Quantitative Understanding of the Longitudinal Relationship between Short-term MRI outcomes and Long-term Clinical Outcomes Measures in Multiple Sclerosis

Matthew Wiens[1], Amir-Hadi Maghzi[2], Naveen Mangal[2], Kumar Kandadi Muralidharan[2], Himanshu Naik[2], Jonathan French[1], Francois Gaudreault[2]
[1]*Metrum Research Group, Tariffville, CT, USA,* [2]*Biogen Inc., Cambridge, MA, USA*

Abstract

Objectives: Multiple sclerosis is a relapsing-remitting autoimmune disease which is characterized by distinct episodes of acute neurological worsening (relapse) and formation of new lesions on brain magnetic resonance imaging. The relation of relapse rate and formation of new brain lesions on MRI is not fully established.

The objective of this analysis was to perform a model-based meta-analysis with individual level data of the placebo arm across multiple clinical trials to inform the underlying trends and quantify the variability within the disease process and across MRI endpoints. The primary predictive goal is making long-term predictions (e.g. 2 year relapse rates) from short-term MRI endpoints and clinical data to guide evaluation and decision making for future phase II trials of novel compounds in MS.

Methods: Data for this analysis included patient-level data from eleven studies, with durations up to 2 years, in both secondary progressive MS (SPMS) and relapsing-remitting MS (RRMS) patients. The generative model was structured with latent variables linking observed longitudinal outcomes, analogous to an item response theory model. The latent variable represented an unobservable disease state which was predicted by covariates and population parameters and observed with uncertainty by longitudinal outcome measures. Gaussian processes were used as a non-parametric method to describe each individual's disease activity over time and creates temporal correlation across endpoints.

We focused on out-of-sample prediction to measure the predictive performance on new data and demonstrate that the model is generalizable. The test arm consisted of half of the patients with >1.75 years of data; the data for these patients were truncated to 6 months of data for the model fitting, with the data after 1 year used for predictive evaluation. The markers of clinical disease activity were annualized relapse rate (ARR) and the MRI endpoints included: new/newly enlarging (N/NE) T2 lesions, T1 gadolinium enhancing (Gd+ T1) lesions and T2 lesion volume.

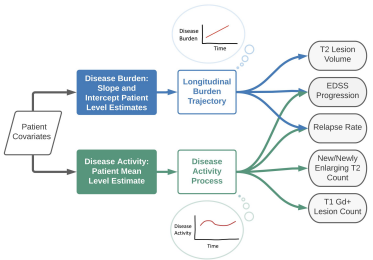
Results: The analysis set comprised 1999 patients, with 27,091 total MRI observations. In the holdout set, the predicted year 2 ARR between the 4th to 1st quartile had a ratio of 4.3 (0.883 to 0.205), compared to an observed ratio of 6.67 (0.859 to 0.127). Patients with different types of MS (RRMS vs SPMS) had substantially different predicted year 2 ARR; for RRMS the mean ARR was 0.558 and a 95% CI of 0.213-1.31 and for SPMS the mean ARR was 0.269 with a 95% CI of 0.134 - 0.597, despite disease population not being included as a covariate in the model. Simulations of patients with a range of baseline to 6 months measurements showed the long-term predictive power, for example comparing a patient with zero Gd+ T1 lesions and zero N/NE T2 lesions at month 6 to a patient with one of each type changed the simulated year 2 ARR from 0.19 to 0.33.

Conclusions: A latent variable model with non-parametric functions of the disease trajectory is able to predict long term disease activity from short term data. This model allows for future incorporation of medication effects, both from historical trials and compounds under development.

Methods

This model-based meta analysis used individual patient level data from the placebo and standard of care arms from eleven clinical trials in both SPMS (two studies) and RRMS (nine studies) to jointly model four longitudinal endpoints and a binary endpoint. Unlike typical meta-analyses which use arm level data, using patient level data in this modeling allows for predictions at an individual level for future trials, inferences about covariates, and avoids the ecological fallacy. Individual disease trajectories were modeled to describe inter-patient variability. The model was set up in a Bayesian framework, allowing for natural incorporation of uncertainty in the model parameters and predictions.

Figure 1: The arrows indicate the structural relationships between different data, parameters, and outcomes in the model. The disease burden latent variable flows across the top of the diagram, and the disease activity latent variable flows across the bottom.



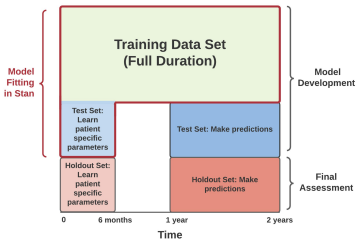
A Bayesian network model, which in the current form shares characteristics with a Generalized Linear Mixed Model (GLMM), was fit to link together the covariates and outcomes. It used two latent variables; one latent variable corresponded to total disease burden and the other corresponded to short term disease activity [1].

The disease burden latent variable was estimated as a linear function of time, with patient specific parameters estimated using a hierarchical model with covariates which partially pools information across patients. The disease activity latent variable was composed of a hierarchial mean level, with an additive Gaussian process [2]. The Gaussian process introduced a non-linear trajectory, which allowed for periods of greater activity and periods of lower activity, and is interpreted as a patient specific trajectory of inflammatory MS activity. Covariates included for both latent variables were age, sex, and if a patient was Japanese.

The typical level of disease burden, the typical rate of change of disease burden, and typical average disease activity all depended on the covariates, explaining some inter-patient and inter-trial variability. Model evaluation focused on out-of-sample predictive performance. We used a test set of patients where their outcome data after 6 months was held out and their year-two data used for model evaluation. A flexible model could fit the observed data well, but also overfit the individual trajectories to the noisy observed data and then predict poorly into a time period when data is unobserved. Out-of-sample prediction for model evaluation validates the ability of the model to make accurate predictions in new circumstances.

The latent variable model was implemented in Stan, version 2.21 [3], using the default No-U-Turn-Sampler with Hamiltonian Monte Carlo method, through the Rstan interface [4].

Figure 2: Data is split into 3 sets to evaluate predictive power on data not used to fit the model. 20% of patients in studies with >100 patients are in the green final holdout set. Of the remaining patients, 50% of the patients with a duration of 1.75 years or longer are in the blue test set.



Key Components of the Statistical Model

The disease burden was defined as:

$$f_i(t) = f_{0,i} + f_{s1,i}\,t$$
$$\log(\text{T2 Volume}_i(t)) \sim \mathcal{T}(f_i(t)), \sigma_1, \text{df} = 5)$$
$$f_{0,i}(t) \sim \mathcal{N}(\mathbf{x}_i\beta_1, \omega_1)$$
$$f_{s1,i}(t) \sim \mathcal{N}(\mathbf{x}_i\beta_2, \omega_2)$$

The disease activity (acute, inflammatory aspects of MS) was defined as:

$$g_i(t) = g_{0,i} + \text{GP}(t; 0, K(\alpha, \rho))$$
$$\text{T1}_i(t) \sim \text{Neg-Bin}(\exp(g_i(t)), \phi_1)$$
$$\text{New T2}_i(t) \sim \text{Neg-Bin}(\exp(\kappa_0 + \kappa_1 \times g_i(t)), \phi_2)$$
$$\text{Hazard of Relapse}_i(t) = h_0 \times \exp(\kappa_2 + \kappa_3 \times g_i(t))$$
$$\text{Hazard of EDSS Progression}_i(t) = h_1 \times \exp(\kappa_4 + \kappa_5 \times g_i(t))$$

- $f_i(t)$ is the latent process for patient i (and directly relates to T2 lesion volume).
- $g_i(t)$ is the short-term disease activity process for patient i (and directly relates to Gd+ T1 lesion counts and new T2 lesions).
- $f_{0,i}$ is an intercept random effect, with a normal distribution and covariate effects.
- $f_{s1,i}$ is a slope random effect, with a normal distribution and covariate effects. Note that the slope was allowed to be negative.
- $g_{0,i}$ is a random effect for the expected short-term disease activity, with a normal distribution and covariate effects.
- $\mathcal{T}$ is the student-T distribution, parameterized with location, scale, and degrees of freedom.
- κ are coefficients relating the latent variables to observed outcomes.
- h_0 and h_1 are baseline hazard rates for relapses, and EDSS progression, respectively. For computational purposes, hazard was averaged over time intervals (typically 60 days), with the hazard evaluated at the midpoint of the interval.
- ϕ are dispersion parameters for the negative-binomial distribution.
- $\mathbf{x}_i$ is the covariate vector for patient i, with an included intercept term.
- β are covariate effects on the random effects.
- ω are standard deviations of the random effects.
- t is the study day.
- GP is a gaussian process, with mean zero, evaluated at times t with kernel function K. The kernel function had its own parameters which depend on the choice of kernel function. The kernel parameters were not patient specific.

Results

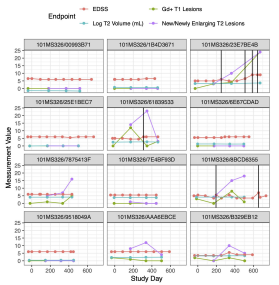
Model validation was done with two goals in mind. First, to show the accuracy of the model on held out data that was not used to fit the model (Table 2). Second, when applying the model to new circumstances and data that were not used to fit the model, to assess whether the model produced reasonable inferences that conform to general clinical understanding and previous work (Figures 5 and 6). Finally, we applied the model to make predictions about different circumstances relevant to trial design (Figures 4 and 6).

Data

Table 1: Description of the 11 studies used in the modeling. Duration is summarized as mean (sd). There were 9 RRMS and 2 SPMS studies included. The studies included both phase 2 studies of a duration of approximately half a year, and phase 3 studies of approximately 2 years in duration. Some of the studies were placebo controlled and others only used active controls.

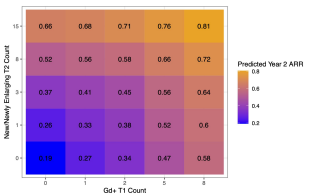
Study	Study Start Year	Study Population	n	Duration in study (years)
AFFIRM (C-1801)	2001	RRMS	314	1.84 (0.443)
SENTINEL (C-1802)	2002	RRMS	595	1.79 (0.580)
C-1803	2003	RRMS	55	0.373 (0.0539)
C-1900	2004	RRMS	65	0.438 (0.0824)
DEFINE (109MS301)	2007	RRMS	409	1.48 (0.589)
CONFIRM (109MS302)	2007	RRMS	363	1.50 (0.559)
ADVANCE (105MS301)	2009	RRMS	500	0.879 (0.157)
101MS203	2010	RRMS	47	0.366 (0.0679)
ASCEND (101MS326)	2011	SPMS	449	1.66 (0.658)
109MS305	2013	RRMS	113	0.452 (0.0467)
109MS308	2015	SPMS	30	0.184 (0.0913)

Figure 3: Endpoint trajectories in 12 sample patients. Each panel shows how the four endpoints (colored lines) change over time in 12 sample patients. The vertical black lines indicate when the patient has a relapse. These patients show the general trend of disease activity (Gd+ T1 lesions, N/NE T2 lesions, and relapses) being correlated in time and across patients. However, the trajectories were heterogenous and within a two year period did not show consistent trends over time.



Simulations

Figure 4: Simulations of year 2 ARR based on different amounts of lesion activity at baseline, for a hypothetical patient with reference (mean) covariate values. Observations at baseline had a long term effect on the ARR predictions by informing the distribution of average disease activity. There was a large effect from varying the baseline lesion counts, from an ARR of 0.19 to 0.81 within the range considered.



Model Predictions and Validation

Table 2: Out of sample prediction results demonstrate predictive power. The patients' data (blue boxes in Figure 2) used to make predictions was truncated at six months, and the predictions were made on the second year of relapse data (dark blue). Here, the patients are grouped into quartiles based on the predicted ARR. This shows the predictive power of short term data to predict long term outcomes, because the ARR of the fourth quartile was 4.5 times the ARR of the first quartile.

Predicted ARR Quartile	Estimated Average ARR	Observed ARR
1	0.205	0.127
2	0.333	0.360
3	0.508	0.538
4	0.883	0.859

Figure 5: The model demonstrates clinically relevant emergent properties. Specifically, the model was also used to make population level predictions and explore the consequences of the modeled relationships. Each density curve is the predicted distribution of ARR for the patients diagnosed with each type of MS, and the tick marks are the observed ARR in the studies with a two year duration. MS type is not included in the model, but based on observed data the model differentiates the two populations, based on the distribution of predicted relapse rates, with the SPMS population having lower relapse rates. This suggests that the model is reflecting underlying biological differences between the two populations. In addition, larger T2 lesion volume was associated with lower relapse rates, reflecting a connection between disease severity and a transition to less active forms of MS.

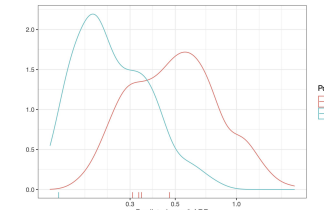
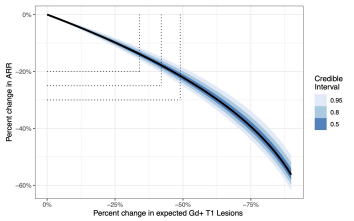


Figure 6: To understand the modeled effect of reducing Gd+ T1 lesions (without changing other disease characteristics), the modeled relationship between a potential change in the Gd+ T1 lesions and the predicted change in ARR (over the same time period) was compared. The dashed lines are estimated effects from Sormani 2013 et al. [5]. The intervals are credible intervals which incorporate parameter uncertainty in the model. This model predicts smaller ARR effects than the Sormani meta-analysis, which may reflect differences in the mechanism different T1 Gd+ lesions (patients with naturally different T1 Gd+ lesion counts vs therapeutic effects), the incorporation of time effects into this model, and the difference between modeling individual data compared with summary data. The model also can make predictions of the relationship at across different time ranges for each endpoint and comparisons between the other modeled endpoints.



Conclusion

- A Bayesian model was developed that characterized placebo data across studies, populations, and drugs at an individual level.
- The model showed predictive validity at the individual level when predicting 2-year ARR from 6 months of data. Further analyses suggested additional data can improve predictive performance.
- This model can be used to predict long term effects from limited and/or short term available data.
- Future planned analyses include:
  - Incorporation of treatment arms to estimate effect sizes and compare treatments
  - Predict efficacy of compounds in development
  - Optimization of planned trial designs
  - Link this model to other biomarker data (e.g. neurofilament light chains)

References

[1] Hickey, G.L., Philipson, P., Jorgensen, A. and Kolamunnage-Dona, R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med. Res. Methodol.* **16** (2016):117.

[2] C. E. Rasmussen C. K. I. Williams. *Gaussian Processes for Machine Learning* (the MIT Press, 2006).

[3] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles* **76** (2017):1–32.
URL `https://www.jstatsoft.org/v076/i01`

[4] Stan Development Team. RStan: the R interface to Stan (2020). R package version 2.21.2.
URL `http://mc-stan.org/`

[5] Sormani, M.P. and Bruzzi, P. MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. *Lancet Neurol.* **12** (2013):669–676.

Acknowledgements

This work was funded by Biogen Inc.