Predictive Powered Inference for Healthcare; Relating Optical Coherence Tomography Scans to Multiple Sclerosis Disease Progression

Jacob Schultz JSCHULTZ@PDX.EDU
[1] *Fariborz Maseeh Department of Mathematics and Statistics*
Portland State University, Portland, Oregon USA
Angeliki Filippatou AFILIPP5@JHMI.EDU
[2] *Department of Neurology, Johns Hopkins University*
Baltimore, Maryland, USA
Brenna McCormack[2] BMCCORM9@JH.EDU
Kathryn C. Fitzgerald[2] FITZGERALD@JHMI.EDU
Shuwen Wei SWEI14@JHU.EDU
[3] *Whiting School of Engineering, Johns Hopkins University*
Baltimore, Maryland USA
Evan Johnson[2] KJOHN224@JHMI.EDU
Grigorios Kalaitzidis GRIGORIOS.KALAITZIDIS@BMC.ORG
Boston University Medical Campus, Boston University
Boston, Massachusetts USA
Elena Vasileiou ELENI.VASILEIOU@MOUNTSINAI.ORG
Department of Neurology, Mount Sinai Hospital
New York, New York USA
Elias Sotirchos[2] ESS@JHMI.EDU
Shiv Saidha[2] SSAIDHA2@JHMI.EDU
Peter Calabresi[2] PCALABR1@JHMI.EDU
Jerry Prince[3] PRINCE@JHU.EDU
Bruno M. Jedynak[1] BRUNO.JEDYNAK@PDX.EDU

Abstract

Prediction-powered inference (PPI and PPI++) is a recently developed statistical method for computing confidence intervals and tests. It combines observations with machine-learning predictions. We use this technique to measure the association between the thickness of retinal layers and the time from the onset of Multiple Sclerosis (MS) symptoms. Further, we correlate the former with the Expanded Disability Status Scale, a measure of the progression of MS. In both cases, the confidence intervals provided with PPI++ improve upon standard statistical methodology, showing the advantage of PPI++ for answering inference problems in healthcare.

1. Introduction

The recently formulated idea (Angelopoulos et al., 2023a,b) of prediction-powered inference (PPI and PPI++) describes how predictions, e.g., machine-learning predictions, can be

combined with observations for providing reduced-size confidence intervals (CIs) and more powerful tests. Suddenly, machine learning is not only a technology that allows for better predictions; it is also a powerful tool for doing better science, allowing for better estimation, more powerful statistical tests, and reducing the sample size in scientific experiments. What are the consequences for the health sciences and healthcare? PPI and PPI++ could dramatically improve scientific experiments measuring disease prevention, diagnosis, treatment, amelioration, or cures. Is it trivial to apply the prediction-powered inference technique to a healthcare problem? It is not since the statistical settings in the aforementioned papers are statistical "plain vanilla" cases of estimating a mean, a median, or a parameter in simple or ridge regression. In contrast, the statistical settings used in healthcare involve control groups, random effects, covariates, and missing data, to cite some challenges. This paper aims to illustrate the use of PPI++ in Multiple Sclerosis (MS) and stimulate the use of this methodology in healthcare.

We now describe the challenge related to the research on multiple sclerosis that we plan to tackle. MS is a devastating neurological disease that affects millions of people worldwide. MS is a disorder of the central nervous system involving white matter and gray matter that is more common in women and typically has its first symptoms in individuals from 20 to 40 years old. Thus, it typically strikes at a time of otherwise maximum productivity, causing physical disabilities, including fatigue, numbness, weakness, gait deficits, postural instability, cognitive dysfunction, and pain. There is no cure for MS, but disease-modifying therapies do exist, and choosing the best disease-modifying therapy is challenging. A better understanding of the fundamental mechanisms of MS would help select existing therapies and develop new ones.

Magnetic resonance imaging (MRI) is used in MS diagnosis, monitoring the development of new or enlarging T2 lesions, and in population studies for clinical research, but is of limited practical use in tracking neurodegeneration in individuals. Clinical testing and MRI are the dominant methods for diagnosing and monitoring MS; in fact, an MS diagnosis should not be made without an MRI unless it is not available. Whereas MRI measurements of inflammation—i.e., white matter lesion load—correlate only modestly to disability progression, MRI measurements of neurodegeneration correlate well with disability progression. The respective roles of white matter lesions and grey matter loss in MS are still a subject of debate.

Optical coherence tomography (OCT) has emerged as a promising imaging modality for assessing MS. MS has been primarily associated with demyelinating white matter lesions in the cerebrum, brain stem, and cervical spine, but widespread neuronal and axonal losses are also expected in MS. Vision has long ago been identified as one of the twelve domains of disability in persons with MS (PwMS), and OCT imaging of the retina has become an important tool in the study of and clinical management of MS. OCT imaging has revealed (Feng et al., 2013) that thinning of both the macular retinal nerve fiber layer (RNFL) and the macular ganglion cell/inner plexiform (GCIP) layer is correlated with white matter brain atrophy. In fact, retinal measures such as RNFL and GCIP thickness derived from OCT reflect a global aspect of the MS disease process, correlating with disability measures and gray matter atrophy. Also, the rates of retinal atrophy are also differentially modulated by different disease-modifying therapies used to treat MS. It is hypothesized that this thinning could be caused by retrograde degeneration of the neuronal cells due to demyelination of

the optic nerve (Shindler et al., 2008). However, macular imaging has revealed potential atrophy of deeper retinal layers, which implies that MS might independently target the retina or that trans-synaptic degeneration occurs. If RNFL and GCIP atrophy is thought to be caused by optic nerve demyelination and retrograde degeneration of its constituent axons (Saidha et al., 2013), atrophy in deeper layers such as the inner nuclear layer (INL) and outer nuclear layer (ONL) may be the result of primary retinal pathological mechanisms and could be analogous to or associated with early gray matter loss in MS (Pietroboni et al., 2019). Therefore, learning the precise timing of the atrophy of each layer could help better refine these hypotheses.

In this paper, we present a pipeline of algorithms using PPI and PPI++, ultimately providing CIs and tests for the timing of each retinal layer's degeneration with respect to the onset of symptoms for PwMS, correcting for degeneration due to aging. We also present CIs for the correlation between the thickness of a collection of retinal layers and the Expanded Disability Status Scale (EDSS), which assesses the severity of MS.

Generalizable Insights about Machine Learning in the Context of Healthcare

1. We show examples in healthcare where confidence intervals are improved using machine learning predictions;
2. We gently introduce the reader to prediction-powered inference (PPI) and PPI++. While the presentations in Angelopoulos et al. (2023a,b) require a high level of sophistication in statistical science due to its generality, we provide use cases in healthcare requiring only a basic knowledge of statistics;
3. We show how PPI and PPI++ can be applied to longitudinal patient data, controlling for the effect of aging;
4. We also show how PPI and PPI++ can be applied for computing a CI for a Pearson correlation coefficient;
5. We discuss under which conditions these techniques actually improve compared to traditional CIs and tests.

2. Methods

2.1. PPI and PPI++ for the estimation of an expected value

We present the PPI and PPI++ estimators of a mean found in Angelopoulos et al. (2023b). This subsection contains no novelty, but it will make the paper self-contained and help establish some necessary notation. Notate $x \in \mathcal{X}$, a feature vector, and $y \in \mathbb{R}$, an outcome. The set $\mathcal{X}$ is arbitrary. The data is composed of two samples:

1. labeled (or paired): (x_i, y_i), independent with same distribution, $x_i \in \mathcal{X}, y_i \in \mathbb{R}, i = 1 \dots n$. In our data, y_i is the observed retinal thickness x_i years after onset of MS for patient i.

2. unlabeled (or unpaired): $(\tilde{x}_i)$, independent with the same distribution, $\tilde{x}_i \in \mathcal{X}, i = 1 \dots N$. x_i and $\tilde{x}_i$ are independent and have the same marginal distribution. In our data, the retinal thickness for patient i at $\tilde{x}_i$ years after onset is missing.

Moreover, we assume we can access a (machine learning) predictor (or regressor), a function $f : \mathcal{X} \to \mathbb{R}$, which has been trained to predict the outcome. In our data, f predicts the missing retinal thickness of the ith patient $\tilde{x}_i$ years after onset of MS. A critical point is that f can be trained using the feature vectors x and $\tilde{x}$ above (unlabeled data) and external data, labeled or not, but cannot depend on the observed outcome y in the labeled data above. Indeed, the observed outcome will be used to *rectify* or de-bias the predictor f. We ask, can we benefit from this predictor for estimating the parameter $\theta = E[y_1]$? We start with the baseline estimator for θ

$$\hat{\theta}_B = \frac{1}{n} \sum_{i=1}^{n} y_i \quad . \tag{1}$$

Note that this estimator is unbiased $E[\hat{\theta}_B] = \theta$ and with variance $V[\hat{\theta}_B] = \frac{\sigma^2}{n}$ where σ^2 is the variance of y_1. The prediction-powered inference estimator is

$$\hat{\theta}_{PP} = \frac{1}{N} \sum_{i=1}^{N} f(\tilde{x}_i) + \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i)) \quad . \tag{2}$$

The first term is the pure machine learning estimator of θ since f is a predictor of the outcome. Only the unlabeled data is used. The second term is a rectifier, which uses the labeled data to correct the bias of f. A straightforward calculation shows that

$$E[\hat{\theta}_{PP}] = \theta \text{ and } V[\hat{\theta}_{PP}] = \frac{\tau^2}{N} + \frac{1}{n}(\sigma^2 + \tau^2 - 2\rho\sigma\tau) \quad , \tag{3}$$

with $\tau^2 = V[f(x_1)]$, and ρ is the correlation between y_1 and $f(x_1)$. Since both $\hat{\theta}_B$ and $\hat{\theta}_{PP}$ are unbiased, the lower variance estimator is preferred. The PPI estimator has a smaller variance than the baseline estimator when the following two conditions are met:

1. N is large, which is common since unlabeled data is easier to obtain than labeled and
2. the variance of the difference between the predictor and the outcome is small.

These conditions are not always fulfilled. The PPI++ is an improved estimator. It is a linear combination of the PPI and baseline, guaranteed to be at least as good as both of them under some mild regularity assumptions. Specifically, consider the family of estimators for $\lambda \in \mathbb{R}$,

$$\hat{\theta}(\lambda) = \lambda \hat{\theta}_{PP} + (1 - \lambda)\hat{\theta}_B \quad . \tag{4}$$

Using linearity, these estimators are unbiased. Their variance is a convex quadratic function of λ which can be minimized analytically and yields the PPI++ estimator

$$\hat{\theta}_{PP++} = \hat{\theta}(\lambda^*), \lambda^* = \frac{N}{n+N} \frac{\rho\sigma}{\tau} \quad , \tag{5}$$

with

$$E[\hat{\theta}_{PP++}] = \theta \text{ and } V[\hat{\theta}_{PP++}] = \frac{\sigma^2}{n} \left(1 - \rho^2 \frac{N}{n(n+N)}\right) \quad . \tag{6}$$

Note that if the predictor is uncorrelated with the outcome ($\rho = 0$) then $\hat{\theta}_{PP++} = \hat{\theta}_B$ but in all the other cases when there is some unlabeled data available ($N \geq 1$), θ_{PP++} is a more powerful estimator than the baseline estimator, even when the predictor is negatively correlated with the outcome. An actual estimator is obtained by replacing the parameters σ, τ, ρ with their empirical estimates using the labeled and unlabeled data.

2.2. PPI++ for the estimation of a mean value with longitudinal data

Consider now a special case: we have access to time-series data of real-valued measurements $z = (z_{ij})$ for subject i at times t_j, and the data is irregular, not all the subjects being observed at the same time. We are interested in a particular time, say s, and estimating θ, the population expected value of the measurement at that time. Then, we organize the data as follows:

1. labeled: subjects having an observation at time s, $(x_i = (z_{ij}, t_j \neq s), y_i = z_{ij}, t_j = s$
2. unlabeled: subjects not having an observation at time s, $\tilde{x}_i = (z_{ij})$

We can then proceed as follows: for the sake of training, remove all points z_{ij} in the data set z at time s, or in a small interval around s, providing $(x, \tilde{x})$. Use this training set to train a predictor f for the measurement value at time s. Then, use the predictions at time s for all the subjects, labeled and unlabeled for computing $\hat{\theta}_{PP++}$. In this case, this estimator allows for using the subjects not having an observation at time s to contribute, together with the ones having an observation at time s, to estimate the population mean at time s. This will be demonstrated in a healthcare setting in Section 5.

2.3. PPI++ for the estimation of a correlation

Consider now the case of estimating a correlation, say a correlation between real-valued measurements notated respectively y and u. We also have access to covariates notated $x \in \mathcal{X}$. In one sample (labeled), we have access to both y and u, while in another sample (unlabeled), we have access to u but not y. In all cases, we have access to the covariate x. The data is as follows:

1. labeled: (x_i, u_i, y_i), independent with the same distribution, $x_i \in \mathcal{X}$, $u_i, y_i \in \mathbb{R}$, $i = 1 \ldots, n$;
2. unlabeled: $(\tilde{x}_i, \tilde{u}_i)$, independent with the same distribution, $\tilde{x}_i \in \mathcal{X}$, $\tilde{u}_i \in \mathbb{R}$, $i = 1 \ldots, N$. (x_i, u_i) and $(\tilde{x}_i, \tilde{u}_i)$ are independent with the same distribution;

As previously, we assume that we have a trained classifier $f : \mathcal{X} \to \mathbb{R}$, $x \mapsto f(x)$, predicting the outcome y. The Pearson correlation coefficient between the measurements y and u is notated $\hat{\rho}(y, u)$, where, with a slight abuse of notation, we notate y the vector (y_i), and u the vector (u_i), $i = 1 \ldots n$. A CI for $\hat{\rho}(y, u)$ can be then be obtained by computing the Fisher transform (inverse hyperbolic tangent transform) $\hat{F}_B = F(\hat{\rho}(y, u))$, where

$$F(\rho) = \operatorname{arctanh}(\rho) = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \quad . \tag{7}$$

Then, for n large enough, $F(\hat{\rho}(y,u))$ is approximately Normally distributed with mean $F(\rho)$ and variance $(n-3)^{-1}$. This is a standard result (Fisher, 1915). A CI for $F(\rho)$ is then constructed using the asymptotic Normal distribution and applying the hyperbolic tangent (the inverse of F) to each bound. This method is the most commonly used for obtaining a CI for a correlation. Let us now use PPI to improve this CI. Consider first the PPI estimator

$$\hat{F}_{PP} = F(\hat{\rho}(f(\tilde{x}),\tilde{u})) + F(\hat{\rho}(y,u)) - F(\hat{\rho}(f(x),u)) \quad , \tag{8}$$

where $f(\tilde{x})$ is the vector $(f(\tilde{x}_i)), i = 1 \dots N$ and $f(x)$ is the vector $(f(x_i)), i = 1 \dots n$. The first term is the estimator provided by the machine learning predictor using the unlabeled data. The last two terms define the rectifier, correcting the bias of the first term. Indeed, since (x,u) has the same distribution as $(\tilde{x},\tilde{u})$, $(f(x),u)$ has the same distribution as $(f(\tilde{x}),\tilde{u})$ and thus $\hat{F}_{PP}$ is unbiased of $F(\rho)$. Moreover, the independence assumptions guarantee that it is asymptotically Normal, using the same standard arguments as the baseline estimator. Unlike the cases in the last two subsections, the variance terms necessary to obtain a CI cannot be computed using empirical estimates. Instead, we use the bootstrapping method. Specifically, for computing the variance of $F(\hat{\rho}(f(\tilde{x}),\tilde{u}))$, we draw with replacement N samples from the data $(f(\tilde{x}),\tilde{u})$ and compute their correlation ρ_1. This is repeated B times (e.g. $B = 1,000$). The empirical variance of $(F(\rho_j)), j = 1 \dots B$ provides the desired estimator of the variance of $F(\hat{\rho}(f(\tilde{x}),\tilde{u}))$. A similar procedure is applied for the other terms in the variance of $\hat{F}_{PP}$. A CI for $\hat{F}_{PP}$ is then computed using the standard procedure for Normal distribution, which is then inverted to provide a CI for ρ. As for estimating a mean, there are cases where the CI for $\hat{F}_{BB}$ is larger than the baseline interval $\hat{F}_B$. Thus, mimicking the construction in (4), we consider the family of estimators, indexed by a parameter $\lambda \in \mathbb{R}$,

$$\hat{F}(\lambda) = \lambda \hat{F}_{PP} + (1-\lambda)\hat{F}_B \quad . \tag{9}$$

By linearity, these estimators are unbiased, and their variance can be minimized, obtaining the PPI++ estimator $\hat{F}(\lambda^*)$, with

$$\lambda^* = \frac{N}{n+N}\frac{\tilde{\rho}\tilde{\sigma}}{\tilde{\tau}} \quad , \tag{10}$$

where

$$\tilde{\sigma}^2 = V[F(\hat{\rho}(y,u))], \tilde{\tau}^2 = V[F(\hat{\rho}(f(x),u))], \rho = Corr(F(\hat{\rho}(y,u)), F(\hat{\rho}(f(x),u))) \quad . \tag{11}$$

Thus,

$$E[\hat{F}_{PP++}] = F(\rho) \text{ and } V[\hat{F}_{PP++}] = \frac{\tilde{\sigma}^2}{n}\left(1 - \tilde{\rho}^2\frac{N}{n(n+N)}\right) \quad . \tag{12}$$

The standard procedure for a Normal distribution is then applied to obtain a CI around $F(\rho)$, which is then inverted, applying the hyperbolic tangent on the bound of the interval to obtain a PPI++ CI for ρ. This method will be demonstrated in Section 5.

3. Related Work

PPI and PPI++ are closely related to other topics in the statistical literature. For example, it is well-known in regression analysis that adding covariates that are predictive of the

outcome but not of the treatment can increase the precision of the inference. There are also close connections with missing data analysis and the design of experiments with multiple measurements; see the related work sections in Angelopoulos et al. (2023a,b). In Section 2.1, bootstrapping techniques can be used to estimate the CI in case the Normal theory does not hold. In section 2.3, permutation tests are an alternative to the Fisher transformation. Note also that the Pearson correlation coefficient can be replaced with Spearman's rank correlation coefficient for added robustness to outliers.

4. Cohort

The cohort comprises over 1,000 PwMS and over 200 healthy controls (HC). The participants are monitored with OCT and the Expanded Disability Status Scale (EDSS). This work focuses on the OCT measurements and their correlation with the EDSS. An estimated year of onset of the MS symptoms is obtained with a patient questionnaire for each subject in the MS group.

We now describe the OCT data and processing pipeline. More than 16,000 scans were acquired longitudinally with Cirrus HD OCT. The processing aims to segment the neuronal layers that make up the macula, which is the central region of the retina, responsible for central vision. Specifically, these neuronal layers are defined by nine boundaries, partitioning an OCT image into ten regions. Traversing from within the eye outwards, these regions are 1) vitreous humor, 2) retinal nerve fiber layer (RNFL), 3) ganglion cell layer and inner plexiform layer (GCL-IPL), 4) inner nuclear layer (INL), 5) outer plexiform layer (OPL), 6) outer nuclear layer (ONL), 7) inner segment (IS), 8) outer segment (OS), 9) retinal pigment epithelium (RPE) complex, and 10) choroid areas. Within some boundaries exist extracellular membranes. Specifically, boundaries associated with the vitreous humor to RNFL, the ONL to IS, and the RPE to choroid contain the inner limiting membrane (ILM), the external limiting membrane (ELM), and Bruch's membrane (BrM), respectively. Figure 1, from Lang et al. (2013) shows an image (A scan and B scan) obtained by the OCT scanner together with a fundus image. It also shows, superimposed, the neuronal layers that constitute the retina. Based on previous results and to simplify the presentation, we have limited the analysis to five neuronal layers known to be the most affected by MS: RNFL, GCL-IPL, INL, ONL, and RPE.

4.1. Segmentation

The retinal layers were segmented using the algorithm described in Lang et al. (2013). In this method, a random forest classifier is built to segment eight retinal layers in macular cube images acquired by OCT. The random forest classifier learns the boundary pixels between layers, producing an accurate probability map for each boundary, which is then processed to finalize the boundaries using a max flow/min cut algorithm.

4.2. ComBat harmonization

The ComBat algorithm Johnson et al. (2007) was used to correct subtle changes in software and hardware over time. ComBat is an empirical Bayes algorithm initially developed for

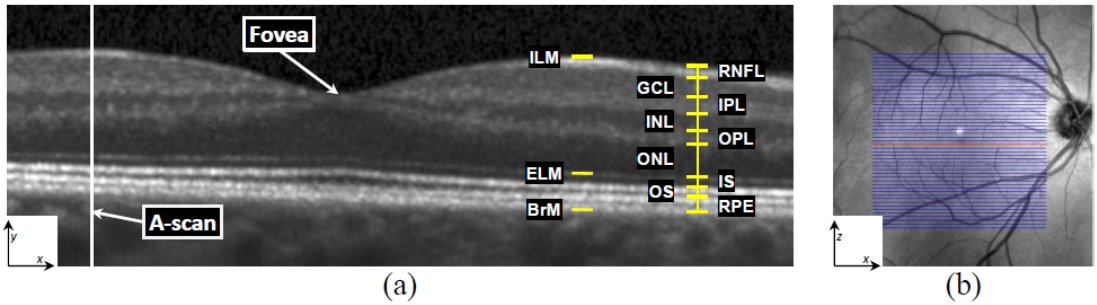

Figure 1: (a) A typical retinal OCT image (B-scan) enlarged with the layers labeled on the right-hand side. Every B-scan consists of a set of vertical scan lines (A-scans). The fovea is characterized by a depression in the retina's surface where the top five (inner) layers are absent. (b) A fundus image with lines overlaid representing the locations of every B-scan within a volume. The red line corresponds to the B-scan in (a).

adjusting batch effects in microarray expression data and later extended to MRI in Fortin et al. (2018).

4.3. Cohort Selection and Data Extraction

The raw data was filtered according to the following procedure. Subjects with no recorded diagnosis (e.g., MS vs. control) were removed, as were observations where the patient's age was not recorded. Only data collected on CIRRUS machines were kept to eliminate a possible source of variance. A few subjects with a first-visit age of less than 18 were also removed.

	MS	Control
Number of Subjects	1030	225
Years In Study	4.70(3.68)	2.16(3.00)
Number Of Visits	4.75(3.99)	3.08(2.61)
Age	45.7(11.9)	36.5(12.9)
Percent Male	24.0	33.8
Est. Years Sick	12.76(9.60)	-
EDSS	2.88[2.5](2.04){2.5}	-

Table 1: Summary of filtered data. Entries employing brackets and/or parentheses follow the format $mean[median](sd)\{\text{Interquartile range}\}$

In most cases, both eyes were measured during the same visit, producing a dataset where pairs of points appear for a given subject at a given time. Since the thinning of the retinal layers may not occur simultaneously in both eyes, these paired measurements were averaged to obtain an analysis at the subject level and not at the eye level. Therefore, the precise interpretation of all thickness measurements used below is "the mean retinal layer thickness of the participant's two eyes for some point in time." All observations where only one eye was measured were removed from the data set.

Optic neuritis (ON) refers to the acute onset of inflammation in the optic nerve . About 50% of PwMS develop optic neuritis and it is often the first sign of MS (Petzold et al., 2022). In this study, we have chosen to exclude observations after the first occurrence of ON. Specifically, all diagnosed ON events were recorded, allowing for the removal of each subject's data during and after their first ON event. OCT scans taken after ON events were removed due to the confounding effects of swelling followed by rapid degeneration in one eye that is often dyssynchronous with the insidious clinical worsening seen in people with MS .

5. Analysis of OCT Data

5.1. Computing predictions

Prediction-powered inference requires choosing a machine-learning prediction method. In our case, the training set is made of time-indexed snapshots, each of dimension five, corresponding to the number of retinal layers. The time is the age of the subject. During inference, one or a few time-indexed snapshots are provided for a given subject. One needs to predict the value of the retinal thickness of each of the five layers at a prescribed age. Many methods are available for inference with longitudinal sequences and Sheetal et al. (2023) provides a useful review of them. We have decided to use ordinary differential equations (ODE). Such a model is fully described by a vector field, here, a smooth mapping $f : \mathbb{R}^5 \to \mathbb{R}^5$ together with initial conditions for each subject

$$\dot{x}_i = f(x_i), x_i(0) = u_i \quad , \tag{13}$$

where i is the index of a subject, $x_i(t) \in \mathbb{R}^5$ and t is the age of the subject. When the vector field and the initial condition are learned, the inference consists of integrating the ODE up to the desired time (or age). The recently developed ODE-RKHS method Lahouel et al. (2022) uses Reproducing Kernel Hilbert Spaces for modeling vector fields and the penalty method of optimization. It shows remarkable performances with noisy data, and asymptotic guarantees are provided. The algorithm is demonstrated on the OCT data on the right of Figure 2. Note that this method uses the data in a multivariate way, such that the prediction for each retinal layer uses the data from all the layers.

5.2. Incorporating Control Subjects Using A Mixed Effect Model

It is a fact that MS subjects experience a thinning of certain layers of the retina, predominantly the GCL-IPL layer, compared to controls, even after correcting for the effect of aging. However, it is unknown how early this phenomenon occurs within the course of MS and how much MS-driven thinning occurs at different points in the disease progression. We

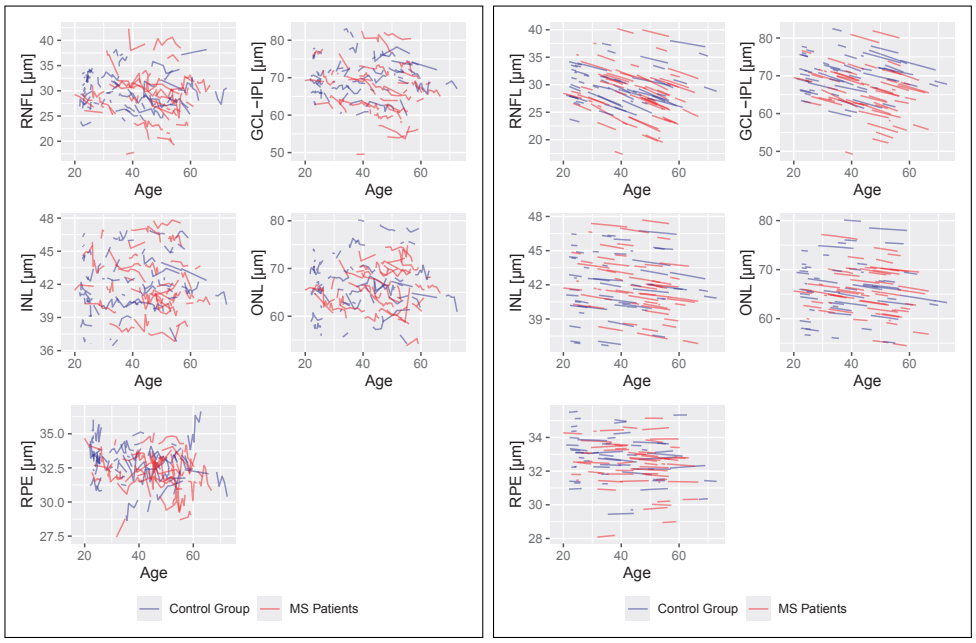

Figure 2: **Left:** Longitudinal data by subject. Each panel shows the thickness of a retinal layer in microns (μm): RNFL, GCL-IPL, INL, ONL, and RPE. 55 randomly selected MS patients are shown alongside all 55 Control subjects. **Right:** Predictions computed with the ODE-RKHS algorithm over the observation time for each subject of the left panel.

aim to answer these questions using OCT data from MS patients that has been age-adjusted by a control group.

The effect of aging on the various retinal layers absent of MS can be measured by finding the overall thickness trends in the control group. Since the data is longitudinal, a reasonable model for this effect in the control population is a mixed-effect linear model with subject-level random component. We hypothesize each of the layers' slope (fixed effect) is negative since retinal layers thin with aging. Figure 3 shows the layer thickness for all the subjects in the control group. Table 2 shows that this thinning is strongly significant in the studied sample for the RNFL and GCL-IPL. Let us fix a number of years, notated s,

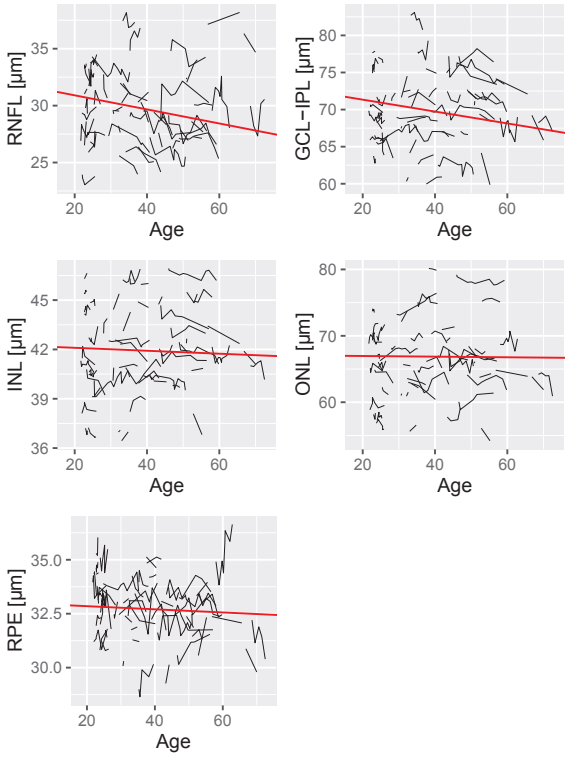

Figure 3: Retinal layer thickness for the control group in microns is shown in black, and linear fixed effects are shown in red.

after the self-reported onset of symptoms and consider $s = 1, 5$, and 10. Consider $\mathcal{M}(s)$, the sub-population of MS subjects having an OCT exam s years after onset, plus or minus one year. Let us consider the GCL-IPL layer. The same analysis will then be repeated with the four other layers. We notate $y_i(t)$ the GCL-IPL layer thickness at age t of a subject in $\mathcal{M}(s)$. To correct the effect of aging, we replace $y_i(t)$ with

$$z_i(t) \leftarrow y_i(t) - g(t) \quad , \tag{14}$$

where $g(t)$ is the expected thickness of a control subject at age t, shown in red in Figure 3.

Layer	Intercept	Slope	Slope P-value
RNFL	32.14	**-0.0618**	0.0021
GCL-IPL	72.94	**-0.0801**	0.0007
INL	42.28	-0.0090	0.493
ONL	67.06	-0.0048	0.857
RPE	32.99	-0.0073	0.493

Table 2: Fitted coefficients for the fixed effect on the control subjects. Significant coefficients at the 5% level are shown in bold

Notate t_i the age at which subject i belongs to $\mathcal{M}(s)$. We are interested in estimating $\theta(s)$, the population average of $z_i(t_i)$, that is the expected layer thickness s years after the beginning of the MS symptoms, compared to controls of the same age. Also, we want to test whether $\theta(s) = 0$ or not, where a nonzero value indicates a difference between the MS and control groups.

5.3. Modeling Age-Adjusted Retinal Thicknesses In MS Patients

5.3.1. Applying PPI++ To Retinal Layer Thickness Modeling

The onset of MS is only recorded at the one-year resolution in our data set, which forces our "time since onset" variable to also be at a one-year resolution. To account for possible rounding errors in these dates, a one-year window around s was established for analysis. Therefore our analysis at 5 years after onset includes data $[4, 6]$ years after onset, and similar for $s = 1$, 10. The baseline predictor in (1) is therefore the mean thicknesses of all subjects who were observed in the $[s-1, s+1] = \mathcal{S}$ window. There is a possibility of subjects being sampled multiple times in this window, so only the observation closest to s was used for each subject.

For prediction-powered inference, predictions are also needed at these points. However, any prediction must be generated independently of the data being analyzed, so the prediction method outlined in Section 5.1 was only trained on data outside of $\mathcal{S}$. Moreover, at least two observations are required to be used in training, and at least one observation is needed to generate the predictions. Therefore, the PPI++ parameters n and N can be computed this way:

- n: The number of subjects with at least one point in $\mathcal{S}$ and at least one outside of $\mathcal{S}$;
- N: The number of subjects with no points in $\mathcal{S}$ and at least one point outside of $\mathcal{S}$.

The $N + n$ predictions and n observations are used with equations (5) and (6) to obtain thickness point estimates and variances, which can be used to generate the CIs shown in the next section. The values of N, n, and λ^* are shown in Table 3. In all cases, N is sufficiently large and the predictor is sufficiently accurate (λ^* is sufficiently large in absolute value), so both PPI and PPI++ CIs will be tighter than their baseline counterparts.

s	Layer	Estimate	95% CI LB	95% CI UB	CI Width	λ^*	N	n
1	RNFL	**-0.491**	-0.886	-0.097	0.788	0.724	497	126
5	RNFL	**-1.497**	-1.843	-1.151	0.692	0.661	423	204
10	RNFL	**-1.972**	-2.315	-1.628	0.686	0.704	466	200
1	GCL-IPL	**-3.022**	-3.554	-2.490	1.065	0.727	497	126
5	GCL-IPL	**-4.060**	-4.596	-3.523	1.072	0.658	423	204
10	GCL-IPL	**-4.735**	-5.272	-4.199	1.073	0.687	466	200
1	INL	0.074	-0.139	0.287	0.426	0.783	497	126
5	INL	0.032	-0.197	0.260	0.457	0.663	423	204
10	INL	-0.107	-0.336	0.123	0.459	0.700	466	200
1	ONL	0.140	-0.324	0.604	0.928	0.820	497	126
5	ONL	-0.027	-0.495	0.441	0.937	0.663	423	204
10	ONL	**-0.488**	-0.942	-0.035	0.907	0.681	466	200
1	RPE	**-0.387**	-0.535	-0.239	0.297	0.768	497	126
5	RPE	**-0.317**	-0.441	-0.193	0.248	0.627	423	204
10	RPE	**-0.266**	-0.387	-0.146	0.241	0.663	466	200

Table 3: Summary of PPI++ age-adjusted retinal thickness analysis with $\alpha = 5\%$ significant estimates in bold

5.3.2. Thickness Modeling Results

Modeled thicknesses are shown and in Figure 4, comparing the conventional mean calculation and both types of PPI. There is an improvement between the CI obtained with the conventional (or baseline) and PPI estimates (PPI or PPI++). They are narrower but still statistically valid. We can also see a subtle but clear improvement in the widths of the CIs for the PPI++ versus the PPI. As previously stated, the PPI++ method provides smaller CIs than the other methods, but this example illustrates that the difference can be significant in some cases and minimal in other situations.

All methods indicate that the RNFL, GCL-IPL, and RPE intervals are below the horizontal line at zero, indicating that these layers are significantly thinner in MS patients than in Control patients 1, 5, and 10 years after the onset of MS. In the case of GCL-IPL, five years after onset, MS patients' measurements are about 4 microns thinner than those of control subjects of the same age. The details are provided in Table 3, with significance shown in bold face. MS also impacts the RNFL layer. A significant difference is also observed for the RPE complex, five years after onset. However, the average difference of less than half a micron might not be easily detectable in individual subjects and thus might not be clinically relevant.

We now describe PPI++ results and evaluate them at 1, 5, and 10 years after the onset of MS, which can be seen in the bottom right of Figure 4. Here, we see the difference between the MS and control groups growing with disease progression in both RNFL and

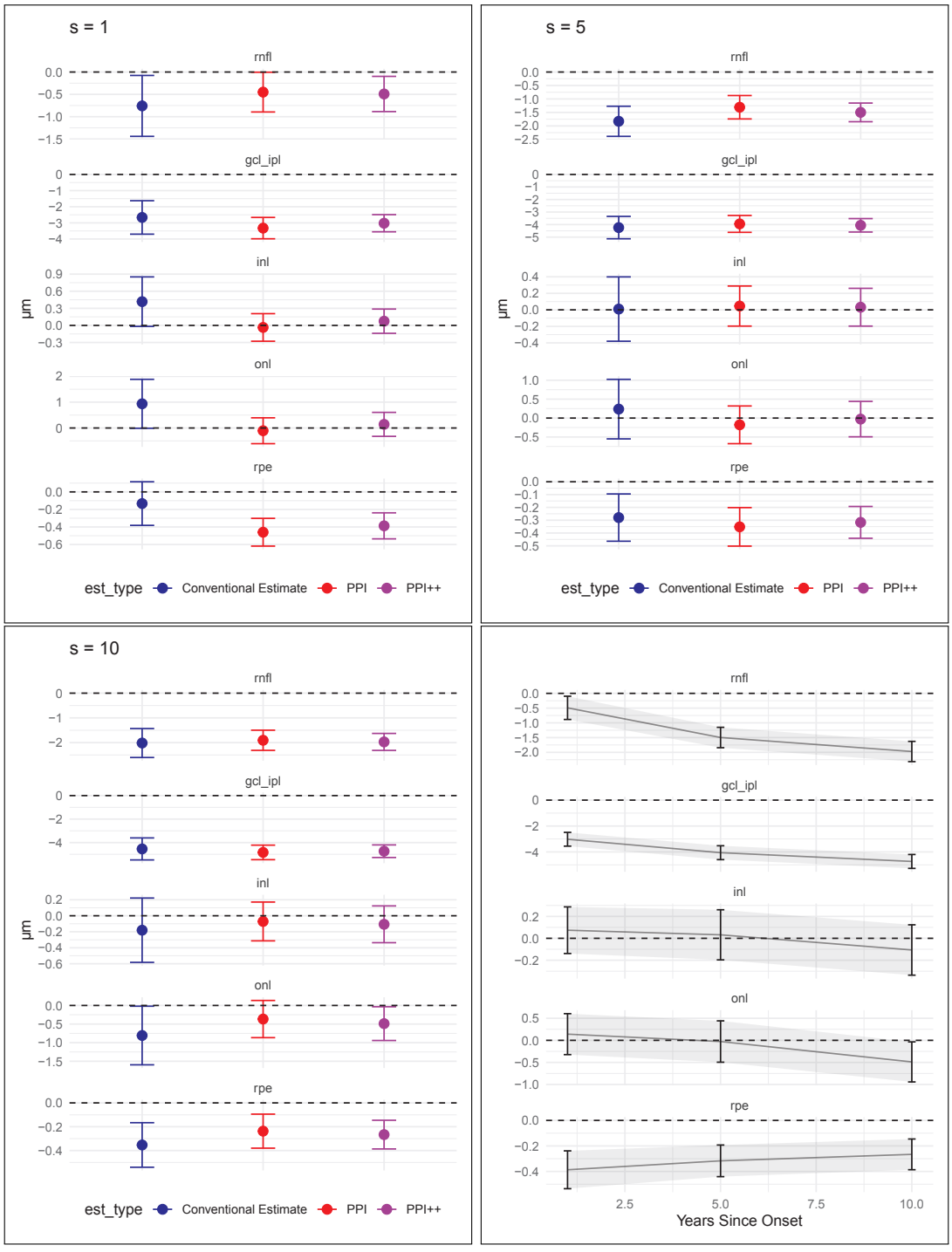

Figure 4: **Top Left:** Modeled age-adjusted (MS - Control) thicknesses in microns [0,2] years after onset of MS ($s = 1$) with 95% CI's **Top Right:** Similar, for $s = 5$ **Bottom Left:** Similar, for $s = 10$ **Bottom Right:** PPI++ estimated thicknesses for $s = 1, 5, 10$ with 95% CI's .

s	Estimate	95% CI LB	95% CI UB	CI Width	λ^*	N	n
1	-0.090	-0.255	0.080	0.334	0.048	6	112
5	**-0.246**	-0.362	-0.123	0.240	0.134	16	217
10	**-0.202**	-0.322	-0.075	0.248	0.257	66	215

Table 4: Summary of PPI++ Correlation Analysis Between EDSS and GCL-IPL

GCL-IPL. The RPE measurements are only slightly below the zero mark. The INL and ONL layers are not found to be impacted by MS.

5.4. Modeling The Correlation Between Retinal Thicknesses And EDSS Functional System Scores In MS Patients

5.4.1. Applying The PPI++ Correlation Estimator

Correlations were modeled using the PPI++ method outlined in Section 2.2 with $B = 2000$ bootstrap iterations. The data used was the same OCT data, predictions, and $\mathcal{S}$ definition used in Section 5.2. We added the Extended Disability Score (EDSS). The goal was to investigate the correlation between the EDSS and the GCL-IPL retinal layer thickness. Analyses for the other four retinal layers were also carried out and will be presented elsewhere. Predictions on missing EDSS data have not been made, so the availability of this data limits any correlation calculations. The critical numbers n and N for carrying on the PPI and PPI++ analysis are computed the same way as in Section 5.3.1 with the additional restriction that the EDSS score must also be available within $\mathcal{S}$ for all $n + N$ subjects.

The sample sizes are given in Table 4. The above data constraints forced N to be alarmingly small, especially one year after the onset of symptoms ($s = 1$) time-point. Thus, the PPI CIs will not be useful. However, the PPI++ estimate approaches the conventional mean by Equation 4. The small-N problem is covered in detail in the Discussion (Section 6.2).

5.4.2. Correlation Modeling Results

We expect the retinal layers to thin and the EDSS score to increase with MS disease progression. Correlation estimates and 95% CI's are given in Figure 5 for the baseline, PPI, and PPI++ methods. The thickness of the GC-IPL layer is negatively correlated with the EDSS score at 5 and 10 years after onset. This correlation is not significant at one year after the onset of the symptoms of MS. These findings are obtained with the baseline CIs. They are confirmed and slightly improved (smaller p-value due to tighter CI) with the PPI++ intervals. Note that the PPI intervals are not helpful since N is too small. Also, the sparsity of the measurements (EDSS not always available in the time window of interest) did not allow for the PPI++ to improve the baseline CIs significantly.

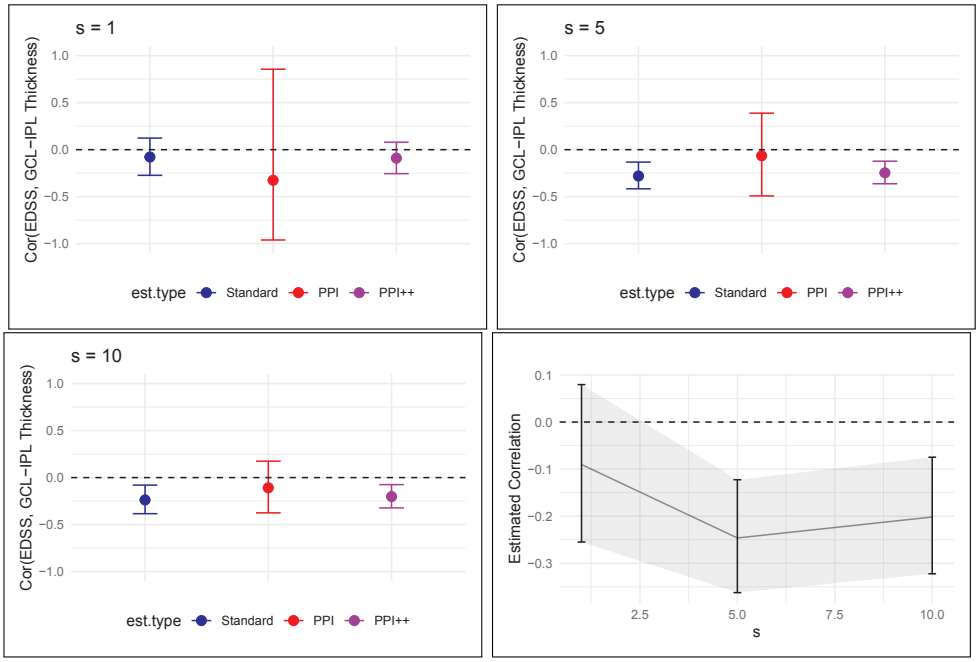

Figure 5: **Left:** Estimated Correlation Between EDSS And GCL-IPL Thickness [0,2] Years After Onset of MS ($s = 1$) With 95% CI's **Top Right:** Similar, for $s = 5$ **Bottom Left:** Similar, for $s = 10$ **Bottom Right:** PPI++ Estimated Correlations Between GCL-IPL And EDSS Scores 1, 5, and 10 Years After Onset of MS With 95% CI's

6. Discussion

6.1. Scientific Results

Prediction-powered inference applied to age-adjusted retinal thickness showed statistically significant thinning in the MS population without an episode of ON compared to the control group, with GCL-IPL being impacted the most. In the cases of RNFL and GCL-IPL, these differences became more pronounced as the disease progresses. The RPE complex was consistently thinner with a sub-micron effect size throughout disease progression. The estimated thinning in RNFL is 0.17 microns per year more than controls (0.06 in average) and GCL-IPL is 0.21 microns per year more than controls (0.08 in average) in PwMS without ON. Extrapolating these results, it is plausible that the thinning of the GCL-IPL due to MS occurs before the onset of the symptoms of MS. The correlation between the GCL-IPL and the EDSS is significantly negative at 5 and 10 years after onset.

This confirms the findings in Rothman et al. (2019), assessing that lower baseline total retinal thickness measured by OCT significantly predicts higher disability at ten years, even after accounting for baseline disability status. The thickness results are consistent with the findings in Sotirchos et al. (2020). However, this later work suggests that a more refined analysis would require distinguishing between progressive and relapsing-remitting MS. In Martinez-Lapiscina et al. (2016), the authors study the peripapillary region of the

retina, an elliptical annulus that extends from the optic disc boundary and is disjointed from the macula. They provide evidence of the usefulness of monitoring pRNFL thickness by OCT for predicting the risk of disability worsening measured by EDSS with time in PwMS. This finding is consistent with our analysis. It would remain to find out whether pRNFL or (macular) GC-IPL would be more correlated with a worsening of MS disability measured by EDSS. Also, both Saidha et al. (2015) and Cordano et al. (2022) find that OCT measurements positively correlate with the cortical gray matter volume. We plan to verify this in the near future with our cohort and the PPI++ methodology.

6.2. General Remarks On Implimenting PPI++

PPI++ is an optimized linear combination of a conventional estimate (e.g. a sample mean) and a PPI estimate, which is driven by predictions that have been bias-corrected by observations. While PPI++ will provably produce predictions at least as good as either method, it is useful to investigate the amount of improvement as a function of data quantity. In Figure 4, CIs improve noticeably when moving from the conventional estimate to PPI, but PPI++ shows little improvement compared to PPI. In contrast, Figure 5 showcases a scenario where PPI does worse than the conventional estimate. In the case of correlations, we have strict data requirements, leading to many subjects being excluded from analysis and making the number of unlabeled data very small. The PPI++ estimate is then similar to the conventional estimate.

However, and most importantly, the PPI++ CI can improve upon the baseline CI only when the predictions correlate with the outcome. The larger the correlation (positive or negative), the tighter the CI.

In summary, for deciding whether or not to apply the PPI and PPI++ techniques in the context of healthcare, the following considerations and limitations apply:

1. The PPI++ CI is strictly superior to the PPI one and, therefore, should always be preferred;
2. Data both labeled and unlabeled need to be available;
3. Enough unlabeled data is necessary for the PPI++ CI to be tighter than the baseline one;
4. The machine learning prediction algorithm should be trained after hiding the labeled data used in the rectifier.
5. The absolute value of the correlation between the predictions and the outcome should be used when choosing among several prediction algorithms;

6.3. Future Work and Concluding Remarks

We have used a prediction algorithm that has performed well with noisy longitudinal data. However, other algorithms are available. We plan to compare and combine prediction algorithms to improve the CIs further. In the context of MS, it is important to include MRI measurements, OCT Angiography, and visual acuity. We plan to extend the current analysis in this direction.

We have made two contributions to PPI and PPI++ to help make these techniques more useful in healthcare. We have shown how to use a control group when estimating a population mean and proposed a PPI/PPI++ CI for a correlation. However, much more is needed, including mixed effect models and survival analysis.

In conclusion, the setting presented in this paper provides further examples of how machine learning can be useful in healthcare science.

Acknowledgements

We would like to thank our three anonymous reviewers for their helpful comments. This work was partially supported by the NIH under NEI grant R01-EY032284 (PI: J.L. Prince, co-PI: B.M. Jedynak). The grant R01NS082347 supported data acquisition. The work at Portland State University was funded in part by NSF RTG 2136228.

Disclosures

Dr. Saidha has received consulting fees from Medical Logix for the development of CME programs in neurology; has served on scientific advisory boards for Biogen, Clene, Genentech, Impaact Bio, Horizon Therapeutics, Amgen, Novartis Pharmaceuticals Corporation, and ReWind Therapeutics; has been a consultant for Genentech, Biogen, InnoCare Pharma, JuneBrain, Kiniksa, LAPIX Therapeutics, Setpoint Medical and Novartis Pharmaceuticals Corporation; is the Principal Investigator of Investigator-initiated studies funded by Biogen, Genentech, and Novartis Pharmaceuticals Corporation; has received support from the Race to Erase MS foundation; has received equity compensation for consulting from JuneBrain; was the site Investigator of trials sponsored by Clene and MedDay; and is the site Investigator of trials sponsored by LAPIX Therapeutics and Novartis Pharmaceuticals Corporation.

Dr. Calabresi has received consulting fees from Novartis and Lilly for serving on SABs, and is PI on a grant to JHU from Genentech.

Kathryn Fitzgerald received research funding from the NIH, National Multiple Sclerosis Society, the Department of Defense Congressionally Directed Medical Research Program, and the International Progressive Multiple Sclerosis Alliance. She is a member of the Data and Safety Monitoring Board for A Trial of Bile Acid Supplementation in Patients With Multiple Sclerosis, Comparative Effectiveness Trial of COVID-19 Testing Modalities (C-FORWARD), and VIRTual vs UsuAL in-office care for MS (VIRTUAL-MS).

Author Contributions

An author contribution table is given below:

Contribution	JS	BJ	JP	PC	SS	AF	BM	KF	SW	EJ	GK	EV	ES
Analysis development	✓	✓											
Data Collection				✓	✓	✓	✓				✓	✓	✓
Provided data or tools		✓						✓					
Analysis	✓	✓											
Writing and editing	✓	✓	✓	✓	✓								✓
Provided funding		✓	✓	✓	✓								
Data selection							✓	✓	✓	✓		✓	

JS: Jacob Schultz, BJ: Bruno Jedynak, JP: Jerry Prince, PC: Peter Calabresi, SS: Shiv Saidha, AF: Angeliki Filippatou, BM: Brenna McCormack, KF: Kathryn C. Fitzgerald, SW: Shuwen Wei, EJ: Evan Johnson, GK: Grigorios Kalaitzidis, EV: Elena Vasileiou, ES: Elias Sotirchos

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Christian Cordano, Bardia Nourbakhsh, Hao H Yiu, Nico Papinutto, Eduardo Caverzasi, Ahmed Abdelhak, Frederike C Oertel, Alexandra Beaudry-Richard, Adam Santaniello, Simone Sacco, et al. Differences in age-related retinal and cortical atrophy rates in multiple sclerosis. *Neurology*, 99(15):e1685–e1693, 2022.

Liguo Feng, Jie Shen, Xiaohong Jin, Jiuke Li, and Yumin Li. The evaluation of the retinal nerve fiber layer in multiple sclerosis with special-domain optical coherence tomography. *Ophthalmologica*, 230(3):116–120, 2013.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167:104–120, 2018.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Kamel Lahouel, Michael Wells, Victor Rielly, Ethan Lew, David Lovitz, and Bruno M Jedynak. Learning nonparametric ordinary differential equations from noisy data. *arXiv preprint arXiv:2206.15215*, 2022.

Andrew Lang, Aaron Carass, Matthew Hauser, Elias S Sotirchos, Peter A Calabresi, Howard S Ying, and Jerry L Prince. Retinal layer segmentation of macular oct images using boundary classification. *Biomedical optics express*, 4(7):1133–1152, 2013.

Elena H Martinez-Lapiscina, Sam Arnow, James A Wilson, Shiv Saidha, Jana Lizrova Preiningerova, Timm Oberwahrenbrock, Alexander U Brandt, Luis E Pablo, Simone Guerrieri, Ines Gonzalez, et al. Retinal thickness measured with optical coherence tomography and risk of disability worsening in multiple sclerosis: a cohort study. *The Lancet Neurology*, 15(6):574–584, 2016.

Axel Petzold, Clare L Fraser, Mathias Abegg, Raed Alroughani, Daniah Alshowaeir, Regina Alvarenga, Cécile Andris, Nasrin Asgari, Yael Barnett, Roberto Battistella, et al. Diagnosis and classification of optic neuritis. *The Lancet Neurology*, 21(12):1120–1134, 2022.

Anna M Pietroboni, Laura Dell'Arti, Michela Caprioli, Marta Scarioni, Tiziana Carandini, Andrea Arighi, Laura Ghezzi, Giorgio G Fumagalli, Milena A De Riz, Paola Basilico,

et al. The loss of macular ganglion cells begins from the early stages of disease and correlates with brain atrophy in multiple sclerosis patients. *Multiple Sclerosis Journal*, 25(1):31–38, 2019.

Alissa Rothman, Olwen C Murphy, Kathryn C Fitzgerald, Julia Button, Eliza Gordon-Lipkin, John N Ratchford, Scott D Newsome, Ellen M Mowry, Elias S Sotirchos, Stephanie B Syc-Mazurek, et al. Retinal measurements predict 10-year disability in multiple sclerosis. *Annals of Clinical and Translational Neurology*, 6(2):222–232, 2019.

Shiv Saidha, Elias S Sotirchos, Jiwon Oh, Stephanie B Syc, Michaela A Seigo, Navid Shiee, Chistopher Eckstein, Mary K Durbin, Jonathan D Oakley, Scott A Meyer, et al. Relationships between retinal axonal and neuronal measures and global central nervous system pathology in multiple sclerosis. *JAMA neurology*, 70(1):34–43, 2013.

Shiv Saidha, Omar Al-Louzi, John N Ratchford, Pavan Bhargava, Jiwon Oh, Scott D Newsome, Jerry L Prince, Dzung Pham, Snehashis Roy, Peter Van Zijl, et al. Optical coherence tomography reflects brain atrophy in multiple sclerosis: a four-year study. *Annals of neurology*, 78(5):801–813, 2015.

Abhishek Sheetal, Zhou Jiang, and Lee Di Milia. Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers. *Applied Psychology*, 72(3):1339–1364, 2023.

Kenneth S Shindler, Elvira Ventura, Mahasweta Dutt, and Abdolmohamad Rostami. Inflammatory demyelination induces axonal injury and retinal ganglion cell apoptosis in experimental optic neuritis. *Experimental eye research*, 87(3):208–213, 2008.

Elias S Sotirchos, Natalia Gonzalez Caldito, Angeliki Filippatou, Kathryn C Fitzgerald, Olwen C Murphy, Jeffrey Lambe, James Nguyen, Julia Button, Esther Ogbuokiri, Ciprian M Crainiceanu, et al. Progressive multiple sclerosis is associated with faster and specific retinal layer atrophy. *Annals of neurology*, 87(6):885–896, 2020.