), is characterized by discrete episodes of focal inflammation, primarily causing demyelination, and to a lesser extent, axonal loss, visible on magnetic resonance imaging (MRI) of the brain or spinal cord as T2-hyperintense lesions [6]. These lesions are typically also gadolinium-enhancing (Gad) for the first two to eight weeks due to an associated breakdown of the blood brain barrier (BBB) [7]. Subsequently, T2 lesions can either slowly expand, remain static, decrease in size, and sometimes become unapparent on conventional MRI. About one in five to ten new lesions is associated with the onset of clinical symptoms or signs [8], defining a clinical relapse. Recovery from a relapse occurs over weeks, and can sometimes leave residual disability [9]. After 10-15 years, a large proportion of untreated RRMS patients transition to secondary progressive MS (SPMS), which is characterized by slow progression of disability independent of relapse activity [10].

Categorically distinct from the RRMS-SPMS spectrum, primary progressive MS (PPMS) affects around 10% of individuals [11]. Whereas SPMS follows a relapsing-remitting disease onset, PPMS is defined by onset with disability progression independent of relapse activity (PIRA) [12]. Because of their overlapping phenotypes, PPMS and SPMS are often collectively referred to as progressive MS (PMS).

It is now well recognized that the two main disease manifestations consisting of

episodic inflammatory activity (clinical relapses, new/enlarging T2 lesions, and Gad lesions) and slow disability progression, can co-occur in all MS subtypes, thus hinting at a pathophysiological continuum. Arguably, the most accepted hypothesis to date implicates common immune-mediated underpinnings for all subtypes that results in variable inflammatory and neurodegenerative processes, the latter of which is believed to translate clinically to disability progression [13]. Despite this overlap, peripherally administered immune therapies successful in suppressing the acute/subacute inflammatory manifestations of RRMS have largely been unsuccessful in slowing disability progression in PMS clinical trials [14–20]. Only two immune therapies, ocrelizumab [21] and siponimod [22], have demonstrated efficacy in slowing disability progression, and both have a modest effect.

Clearly, novel therapeutic targets are needed to better treat progression. However, there exists another, parallel path, to hasten development of therapeutics for progression. Ideally, we would want treatments that are effective for slowing disability progression, at least in a sub-group of individuals, to be easily identified in clinical trials. However, this task has so far been challenging. Solving this identification problem could rapidly improve access to disease modifying therapies (DMTs).

Historically, the strategy used in RRMS trials has been to perform relatively short and small phase 2 trials with an MRI biomarker as endpoint (such as suppression of new/enlarging T2 lesions and Gad lesions). These surrogate markers of activity are more sensitive to the underlying inflammatory process than the clinical event of a relapse, and therefore enable identification of efficacious medications with fewer patients in a shorter amount of time. This establishes proof-of-concept and finds the optimal dose, before proceeding to longer, more expensive phase 3 trials where clinical endpoints play a more important role. However, the absence of an accepted analogous MRI biomarker for PMS precludes using this strategy, and the clinical outcomes measuring disability progression occur too slowly to enable detection of a significant effect in a short clinical trial of one to two years. Novel solutions are therefore needed to speed up the drug-development process

for PMS.

In the absence of a sensitive biomarker for progression, one strategy to improve a clinical trial's ability to detect a significant treatment effect is to increase it's statistical power. To do so while keeping the trial duration and the sample size at a minimum, one can increase the effect size by identifying a sub-population that is expected to be more responsive to treatment. Indeed, it is often the case that medications are more effective in some individuals than others. Predicting who are the most responsive individuals and preferentially enrolling them in a clinical trial increases the expected effect size and therefore the power of a study. This method has been called *predictive enrichment* [23]. A drug proven to be efficacious in an enriched trial can later be tested in a larger and longer trial on a sub-group predicted to be less responsive. This step-wise approach prevents efficacious medications from having their effect diluted in early clinical trials due to inclusion of a population that is too heterogeneous, while still striving to provide access to a broad population.

Recently, Bovis *et al.* [24] used survival modeling to successfully predict a more responsive sub-group of RRMS patients to laquinimod, a medication whose average treatment effect (ATE) in the original phase 3 studies was insufficient for drug approval. Doing so in the PMS population remains an open problem. Machine learning (ML) provides ample strategies to tackle this task. Artificial neural networks are flexible architectures that can learn arbitrarily complex non-linear functions mapping input features to the outcome of interest [25], and therefore have a theoretical advantage over classical statistical models, particularly in the higher-dimensional setting. Deep learning (DL), characterized by deeper networks with more hidden layers, has already been used to accurately predict future confirmed disability progression (CDP) using baseline brain MRI [26].

1.1 Review of Relevant Literature

1.1.1 Treatment Effect Estimation Using Machine Learning

To enrich clinical trials with more treatment-responsive individuals, the machine learning task is best framed as a causal inference problem. Two of the most influential frameworks for causal inference include Pearl's structural causal model [27] and Newman-Rubin's potential outcome framework [28]. In this work, we will use the latter. The causal estimand of interest is the individual treatment effect (ITE), defined as the difference between a person's outcome (in this case disability progression) on treatment and their outcome on placebo. Because an individual is only given one of the two treatments, one of these potential outcomes is unobserved, which makes the ITE unobservable. This is known as the *Fundamental Problem of Causal Inference* [29].

Traditionally, ML has focused on modeling the relationship between an observation, such as a disease outcome, and input features such as individual-level characteristics. Adapting machine learning methods to causal inference is therefore non-trivial, because the ITE is unobservable. As a result, most of the recent work on personalized treatment effect estimation in machine learning has focused on a related causal estimand, the conditional average treatment effect (CATE).

CATE is defined as the expected treatment effect of a group of individuals defined by specific features. Given an individual characterized by a set of features, one can therefore estimate the expected effect for people with the same features and use this as a personalized (but population-based) estimate for the individual's treatment effect. Because CATE is identifiable from observed data under certain assumptions, several machine learning frameworks have emerged for CATE estimation. Two key assumptions are needed for identifiability of CATE: *unconfoundedness* (the potential outcomes are independent of treatment given the observed covariates), and *positivity*, or *overlap* (the probability of being assigned either treatment is non-zero given all possible co-variates

[28]. These assumptions hold in the randomized controlled setting examined in this work.

It is important to note that CATE is not equivalent to ITE, even though they are generally correlated [30]. Estimating ITE is in some ways more challenging, and requires counterfactual logic [27] that lies beyond the scope of this thesis. Moreover, estimating CATE is appropriate here, given that the estimator will be evaluated at the group-level in its ability to identify a more responsive group of individuals for randomization in clinical trials.

The uplift modeling literature, born out of the need to target customers most likely to respond to marketing interventions, has contributed numerous machine learning approaches for CATE estimation. A recent survey of uplift modeling classifies approaches into three types [31]. The first, called the *Two Model* approach, learns separate models for the outcome on treatment and control by training one model on the sample that received treatment and the other on the sample that received control. The *Two Model* CATE estimator is then the difference of the predicted outcome on treatment (using the treatment model) minus the predicted outcome on control (using the control model). Several extensions exist to reduce model bias and/or variance (e.g. [32]) and to correct for confounding (e.g. [33]). A second approach, the *Class Transformation* approach, exploits the fact that even the single observed outcome narrows the space of possible effect sizes (e.g. an individual observed to have the best possible outcome on no medication cannot have a positive effect from a treatment). The observed outcome can therefore be used to compute a "transformed outcome" that, in expectation, and under certain assumptions, equals CATE. This method has primarily been used with binary outcomes, but was extended to continuous outcomes [34]. The third approach maximizes the heterogeneity of treatment effects between sub-groups of a population, and has been primarily applied to decision trees. For example, causal trees can learn to assign individuals to different leaves in order to maximize the heterogeneity of CATE estimates between leaves ([35]). All three approaches have been widely used, but the *Two Model Approach* is arguably the most common due to its simplicity and flexibility.

1.1.2 Predictors of Future Disability Progression

Despite the absence of an accepted biomarker endpoint for clinical trials, several biomarkers that can be measured at a baseline (pre-treatment) visit have been shown to predict future disability progression with variable accuracy [36]. These could potentially be used as part of a CATE estimator to isolate a group more likely to respond to an investigational treatment.

Traditional biomarkers of inflammatory activity used in RRMS trials are easily measurable, and remain valuable for predicting progression. Baseline T2 lesion burden [37–40], and the presence of lesions in particular anatomic locations such as the spinal cord and infratentorially [39, 41], have been found to correlate with future clinical disability and disability progression, at least modestly. Although clinical trials rely on brain imaging, the adoption of spinal cord imaging has lagged behind, in part due to technical and analytical difficulties. Lesions in the cortical grey matter have also been shown to be moderately predictive of future disability progression [42, 43]. As for Gad lesions, some authors [44] have found modest correlations with future disability at least 2 years from baseline, but others [39] have not.

More recently, biomarkers have emerged from the observation that SPMS is associated with a shift from the episodic translocation of peripheral immune cells into the CNS that is seen in RRMS, towards a form of chronic, compartmentalized inflammation in the CNS. One example is a sub-population of chronically active lesions associated with slow demyelination and axonal loss, called slowly expanding lesions (SELs). These can be identified by sequential T2/T1-weighted MRIs acquired over time [45]. Activated iron-laden microglia/macrophages are seen at the edge of around 40% of SELs [46], and can be seen on susceptibility-weighted MRI sequences as paramagnetic rim lesions (PRLs). Both SELs and PRLs have been associated with future disability progression [47], especially when present in combination [48]. Nevertheless, SELs and PRLs face several issues limiting their use in clinical practice, including the need for serial imaging for SEL detection,

absence of an accepted threshold for considering the number of lesions abnormal, and the general requirement for offline processing to detect both SELs and PRLs.

Biomarkers that are representative of neurodegeneration have also been studied. Brain volume has been consistently shown to correlate with future disability progression [49], which explains why it is more frequently used as a surrogate marker of diasability progression in clinical trials for PMS. However, it's ability to predict future disability progression remains modest at best. If specific structures are considered, the grey matter (GM) volume or fraction [50], and specifically thalamic volume [51], appear to be particularly predictive future disability progression [52]. Other studies examining microstructural alterations in GM according to diffusivity [53], and covariance patterns in the volume of different GM regions [54], have corroborated this association. Finally, subtle abnormalities in normal appearing white matter (NAWM) not visible on T2/FLAIR sequences can be detected by the magnetization transfer ratio, and this was shown to be predictive of future disability progression [55].

While difficult to classify due to an incomplete understanding of its biological underpinnings, an intermediate pathology between NAWM and focal white matter lesions (FWMLs), termed diffusely abnormal white matter (DAWM), has long been viewed as a marker of progression in MS [56]. Although Vertinsky *et al.* [57] did not find that baseline DAWM predicted future disability progression in an RRMS cohort, the volume of conversion of DAWM into FWMLs over time was shown to be associated with progression in both a RRMS and a SPMS cohort [58]. More work is needed to clarify the role of DAWM, particularly at baseline, in predicting future disability progression.

Other non-MRI biomarkers are worth mentioning. Neurofilament light chain (NfL), a marker of neuronal injury released into the cerebrospinal fluid and typically measured in the serum, has been shown to predict disability progression [59–64]. However, use of NfL as a biomarker in clinical trials has been hampered by short-term fluctuations in titers, and its association with age, comorbidities and treatment [65]. Moreover, other studies have not found an association with disability progression [66–69]. This discrepancy is

hypothesized to be due to NfL being more associated with acute inflammatory activity, and therefore only correlated with future disability progression when it is the result of such activity, as opposed to PIRA. Indeed, many of the studies showing a positive association studied RRMS patients at higher risk for acute inflammation, whereas the ones showing no association studied populations either on immune therapy or at higher risk for progressive disease [69]. One of the studies reaching the latter conclusion compared NfL to another biomarker of neuronal injury, serum glial fibrillary acidic protein (GFAP) [69]. Contrary to NfL, they showed that GFAP was correlated with disability progression, particularly in the subset of patients at low risk for acute inflammatory activity. More work is therefore warranted to evaluate the predictive role of GFAP. Optical coherence tomography (OCT) has also been used to predict future disability progression using retinal thickness [70, 71]. Unfortunately, OCT suffers from several confounders and technical difficulties that have so far limited its practical utility. Finally, positron emission tomography (PET) has shown promise using radioligands for translocator protein (TSPO), a marker of activated microglia/macrophages. Increased uptake of TSPO in NAWM was found to be associated with future disability progression [72]. However, these tracers are not specific to MS pathology and there are risks associated with exposing patients to repeated radiation for the purpose of disease monitoring.

Few studies have looked at the predictive value on disability progression of a large number of clinical variables in aggregate. This includes age, sex, ethnicity, education, past medical history and disability scores, with or without the inclusion of scalar MRI-derived metrics. For example, Pellegrini *et al.* [73] showed, using classical machine learning, that the predictive value of a range of clinical predictors was limited, with a top C-index achieved of 0.65. Nonetheless, they did not evaluate more expressive models such as deep neural networks, and did not evaluate their models on treated cohorts. Stühler *et al.* [74] took a Bayesian approach to predictive modeling and achieved similar performance in terms of C-index, while using relatively few clinical/demographic variables.

It is difficult to extrapolate all these findings to the task of predictive enrichment

for clinical trials, since most aforementioned studies studied prediction of progression off medication or on placebo (whereas estimating CATE requires modeling prognosis on both treatment and control), reported different outcome metrics, studied different MS sub-types (with or without consideration for relapse-associated disability worsening, which differs from PIRA), and measured the strength of the association between their biomarker and an future disability over a variable time horizon (varying from 1 to 20 years). Many of these studies also have significant methodological limitations, including insufficient details about model specification, optimization and validation procedure, as well as improper use of certain evaluation metrics (such as measuring accuracy in the setting of class imbalance). Importantly, very few evaluated their model in an external validation cohort, and combined with very small training sets, there is a significant risk that their models could overfit the training data and would not generalize to new individuals.

Finally, there may be complex interactions between many of the previously mentionned biomarkers that together would be more predictive of future disability progression and treatment effect. Deep artificial neural networks, on top of their expressive power, learn a hidden representation of the input that is predictive of the outcome of interest, thus in many cases alleviating the need for feature selection. To the best of our knowledge, a DL framework for data-driven prediction of disability progression and CATE estimation using as input a multitude of common MRI-derived markers and clinical features has not yet been developed.

1.2 Rationale and Objectives

There has been significant progress in identifying medications to treat the acute inflammatory manifestations characteristic of RRMS. However, drug development targeted at slowing disabliity progression, which is most prominent in PMS but also affects patients with RRMS, is stagnant. To circumvent the difficulty in identifying efficacious treatments for PMS, increasing the efficiency of clinical trials is paramount. The overarching goal

of this work is to develop a method for predictive enrichment of clinical trials using DL to increase the statistical power of short, proof-of-concept clinical trials, thus improving the chance that efficacious treatments are identified early in the development process and made accessible to patients.

This primary objective is addressed in Chapter 2, where we present a *Two Model* CATE estimator parametrized by a deep neural network. This model predicts the treatment effect on disability progression for an individual given readily available baseline clinical information and scalar MRI metrics (i.e. lesional and volumetric) obtained at a pre-treatment baseline visit. Its utility for predictive enrichment is explored through a sample size estimation experiment by using its predictions to rank individuals in terms of predicted effect.

Secondary objectives of this work, also addressed in Chapter 2, included gaining insight into the predictors of treatment effect by examining differences in baseline features between more responsive and less responsive individuals, estimating the generalization error for two treatments with different mechanisms of action, and benchmarking the proposed DL model against other, less expressive models.

Chapter 2

Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning

---

Published in Falet, J.-P. R. *et al.* Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nature Communications* **13,** 5645 (1 2022)

This manuscript addresses the primary objective of this thesis by setting out to estimate the treatment effect on disability progression and determining the impact on statistical power when using such an estimator for predictive enrichment. This work also addresses the secondary objectives by providing a comparative analysis of baseline features differentiating more responsive from less responsive individuals, finds that a model trained on one type of medication (anti-CD20 monoclonal antibodies (anti-CD20-Abs)) can generalize to another medication with a different mechanism of action (laquinimod), and shows that the proposed DL model outperformes less expressive models or models that consider fewer input features. Supplementary information published with this manuscript is included in Appendix A.

Abstract

Disability progression in multiple sclerosis remains resistant to treatment. The absence of a suitable biomarker to allow for phase 2 clinical trials presents a high barrier for drug development. We propose to enable short proof-of-concept trials by increasing statistical

power using a deep-learning predictive enrichment strategy. Specifically, a multi-headed multilayer perceptron is used to estimate the conditional average treatment effect (CATE) using baseline clinical and imaging features, and patients predicted to be most responsive are preferentially randomized into a trial. Leveraging data from six randomized clinical trials ($n = 3,830$), we first pre-trained the model on the subset of relapsing-remitting MS patients ($n = 2,520$), then fine-tuned it on a subset of primary progressive MS (PPMS) patients ($n = 695$). In a separate held-out test set of PPMS patients randomized to anti-CD20 antibodies or placebo ($n = 297$), the average treatment effect was larger for the 50% (HR, 0.492; 95% CI, 0.266-0.912; $p = 0.0218$) and 30% (HR, 0.361; 95% CI, 0.165-0.79; $p = 0.008$) predicted to be most responsive, compared to 0.743 (95% CI, 0.482-1.15; $p = 0.179$) for the entire group. The same model could also identify responders to laquinimod in another held-out test set of PPMS patients ($n = 318$). Finally, we show that using this model for predictive enrichment results in important increases in power.

2.1 Introduction

Several disease modifying therapies have been developed for the treatment of the focal inflammatory manifestations of RRMS (clinical relapses and lesion activity) using the strategy of performing relatively short and small phase 2 trials with a MRI endpoint. These were meant to establish proof-of-concept and find the optimal dose, before proceeding to longer, more expensive phase 3 trials. In contrast to focal inflammatory manifestations, the absence of analogous MRI endpoints for disability progression independent of relapses has hampered progress in developing drugs for this aspect of the disease. Progressive biology predominates in progressive forms of multiple sclerosis, but is increasingly appreciated to be important in RRMS [1]. Although brain atrophy has been used as a biomarker of progression in phase 2 trials of progressive disease, its ability to predict the effect on disability progression in subsequent phase 3 clinical trials remains uncertain. As proceeding directly to large, phase 3 trials is expensive and risky, most programs that

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* [2]. A drug proven to be efficacious in a trial enriched with predicted responders can later be tested more confidently in a population predicted to be less responsive. This sequence prevents efficacious medications from having their effect diluted in early clinical trials due to inclusion of a population that is too heterogeneous, while still allowing for broadening of indication criteria. It also improves the balance of risks and benefits for participants, since those who are unlikely to benefit from a drug would not be exposed to it and therefore would not experience potential adverse effects. A relevant application of predictive enrichment was described by Bovis *et al.* [3], who used Cox proportional hazards (CPH) models to successfully 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.

Deep learning is a highly expressive and flexible type of machine learning that can potentially uncover complex, non-linear relationships between baseline patient characteristics and their responsiveness to treatment. However, contrary to traditional machine learning problems where a mapping between features and targets is learned from a sample of observations, the target in a treatment response (or treatment effect) task is not directly observable. Adaptations to machine learning frameworks must therefore be made in order to frame the problem through the lens of causal inference (reviewed in detail in the survey on uplift modeling by Gutierrez & Gérardy [4]). Arguably some of the most popular methods have been tree-based approaches [5] which model treatment effect directly, and meta-learning approaches [6] which decompose the treatment effect estimation problem into simpler problems that can be tackled using traditional machine learning models. In a recent paper, Durso-Finley *et al.* [7] presented a meta-learning approach for the estimation of treatment effect (as measured by suppression of new/enlarging T2-lesions) in RRMS using baseline brain MRI and clinical variables.

In this work, we present a new deep learning framework to estimate an individual's treatment effect using readily available clinical information (demographic characteristics and clinical disability scores) and scalar MRI metrics (lesional and volumetric) obtained at the screening visit of a clinical trial. This approach, based on an ensemble of multi-headed multilayer perceptrons (MLPs), can identify more responsive individuals to both anti-CD20-Abs and laquinimod better than alternative strategies. We demonstrate how using this model for predictive enrichment could greatly improve the feasibility of short proof-of-concept trials studying the effect of novel treatments for progression, thus accelerating therapeutic advances.

2.2 Results

2.2.1 Datasets

Data were pooled from six randomized clinical trials ($n = 3,830$): OPERA I [8], OPERA II [8], BRAVO [9], ORATORIO [10], OLYMPUS [11], and ARPEGGIO [12] (ClinicalTrials.gov numbers, NCT01247324, NCT01412333, NCT00605215, NCT01194570, NCT00087529, NCT02284568, respectively). OPERA I/II, and BRAVO were RRMS trials which compared ocrelizumab with subcutaneous interferon beta-1a (IFNb-1a), and laquinimod with both intramuscular IFNb-1a and placebo, respectively. ORATORIO, OLYMPUS, and ARPEGGIO were placebo-controlled PPMS trials which studied ocrelizumab, rituximab, and laquinimod, respectively.

The dataset is divided into three subsets for different phases of training and evaluation. The first subset ($n = 2,520$) contains data from the three RRMS trials, and is used for pre-training the MLP to learn predictors of treatment effect under the RRMS condition (for details, see Section 2.4, Methods). This pre-training phase falls under the umbrella of transfer learning, a deep learning strategy that is used to transfer knowledge acquired from a related task to a task with fewer samples in order to improve learning on the latter

[13]. Importantly, the RRMS dataset is only used for pre-training and does not take part in final model evaluation, since this study is focused on the challenge of improving the efficiency of clinical trials for progressive MS. The second subset consists of two PPMS trials ($n = 992$): OLYMPUS and ORATORIO. This subset is divided into a 70% training set ($n = 695$) which is used to fine-tune the pre-trained MLP to estimate treatment effect to anti-CD20-Abs, and the remaining 30% ($n = 297$) is held out as a test set to estimate the generalization error of the fully trained model. The third subset contains PPMS data from the trial ARPEGGIO ($n = 318$), which is also held out as a second test set.

Mean and standard deviation for the baseline features and the outcome metrics in the PPMS subset are shown in Table 2.1, separated by treatment arm (the same statistics for the RRMS subset are shown in Supplementary Table A.1). The groups are comparable for all features except for disease duration which is shorter in ORATORIO, and Gad count and T2 lesion volume, which are greater in ORATORIO. This may be due to ORATORIO's inclusion criteria, which had a maximum time from symptom onset, and to inter-trial differences in automatic lesion segmentation, which are accounting for using a scaling procedure explained in Section 2.4.1. Some heterogeneity exists between the outcomes of each trial when looking at the placebo arms, which on average have a smaller restricted mean survival time (RMST) at 2 years in ARPEGGIO and OLYMPUS compared to ORATORIO, indicating more rapid disability progression on the Expanded Disability Status Scale (EDSS).

Table 2.1: Baseline features and outcomes per treatment arm.

	Ocrelizumab ORATORIO $n = 436$	**Rituximab** OLYMPUS $n = 212$	**Laquinimod** ARPEGGIO $n = 186$	**Placebo** ORATORIO $n = 225$	OLYMPUS $n = 119$	ARPEGGIO $n = 132$
Demographics:						
Age (years)	44.50 (7.90)	49.54 (9.01)	46.35 (6.62)	44.41 (8.40)	49.89 (8.68)	46.70 (7.16)
Sex (% male)	51.61	48.11	56.45	47.56	43.70	50.76
Height (cm)	170.20 (9.61)	170.77 (9.30)	172.11 (9.41)	170.20 (9.57)	169.87 (8.90)	171.23 (9.73)
Weight (kg)	72.35 (17.26)	78.13 (16.37)	75.25 (15.40)	72.51 (15.24)	77.60 (17.13)	73.20 (16.21)
Disease duration (years)	6.56 (3.77)	9.03 (6.25)	8.12 (6.07)	6.01 (3.38)	8.59 (6.81)	7.41 (5.23)
Disability Scores:						
EDSS	4.69 (1.18)	4.79 (1.36)	4.49 (0.98)	4.65 (1.16)	4.58 (1.41)	4.46 (0.91)
FSS-Bowel and Bladder	1.14 (0.85)	1.42 (0.95)	1.27 (0.95)	1.14 (0.91)	1.21 (0.94)	1.16 (0.88)
FSS-Brainstem	0.88 (0.91)	0.75 (0.90)	1.01 (0.92)	0.89 (0.93)	0.61 (0.81)	0.98 (0.95)
FSS-Cerebellar	2.11 (0.98)	2.03 (1.12)	2.11 (0.83)	2.14 (0.89)	1.99 (1.10)	2.10 (0.89)
FSS-Cerebral	0.91 (0.88)	1.30 (0.84)	0.93 (0.91)	0.91 (0.82)	1.24 (0.89)	0.86 (0.88)
FSS-Pyramidal	2.87 (0.62)	2.69 (0.82)	2.92 (0.55)	2.83 (0.65)	2.82 (0.78)	2.85 (0.66)
FSS-Sensory	1.58 (1.04)	1.48 (0.99)	1.73 (1.04)	1.53 (1.07)	1.52 (1.11)	1.74 (1.01)
FSS-Visual	0.79 (0.87)	0.86 (1.04)	0.92 (1.30)	0.71 (0.82)	0.91 (1.05)	0.79 (1.10)
Mean T25FW (sec)	13.93 (18.44)	11.74 (14.56)	9.61 (8.85)	11.71 (12.35)	11.01 (13.65)	9.68 (7.54)
Mean 9HPT dominant (sec)	34.09 (33.99)	28.80 (17.60)	28.57 (12.37)	31.67 (21.50)	27.22 (10.22)	28.22 (12.15)
Mean 9HPT non-dominant (sec)	36.05 (38.50)	31.88 (24.99)	31.44 (18.04)	37.51 (40.29)	30.95 (17.50)	29.04 (12.16)
MRI metrics:						
Gad count	1.23 (5.36)	0.63 (2.47)	0.27 (0.81)	0.56 (1.47)	0.47 (1.14)	0.45 (1.84)
T2 lesion volume (mL)	12.45 (14.92)	8.44 (10.50)	5.86 (9.11)	11.33 (13.27)	8.57 (11.66)	5.96 (8.65)
Normalized brain volume (L)	1.46 (0.08)	1.20 (0.12)	1.46 (0.10)	1.47 (0.09)	1.21 (0.12)	1.46 (0.11)
Outcome:						
Slope (EDSS change / yr)*	0.22 (0.53)	0.27 (0.65)	0.32 (0.77)	0.27 (0.71)	0.39 (0.63)	0.28 (0.64)
RMST (at 2 years)†	1.92	1.89	1.69	1.91	1.87	1.72

Values in brackets are standard deviations, unless otherwise specified.
* Slope is based on the coefficient of regression from a linear regression model that is fit on an individual's EDSS values over time, as described in Section 2.4.2.
† RMST calculated at 2 years using time to 24-week confirmed disability progression on the EDSS.
RMST=Restricted mean survival time; EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion.

2.2.2 Predicting response to anti-CD20 monoclonal antibodies

As described in Section 2.4 (Methods), we train an ensemble of multi-headed MLPs to predict the change in EDSS over time (obtained by fitting a linear regression model to an individual's EDSS values recorded over time and taking the slope of the regression to be the prediction target) on both anti-CD20-Abs and placebo. These two predictions are then subtracted to obtain an estimate of the CATE for each individual, given their baseline features. The CATE estimate is used to infer an individual's treatment effect, as explained in Section 4.3.

The fully trained model is then evaluated on the held-out anti-CD20-Abs test set (30% of the dataset; $n = 297$). A histogram of predictions on this test set is shown in Supplementary Fig. A.1. The model's ability to rank response is assessed using an average difference curve, AD(c), which is described by Zhao *et al.* [14] and is well suited for measuring performance in predictive enrichment. Our implementation measures the ground-truth average difference in RMST (calculated at 2 years from time to CDP at 24 weeks (CDP)) between anti-CD20-Abs and placebo for individuals predicted to respond more than a certain threshold, as a function of this threshold. The AD(c) curve for our model, shown in Fig. 2.1, appropriately increases as a sub-group that is predicted to be more and more responsive is selected. The AD_{wabc}, a metric derived from the area under the AD(c) curve in Supplementary Methods A.3, provides a measure of how well the model can rank individuals on the basis of their responsiveness to treatment. Larger positive AD_{wabc} values indicate better performance. The AD_{wabc} in this case is positive, relatively large (0.0565), and nearly monotonic (Spearman r correlation coefficient 0.943), demonstrating the ability for the model to rank response to anti-CD20-Abs.

Kaplan-Meyer curves of the ground-truth time-to-CDP for predicted responders in the test set are shown in Fig. 2.2 for two predictive enrichment thresholds (selecting the 50% or the 30% that are predicted to be most responsive). The Kaplan-Meyer curves for corresponding non-responder groups (the 50% and 70% predicted to be least responsive) are also shown. Compared to the entire test set, whose HR is 0.743 (95% CI, 0.482-1.15; $p = 0.179$), predictive enrichment leads to a HR of 0.492 (95% CI, 0.266-0.912; $p = 0.0218$) and 0.361 (95% CI, 0.165-0.79; $p = 0.008$) when selecting the 50% and 30% most responsive, respectively. The corresponding non-responder groups have a HR of 1.11 (95% CI, 0.599-2.05; $p = 0.744$) and 0.976 (95% CI, 0.578-1.65; $p = 0.925$) when selecting the 50% and 70% least responsive, respectively. This heterogeneity suggests that a significant part of the trend for an effect at the whole-group level may be explained by a small proportion of more responsive patients.

Of ocrelizumab and rituximab, only the former had a significant effect in a phase 3 trial

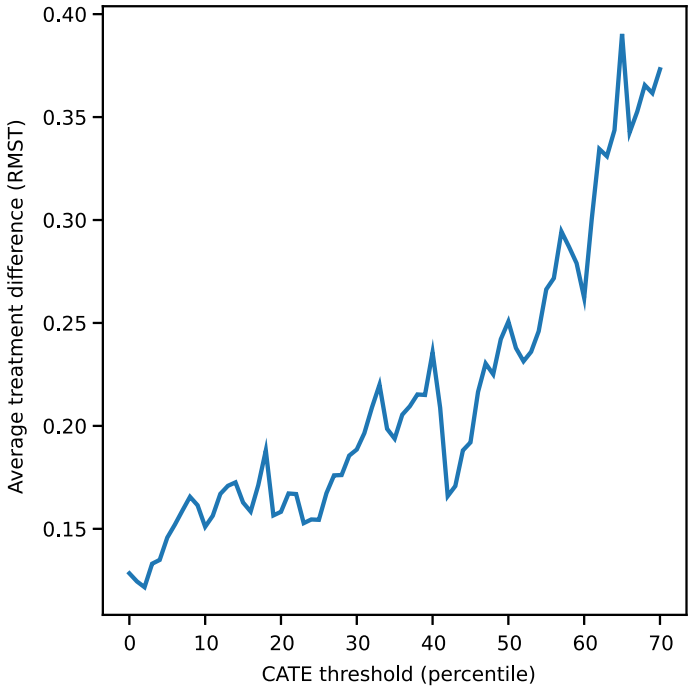


Figure 2.1: Average treatment difference curve for the anti-CD20-Abs held-out test set. Represents the difference in the ground-truth restricted mean survival time (RMST), calculated at 2 years using time-to-CDP24, between anti-CD20-Abs and placebo, among predicted responders defined using various thresholds. The conditional average treatment effect (CATE) percentile threshold is the minimum CATE (expressed as a percentile among all CATE estimates in the test set) that is used to define an individual as a responder (i.e. a threshold of 0.7 means the 30% predicted to be most responsive are considered responders)

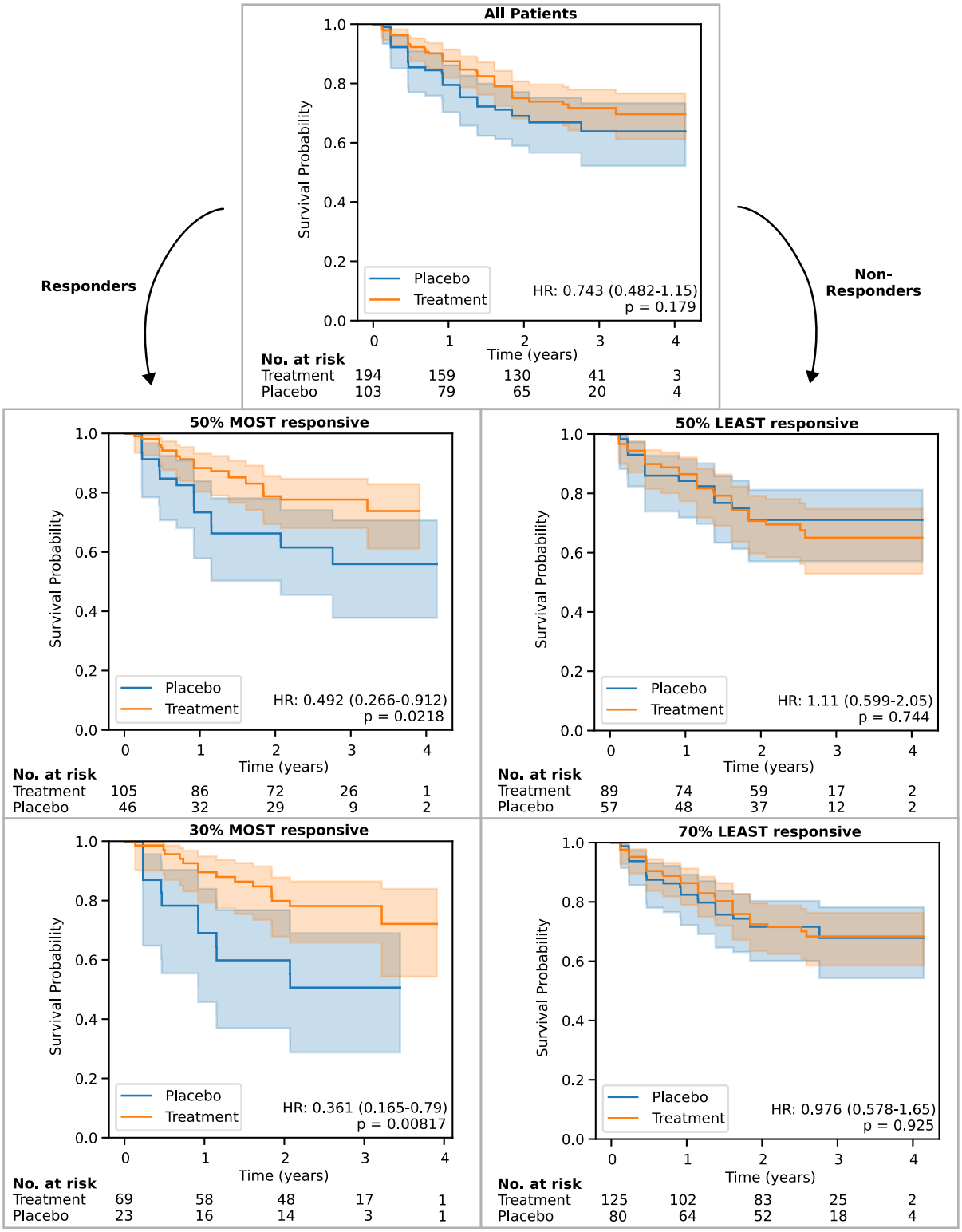


Figure 2.2: Kaplan-Meyer curves +/- 95% confidence intervals (CI) for predicted responders and non-responders to anti-CD20-Abs in the held-out test set, defined at two thresholds of predicted effect size. These are compared to the whole group (top). The placebo group is displayed in blue, and the treatment (anti-CD20-Abs) group is displayed in orange. Survival probability is measured in terms of time-to-CDP24 using the EDSS. p values are calculated using log-rank tests. 95% CIs are estimated using Greenwood's Exponential formula.

(ORATORIO), and it is the only drug approved in PPMS. We therefore verified whether the model's enrichment capabilities are maintained within the ORATORIO subgroup ($n = 188$) of the test set, which has HR of 0.661 (95% CI 0.383-1.14, $p = 0.135$). If selecting the 50% ($n = 96$) and 30% ($n = 57$) predicted to be most responsive, the HR reduces to 0.516 (95% CI, 0.241-1.1; $p = 0.084$) and 0.282 (95% CI, 0.105-0.762; $p = 0.0082$), respectively. The corresponding 50% and 70% predicted to be least responsive have a HR of 0.849 (95% CI, 0.385-1.87; $p = 0.685$) and 0.915 (95% CI, 0.471-1.78; $p = 0.791$), respectively.

We then considered specific demographic subgroups to understand their effect on model performance. For men, the model achieved a AD_{wabc} of 0.0405, while for women the model performs better ($AD_{wabc} = 0.0844$). For those with an age < 51, the AD_{wabc} of 0.0353 is lower than for those with an age $>= 51$ ($AD_{wabc} = 0.0661$). For those with a disease duration < 5, the model performs less well than on those with a disease duration $>=5$ ($AD_{wabc} = 0.0385$ compared to 0.0117). Finally, the model performs better for those with an EDSS < 4.5 ($AD_{wabc} = 0.069$) than for those with an EDSS of $>= 4.5$ ($AD_{wabc} = 0.0451$).

Group characteristics for the predicted responders and non-responders, defined at the 50th and 70th percentile thresholds, are shown in Table 2.2. We observe enrichment across a broad range of input features in the responder sub-groups: younger age, shorter disease duration, higher disability scores, and more lesional activity (particularly T2 lesion volume). The largest effect on the Functional Systems Scores (FSS) was seen in Cerebellar and Visual sub-scores, while FSS-Bowel and Bladder, Brainstem, Cerebral, Pyramidal, and Sensory did not reach statistical significance ($p < 0.05$). Timed 25-foot walk (T25FW) was significantly different only for the 70th percentile threshold. Normalized brain volume was the only baseline MRI feature which did not differ significantly between the two groups at either threshold.

Table 2.2: Group statistics for predicted responders and non-responders to anti-CD20-Abs at the 50th and 70th percentile thresholds, in the held-out test set.

	50th percentile threshold*				70th percentile threshold*			
	Responders	Non-responders	Effect size (95% CI)†	*p* value‡	Responders	Non-responders	Effect size (95% CI)†	*p* value‡
Trial contribution:								
OLYMPUS	55	54			35	74		
ORATORIO	96	92			57	131		
Demographics:								
Age (years)	45.20 (8.58)	47.84 (7.89)	-2.64 (-4.53, -0.76)	0.006	44.59 (9.05)	47.36 (7.87)	-2.77 (-4.93, -0.61)	0.013
Sex (% male)	47.02	50.68	0.86 (0.53, 1.40)	0.562	45.65	50.24	0.83 (0.49, 1.40)	0.530
Height (cm)	170.05 (10.56)	170.55 (8.80)	-0.50 (-2.72, 1.71)	0.657	169.78 (10.29)	170.52 (9.47)	-0.74 (-3.23, 1.75)	0.560
Weight (kg)	76.17 (18.93)	72.96 (13.77)	3.21 (-0.56, 6.98)	0.096	75.68 (20.07)	74.10 (14.87)	1.58 (-3.04, 6.20)	0.502
Disease duration (years)	6.07 (4.14)	8.72 (5.45)	-2.65 (-3.76, -1.54)	<0.001	5.79 (4.15)	8.09 (5.19)	-2.30 (-3.41, -1.19)	<0.001
Disability Scores:								
EDSS	4.87 (1.18)	4.52 (1.23)	0.34 (0.07, 0.62)	0.015	5.07 (1.14)	4.53 (1.21)	0.54 (0.25, 0.83)	<0.001
FSS-Bowel and Bladder	1.25 (0.93)	1.11 (0.80)	0.14 (-0.05, 0.34)	0.157	1.27 (0.98)	1.15 (0.82)	0.12 (-0.11, 0.35)	0.315
FSS-Brainstem	0.82 (0.93)	0.79 (0.87)	0.04 (-0.17, 0.24)	0.726	0.90 (0.95)	0.77 (0.88)	0.13 (-0.10, 0.36)	0.265
FSS-Cerebellar	2.38 (0.97)	1.78 (1.05)	0.60 (0.37, 0.83)	<0.001	2.57 (0.81)	1.86 (1.08)	0.71 (0.48, 0.93)	<0.001
FSS-Cerebral	1.07 (0.83)	1.05 (0.89)	0.02 (-0.18, 0.22)	0.848	1.13 (0.84)	1.04 (0.87)	0.09 (-0.12, 0.30)	0.404
FSS-Pyramidal	2.75 (0.69)	2.90 (0.58)	-0.14 (-0.29, 0.00)	0.052	2.77 (0.76)	2.85 (0.58)	-0.08 (-0.26, 0.10)	0.382
FSS-Sensory	1.55 (1.06)	1.64 (1.02)	-0.08 (-0.32, 0.15)	0.488	1.56 (1.00)	1.61 (1.06)	-0.05 (-0.30, 0.20)	0.703
FSS-Visual	1.04 (1.04)	0.43 (0.62)	0.62 (0.42, 0.81)	<0.001	1.28 (1.07)	0.50 (0.71)	0.78 (0.54, 1.02)	<0.001
Mean T25FW (sec)	13.55 (17.61)	10.75 (11.08)	2.80 (-0.55, 6.15)	0.103	15.95 (21.79)	10.48 (9.82)	5.47 (0.77, 10.17)	0.024
Mean 9HPT dominant (sec)	32.62 (26.89)	26.70 (10.24)	5.92 (1.29, 10.55)	0.013	36.01 (33.25)	26.88 (9.89)	9.13 (2.12, 16.15)	0.012
Mean 9HPT non-dominant (sec)	37.33 (31.11)	26.97 (9.32)	10.36 (5.14, 15.58)	<0.001	42.39 (38.33)	27.68 (9.33)	14.71 (6.68, 22.75)	<0.001
MRI metrics:								
Gad count	1.62 (3.14)	0.16 (0.48)	1.46 (0.95, 1.97)	<0.001	1.90 (3.64)	0.46 (1.27)	1.44 (0.67, 2.22)	<0.001
T2 lesion volume (mL)	13.09 (12.85)	7.72 (10.17)	5.37 (2.73, 8.01)	<0.001	14.31 (14.22)	8.72 (10.27)	5.59 (2.33, 8.85)	<0.001
Normalized brain volume (L)	1.37 (0.16)	1.38 (0.16)	-0.02 (-0.05, 0.02)	0.367	1.35 (0.16)	1.38 (0.16)	-0.03 (-0.07, 0.01)	0.107

Values in brackets are standard deviations, unless otherwise specified.
*Percentile threshold for defining responders. The 50th percentile defines responders as the top 50% who are predicted to be most responsive, while the 70th percentile defines them as the top 30%. The non-responders are those who fall below the percentile threshold.
†Effect size is the average difference between responders and non-responders for all covariates except for "sex" which is an odd's ratio (OR).
‡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. Exact p-values for the 50th percentile threshold: Disease duration, $p = 4.39 \times 10^{-6}$; FSS-Cerebellar, $p = 6.42 \times 10^{-7}$; FSS-Visual, $p = 2.18 \times 10^{-9}$; Mean 9HPT non-dominant, $p = 1.36 \times 10^{-4}$; Gad count, $p = 8.72 \times 10^{-}8$; T2 lesion volume, $p = 8.57 \times 10^{-5}$. Exact p-values for the 70th percentile threshold: Disease duration, $p = 7.04 \times 10^{-5}$; EDSS, $p = 3.03 \times 10^{-4}$; FSS-Cerebellar, $p = 2.61 \times 10^{-9}$; FSS-Visual, $p = 3.38 \times 10^{-9}$; Mean 9HPT non-dominant, $p = 4.82 \times 10^{-4}$; Gad count, $p = 3.69 \times 10^{-4}$; T2 lesion volume, $p = 9.59 \times 10^{-4}$.
EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion.

2.2.3 Predicting response to laquinimod

To determine whether the same model trained on the anti-CD20-Abs dataset could be predictive of treatment response to a medication with a different mechanism of action,

and to provide a second validation for the model trained on the single 70% training set in the first anti-CD20-Abs experiment, we tested it on data from ARPEGGIO ($n = 318$). The model trained on the anti-CD20-Abs training dataset also generalized to this second test set, as shown by a positive $AD_{wabc} = 0.0211$. From the whole-group HR of 0.667 (95% CI: 0.369-1.2; $p = 0.933$), selecting the 50% and the 30% predicted to be most responsive yields a HR of 0.492 (95% CI 0.219-1.11; $p = 0.0803$) and 0.338 (95% CI, 0.131-0.872; $p = 0.0186$), respectively. The corresponding 50% and 70% predicted to be least responsive have a HR of 0.945 (95% CI, 0.392-2.28; $p = 0.901$) and 0.967 (95%CI, 0.447-2.09; $p = 0.933$), respectively. The Kaplan-Meyer curves for these predicted subgroups are shown in Supplementary Fig. A.2.

Group characteristics for predicted responders are shown in Supplementary Table A.2. Groupwise differences are largely similar to those obtained on the anti-CD20-Abs dataset, with a few exceptions. In the laquinimod dataset, a significantly greater FSS-Bowel and Bladder and smaller normalized brain volume (NBV) are observed (whereas these did not reach the same level of significance in the anti-CD20-Abs test set), and the difference in T25FW is not statistically significant ($p < 0.05$). A smaller NBV was found in the responder group, but this only reached significance at the 50th percentile threshold. Nonetheless, the direction of the effect for these differences is concordant between the two test sets.

2.2.4 Comparison to baseline models

The performance of the non-linear model described in this paper is compared to numerous other baseline models in Table 2.3, as measured by the AD_{wabc} on the anti-CD20-Abs test set and on the laquinimod dataset. Scatter plots of the metrics obtained on both test sets are also provided in Supplementary Fig. A.3-A.4. All models were trained using the same procedure, on the same dataset, and with the same regression target. The MLP outperforms all other baselines on this metric, but some models (such as a linear

regression model with L2 regularization (ridge regression) and a CPH model) compare favorably on one of the two datasets. Without pre-training on the RRMS dataset, the performance of the MLP is still strong but inferior to the fine-tuned model. All single feature models are inferior to the MLP and CPH models except for the T2 lesion volume / disease duration model which falls between the these two models in terms of performance on the anti-CD20-Abs test set. We also tested a prognostic MLP which is only trained to predict progression on placebo, and which uses this prediction in place of the CATE estimate (assumes that more rapid progression leads to greater potential for treatment effect). This model's performance on the anti-CD20-Abs test set falls between that of the CPH model and the T2 lesion volume / disease duration model.

Table 2.3: Comparison of model performance (measured by AD_{wabc}) on the held-out test set of patients from ORATORIO and OLYMPUS (anti-CD20-Abs), and on the held-out dataset from ARPEGGIO (laquinimod)

	Anti-CD20-Abs	**Laquinimod**
Single feature*:		
Negative disease duration	0.0225	0.0114
Negative age	0.0067	-0.0287
Negative EDSS	0.0264	0.0074
Negative 9HPT dominant hand	-0.0109	0.0023
Negative 9HPT non-dominant hand	-0.0012	-0.0006
Negative T25FW	0.0033	0.0020
T2 lesion volume	0.0167	-0.0051
Gad count	0.0021	NaN‡
Feature / disease duration ratio†:		
Age / disease duration	0.0268	0.0138
EDSS / disease duration	0.0021	0.0020
9HPT dominant hand / disease duration	0.0238	0.0146
9HPT non-dominant hand / disease duration	0.0179	0.0098
T25FW / disease duration	0.0257	0.0049
T2 lesion volume / disease duration	0.0432	0.0164
Gad count / disease duration	0.0030	NaN‡
Regression model using all features:		
MLP (our model)	0.0565	0.0211
MLP (no pre-training§)	0.0486	0.019
MLP (prognostic model¶)	0.0408	0.0170
Ridge Regression	0.0227	0.0194
Survival model using all features:		
CPH	0.0305	0.0031

*The value of the feature is taken to be the CATE estimate for an individual. For example, the "T2 lesion volume" model uses the value of an individual's T2 lesion volume as the CATE estimate for that individual, such that a larger baseline volume predicts a larger treatment effect. A "negative" feature implies that the CATE estimate is the negative of the value of the feature. For example, the "negative disease duration" model predicts a larger treatment effect with shorter disease duration.
†The value of the feature divided by the disease duration is taken to be the CATE estimate for an individual. For example, the "EDSS / disease duration" model predicts a larger treatment effect with a more rapid historical rate of change in the EDSS over time.
‡Value for AD_{wabc} could not be computed due to low variance in values for Gad lesions in the laquinimod dataset.
§This MLP was trained without pre-training on the RRMS dataset.
¶The value of the predicted slope of disability progression on the placebo arm is used as the CATE estimate. In other words, a patient predicted to progress more rapidly on placebo (worse prognosis) predicts a larger treatment effect.
EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion; MLP = Multi-layer perceptron.

In OLYMPUS, Hawker *et al.* [11] identify a cutoff of age < 51 years and Gad lesion count > 0 at baseline as predictive of treatment effect. Using their definition, 21.9% and 11.3% of the patients in the the anti-CD20-Abs and laquinimod datasets, respectively, would be classified as responders. This is more restrictive than our most restrictive threshold which selects the 30% predicted to be most responsive. The HR for these predicted responders is 0.91 (95% CI, 0.392-2.11; $p = 0.831$) and 0.305 (95% CI, 0.0558-1.67; $p = 0.147$) for the anti-CD20-Abs and the ARPEGGIO patients, respectively. For both datasets, these effect size estimates do not reach statistical significance ($p < 0.05$). The effect size estimate for the anti-CD20-Abs dataset is also smaller compared to that obtained with our predictive enrichment method when selecting the 30% most responsive individuals. This binary cutoff is therefore generally inferior to our approach.

Finally, we compared our approach to the traditional phase 2 approach which typically uses an MRI-based surrogate outcome (brain atrophy being the most common) which is thought to be correlated with the clinical outcome of interest but that is more sensitive to the underlying biological processes or that has a lower variance, in order to increase a study's statistical power. For example, suppose our anti-CD20-Abs test set ($n = 297$) was a small phase 2 trial testing anti-CD20-Abs with brain atrophy as the primary outcome. Measuring brain atrophy at the 48 week MRI for the anti-CD20-Abs, the mean difference between the treatment arms is 0.066 (95% CI, -0.397 to 0.529; $p = 0.7786$). Looking at ORATORIO patents separately, since ORATORIO was the only positive trial in the anti-CD20-Abs dataset, the mean difference is 0.110 (95% CI, -0.352 to 0.572; $p = 0.6379$). Brain atrophy would therefore not have been able to detect a significant effect for ocrelizumab or for anti-CD20-Abs.

2.2.5 Simulating a phase 2 clinical trial enriched with predicted responders

To understand the effect of enriching a future clinical trial studying novel B-cell depleting agents, we simulated both a one and a two-year randomized clinical trial using populations enriched with predicted responders and estimated the sample size that would be needed to detect a significant effect under these conditions. To do so, we first used our model to predict responders to anti-CD20-Abs, defined using a range of thresholds (from including all individuals to including only the top 30% who are predicted to be most responsive). We then fit a CPH model to the ground-truth time-to-CDP24 for the responder group obtained at each threshold in order to estimate their corresponding HR. We then used the observed one-year and two-year CDP24 event rates in each responder group to calculate the sample size needed to detect a significant effect during a one-year or a two-year trial, respectively. This analysis is shown in Table 2.4.

Table 2.4: Estimated sample size for a one or two-year placebo-controlled randomized clinical trial of anti-CD20-Abs, using different degrees of predictive enrichment.

Percentile threshold*	**CDP control†**	**CDP treatment†**	**HR (95% CI)‡**	**Sample size estimate§**	**Number screened¶**
Two-year trial:					
0	0.30	0.24	0.74 (0.48-1.15)	1374	1374
10	0.31	0.24	0.72 (0.46-1.13)	1133	1259
20	0.30	0.22	0.70 (0.43-1.13)	1019	1274
30	0.29	0.22	0.67 (0.40-1.12)	812	1160
40	0.30	0.21	0.59 (0.33-1.03)	464	773
50	0.33	0.20	0.49 (0.27-0.91)	245	490
60	0.36	0.22	0.51 (0.26-0.98)	251	628
70	0.39	0.19	0.36 (0.17-0.79)	111	370
One-year trial:					
0	0.20	0.12	0.74 (0.48-1.15)	2435	2435
10	0.21	0.12	0.72 (0.46-1.13)	1988	2209
20	0.20	0.11	0.70 (0.43-1.13)	1796	2245
30	0.22	0.11	0.67 (0.40-1.12)	1346	1923
40	0.25	0.11	0.59 (0.33-1.03)	710	1183
50	0.26	0.11	0.49 (0.27-0.91)	371	742
60	0.31	0.12	0.51 (0.26-0.98)	365	913
70	0.30	0.10	0.36 (0.17-0.79)	171	570

*Percentile threshold for randomization. The 0th percentile represents an unenriched population, while the 70th percentile leads to inclusion of only the top 30% who are predicted to be most responsive.
†Proportion of CDP24 events for the responder groups corresponding to each percentile threshold.
‡HR for time-to-CDP24 for the responder groups corresponding to each percentile threshold.
§Sample size estimates are calculated using a desired power of 80% and $\alpha = 0.05$, assuming a 2:1 treatment to control randomization ratio. Calculations are based on the one or two-year CDP24 rate and one or two-year HR of responder groups in the anti-CD20-Abs dataset.
¶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).

Using the 50th percentile as a threshold for randomization in a two-year long trial as an example, a total of 490 individuals would be screened and the top 50% who are predicted to be most responsive would be randomized ($n = 245$). This leads to a six-fold reduction in the number of patients that need to be randomized while screening almost

three times less patients compared to the scenario where all participants are randomized into a two-year study ($n = 1374$).

2.3 Discussion

This work addresses the lack of a sufficiently predictive biomarker of treatment response for progression in multiple sclerosis, which has hampered progress by preventing efficient phase 2 clinical trials. We describe a deep learning solution to increasing the efficiency of early proof-of-concept clinical trials based on a multi-headed MLP architecture designed for CATE estimation. This approach can consistently identify and rank treatment effect among patients exposed to anti-CD20-Abs, and could reduce by several fold the sample size required to detect an effect in a short one or two-year long trial. We validate our model using a dataset composed of patients exposed to anti-CD20-Abs, and a second dataset of patients exposed to laquinimod. We demonstrate that a model trained to predict response to anti-CD20-Abs can also generalize to laquinimod, a medication with a very different mechanism of action, suggesting that there exists disease-agnostic predictors of response.

The model's predicted responders were enriched in numerous baseline features, including a younger age, shorter time from symptom onset, higher disability scores, and more lesion 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 [11]. Signori *et al.* [15] also found that younger age and the presence of Gad were associated with greater treatment effect in RRMS. In a study by Bovis *et al.* [3], a response scoring function obtained via CPH models in RRMS also identified Gad lesions and a higher normalized brain volume as predictive of treatment effect, although older age was found to be more predictive in the combination they studied.

In our experiments, a non-linear model (MLP) outperformed other linear (and log-linear) baselines, suggesting that complex relationships exist between the baseline features

and treatment effect. Nonetheless, a prognostic model (that predicts response to a medication solely based on the prediction of progression on placebo) also performed well, suggesting that poor prognosis is also predictive of treatment effect. A prognostic model could therefore be helpful in cases where drugs with very different mechanisms of action (e.g. targetting remyelination, or neurodegeneration) are being tested, in which case a model trained to predict treatment effect on an anti-inflammatory drug might perform less well than a prognostic model.

Interestingly, despite a balanced dataset with respect to gender, our model was better at identifying responders in women compared to men. We also noted that the model performed better in individuals ≥ 51, disease duration < 5 years, and/or an EDSS < 4.5. These findings suggest further studies are needed to determine whether and why predictors of response might differ depending on the stage of disease and sex.

Predictive enrichment is not the only approach to increase the efficiency of clinical trials in PPMS. However, the traditional approach of using a potential surrogate marker (such as brain atrophy) as part of a phase 2 study did not succeed in identifying a significant effect in our experiments, and may therefore limit early identification of effective therapies. Although used in phase 2 trials as a primary outcome, several studies on PPMS [16], RRMS [17], and SPMS [18] suggest no to modest correlation with clinical disability progression based on EDSS even after four to eight years of follow-up.

Another strategy could have been to infer from an RRMS trial that a drug might be effective for treating disability progression in a PMS trial. For example, ocrelizumab and siponimod were first found to be efficacious in the RRMS population in OPERA I/II [8] and BOLD [19], respectively, before being tested in the PPMS trial ORATORIO [10] and the SPMS trial EXPAND [20], respectively. In these cases, there were other reasons to believe that the drugs might be effective for treating progressive biology, but the predictive value of finding an initial effect on inflammatory biology remains of interest. From a predictive enrichment standpoint, baseline T2 lesion burden has been found to correlate with future disability and disability progression, at least modestly [21–24]. Evidence is

less robust for Gad lesions, since some authors [25] have demonstrated modest correlations with future disability at least 2 years from baseline, while others [23] have not. In our experiments, a treatment effect estimation model based on either Gad count or T2 lesion volume alone performed poorly. Only the rate of accumulation of T2 lesions over time (measured from the time of symptom onset) was predictive. Even if the inflammatory hypothesis was correct, a predictive enrichment strategy would be more efficient than awaiting the results of a RRMS study testing the same drug, particularly given that the power of a follow-up PPMS study is likely to be insufficient, as shown by the small proportion of responders to anti-CD20-Abs in our experiments, the dramatic difference in effect size between the inflammatory and progression-related outcomes, and the numerous examples of effective drugs for RRMS that had no identifiable effect on slowing disability progression in PMS [11, 12, 26–30].

Finally, the Food and Drug Administration has published a guidance document with suggestions regarding the design of predictively enriched studies [31]. One approach might be to first conduct a small trial of a short duration as a proof of concept in patients predicted to be highly responsive. If a significant effect is detected, a larger/longer follow-up study with a more inclusive (less enriched) population can be attempted with more confidence. It is also possible that, on the basis of a strong effect in the enriched responder group, the proof of concept would be sufficient for drug approval to be granted for the un-enriched population, given the significant unmet need and irreversible consequences of disability progression. To limit the risk that the predictive model is found to be inaccurate on the study population, stratified randomization can be used by having two parallel groups: the primary group (which would be adequately powered to detect an effect) would be an enriched responder group, while the secondary group would randomize predicted non-responders. Although the non-responder group would not be powered to detect an effect, it would provide a rough estimate of the effectiveness of the drug in this group and help guide design decisions for follow-up trials. The two groups could also be merged in a pre-planned analysis, to provide an estimate of the effect in the combined population.

Limitations of this work include the choice of model. Interpretability of black-box algorithms such as neural networks (reviewed elsewhere [32]) remains an area of active research. Although our MLP outperformed linear baselines, MLPs are more difficult to train and at higher risk of overfitting. Moreover, we made heavy use of several regularization schemes to prevent this. Our hyperparameter tuning procedure is also one of many that can be designed. Next, we used MRI-derived lesion and volumetric measures computed during the individual clinical trials, which could potentially ignore more subtle predictive features found within the MRI voxel-level data. Learning these features in a data-driven fashion through convolutional neural networks is the subject of ongoing work, but this can easily be appended to our MLP architecture. Regarding generalization to novel drug targets, more data is needed from drugs with diverse mechanisms of action to fully grasp the extent to which predictors of anti-inflammatory drugs are applicable to other drug classes, including neurodegenerative targets. Finally, it remains unknown if patients for whom our model predicted minimal effect over two to four years could benefit after longer periods of administration. Answering this question would require longer-term observational data.

2.4 Methods

The study protocol was originally approved by the McGill University Health Center's Research Ethics Board - Neurosciences-Psychiatry (IRB00010120) and then transferred and approved by the McGill University Faculty of Medicine and Health Sciences Institutional Review Board (A03-M14-22A).

2.4.1 Data

Data is taken from six different randomized clinical trials ($n = 3,830$): OPERA I [8], OPERA II [8], BRAVO [9], ORATORIO [10], OLYMPUS [11], and ARPEGGIO [12] (ClinicalTrials.gov numbers, NCT01247324, NCT01412333, NCT00605215, NCT01194570,

NCT00087529, NCT02284568, respectively). Informed consent and participant compensation (if any) was handled by the individual clinical trials. 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. Therefore, it is important to appreciate that the data included in our work are not an exact reproduction of those used in the clinical trials.

All clinical/demographic and MRI features that were consistently recorded as part of all 6 clinical trials (total of 19 features) were used to train our model. Values were recorded at the baseline visit (immediately before randomized treatment allocation), and are a combination of binary (sex), ordinal (EDSS, FSS), discrete (Gad count), and continuous variables (age, height, weight, disease duration, T25FW, 9-hole peg test (9HPT), T2 lesion volume, Gad count, and NBV). Disease duration was estimated from the time of symptom onset.

Lesion segmentation and volumetric measurements are derived from ground-truth lesion masks, which were generated independently (by an image analysis centre outside of this study) during the course of each clinical trial. A fully manual or a semi-automatic segmentation strategy was used during clinical trial analysis for each trial. This analysis began with automated segmentation and was followed by manual correction by experts. The resulting segmentation masks are the best available approximation to ground truth, but would not be expected to be identical between each expert and reading centre in part due to differences in the approach to lesion segmentation between reading centres (school effects). To account for any difference between the trial sites' segmentation pipelines and improve model optimization dynamics [33], we scaled the segmentation-based metrics into a common reference range. To do so, we first isolated the subset of samples that fulfilled the intersection of inclusion criteria for all trials. Then, we scaled all MRI metrics such that their range from -3 SD to +3 SD matches that of a reference trial (in the same interval of ± 3 SD) obtained from the training set. The reference trials were selected on the basis of sample size (ORATORIO for the PPMS trials, and OPERA I/II for the RRMS

trials). The range was clamped at ±3 SD for the scaling to be robust to extreme outliers.

The following right-skewed distributions were log-transformed: NBV, T2 lesion volume, T25FW, and 9HPT. Gad counts were binned into bins of 0, 1, 2, 3, 4, 5-6, 7-9, 10-14, 15-19 and 20+ lesions. Finally, to improve convergence during gradient descent, all non-binary features were standardized by subtracting the mean and dividing by the standard deviation, both calculated from the training dataset [33].

2.4.2 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 use CDP because it is a more robust indication that disability accrual will be maintained after 5 years [34]. CDP is most commonly based on the EDSS, a scale going from 0 (no disability) to 10 (death), in discrete 0.5 increments (except for a 1.0 increment between 0.0 and 1.0). A CDP event is defined as a 24-week sustained increase in the EDSS of 0.5 for baseline EDSS values > 5.5, of 1.5 for a baseline EDSS of 0, and of 1.0 for EDSS values in between. 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 the EDSS scores at different rates [35].

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 [36]), we chose not to model time-to-CDP because of limitations inherent in this metric. As outlined by Healy *et al.* [37], CDP reflects not only the rate of progression but also the baseline stage of the 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 Methods A.2 for details) and take its slope to be the outcome label that our MLP uses for training. One advantage of the slope outcome over time-to-CDP is that it can be modeled using any type of regression model. We revert to using time-to-CDP for model evaluation to facilitate comparison with treatment effect survival metrics reported in the original clinical trial publications.

2.4.3 Treatment effect modeling

To enrich clinical trials with individuals predicted to have an increased response to treatment, it is helpful to begin with the definition of ITE according to the Neyman/Rubin Potential Outcome Framework [38]. Let the ITE for individual i be τ_i, then

$$\tau_i \coloneqq Y_i(1) - Y_i(0)\,, \tag{2.1}$$

where $Y_i(1)$ and $Y_i(0)$ represent the outcome of individual i when given treatment and control medications, respectively. The Fundamental Problem of Causal Inference [39] states 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 ATE and the CATE in Supplementary Methods A.1. Briefly, ATE represents the average effect when considering the entire population, while CATE considers a sub-population characterized by certain characteristics (e.g. 40 year-old women with 2 Gad lesions at baseline). We use CATE estimation to frame the problem of predicting treatment response for individuals.

The best estimator for CATE is conveniently also the best estimator for the ITE in terms of mean squared error (MSE) [6]. Several frameworks have been developed to model 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 [6]. For a broader survey of methods, see the survey on uplift modeling by Gutierrez & Gérardy [4] (the uplift literature has contributed extensively to the field of causal inference, particularly when dealing with randomized experiments from an econometrics perspective).

In this work, an MLP was selected as the base model due to its high expressive power and flexibility to be integrated into larger end-to-end-trainable neural networks consisting of different modules (such as convolutional neural networks). We used a multi-headed architecture, with a common trunk and two output heads: one for 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, the CATE estimate $\hat{\tau}(x)$ given a feature vector x can be computed as:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)\,. \tag{2.2}$$

We use $\hat{\tau}(x)$ as the predicted treatment effect for an individual with characteristics x. Note that we multiplied all $\hat{\tau}(x)$ values by -1 in this paper to simplify interpretation in Section 2.2 (Results), such that a positive effect indicates improvement, while a negative effect indicates worsening on treatment.

This multi-headed approach can be seen as a variant of the T-Learner described for example by Künzel *et al.* [6], except that the two base models in our case share weights in the common trunk. Our network is similar to that conceptualized by Alaa Alaa *et al.* [40], but without the propensity network used to correct for any conditional dependence between the treatment allocation and the outcome given the input features, since our dataset comes from randomized data.

To decrease the size of the hyperparameter search space, we fixed the number of layers

and only tuned the layer width. We used one common hidden layer and one treatment-specific hidden layer. Additional common or treatment-specific layers could be used if necessary, but given the low dimensionality of our feature-space and the relatively small sample size, the network's depth was kept small to avoid over-fitting. The inductive bias behind our choice of using a multi-headed architecture is that disability progression can have both disease-specific and treatment-specific predictors of disability progression, 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 at hidden layers for non-linearity.

2.4.4 Training

The model was trained in two phases, depicted in Fig. 2.3. In the first phase, a 5-headed MLP was pre-trained on an RRMS dataset to predict the slope outcome on each treatment arm. In the second phase, the parameters of the common layers were frozen, and the output heads were replaced with two new randomly initialized output heads for fine-tuning on the PPMS dataset to predict the same outcome.

Optimization was done using mini-batch gradient descent with momentum. To prevent overfitting, the validation loss was monitored during 4-fold cross-validation (CV) to early-stop model training at the epoch with the lowest MSE, up to a maximum of 100 epochs. Dropout and L2 regularization were used, along with a max-norm constraint on the weights [41], to further prevent overfitting.

Mini-batches were sampled in a stratified fashion to preserve the proportions of participants receiving active treatment and placebo. Backpropagation was done using the MSE calculated at the output head that corresponds to the treatment that the patient was allocated to, t_i (the output head with available ground-truth). The squared errors from each output head were then weighted by $n_s/(m * n_t)$, where n_s represents the total

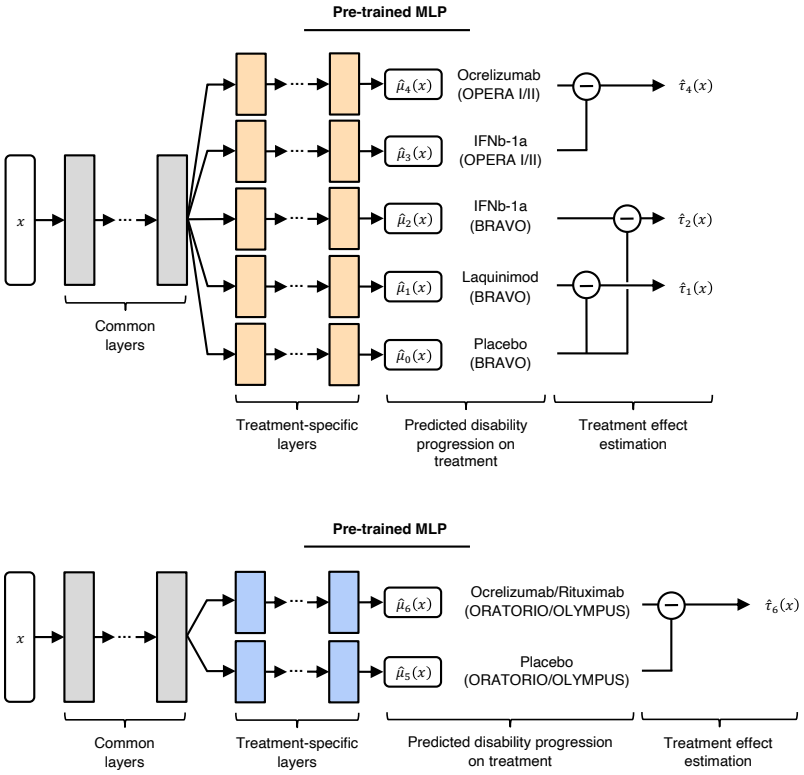


Figure 2.3: Multi-headed multilayer perceptron (MLP) architecture. The MLP was first pre-trained on a relapsing-remitting multiple sclerosis dataset (top), followed by fine tuning on a primary progressive multiple sclerosis dataset (bottom). Subtraction symbols indicate which treatment and control are being subtracted for the CATE estimate. Grey-colored layers indicate the common layers that are transferred from the pre-trained MLP to the fine-tuning MLP, at which point their parameters are frozen and only the parameters of the blue-colored layers are updated. The orange-colored layers are discarded after the pre-training step. x: Feature vector. $\hat{\tau}_t(x)$: CATE estimate for treatment t given feature vector x. $\hat{\mu}_t(x)$: predicted potential outcome on treatment t. IFNb-1a = Interferon beta-1a.

number of participants in the training split, n_t represents the number of participants in the treatment arm corresponding to the output head of interest, and m represents the total number of treatment arms. This compensates for treatment allocation imbalance in the 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 mean of its members' predictions, and is used for inference on the unseen test set.

A random search was used to identify the hyperparameters with the best validation

performance (learning rate, momentum, L2 regularization coefficient, hidden layer width, max norm, dropout probability). We used CV aggregation, or crogging [42], to improve the generalization error estimate using our validation metrics. Crogging involves aggregating all validation set predictions (rather than the validation metrics) and computing one validation metric for the entire CV procedure. The best model during hyperparameter tuning was selected during CV on the basis of two validation metrics: the MSE of the factual predictions, and the AD_{wabc} (described in detail in Supplementary Methods A.3). We combine both validation metrics during hyperparameter tuning by choosing the model with the highest AD_{wabc} among all models that fall within 1 SD of the best performing model based on the MSE loss. The SD of the best performing model's MSE is calculated from the loss values obtained in the individual CV folds.

2.4.5 Baseline models

The performance of the multi-headed MLP was compared to ridge regression and CPH models. Both models were used as part of a T-learner configuration (as defined by Künzel *et al.* [6]). Hyperparameter tuning was done on the same folds and with the same metrics as for the MLP.

2.4.6 Statistical Analysis

Hazard ratios were calculated using CPH models and associated p-values from log-rank tests. Sample size estimation for CPH assumes a two-sided test and was based on Rosner [43], as implemented by the Lifelines library (version 0.27.0) [44].

2.4.7 Software

All experiments were implemented in Python 3.8 [45]. MLPs were implemented using the Pytorch library (version 1.7.1) [46]. Scikit-Learn (version 0.24.2) [47] was used for the implementation of ridge regression, while Lifelines (version 0.27.0) [44] was used for

CPH. For reproducibility, the same random seed was used for data splitting and model initialization across all experiments.

2.5 Data Availability

Data used in this work was obtained from the following clinical trials (OPERA I [8], OPERA II [8], BRAVO [9], ORATORIO [10], OLYMPUS [11], and ARPEGGIO [12], with ClinicalTrials.gov numbers NCT01247324, NCT01412333, NCT00605215, NCT01194570, NCT00087529, NCT02284568, respectively), and are not publicly available. Access requests should be forwarded to the relevant data controllers.

2.6 Code Availability

Code necessary to reproduce the proposed methodological framework can be accessed publicly on the following GitHub repository: https://github.com/jpfalet/ms-predictive-enrichment. This repository does not contain dataset-specific code, since the data we used is not publicly available.

2.7 Acknowledgments

The authors are grateful to the companies who generously provided the clinical trial data that made this work possible: Biogen, BioMS, MedDay, Novartis, Roche / Genentech, and Teva.

This investigation was supported by an award from the International Progressive Multiple Sclerosis Alliance (award reference number PA-1412-02420), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-05471, Arbel, T.), the Canada Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program (Arbel, T.), a technology maturation grant from Mila - Quebec AI Institute (Arbel, T.), an endMS Personnel Award from the Multiple Sclerosis Society of Canada (Falet, J.R.), a Canada Graduate Scholarship-Masters Award from the Canadian Institutes of Health Research (Falet, J.R.), and the Fonds de recherche du Québec - 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 (Falet, J.R.).

2.8 Author Contributions

J.-P.R.F., T.A. and D.L.A. designed the study. J.-P.R.F., J.D.-F., B.N., J.S., F.B., M.-P.S., D.P., T.A., and D.L.A. contributed to the methods, analysis, and writing the manuscript. J.-P.R.F. wrote the code for all experiments. T.A. and D.L.A. jointly supervised this work. D.L.A. oversaw data collection.

2.9 Competing Interests

Bovis, F., has received teaching honoraria from Novartis and has received personal compensation for consulting services from Biogen, Eisai and Chiesi. Sormani, M.P., has received personal compensation for consulting services and for speaking activities

from Merck, Teva, Novartis, Roche, Sanofi Genzyme, Medday, GeNeuro, and Biogen. Precup, D., works part-time for DeepMind. Arnold, D.L., reports consulting fees from Biogen, Celgene, Frequency Therapeutics, Genentech, Merck, Novartis, Race to Erase MS, Roche, and Sanofi-Aventis, Shionogi, Xfacto Communications, grants from Immunotec and Novartis, and an equity interest in NeuroRx. The remaining authors report no competing interests.

References

1. Kappos, L., Wolinsky, J. S., Giovannoni, G., Arnold, D. L., Wang, Q., *et al.* Contribution of Relapse-Independent Progression vs Relapse-Associated Worsening to Overall Confirmed Disability Accumulation in Typical Relapsing Multiple Sclerosis in a Pooled Analysis of 2 Randomized Clinical Trials. *JAMA Neurology* **77,** 1132–1140 (2020).

2. Temple, R. Enrichment of clinical study populations. *Clinical pharmacology and therapeutics* **88,** 774–778 (2010).

3. Bovis, F., Carmisciano, L., Signori, A., Pardini, M., Steinerman, J. R., *et al.* Defining responders to therapies by a statistical modeling approach applied to randomized clinical trial data. *BMC Medicine* **17,** 1–10 (2019).

4. Gutierrez, P. & Gérardy, J.-Y. Causal Inference and Uplift Modelling: A Review of the Literature. *Proceedings of The 3rd International Conference on Predictive Applications and APIs, PMLR* **67** (eds Hardgrove, C., Dorard, L., Thompson, K. & Douetteau, F.) 1–13 (2017).

5. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 7353–7360 (2016).

6. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4156–4165 (2019).

7. Durso-Finley, J., Falet, J.-P. R., Nichyporuk, B., Arnold, D. L. & Arbel, T. Personalized Prediction of Future Lesion Activity and Treatment Effect in Multiple Sclerosis from Baseline MRI. *Medical Imaging with Deep Learning, PMLR* **172** (2022).

8. Hauser, S. L., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H.-P., *et al.* Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *New England Journal of Medicine* **376,** 221–234 (2017).

9. Vollmer, T. L., Sorensen, P. S., Selmaj, K., Zipp, F., Havrdova, E., *et al.* A randomized placebo-controlled phase III trial of oral laquinimod for multiple sclerosis. *Journal of Neurology* **261,** 773–783 (2014).

10. Montalban, X., Hauser, S. L., Kappos, L., Arnold, D. L., Bar-Or, A., *et al.* Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *New England Journal of Medicine* **376,** 209–220 (2017).

11. Hawker, K., O'Connor, P., Freedman, M. S., Calabresi, P. A., Antel, J., *et al.* Rituximab in patients with primary progressive multiple sclerosis: results of a randomized double-blind placebo-controlled multicenter trial. *Annals of Neurology* **66,** 460–471 (2009).

12. Giovannoni, G., Knappertz, V., Steinerman, J. R., Tansy, A. P., Li, T., *et al.* A randomized, placebo-controlled, phase 2 trial of laquinimod in primary progressive multiple sclerosis. *Neurology* **95,** e1027–e1040 (2020).

13. Weiss, K., Khoshgoftaar, T. M. & Wang, D. D. A survey of transfer learning. *Journal of Big Data* **3,** 1–40 (2016).

14. Zhao, L., Tian, L., Cai, T., Claggett, B. & Wei, L. J. Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association* **108,** 527–539 (2013).

15. Signori, A., Schiavetti, I., Gallo, F. & Sormani, M. P. Subgroups of multiple sclerosis patients with larger treatment benefits: a meta-analysis of randomized trials. *European Journal of Neurology* **22,** 960–966 (2015).

16. Ingle, G. T., Stevenson, V. L., Miller, D. H. & Thompson, A. J. Primary progressive multiple sclerosis: a 5-year clinical and MR study. *Brain* **126,** 2528–2536 (2003).

17. Fisher, E., Rudick, R. A., Simon, J. H., Cutter, G., Baier, M., *et al.* Eight-year follow-up study of brain atrophy in patients with MS. *Neurology* **59,** 1412–1420 (2002).

18. Turner, B., Lin, X., Calmon, G., Roberts, N. & Blumhardt, L. D. Cerebral atrophy and disability in relapsing-remitting and secondary progressive multiple sclerosis over four years. *Multiple Sclerosis Journal* **9,** 21–27 (2003).

19. Selmaj, K., Li, D. K., Hartung, H. P., Hemmer, B., Kappos, L., *et al.* Siponimod for patients with relapsing-remitting multiple sclerosis (BOLD): an adaptive, dose-ranging, randomised, phase 2 study. *The Lancet Neurology* **12,** 756–767 (2013).

20. Kappos, L., Bar-Or, A., Cree, B. A., Fox, R. J., Giovannoni, G., *et al.* Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *The Lancet* **391,** 1263–1273 (2018).

21. Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkiel, K. A., Benton, C. E., *et al.* Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain* **131,** 808–817 (2008).

22. Tintoré, M., Rovira, A., Río, J., Nos, C., Grivé, E., *et al.* Baseline MRI predicts future attacks and disability in clinically isolated syndromes. *Neurology* **67,** 968–972 (2006).

23. Minneboo, A., Barkhof, F., Polman, C. H., Uitdehaag, B. M., Knol, D. L. & Castelijns, J. A. Infratentorial Lesions Predict Long-term Disability in Patients With Initial Findings Suggestive of Multiple Sclerosis. *Archives of Neurology* **61,** 217–221 (2004).

24. Rudick, R. A., Lee, J. C., Simon, J. & Fisher, E. Significance of T2 lesions in multiple sclerosis: A 13-year longitudinal study. *Annals of Neurology* **60,** 236–242 (2006).

25. Kappos, L., Moeri, D., Radue, E. W., Schoetzau, A., Schweikert, K., *et al.* Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. *The Lancet* **353,** 964–969 (1999).

26. Lublin, F., Miller, D. H., Freedman, M. S., Cree, B. A., Wolinsky, J. S., *et al.* Oral fingolimod in primary progressive multiple sclerosis (INFORMS): a phase 3, randomised, double-blind, placebo-controlled trial. *The Lancet* **387,** 1075–1084 (2016).

27. Mantia, L. L., Vacchi, L., Pietrantonj, C. D., Ebers, G., Rovaris, M., *et al.* Interferon beta for secondary progressive multiple sclerosis. *The Cochrane database of systematic reviews* **1** (2012).

28. Kapoor, R., Ho, P. R., Campbell, N., Chang, I., Deykin, A., *et al.* Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. *The Lancet Neurology* **17,** 405–415 (2018).

29. Rojas, J. I., Romano, M., Ciapponi, A., Patrucco, L. & Cristiano, E. Interferon beta for primary progressive multiple sclerosis. *The Cochrane database of systematic reviews* (2009).

30. Wolinsky, J. S., Narayana, P. A., O'Connor, P., Coyle, P. K., Ford, C., *et al.* Glatiramer acetate in primary progressive multiple sclerosis: results of a multinational, multicenter, double-blind, placebo-controlled trial. *Annals of Neurology* **61,** 14–24 (2007).

31. Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products Guidance for Industry. *US. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER)* (2019).

32. Zhang, Y., Tino, P., Leonardis, A. & Tang, K. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* **5,** 726–742 (2021).

33. LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. in *Neural Networks: Tricks of the Trade: Second Edition* (eds Montavon, G., Orr, G. B. & Müller, K.-R.) 9–48 (Springer Berlin Heidelberg, 2012).

34. Kalincik, T., Cutter, G., Spelman, T., Jokubaitis, V., Havrdova, E., *et al.* Defining reliable disability outcomes in multiple sclerosis. *Brain* **138,** 3287–3298 (2015).

35. Zurawski, J., Glanz, B. I., Chua, A., Lokhande, H., Rotstein, D., *et al.* Time between expanded disability status scale (EDSS) scores. *Multiple sclerosis and related disorders* **30,** 98–103 (2019).

36. Wang, P., Li, Y. & Reddy, C. K. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.* **51** (2019).

37. Healy, B. C., Glanz, B. I., Swallow, E., Signorovitch, J., Hagan, K., *et al.* Confirmed disability progression provides limited predictive information regarding future disease progression in multiple sclerosis. *Multiple sclerosis journal - experimental, translational and clinical* **7** (2021).

38. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

39. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81,** 945 (1986).

40. Alaa, A. M., Weisz, M. & van der Schaar, M. Deep Counterfactual Networks with Propensity-Dropout. *Proceedings of the 34th International Conference on Machine Learning, PMLR* **70** (2017).

41. 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* **15,** 1929–1958 (2014).

42. Barrow, D. K. & Crone, S. F. Crogging (cross-validation aggregation) for forecasting - A novel algorithm of neural network ensembles on time series subsamples. *Proceedings of the International Joint Conference on Neural Networks* (2013).

43. Rosner, B. *Fundamentals of Biostatistics* 6th edition, 807 (Brooks Cole, 2006).

44. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4,** 1317 (2019).

45. Rossum, G. V. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, 2009).

46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019* (eds Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B. & Garnett, R.) 8024–8035 (2019).

47. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

Chapter 3

Discussion and Conclusions

3.1 Discussion

In this work, we address our primary objective by presenting a deep learning framework for estimating CATE on disability progression in MS, and demonstrate significant improvements in statistical power when using the model for predictive enrichment. The proposed approach therefore enables enrichment of short, proof-of-concept clinical trials to speed drug development for progression in MS. Moreover, this work yields additional insights by 1) providing evidence suggesting that more responsive individuals are more active in terms of inflammatory markers (T2 lesion volume and Gad lesions) at baseline, 2) showing that a model trained to predict the effect of an anti-CD20-Abs can also estimate the effect of laquinimod, a drug with a different mechanism of action, suggesting the existence of shared predictors between drug classes, and 3) demonstrate that the proposed DL model outperforms less expressive models, in particular those that make predictions using one or two features.

Further advancing our understanding of what predicts a favorable response could both help understand the pathophysiology of progression, and hint at drug targets to investigate. The authors of a recent systematic review and meta-analysis of twelve clinical trials concluded that the benefit of investigated immune therapies in slowing disability progression is confined to individuals who have signs of inflammatory activity at baseline (presence of Gad lesions and/or relapses in the last 1-2 years) [75]. This

may be because these treatments do not target distinct pathophysiological pathways that could underly slow disability progression. It also highlights the need to distinguish whether an investigational drug is effective at slowing disability accrual by preventing acute inflammatory activity/relapse-associated worsening, or by preventing PIRA [76]. This is imperative, given that many currently licensed DMTs for RRMS targeting the immune system have been shown to improve disability outcomes for both PPMS [77] and SPMS patients [78].

Explaining the predictions of a DL model in a way that is both intelligible to humans and faithful to the statistical relationships in the data is very challenging, but necessary to further our understanding of the predictors of treatment effect. Explanations from different techniques can be at odds with each other, and several reproducability issues impair their practical utility (see [79] for a comparison of saliency mapping methods in the context of medical imaging). Recently, particularly for medical imaging, counterfactual explanations have surfaced as an attractive solution. In short, a counterfactual explanation is one that uses a generated counterfactual image for the case where the model's prediction would have been different, in order to highlight features of the image that affect the prediction. This is most frequently used to explain classifiers. In work I contributed to during my studies, we applied this to the context of MS by training a counterfactual generator of brain MRI images to explain the predictions of a classifier pre-trained to predict whether a patient would have new or enlarging T2 lesions at a future timepoint [3]. Through a qualitative analysis of the difference between the factual and counterfactual images, we could identify changes local to regions where new lesions would appear, and more global changes, some reminiscent of currently studied markers of progression such as DAWM. Extending this to the setting of treatment effect estimation would be valuable.

While this work focused on scalar MRI metrics, there could be more predictive latent features in the images that are not being captured by these metrics. To this end, in published work that I co-authored, we proposed a similar framework to the one presented in Chapter 2 with the main addition being a convolutional neural network to process

the images before branching off into a multi-head MLP [2]. However, this study was focused on predicting the effect on lesion activity, and therefore doing so for disability progression remains an unmet need. One key difficulty in using a high-dimensional input such as an image is the uncertainty that results from lack of overlap between interventional distributions [80]. Overlap uncertainty occurs when there aren't examples of individuals with the same features who received both treatment and control, and this is expected to occur more frequently with higher dimensions due to the *Curse of Dimensionality* [81]. One solution is to first learn a lower dimensional representation of the image from which one can learn a CATE estimator with less overlap uncertainty. However, this in some ways shifts the problem to one of learning a good representation, which in itself is non-trivial. One of the most common ways to infer distributions over latent variables is using a variational autoencoder (VAE). However, vanilla VAEs result in poor detail-preservation and blurring of reconstructed images [82]. This is problematic, because important predictors of treatment effect can consist of small details, such as a specific "texture" in the white matter, or the appearance of the border of a lesion, and capturing these in the representation is essential. In a paper I co-authored, we proposed a hierarchical VAE to learn a latent representation that preserves finer details, which is more suitable for downstream tasks such as marker discovery [4]. That said, more work is needed to integrate this type of approach into treatment estimation frameworks.

Another important consideration for using the proposed predictive enrichment approach is how well the model generalizes to different drug classes. In the worst case, there would be no shared predictors of treatment effect between the drugs that were included in the training set of the model, and the investigational drugs for which the model is to be used for predictive enrichment. However, this is unlikely. For one, it is frequently the case that once a drug is identified as effective, drugs with a similar mechanism are tested with different hypothetical enhancements (e.g. fewer side effects, or more selective binding to the drug target). There is little ground to believe that a model trained on one drug would not generalize to another such drug with the same or a similar mechanism. In support of this

conjecture, we provide evidence suggesting overlap between two different anti-CD20-Abs (rituximab and ocrelizumab) in Chapter 2. Likewise, recent observational evidence has accumulated in favor of a beneficial effect of rituximab on progression [83], similar to that seen with ocrelizumab. In addition, we also provide evidence of generalization between two different classes of medications (anti-CD20-Abs and laquinimod), which target different aspects of the immune system. We can therefore hypothesize that different classes of immunomodulatory drugs share at least some predictors of response. However, whether the predictors of response are shared between immune therapies and drugs targeting non-immune pathways (e.g. neurodegenerative pathways) remains unclear. In this case, we proposed in Chapter 2 prognostic enrichment as an alternative, where enrichment is based on prediction of prognosis rather than treatment effect, because this does not depend on generalization between interventional distributions. Another risk-reduction strategy would be to conduct a stratified randomized controlled trial, where part of the population that would have been excluded on the basis of predictive enrichment is included in a separate stratum of the trial, and subjected to the same analyses as the enriched group. Ultimately, our approach will have to be tested on a variety of drug classes to ascertain the range of investigational situations where it would be applicable.

Another strategy to increase a study's statistical power, which has so far not been discussed in this thesis, is to measure a treatment effect on a clinical outcome that is more sensitive to disease progression. While CDP based on the EDSS has been used as the primary outcome measure in the vast majority of clinical trials, there have been numerous calls for a change in practice. The EDSS is biased towards ambulation function [84], it changes slowly over time, and it is a noisy measurement with high inter-rater [85] and geographical [86] variability. Several composite measures, such as the multiple sclerosis functional composite (MSFC) [87], the EDSS-Plus [88], and others [89, 90], have been proposed that aggregate different disability scores including EDSS, 9HPT, T25FW, Paced Auditory Serial Addition Test (PASAT), Symbols Digit Modalities Test (SDMT), low contrast letter acuity (LCLA), to capture many facets of disability and increase sensitivity

to detect clinical progression. Composite outcomes therefore could represent one part of the solution, but have seen slow uptake in practice, and whether the resultant increase in statistical power would be sufficient to enable short, proof-of-concept clinical trials, remains to be confirmed.

The work presented in Chapter 2 had to compose with an important challenge in using a multi-trial federated dataset, which is that of image harmonization. Even with scalar MRI-derived metrics, pooling data from different clinical trials relies on the assumption that these metrics are drawn from the same distribution. In the case of segmentation-based metrics (such as lesion volume), these were obtained through a gold-standard fully manual or semi-automatic segmentation strategy, which consists of automated segmentation followed by manual correction by experts. There are well known differences in the approach to determining lesion boundaries (school-effects) that can result in substantial differences in lesion volumes between different reading centers. Thus, the lesion masks we used are the best approximation we have to ground truth, but would not be expected to be identical between each expert and reading centre.

As suggested by one reviewer for the manuscript in Chapter 2, one would ideally have access to segmentation metrics obtained through a single pipeline. However, this single pipeline would not necessarily result in "better" segmentation masks, since the optimal way of segmenting lesions is partly subjective, and would be impractical as this type of work is extremely labour intensive and originally cost millions of dollars to perform. The approach we have taken is more practical in that it standardizes the range of input data to account for these school-effects, and therefore only requires access to the lesion counts/volumes that were generated during a clinical trial. If one is to replace the gold-standard segmentation framework by a fully automated pipeline, these school effects are still present and affect the learning process due to the presence of different lesion annotation styles, depending on the source of the annotations. In work I contributed to during my studies, we showed that training a lesion segmentation model that explicitly learns the label style specific to each cohort can help the model focus on the invariant features in the data [5]. Learning the

cohort bias also provides users with the flexibility to generate segmentation labels using a single, arbitrary segmentation style. Future work aimed at evaluating this strategy in the context of treatment effect estimation would be interesting.

While this work focused on applying DL-based treatment effect estimation to enrich clinical trials, the same methodology can be used for precision medicine in the clinic. A patient and doctor can obtain an estimate for the expected treatment effect given the patient's baseline characteristics, and use this to choose one treatment over another. Indeed, much of the prior work on precision medicine using machine learning has focused CATE estimation [91]. Nonetheless, two caveats about estimating CATE for this purpose are worth discussing. First, when the set of possible treatments includes more than two drugs, comparing their efficacy can be tricky, particularly when they were tested in different cohorts. To do so, one must ensure that the distribution over the input variables from different cohorts overlap (otherwise there exists an overlap violation), and care must be taken to ensure that variables were measured in the same way. Second, estimating ITE seems more natural for individual-level decision making. ITE is correlated with CATE, but they are generally not equal [30, 92]. However, due to the *Fundamental Problem of Causal Inference* [29], both ITE estimation and evaluation of ITE estimators is challenging. As briefly mentioned in Chapter 1.1.1, the most natural approach to estimating ITE involves counterfactual logic. As opposed to CATE estimation which provides an estimate for an expected effect given an individual's characteristics (without regards for the specific outcome that has been observed for this individual and without explicit modeling of unobserved variables that also determine the uniqueness of this indivudual's response to treatment), counterfactual inference-based ITE estimation is based on inferring the counterfactual outcome that would have occurred under the same unique conditions that produced the observed outcome for that individual. The ITE is then the subtraction of the inferred counterfactual and the factual outcome. The *abduction-action-prediction* algorithm [27] and *twin-networks* [93] are two methods to infer counterfactuals, both of been recently extended to DL frameworks [94, 95]. These methods typically rely on

having the correct parametric model for the structural equations of the causal model that generates the observed data (whether provided by an expert or learned through causal discovery), which can be difficult to guarantee. Moreover, in general, it is not possible to identify a counterfactual from observational and experimental data. Instead, probabilities for target counterfactuals can be inferred in the partial identifiability setting, and bounds can be obtained for these probabilities [96]. Despite its challenges, this path is worth pursuing in the quest to bring DL to the bedside for precision medicine.

3.2 Conclusions

The work detailed in this thesis addresses an important limitation in the current process used to identify efficacious therapies that slow disability progression in MS. To our knowledge, this is the first successful application of DL-based CATE estimation to enrich PMS clinical trials. Specifically, we showed that a multi-headed MLP can be used to preferentially randomize more responsive individuals as a means of increasing a clinical trial's statistical power, thereby rendering short, proof-of-concept clinical trials a feasible endeavour. In doing so, we showed that responders to anti-CD20-Abs are typically younger, with a shorter disease duration, and have more lesion activity on pre-treatment MRI. We also showed that a model trained on individuals exposed to anti-CD20-Abs could generalize to a medication with a very different mechanism of action, laquinimod, suggesting that there are common predictors for treatment effect across drug classes. Future work aimed at validating this model on datasets composed of a broader range of medications will be helpful to better understand the extent of this overlap. Finally, we found that a DL framework was superior compared to alternative baseline models.

The proposed training procedure and model architecture is highly flexible, and can easily be integrated or extended to a variety of different artificial neural network architectures. A natural extension of this work is to combine our proposed MLP with data-driven feature extraction from MRI images using convolutional neural networks. Doing so might uncover

more subtle predictors of treatment effect that are not captured by traditional scalar MRI metrics, thereby improving model performance.

Finally, this approach is not limited to predictive enrichment of clinical trials, and can be applied to precision medicine in the clinic. This model can be part of a clinical decision support tool to aid personalized treatment decisions early on and avoid the irreversible disability accrual that can result from trying a treatment that turns out to be ineffective for a particular individual. Future work on DL and causal inference aimed at this specific application could therefore yield substantial improvements in clinical care and in the quality of life of patients with MS.

Appendix A

Supplementary Information for Chapter 2

Published in Falet, J.-P. R. *et al.* Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nature Communications* **13,** 5645 (1 2022)

Supplementary Methods

A.1 Treatment Effect Estimation

To enrich clinical trials with individuals predicted to have an increased response to treatment, it is helpful to begin with the definition of individual treatment effect (ITE) according to the Neyman/Rubin Potential Outcome Framework [1]. Let the ITE for individual i be τ_i, then

$$\tau_i \coloneqq Y_i(1) - Y_i(0)\,, \tag{A.1}$$

where $Y_i(1)$ and $Y_i(0)$ represent the outcome of individual i when given treatment and control medications, respectively. The *Fundamental Problem of Causal Inference* [2] states 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 nonetheless be observed at the group level. The average treatment effect (ATE) is defined as the expected difference between both potential outcomes:

$$\text{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]\,. \tag{A.2}$$

Supplementary Equation A.2 is still in terms of unobservable causal quantities, so additional assumptions are needed. While 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, the ATE can identified from the observed outcome Y as follows

$$\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]\,, \tag{A.3}$$

where $T \in \{0, 1\}$ is the treatment allocation. Broadly speaking, the ATE (sometimes formulated as a ratio instead of a difference) is what is 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 $x \in \mathcal{X} \subseteq \mathbb{R}^d$. The conditional average treatment effect (CATE), denoted $\tau(x)$, is defined as:

$$\tau(x) \coloneqq \mathbb{E}[Y(1)|X=x] - \mathbb{E}[Y(0)|X=x]\,, \tag{A.4}$$

which can similarly be rewritten in terms of the observed outcome Y in the context of randomized controlled trials, where $\{(Y(0), Y(1)) \perp\!\!\!\perp T\}|X$:

$$\tau(x) = \mathbb{E}[Y|X=x, T=1] - \mathbb{E}[Y|X=x, T=0] = \mu_1(x) - \mu_0(x)\,. \tag{A.5}$$

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 multi-headed multilayer perceptron (MLP) 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. We use $\hat{\tau}(x_i)$ as the estimate for the treatment effect of an individual, $\hat{\tau}_i$.

A.2 Slope Outcome

We assume that progression is slow over the course of the one to two year duration of a phase 2 or 3 clinical trial such that the Expanded Disability Status Scale (EDSS) value at time t following treatment initiation can be modeled as the linear relationship

$$\text{EDSS} = \beta_0 + \beta_1 t\,, \tag{A.6}$$

where β_0 and β_1 are the regression coefficients. Using the method of ordinary least squares for linear regression, estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found using all available timepoints t. Each patient i has a separate slope of disability progression, $\hat{\beta}_{1,i}$, found by fitting a linear regression model to their own EDSS values. This slope is then used as the ground-truth outcome y_i that we train a neural network to predict:

$$y_i = \hat{\beta}_{1,i}\,. \tag{A.7}$$

To compute the slope, a minimum of two timepoints t must be available for each patient. We also require that the duration between the first and last timepoints be greater than 24 weeks, given that we are evaluating our model's performance using confirmed disability progression at 24 weeks (CDP). Participants who do not fulfill these two requirements are excluded from the dataset. The average number of visits used to compute the slopes was 12.23 (SD 2.86; range 3-24).

Note that the definition of confirmed disability progression (CDP) used in clinical trials depends on the baseline EDSS of the individual. For a CDP event to occur, a participant who has a baseline EDSS of 0 requires an increase in EDSS of 1.5, while a baseline of > 5.5 requires an increase of 0.5. Baseline values in between require an increase of 1.0.

Therefore, in order for our slope outcome to closely resemble the changes in EDSS that are required to reach CDP, we scaled the EDSS values prior to fitting the linear regression models, such that the increase necessary for a CDP event to occur approximately maps to an increase of 1.0 after the scaled transform:

$$f(\text{EDSS}) = \begin{cases} \frac{\text{EDSS}}{1.5}, & \text{if EDSS} \leq 1.5 \\ \text{EDSS} - 0.5, & \text{if } 1.5 < \text{EDSS} \leq 6.0 \\ \frac{\text{EDSS}-6.0}{0.5} + 5.5, & \text{if EDSS} > 6.0 \end{cases} \tag{A.8}$$

We use the scaled values, $f(\text{EDSS})$ in place of the EDSS when fitting the linear regression model. $f(\text{EDSS})$ is plotted in Supplementary Fig. A.5.

A.3 Weighted Average Treatment Difference Curve

Following Zhao *et al.* [3], we define a conditional expectation, $\text{AD}(c)$, which reflects the ATE of a sub-group of patients who are predicted by our model to have a treatment effect greater than a threshold value c:

$$\text{AD}(c) = \mathbb{E}[\, Y(1) - Y(0) \,|\, \hat{\tau}_i \geq c \,]\,. \tag{A.9}$$

The conditional expectation for $Y(1) - Y(0)$ is estimated using the restricted mean survival time (RMST) for the time-to-CDP, truncated at 2 years [4]. By defining the conditional expectation in terms of the RMST instead of the slope outcome used as the target for training the neural network, the $\text{AD}(c)$ better reflects how well our model can identify responders using a survival-based metric, which is ultimately what clinical trials will use.

The $\text{AD}(c)$ behaves as a population selector for predictive enrichment, whereby patients expected to respond with effect size greater than a desirable threshold c can be enrolled in a clinical trial or recommended the medication in a clinical setting.

If patients are ranked accurately according to their predicted responsiveness to the

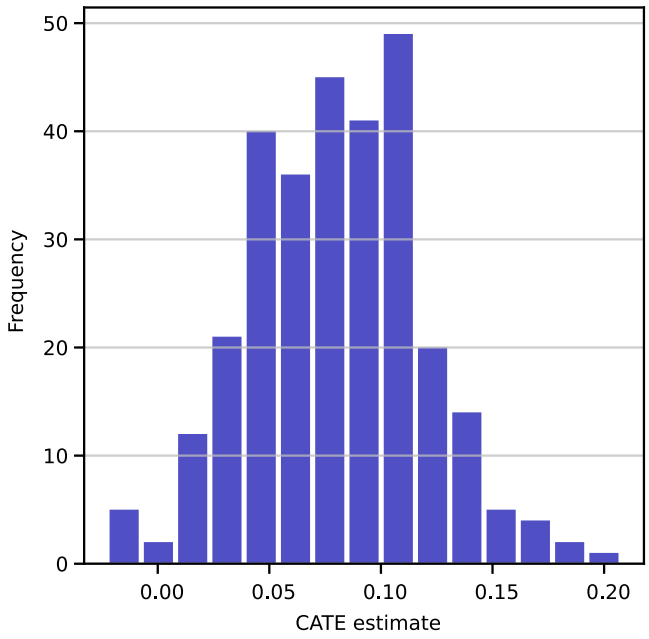
active medication, then the resultant AD(c) curve should have a large area under the curve, $\mathrm{AD}_{\mathrm{auc}}$. The $\mathrm{AD}_{\mathrm{auc}}$ is therefore a useful evaluation metric. We compute the $\mathrm{AD}_{\mathrm{auc}}$ using polygon approximation with operating points every 10 percentiles from 0 until the 70th percentile for better computational efficiency, while we use 1 percentile increments for reporting test metrics and for visualization purposes in this paper. Following Zhao *et al.* [3], we then subtract the effect size of the entire (unenriched) population from the $\mathrm{AD}_{\mathrm{auc}}$ to facilitate the comparison of different models. This metric is called the area between curves, or $\mathrm{AD}_{\mathrm{abc}}$, and can be written as

$$\mathrm{AD}_{\mathrm{abc}} = \mathrm{AD}_{\mathrm{auc}} - \mathrm{AD}(\hat{\tau}_{(0)}) \,, \tag{A.10}$$

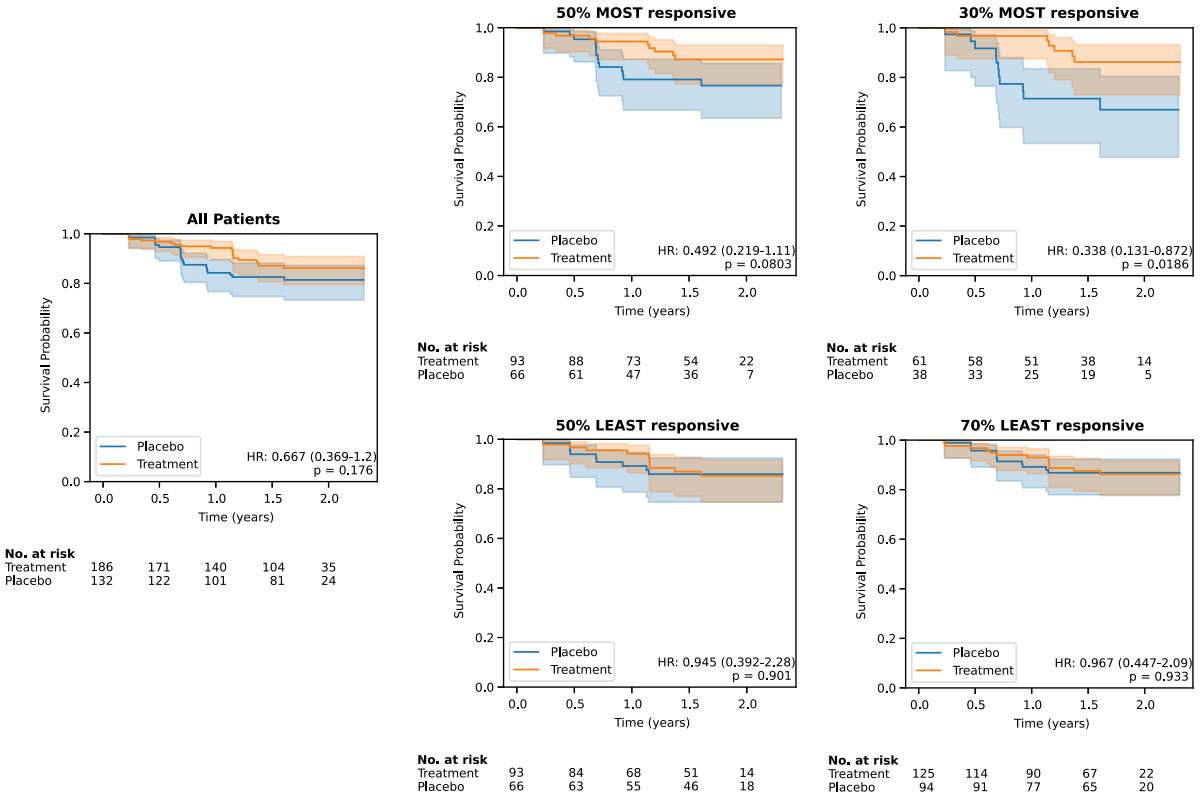
where $\hat{\tau}_{(0)}$ represents the minimum predicted treatment effect in the evaluation set. We further weigh the $\mathrm{AD}_{\mathrm{abc}}$ by multiplying it to a measure of monotonicity to promote a monotonically increasing AD(c), since monotonicity indicates that the model can rank response accurately throughout the range of possible responsiveness. To do so, we use the Spearman's rank correlation coefficient, ρ, calculated between the $\mathrm{AD}_{\mathrm{abc}}$ values and the thresholds c, as the scaling factor for the $\mathrm{AD}_{\mathrm{abc}}$:

$$\mathrm{AD}_{\mathrm{wabc}} = \rho \mathrm{AD}_{\mathrm{abc}} \,. \tag{A.11}$$

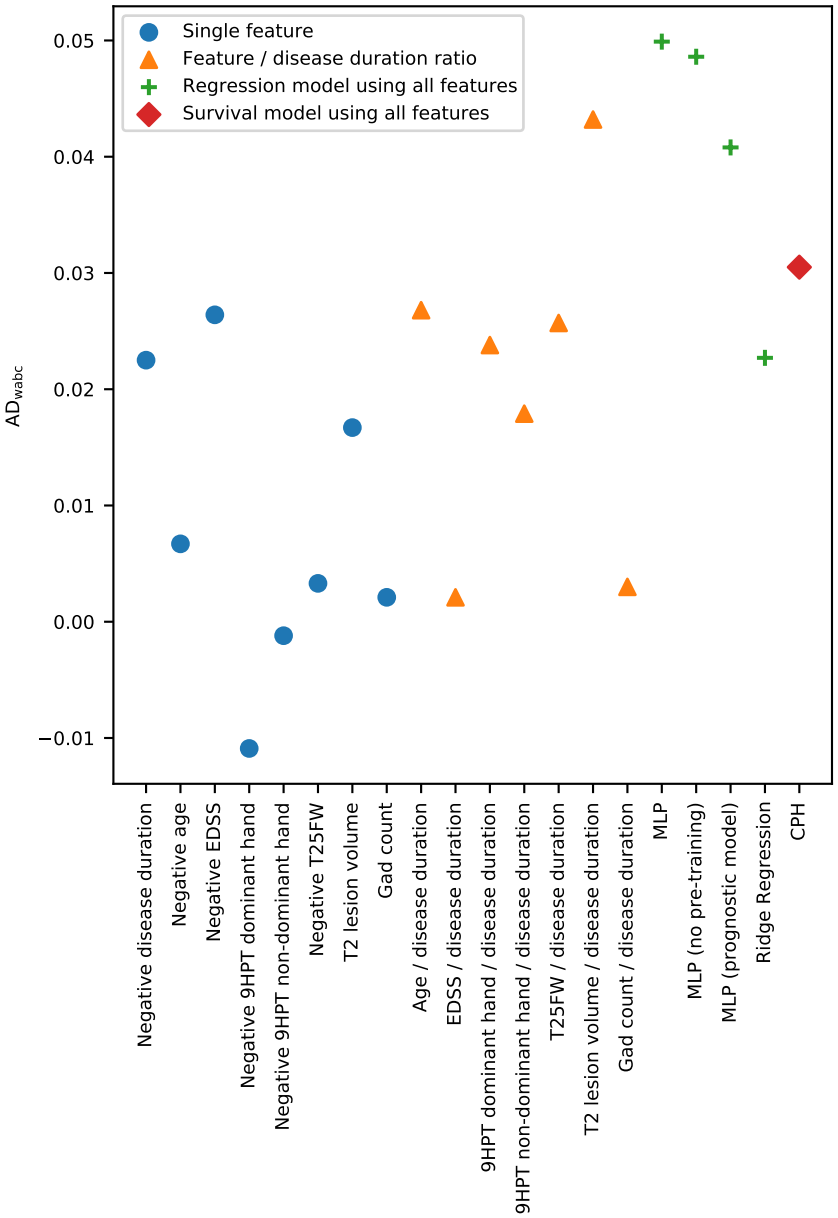
Supplementary Figures



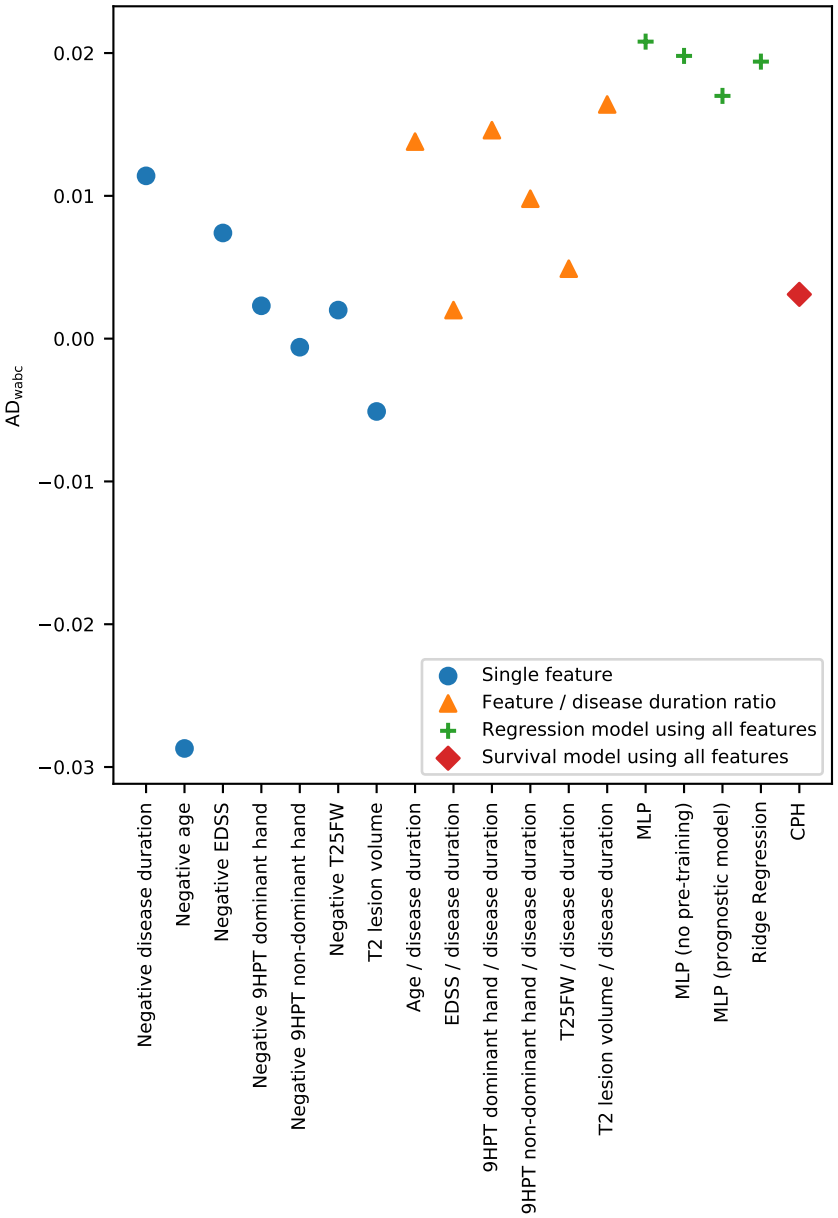
Supplementary Figure A.1: Histogram of CATE estimates for the anti-CD20-Ab test set. Positive numbers indicate a predicted benefit from anti-CD20-Abs over placebo, 0 indicates no predicted benefit, and negative numbers indicate predicted harm.



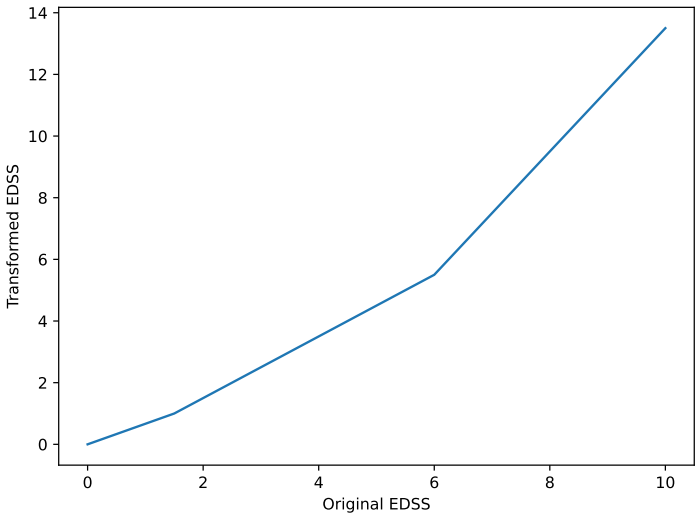
Supplementary Figure A.2: Kaplan-Meyer curves +/- 95% confidence intervals (CI) for predicted responders and non-responders to laquinimod, defined at two thresholds of predicted effect size. These are compared to the whole group (left). The placebo group is displayed in blue, and the treatment (anti-CD20-Abs) group is displayed in orange. Survival probability is measured in terms of time-to-CDP24 using the EDSS. p values are calculated using log-rank tests. 95% CIs are estimated using Greenwood's Exponential formula.



Supplementary Figure A.3: Comparison of model performance (measured by AD_{wabc}) on the held-out test set of patients from ORATORIO and OLYMPUS. Refer to the main text and to Table 3 for details about the implementation of each model. EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion; MLP = Multi-layer perceptron.



Supplementary Figure A.4: Comparison of model performance (measured by AD_{wabc}) on the held-out test set of patients from ARPEGGIO. Refer to the main text and to Table 3 for details about the implementation of each model. EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion; MLP = Multi-layer perceptron.



Supplementary Figure A.5: Expanded Disability Status Scale transformation to account for the baseline-dependent definition of confirmed disability progression.

Supplementary Tables

Supplementary Table A.1: Feature features and outcomes per treatment arm for the relapsing-remitting pre-training dataset.

	Ocrelizumab		**IFNb-1a SC**		**IFNb-1a IM**	**Laquinimod**	**Placebo**
	OPERA I n=320	OPERA II n=335	OPERA I n=295	OPERA II n=329	BRAVO n=412	BRAVO n=407	BRAVO n=422
Demographics:							
Age (years)	37.35 (9.36)	37.44 (8.93)	37.25 (9.54)	37.39 (8.82)	38.02 (9.41)	37.02 (9.19)	37.50 (9.59)
Sex (% male)	35.00	37.01	32.20	31.61	32.28	35.63	28.67
Height (cm)	169.58 (8.91)	169.59 (9.52)	169.40 (9.18)	168.66 (8.81)	168.32 (8.57)	169.12 (8.66)	169.05 (8.64)
Weight (kg)	74.31 (17.42)	76.51 (16.97)	75.25 (17.04)	74.99 (19.00)	69.63 (15.93)	69.50 (15.04)	69.52 (13.66)
Disease duration (years)	6.71 (6.45)	6.58 (5.95)	6.08 (5.79)	6.80 (6.28)	6.93 (5.81)	6.50 (5.80)	6.90 (6.53)
Disability Scores:							
EDSS	2.79 (1.22)	2.68 (1.32)	2.60 (1.26)	2.74 (1.40)	2.63 (1.15)	2.65 (1.24)	2.73 (1.18)
FSS-Bowel and Bladder	0.56 (0.73)	0.64 (0.79)	0.60 (0.79)	0.61 (0.81)	0.52 (0.71)	0.57 (0.76)	0.54 (0.71)
FSS-Brainstem	0.59 (0.81)	0.48 (0.76)	0.57 (0.77)	0.50 (0.79)	0.73 (0.78)	0.78 (0.81)	0.83 (0.82)
FSS-Cerebellar	1.15 (1.02)	1.03 (1.01)	1.00 (0.96)	1.04 (1.01)	1.20 (0.96)	1.21 (1.04)	1.25 (0.99)
FSS-Cerebral	0.50 (0.72)	0.60 (0.81)	0.55 (0.77)	0.65 (0.83)	0.64 (0.76)	0.66 (0.74)	0.70 (0.79)
FSS-Pyramidal	1.71 (1.02)	1.65 (1.05)	1.54 (1.01)	1.54 (1.05)	1.79 (0.96)	1.73 (1.00)	1.75 (0.98)
FSS-Sensory	1.17 (1.00)	1.01 (1.00)	1.04 (0.96)	1.10 (1.01)	0.94 (1.02)	1.04 (1.04)	1.02 (0.99)
FSS-Visual	0.67 (0.84)	0.68 (0.89)	0.72 (0.88)	0.69 (0.91)	0.80 (1.09)	0.79 (1.17)	0.85 (1.25)
Mean T25FW (sec)	7.80 (7.56)	8.19 (11.83)	7.04 (7.14)	7.29 (7.64)	6.31 (5.45)	6.00 (2.89)	6.04 (3.05)
Mean 9HPT dominant hand (sec)	24.47 (17.66)	23.80 (9.09)	23.77 (17.37)	24.52 (13.34)	21.73 (5.87)	21.98 (7.18)	22.83 (17.16)
Mean 9HPT non-dominant hand (sec)	26.85 (23.72)	25.26 (13.02)	24.51 (8.09)	26.31 (19.01)	23.13 (6.00)	23.06 (6.86)	23.87 (12.46)
MRI metrics:							
Gad count	1.76 (4.49)	1.81 (4.51)	1.74 (4.93)	1.96 (5.16)	1.85 (6.86)	1.84 (5.22)	1.47 (5.88)
T2 lesion volume (mL)	10.59 (14.25)	11.28 (15.00)	8.69 (10.13)	10.19 (12.07)	8.86 (10.55)	9.69 (10.38)	7.99 (8.95)
Normalized brain volume (L)	1.50 (0.08)	1.50 (0.09)	1.50 (0.09)	1.50 (0.09)	1.59 (0.08)	1.58 (0.10)	1.59 (0.09)
Outcome:							
Slope (EDSS change / yr)*	-0.01 (0.39)	0.00 (0.58)	0.07 (0.47)	0.09 (0.57)	0.06 (0.72)	0.04 (0.53)	0.14 (0.83)
RMST (at 2 years)†	1.97	1.95	1.93	1.92	1.93	1.93	1.90

Values in brackets are standard deviations, unless otherwise specified.
* Slope is based on the coefficient of regression from a linear regression model that is fit on an individual's EDSS values over time, as described in Section 2.4.2.
† RMST calculated at 2 years using time to 24-week confirmed disability progression on the EDSS.
RMST=Restricted mean survival time; IFNb-1a = Interferon beta-1a; IM = intramuscular; SC = subcutaneous; EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion.

Supplementary Table A.2: Group statistics for predicted responders and non-responders to laquinimod at the 50th and 70th percentile thresholds.

	50th percentile threshold*				70th percentile threshold*			
	Responders	Non-responders	Effect size (95% CI)†	*p* value‡	Responders	Non-responders	Effect size (95% CI)†	*p value*‡
Trial contribution:								
ARPEGGIO	159	159			99	219		
Demographics:								
Age (years)	45.09 (7.68)	47.90 (5.56)	-2.81 (-4.29, -1.33)	<0.001	44.80 (8.26)	47.26 (5.95)	-2.46 (-4.29, -0.64)	0.009
Sex (% male)	53.46	54.72	0.95 (0.60, 1.51)	0.910	55.56	53.42	1.09 (0.66, 1.81)	0.808
Height (cm)	172.12 (9.12)	171.36 (9.95)	0.76 (-1.35, 2.87)	0.479	173.23 (9.49)	171.07 (9.51)	2.15 (-0.11, 4.42)	0.064
Weight (kg)	74.71 (17.78)	74.10 (13.47)	0.61 (-2.88, 4.09)	0.733	76.00 (18.13)	73.68 (14.53)	2.31 (-1.77, 6.40)	0.267
Disease duration (years)	6.89 (5.18)	8.76 (6.12)	-1.87 (-3.12, -0.62)	0.004	6.25 (4.65)	8.54 (6.04)	-2.29 (-3.52, -1.07)	<0.001
Disability Scores:								
EDSS	4.70 (0.94)	4.26 (0.91)	0.44 (0.24, 0.64)	<0.001	4.74 (0.89)	4.36 (0.95)	0.38 (0.17, 0.60)	<0.001
FSS-Bowel and Bladder	1.40 (0.97)	1.05 (0.84)	0.35 (0.15, 0.55)	<0.001	1.38 (0.97)	1.16 (0.89)	0.23 (0.00, 0.45)	0.049
FSS-Brainstem	0.90 (0.90)	1.09 (0.95)	-0.19 (-0.40, 0.01)	0.062	0.89 (0.85)	1.05 (0.96)	-0.16 (-0.37, 0.05)	0.147
FSS-Cerebellar	2.41 (0.73)	1.80 (0.87)	0.61 (0.43, 0.79)	<0.001	2.55 (0.69)	1.90 (0.85)	0.64 (0.46, 0.82)	<0.001
FSS-Cerebral	0.92 (0.92)	0.87 (0.87)	0.05 (-0.15, 0.25)	0.619	0.93 (0.91)	0.89 (0.89)	0.04 (-0.17, 0.26)	0.694
FSS-Pyramidal	2.83 (0.67)	2.95 (0.51)	-0.12 (-0.25, 0.01)	0.075	2.82 (0.67)	2.92 (0.56)	-0.10 (-0.26, 0.05)	0.181
FSS-Sensory	1.76 (1.03)	1.71 (1.02)	0.05 (-0.18, 0.28)	0.664	1.71 (1.08)	1.75 (1.00)	-0.04 (-0.29, 0.21)	0.746
FSS-Visual	1.39 (1.40)	0.35 (0.69)	1.04 (0.80, 1.29)	<0.001	1.63 (1.53)	0.53 (0.86)	1.10 (0.78, 1.43)	<0.001
Mean T25FW (sec)	10.24 (9.75)	9.04 (6.57)	1.21 (-0.63, 3.04)	0.198	10.34 (10.15)	9.32 (7.35)	1.03 (-1.22, 3.27)	0.370
Mean 9HPT dominant (sec)	29.95 (13.32)	26.90 (10.93)	3.04 (0.35, 5.73)	0.027	31.16 (14.55)	27.19 (10.88)	3.98 (0.75, 7.21)	0.017
Mean 9HPT non-dominant (sec)	33.85 (20.63)	27.04 (7.60)	6.81 (3.37, 10.25)	<0.001	36.71 (24.30)	27.61 (8.65)	9.10 (4.12, 14.08)	<0.001
MRI metrics:								
Gad count	0.58 (1.80)	0.11 (0.51)	0.47 (0.18, 0.76)	0.002	0.74 (2.21)	0.17 (0.56)	0.56 (0.12, 1.01)	0.014
T2 lesion volume (mL)	7.77 (10.89)	4.03 (5.80)	3.73 (1.81, 5.66)	<0.001	8.35 (11.85)	4.79 (6.94)	3.55 (1.02, 6.09)	0.007
Normalized brain volume (L)	1.44 (0.10)	1.47 (0.10)	-0.03 (-0.05, -0.01)	0.012	1.44 (0.10)	1.47 (0.10)	-0.02 (-0.05, 0.00)	0.063

Values in brackets are standard deviations, unless otherwise specified.
*Percentile threshold for defining responders. The 50th percentile defines responders as the top 50% who are predicted to be most responsive, while the 70th percentile defines them as the top 30%. The non-responders are those who fall below the percentile threshold.
†Effect size is the average difference between responders and non-responders for all covariates except for "sex" which is an odd's ratio (OR).
‡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. Exact p-values for the 50th percentile threshold: Age, $p = 2.33 \times 10^{-4}$; EDSS, $p = 3.00 \times 10^{-5}$; FSS-Bowel and Bladder, $p = 6.09 \times 10^{-4}$; FSS-Cerebellar, $p = 6.54 \times 10^{-11}$; FSS-Visual, $p = 5.21 \times 10^{-15}$; Mean 9HPT non-dominant, $p = 1.34 \times 10^{-4}$; T2 lesion volume, $p = 1.81 \times 10^{-4}$. Exact p-values for the 70th percentile threshold: Disease duration, $p = 2.90 \times 10^{-4}$; EDSS, $p = 6.28 \times 10^{-4}$; FSS-Cerebellar, $p = 1.40 \times 10^{-11}$; FSS-Visual, $p = 7.02 \times 10^{-10}$; Mean 9HPT non-dominant, $p = 4.71 \times 10^{-4}$.
EDSS = Expanded Disability Status Scale; FSS = Functional Systems Score; T25FW = timed 25-foot walk; 9HPT = 9-hole peg test; Gad = Gadolinium-enhancing lesion.

Supplementary References

1. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

2. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81,** 945 (1986).

3. Zhao, L., Tian, L., Cai, T., Claggett, B. & Wei, L. J. Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association* **108,** 527–539 (2013).

4. Royston, P. & Parmar, M. K. 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* **13,** 1–15 (2013).

Bibliography

1. Falet, J.-P. R., Durso-Finley, J., Nichyporuk, B., Schroeter, J., Bovis, F., *et al.* Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning. *Nature Communications* **13,** 5645 (2022).

2. Durso-Finley, J., Falet, J.-P. R., Nichyporuk, B., Arnold, D. L. & Arbel, T. Personalized Prediction of Future Lesion Activity and Treatment Effect in Multiple Sclerosis from Baseline MRI. *Medical Imaging with Deep Learning, PMLR* **172** (2022).

3. Kumar, A., Hu, A., Nichyporuk, B., Falet, J.-P. R., Arnold, D. L., *et al.* Counterfactual Image Synthesis for Discovery of Personalized Predictive Image Markers. *Medical Imaging Computing and Computer Assisted Intervention Society, Workshop on Medical Image Assisted Biomarkers Discovery (MIABID22), LNCS* **13602,** 113–124 (2022).

4. Hu, A., Falet, J.-P. R., Nichyporuk, B., Shui, C. & Arbel, T. Clinically Plausible Pathology-Anatomy Disentanglement in Patient Brain MRI with Structured Variational Priors. *NeurIPS 2022 Machine Learning for Health Workshop* (2022).

5. Nichyporuk, B., Cardinell, J., Szeto, J., Mehta, R., Falet, J.-P., *et al.* Rethinking Generalization: The Impact of Annotation Style on Medical Image Segmentation. *Machine Learning for Biomedical Imaging* **1** (2022).

6. Filippi, M., Bar-Or, A., Piehl, F., Preziosa, P., Solari, A., *et al.* Multiple sclerosis. *Nature Reviews Disease Primers* **4,** 1–27 (2018).

7. Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., *et al.* Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain* **142,** 1858 (2019).

8. McDonald, W. I., Miller, D. H. & Thompson, A. J. Are magnetic resonance findings predictive of clinical outcome in therapeutic trials in multiple sclerosis? The dilemma of interferon-β. *Annals of Neurology* **36,** 14–18 (1994).

9. Weinshenker, B. G. Natural history of multiple sclerosis. *Annals of Neurology* **36 Suppl** (1994).

10. Lublin, F. D., Reingold, S. C., Cohen, J. A., Cutter, G. R., Sørensen, P. S., *et al.* Defining the clinical course of multiple sclerosis. *Neurology* **83,** 278–286 (2014).

11. Koch, M., Kingwell, E., Rieckmann, P. & Tremlett, H. The natural history of primary progressive multiple sclerosis. *Neurology* **73,** 1996–2002 (2009).

12. Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet. Neurology* **17,** 162–173 (2018).

13. Lassmann, H., Horssen, J. V. & Mahad, D. Progressive multiple sclerosis: pathology and pathogenesis. *Nature Reviews Neurology* **8,** 647–656 (2012).

14. Lublin, F., Miller, D. H., Freedman, M. S., Cree, B. A., Wolinsky, J. S., *et al.* Oral fingolimod in primary progressive multiple sclerosis (INFORMS): a phase 3, randomised, double-blind, placebo-controlled trial. *The Lancet* **387,** 1075–1084 (2016).

15. Hawker, K., O'Connor, P., Freedman, M. S., Calabresi, P. A., Antel, J., *et al.* Rituximab in patients with primary progressive multiple sclerosis: results of a randomized double-blind placebo-controlled multicenter trial. *Annals of Neurology* **66,** 460–471 (2009).

16. Giovannoni, G., Knappertz, V., Steinerman, J. R., Tansy, A. P., Li, T., *et al.* A randomized, placebo-controlled, phase 2 trial of laquinimod in primary progressive multiple sclerosis. *Neurology* **95,** e1027–e1040 (2020).

17. Mantia, L. L., Vacchi, L., Pietrantonj, C. D., Ebers, G., Rovaris, M., *et al.* Interferon beta for secondary progressive multiple sclerosis. *The Cochrane database of systematic reviews* **1** (2012).

18. Kapoor, R., Ho, P. R., Campbell, N., Chang, I., Deykin, A., *et al.* Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. *The Lancet Neurology* **17,** 405–415 (2018).

19. Rojas, J. I., Romano, M., Ciapponi, A., Patrucco, L. & Cristiano, E. Interferon beta for primary progressive multiple sclerosis. *The Cochrane database of systematic reviews* (2009).

20. Wolinsky, J. S., Narayana, P. A., O'Connor, P., Coyle, P. K., Ford, C., *et al.* Glatiramer acetate in primary progressive multiple sclerosis: results of a multinational, multicenter, double-blind, placebo-controlled trial. *Annals of Neurology* **61,** 14–24 (2007).

21. Montalban, X., Hauser, S. L., Kappos, L., Arnold, D. L., Bar-Or, A., *et al.* Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *New England Journal of Medicine* **376,** 209–220 (2017).

22. Kappos, L., Bar-Or, A., Cree, B. A., Fox, R. J., Giovannoni, G., *et al.* Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *The Lancet* **391,** 1263–1273 (2018).

23. Temple, R. Enrichment of clinical study populations. *Clinical pharmacology and therapeutics* **88,** 774–778 (2010).

24. Bovis, F., Carmisciano, L., Signori, A., Pardini, M., Steinerman, J. R., *et al.* Defining responders to therapies by a statistical modeling approach applied to randomized clinical trial data. *BMC Medicine* **17,** 1–10 (2019).

25. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359–366 (1989).

26. Tousignant, A., Lemaître, P., Precup, D., Arnold, D. L. & Arbel, T. Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data. *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, PMLR* **102,** 483–492 (2019).

27. Pearl, J., Glymour, M. & Jewell, N. P. *Causal Inference in Statistics: A Primer* (Wiley, 2016).

28. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

29. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81,** 945 (1986).

30. Vegetabile, B. G. On the Distinction Between "Conditional Average Treatment Effects" (CATE) and "Individual Treatment Effects" (ITE) Under Ignorability Assumptions. *Workshop on the Neglected Assumptions in Causal Inference (NACI) at the 38th International Conference on Machine Learning* (2021).

31. Gutierrez, P. & Gérardy, J.-Y. Causal Inference and Uplift Modelling: A Review of the Literature. *Proceedings of The 3rd International Conference on Predictive Applications and APIs, PMLR* **67** (eds Hardgrove, C., Dorard, L., Thompson, K. & Douetteau, F.) 1–13 (2017).

32. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 4156–4165 (2019).

33. Shalit, U., Johansson, F. D. & Sontag, D. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. *Proceedings of the 34th International Conference on Machine Learning, PMLR* **70** (2017).

34. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* **113,** 7353–7360 (2016).

35. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* **113,** 1228–1242 (2018).

36. Filippi, M., Preziosa, P., Langdon, D., Lassmann, H., Paul, F., *et al.* Identifying Progression in Multiple Sclerosis: New Perspectives. *Annals of Neurology* **88,** 438–452 (2020).

37. Fisniku, L. K., Brex, P. A., Altmann, D. R., Miszkiel, K. A., Benton, C. E., *et al.* Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain* **131,** 808–817 (2008).

38. Tintoré, M., Rovira, A., Río, J., Nos, C., Grivé, E., *et al.* Baseline MRI predicts future attacks and disability in clinically isolated syndromes. *Neurology* **67,** 968–972 (2006).

39. Minneboo, A., Barkhof, F., Polman, C. H., Uitdehaag, B. M., Knol, D. L. & Castelijns, J. A. Infratentorial Lesions Predict Long-term Disability in Patients With Initial Findings Suggestive of Multiple Sclerosis. *Archives of Neurology* **61,** 217–221 (2004).

40. Rudick, R. A., Lee, J. C., Simon, J. & Fisher, E. Significance of T2 lesions in multiple sclerosis: A 13-year longitudinal study. *Annals of Neurology* **60,** 236–242 (2006).

41. Dekker, I., Sombekke, M. H., Balk, L. J., Moraal, B., Geurts, J. J., *et al.* Infratentorial and spinal cord lesions: Cumulative predictors of long-term disability? *Multiple Sclerosis Journal* **26,** 1381–1391 (2020).

42. Calabrese, M., Rocca, M. A., Atzori, M., Mattisi, I., Favaretto, A., *et al.* A 3-year magnetic resonance imaging study of cortical lesions in relapse-onset multiple sclerosis. *Annals of Neurology* **67,** 376–383 (2010).

43. Calabrese, M., Poretto, V., Favaretto, A., Alessio, S., Bernardi, V., *et al.* Cortical lesion load associates with progression of disability in multiple sclerosis. *Brain* **135,** 2952–2961 (2012).

44. Kappos, L., Moeri, D., Radue, E. W., Schoetzau, A., Schweikert, K., *et al.* Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. *The Lancet* **353,** 964–969 (1999).

45. Elliott, C., Wolinsky, J. S., Hauser, S. L., Kappos, L., Barkhof, F., *et al.* Slowly expanding/evolving lesions as a magnetic resonance imaging marker of chronic active multiple sclerosis lesions. *Multiple Sclerosis Journal* **25,** 1915–1925 (2019).

46. Popescu, B. F., Frischer, J. M., Webb, S. M., Tham, M., Adiele, R. C., *et al.* Pathogenic implications of distinct patterns of iron and zinc in chronic MS lesions. *Acta Neuropathologica* **134,** 45 (2017).

47. Elliott, C., Belachew, S., Wolinsky, J. S., Hauser, S. L., Kappos, L., *et al.* Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis. *Brain* **142,** 2787–2799 (2019).

48. Calvi, A., Clarke, M. A., Prados, F., Chard, D., Ciccarelli, O., *et al.* Relationship between paramagnetic rim lesions and slowly expanding lesions in multiple sclerosis. *Multiple Sclerosis Journal* (2022).

49. Sastre-Garriga, J., Ingle, G. T., Rovaris, M., Téllez, N., Jasperse, B., *et al.* Long-term clinical outcome of primary progressive MS: predictive value of clinical and MRI data. *Neurology* **65,** 633–635 (2005).

50. Filippi, M., Preziosa, P., Copetti, M., Riccitelli, G., Horsfield, M. A., *et al.* Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology* **81,** 1759–1767 (2013).

51. Hänninen, K., Viitala, M., Paavilainen, T., Karhu, J. O., Rinne, J., *et al.* Thalamic Atrophy Predicts 5-Year Disability Progression in Multiple Sclerosis. *Frontiers in Neurology* **11,** 606 (2020).

52. Tsagkas, C., Naegelin, Y., Amann, M., Papadopoulou, A., Barro, C., *et al.* Central nervous system atrophy predicts future dynamics of disability progression in a real-world multiple sclerosis cohort. *European Journal of Neurology* **28,** 4153–4166 (2021).

53. Rocca, M. A., Sormani, M. P., Rovaris, M., Caputo, D., Ghezzi, A., *et al.* Long-term disability progression in primary progressive multiple sclerosis: a 15-year study. *Brain* **140,** 2814–2819 (2017).

54. Colato, E., Stutters, J., Tur, C., Narayanan, S., Arnold, D. L., *et al.* Predicting disability progression and cognitive worsening in multiple sclerosis using patterns of grey matter volumes. *Journal of Neurology, Neurosurgery & Psychiatry* **92,** 995–1006 (2021).

55. Santos, A. C., Narayanan, S., Stefano, N. D., Tartaglia, M. C., Francis, S. J., *et al.* Magnetization transfer can predict clinical evolution in patients with multiple sclerosis. *Journal of Neurology* **249,** 662–668 (2002).

56. Seewann, A., Vrenken, H., Valk, P. V. D., Blezer, E. L., Knol, D. L., *et al.* Diffusely abnormal white matter in chronic multiple sclerosis: imaging and histopathologic analysis. *Archives of Neurology* **66,** 601–609 (2009).

57. Vertinsky, A. T., Li, D. K., Vavasour, I. M., Miropolsky, V., Zhao, G., *et al.* Diffusely Abnormal White Matter, T2 Burden of Disease, and Brain Volume in Relapsing-Remitting Multiple Sclerosis. *Journal of Neuroimaging* **29,** 151–159 (2019).

58. Dadar, M., Narayanan, S., Arnold, D. L., Collins, D. L. & Maranzano, J. Conversion of diffusely abnormal white matter to focal lesions is linked to progression in secondary progressive multiple sclerosis. *Multiple Sclerosis Journal* **27,** 208–219 (2021).

59. Uphaus, T., Steffen, F., Muthuraman, M., Ripfel, N., Fleischer, V., *et al.* NfL predicts relapse-free progression in a longitudinal multiple sclerosis cohort study: Serum NfL predicts relapse-free progression. *EBioMedicine* **72,** 103590 (2021).

60. Siller, N., Kuhle, J., Muthuraman, M., Barro, C., Uphaus, T., *et al.* Serum neurofilament light chain is a biomarker of acute and chronic neuronal damage in early multiple sclerosis. *Multiple Sclerosis Journal* **25,** 678–686 (2019).

61. Disanto, G., Barro, C., Benkert, P., Naegelin, Y., Schädelin, S., *et al.* Serum Neurofilament light: A biomarker of neuronal damage in multiple sclerosis. *Annals of Neurology* **81,** 857–870 (2017).

62. Kuhle, J., Plavina, T., Barro, C., Disanto, G., Sangurdekar, D., *et al.* Neurofilament light levels are associated with long-term outcomes in multiple sclerosis. *Multiple Sclerosis Journal* **26,** 1691–1699 (2020).

63. Barro, C., Benkert, P., Disanto, G., Tsagkas, C., Amann, M., *et al.* Serum neurofilament as a predictor of disease worsening and brain and spinal cord atrophy in multiple sclerosis. *Brain* **141,** 2382–2391 (2018).

64. Thebault, S., Abdoli, M., Fereshtehnejad, S.-M., Tessier, D., Tabard-Cossa, V. & Freedman, M. S. Serum neurofilament light chain predicts long term clinical outcomes in multiple sclerosis. *Scientific Reports* **10,** 1–11 (2020).

65. Preziosa, P., Rocca, M. A. & Filippi, M. Current state-of-art of the application of serum neurofilaments in multiple sclerosis diagnosis and monitoring. *Expert Review of Neurotherapeutics* **20,** 747–769 (2020).

66. Cantó, E., Barro, C., Zhao, C., Caillier, S. J., Michalak, Z., *et al.* Association between serum neurofilament light chain levels and long-term disease course among patients with multiple sclerosis followed up for 12 years. *JAMA neurology* **76,** 1359–1366 (2019).

67. Gafson, A. R., Jiang, X., Shen, C., Kapoor, R., Zetterberg, H., *et al.* Serum Neurofilament Light and Multiple Sclerosis Progression Independent of Acute Inflammation. *JAMA Network Open* **5,** e2147588 (2022).

68. Bridel, C., Leurs, C. E., van Lierop, Z. Y., van Kempen, Z. L., Dekker, I., *et al.* Serum neurofilament light association with progression in natalizumab-treated patients with relapsing-remitting multiple sclerosis. *Neurology* **97,** e1898–e1905 (2021).

69. Barro, C., Healy, B. C., Liu, Y., Saxena, S., Paul, A., *et al.* Serum GFAP and NfL levels differentiate subsequent progression and disease activity in patients with progressive multiple sclerosis. *Neurology - Neuroimmunology & Neuroinflammation* **10,** 2382–2391 (2023).

70. Martinez-Lapiscina, E. H., Arnow, S., Wilson, J. A., Saidha, S., Preiningerova, J. L., *et al.* Retinal thickness measured with optical coherence tomography and risk of disability worsening in multiple sclerosis: a cohort study. *The Lancet Neurology* **15,** 574–584 (2016).

71. Berek, K., Hegen, H., Hocher, J., Auer, M., Pauli, F. D., *et al.* Retinal layer thinning as a biomarker of long-term disability progression in multiple sclerosis. *Multiple Sclerosis Journal* **28,** 1871–1880 (2022).

72. Sucksdorff, M., Matilainen, M., Tuisku, J., Polvinen, E., Vuorimaa, A., *et al.* Brain TSPO-PET predicts later disease progression independent of relapses in multiple sclerosis. *Brain* **143,** 3318–3330 (2020).

73. Pellegrini, F., Copetti, M., Sormani, M. P., Bovis, F., de Moor, C., *et al.* Predicting disability progression in multiple sclerosis: Insights from advanced statistical modeling. *Multiple Sclerosis Journal* **26,** 1828–1836 (2020).

74. Stühler, E., Braune, S., Lionetto, F., Heer, Y., Jules, E., *et al.* Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC medical research methodology* **20,** 1–15 (2020).

75. Capanna, M., Signori, A. & Sormani, M. P. Is the effect of drugs in progressive MS only due to an effect on inflammation? A subgroup meta-analysis of randomised trials. *Multiple Sclerosis Journal* **28** (2022).

76. Kappos, L., Wolinsky, J. S., Giovannoni, G., Arnold, D. L., Wang, Q., *et al.* Contribution of Relapse-Independent Progression vs Relapse-Associated Worsening to Overall Confirmed Disability Accumulation in Typical Relapsing Multiple Sclerosis in a Pooled Analysis of 2 Randomized Clinical Trials. *JAMA Neurology* **77,** 1132–1140 (2020).

77. Portaccio, E., Fonderico, M., Iaffaldano, P., Pastò, L., Razzolini, L., *et al.* Disease-modifying treatments and time to loss of ambulatory function in patients with primary progressive multiple sclerosis. *JAMA Neurology* **79,** 869–878 (2022).

78. Lizak, N., Malpas, C. B., Sharmin, S., Havrdova, E. K., Horakova, D., *et al.* Association of sustained immunotherapy with disability outcomes in patients with active secondary progressive multiple sclerosis. *JAMA neurology* **77,** 1398–1407 (2020).

79. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., *et al.* Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence* **3** (2021).

80. Jesson, A., Mindermann, S., Gal, Y. & Shalit, U. *Quantifying ignorance in individual-level causal-effect estimates under hidden confounding* in *Proceedings of the 38th International Conference on Machine Learning, PMLR* (2021).

81. Bellman, R. *Dynamic Programming* (Courier Corporation, 2003).

82. Hu, Z., Yang, Z., Salakhutdinov, R. & Xing, E. P. On unifying deep generative models. *International Conference on Learning Representations* (2018).

83. Alcalá, C., Quintanilla-Bordás, C., Gascón, F., Sempere, Á. P., Navarro, L., *et al.* Effectiveness of rituximab vs. ocrelizumab for the treatment of primary progressive

multiple sclerosis: a real-world observational study. *Journal of Neurology* **269,** 3676–3681 (2022).

84. Hyland, M. & Rudick, R. A. Challenges to clinical trials in multiple sclerosis: outcome measures in the era of disease-modifying drugs. *Current opinion in neurology* **24,** 255–261 (2011).

85. Cohen, M., Bresch, S., Rocchi, O. T., Morain, E., Benoit, J., *et al.* Should we still only rely on EDSS to evaluate disability in multiple sclerosis patients? A study of inter and intra rater reliability. *Multiple Sclerosis and Related Disorders* **54** (2021).

86. Bovis, F., Signori, A., Carmisciano, L., Maietta, I., Steinerman, J. R., *et al.* Expanded disability status scale progression assessment heterogeneity in multiple sclerosis according to geographical areas. *Annals of Neurology* **84,** 621–625 (2018).

87. Cutter, G. R., Baier, M. L., Rudick, R. A., Cookfair, D. L., Fischer, J. S., *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* **122,** 871–882 (1999).

88. Cadavid, D., Cohen, J. A., Freedman, M. S., Goldman, M. D., Hartung, H.-P., *et al.* The EDSS-Plus, an improved endpoint for disability progression in secondary progressive multiple sclerosis. *Multiple Sclerosis Journal* **23,** 94–105 (2017).

89. Zhang, J., Waubant, E., Cutter, G., Wolinsky, J. & Leppert, D. Composite end points to assess delay of disability progression by MS treatments. *Multiple Sclerosis Journal* **20,** 1494–1501 (2014).

90. Goldman, M. D., LaRocca, N. G., Rudick, R. A., Hudson, L. D., Chin, P. S., *et al.* Evaluation of multiple sclerosis disability outcome measures using pooled clinical trial data. *Neurology* **93,** e1921–e1931 (2019).

91. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q. & Tsaftaris, S. A. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science* **9,** 220638 (2022).

92. Mueller, S. & Pearl, J. Personalized Decision Making–A Conceptual Introduction. [preprint] arXiv:2208.09558 (2022).

93. Balke, A. & Pearl, J. in *Probabilistic and Causal Inference: The Works of Judea Pearl* (eds Geffner, H. & Dechter, R.) 237–254 (Association for Computing Machinery, New York, NY, United States, 2022).

94. Pawlowski, N., Coelho de Castro, D. & Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems* **33,** 857–869 (2020).

95. Vlontzos, A., Kainz, B. & Lee, C. Estimating categorical counterfactuals via deep twin networks. [preprint] arXiv:2109.01904 (2022).

96. Zhang, J., Tian, J. & Bareinboim, E. Partial counterfactual identification from observational and experimental data. *Proceedings of the 39th International Conference on Machine Learning, PMLR* **162** (2022).