Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

SUPPLEMENT TO "ANALYSIS OF MULTIPLE SCLEROSIS LESIONS VIA SPATIALLY VARYING COEFFICIENTS"

By Tian Ge, Nicole Müller-Lenke, Kerstin Bendfeldt, Thomas E. Nichols, and Timothy D. Johnson

In this web-based material, we provide some theoretical aspects of the methods used in the paper, and some supplementary figures for a more complete illustration of our methods.

APPENDIX A: THE GIBBS SAMPLER

We begin with the latent variable representation of our model found in the main manuscript in equations (7) and (3). We assume the priors $\pi(\boldsymbol{\alpha}) \propto \mathbf{1}$, $\pi(\gamma) \propto 1$ and $\nu = 0$ (the degrees of freedom of the Wishart prior on the precision, $\boldsymbol{\Sigma}^{-1}$). The prior for $\boldsymbol{\beta}^*$ is the non-zero-centered MCAR prior:

$$\pi\left[\boldsymbol{\beta}^* \mid \boldsymbol{\Sigma}\right] \propto \exp\left\{-\frac{1}{2}\sum_{s_i \sim s_j}\left[\boldsymbol{\beta}^*(s_i) - \boldsymbol{\beta}^*(s_j)\right]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\beta}^*(s_i) - \boldsymbol{\beta}^*(s_j)\right]\right\}.$$

Due to conjugacy, the full conditional posteriors of all model parameters are distributions from which we can easily sample.

Update $z_i(s_j)$: The full conditional of each $z_i(s_j)$ is a truncated normal distribution:

$$\left[z_i(s_j) \mid Y_i(s_j), \eta_i(s_j)\right] \sim \begin{cases} \mathrm{N}\left(\eta_i(s_j), 1\right) \times I\left(z_i(s_j) > 0\right) & : \quad Y_i(s_j) = 1, \\ \mathrm{N}\left(\eta_i(s_j), 1\right) \times I\left(z_i(s_j) \leqslant 0\right) & : \quad Y_i(s_j) = 0, \end{cases}$$

where $I(\cdot)$ is the indicator function. The truncated normal distribution can be sampled using the algorithm by Robert (1995).

Update $\boldsymbol{\beta}^*(s_j)$: The full conditional distribution of $\boldsymbol{\beta}^*(s_j)$ for each s_j is a multivariate normal distribution. Let $\xi_i(s_j) = z_i(s_j) - w(s_j)\gamma$. Define

$$\boldsymbol{V}_\beta(s_j) = \left[n(s_j)\boldsymbol{\Sigma}^{-1} + \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1},$$

and

$$\boldsymbol{\mu}_\beta(s_j) = \boldsymbol{V}_\beta(s_j) \left[\boldsymbol{\Sigma}^{-1} \sum_{s_r \in \partial s_j} \boldsymbol{\beta}^*(s_r) + \sum_{i=1}^{N} \xi_i(s_j) \boldsymbol{x}_i \right].$$

Then

$$[\boldsymbol{\beta}^*(s_j) \mid \boldsymbol{\beta}^*(s_r), s_r \in \partial s_j, \boldsymbol{\Sigma}, \xi_i(s_j)] \sim \text{MVN}\left(\boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta\right).$$

Update γ: The full conditional of γ is also multivariate normal. Let $\zeta_i(s_j) = z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^*(s_j)$. Define

$$V_\gamma = \left[N \sum_{j=1}^{M} w^2(s_j) \right]^{-1},$$

and

$$\mu_\gamma = V_\gamma \left[\sum_{i=1}^{N} \sum_{j=1}^{M} w(s_j) \zeta_i(s_j) \right].$$

Then

$$[\gamma \mid \zeta_i(s_j)] \sim \text{MVN}\left(\mu_\gamma, V_\gamma\right).$$

Update $\boldsymbol{\Sigma}^{-1}$: The full conditional of $\boldsymbol{\Sigma}^{-1}$ is a Wishart distribution. Let

$$\boldsymbol{S} = \left[\boldsymbol{I} + \sum_{s_i \sim s_j} \left[\boldsymbol{\beta}^*(s_i) - \boldsymbol{\beta}^*(s_j)\right] \left[\boldsymbol{\beta}^*(s_i) - \boldsymbol{\beta}^*(s_j)\right]^{\mathrm{T}} \right].$$

Then

$$\left[\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{\beta}^*\right] \sim W\left(M - 1 + \nu, \boldsymbol{S}^{-1}\right).$$

APPENDIX B: LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV)

To assess the predictive capabilities of MRI imaging along with the five covariates age, gender, disease duration, EDSS and PASAT under our model, we use a cross-validation approach. Leaving one subject out at a time, the posterior is estimated from the remaining subjects and a prediction is made

on the subtype of the subject left out. The procedure is repeated for each subject and the classification rate based on all the subjects is computed. To avoid sampling from the posterior distribution for each leave-one-out situation, an importance sampling approach originally proposed by Gelfand, Dey and Chang (1992) is used to reduce the computation of LOOCV.

Let $\boldsymbol{y}_i = [y_i(s_1), \cdots, y_i(s_M)]^{\mathrm{T}}$ denote the binary data of all voxels from subject i. $\mathcal{D}_N = \{\boldsymbol{y}_i, g_i\}_{i=1}^N$ consists of the data from all the N subjects where g_i is a sample realization of the subtype indicator $G_i \in \{1, \cdots, G\}$. G is the number of subtypes. Let $\mathcal{D}_{-\ell} = \{\boldsymbol{y}_i, g_i\}_{i \neq \ell}$ denote the data with subject ℓ left out. The goal is to make inference on the subtype g_i given the binary image, the covariates, and the observed data $\mathcal{D}_{-i}$ for each subject i. Let $\boldsymbol{\Theta}$ denote all model parameters. It can be shown that the LOOCV predictive probabilities of g_i for $i = 1, \cdots, N$ is given by

$$\Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) = \frac{q_g Q_{gg_i}}{\sum_{g'=1}^{G} q_{g'} Q_{g'g_i}}, \tag{S1}$$

where $q_g = \Pr(G_i = g), g = 1, \cdots, G$ is the prior probability for the subtype g and

$$Q_{gg_i} = \int \frac{\Pr(\boldsymbol{y}_i \mid G_i = g, \boldsymbol{\Theta})}{\Pr(\boldsymbol{y}_i \mid G_i = g_i, \boldsymbol{\Theta})} \pi(\boldsymbol{\Theta} \mid \mathcal{D}_N) \mathrm{d}\boldsymbol{\Theta}.$$

A proof of equation (S1) is given below. The algorithm to estimate the LOOCV predictive probabilities is as follow:

- Run the Gibbs sampler for K iterations after burn-in to obtain posterior draws $\boldsymbol{\Theta}^{(k)} \sim \pi(\boldsymbol{\Theta} \mid \mathcal{D}_N)$, for $k = 1, \cdots, K$.

- For each subject i and each subtype g, compute

$$\widehat{Q}_{gg_i} = \frac{1}{K} \sum_{k=1}^{K} \frac{\Pr\left(\boldsymbol{y}_i \mid G_i = g, \boldsymbol{\Theta}^{(k)}\right)}{\Pr\left(\boldsymbol{y}_i \mid G_i = g_i, \boldsymbol{\Theta}^{(k)}\right)},$$

  where $\Pr\left(\boldsymbol{y}_i \mid G_i = g, \boldsymbol{\Theta}^{(k)}\right) = \prod_{j=1}^{M} \left[\Pr\left(Y_i(s_j) = 1 \mid G_i = g, \boldsymbol{\Theta}^{(k)}\right)\right]^{y_i(s_j)} \times$ $\left[\Pr\left(Y_i(s_j) = 0 \mid G_i = g, \boldsymbol{\Theta}^{(k)}\right)\right]^{1 - y_i(s_j)}$.

- The estimate of the predictive probability of G_i is given by

$$\widehat{\Pr}(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) = \frac{q_g \widehat{Q}_{gg_i}}{\sum_{g'=1}^{G} q_{g'} \widehat{Q}_{g'g_i}}.$$

And the estimate of G_i is:

$$\widehat{g}_i = \arg\max_g (q_g \widehat{Q}_{gg_i}).$$

The $G \times G$ LOOCV confusion matrix $\boldsymbol{C} = \{c_{gg'}\}$ is then defined as

$$c_{gg'} = \frac{\sum_{i=1}^{N} I_g(g_i) I_{g'}(\widehat{g}_i)}{\sum_{i=1}^{N} I_g(g_i)},$$

where $I_u(v)$ is an indicator function. $I_u(v) = 1$ if $u = v$ and $I_u(v) = 0$ otherwise. Then the overall and the average correct classification rates are respectively given by $c_o = \frac{1}{N}\sum_{i=1}^{N} I_{g_i}(\widehat{g}_i)$ and $c_a = \frac{1}{G}\sum_{g=1}^{G} c_{gg}$.

Proof of equation (S1). The LOOCV posterior predictive probabilities of g_i is:

$$\begin{aligned}
&\Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) \\
&= \int \Pr(G_i = g \mid \boldsymbol{y}_i, \boldsymbol{\Theta}) \pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) \mathrm{d}\boldsymbol{\Theta} \\
&= \int \Pr(G_i = g \mid \boldsymbol{y}_i, \boldsymbol{\Theta}) \frac{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i})} \pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i}) \mathrm{d}\boldsymbol{\Theta} \\
&= \int \Pr(G_i = g \mid \boldsymbol{y}_i, \boldsymbol{\Theta}) \frac{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i})} \pi(\boldsymbol{\Theta} \mid \mathcal{D}_N) \mathrm{d}\boldsymbol{\Theta}.
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i})} &= \frac{\pi(\boldsymbol{\Theta}, \boldsymbol{y}_i, \mathcal{D}_{-i}) \pi(\boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i})}{\pi(\boldsymbol{y}_i, \mathcal{D}_{-i}) \pi(\boldsymbol{\Theta}, \boldsymbol{y}_i, G_i = g_i, \mathcal{D}_{-i})} \\
&= \frac{\Pr(G_i = g_i \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\Pr(G_i = g_i \mid \boldsymbol{y}_i, \boldsymbol{\Theta})}.
\end{aligned}$$

This implies

$$\begin{aligned}
\frac{\Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\Pr(G_i = g_i \mid \boldsymbol{y}_i, \mathcal{D}_{-i})} &= \int \frac{\Pr(G_i = g \mid \boldsymbol{y}_i, \boldsymbol{\Theta})}{\Pr(G_i = g_i \mid \boldsymbol{y}_i, \boldsymbol{\Theta})} \pi(\boldsymbol{\Theta} \mid \mathcal{D}_N) \mathrm{d}\boldsymbol{\Theta} \\
&= \int \frac{\Pr(\boldsymbol{y}_i \mid G_i = g, \boldsymbol{\Theta}) q_g}{\Pr(\boldsymbol{y}_i \mid G_i = g_i, \boldsymbol{\Theta}) q_{g_i}} \pi(\boldsymbol{\Theta} \mid \mathcal{D}_N) \mathrm{d}\boldsymbol{\Theta} \\
&:= \frac{q_g}{q_{g_i}} Q_{gg_i}.
\end{aligned}$$

Since $\sum_{g=1}^{G} \Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) = 1$, if follows that

$$\frac{1 - \Pr(G_i = g_i \mid \boldsymbol{y}_i, \mathcal{D}_{-i})}{\Pr(G_i = g_i \mid \boldsymbol{y}_i, \mathcal{D}_{-i})} = \sum_{g \neq g_i} \frac{q_i}{q_{g_i}} Q_{g g_i}.$$

Therefore

$$\Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) = \frac{q_g Q_{g g_i}}{\sum_{g'=1}^{G} q_{g'} Q_{g' g_i}}.$$

□

APPENDIX C: NAÏVE BAYESIAN CLASSIFIER

A naïve Bayesian classifier is used to predict the subtype of the held-out subject and is compared to the classification results of the Bayesian spatial model. The naïve Bayesian classifier assumes all features are mutually independent. For the present study, this assumption implies all voxels are treated independently. The predictive probability is

$$\begin{aligned} \Pr(G_i = g \mid \boldsymbol{y}_i, \mathcal{D}_{-i}) &\propto \Pr(\boldsymbol{y}_i \mid G_i = g, \mathcal{D}_{-i}) \Pr(G_i = g) \\ &= q_g \prod_{j=1}^{M} \Pr\left(y_i(s_j) \mid G_i = g, \mathcal{D}_{-i}\right), \end{aligned}$$

where the second equality follows from the independence assumption. At each voxel s_j, assume $\Pr\left(Y_i(s_j) = 1 \mid G_i = g\right) = p_g(s_j)$, and $p_g(s_j)$ is assigned Jeffrey's prior: Beta$(0.5, 0.5)$. Then the posterior distribution of $p_g(s_j)$ is also beta distributed. The probability $\Pr\left(y_i(s_j) \mid G_i = g, \mathcal{D}_{-i}\right)$ is then estimated by the posterior mean of $p_g(s_j)$:

$$\widehat{\Pr}\left(Y_i(s_j) = 1 \mid G_i = g, \mathcal{D}_{-i}\right) = \frac{0.5 + \sum_{\ell \neq i} y_\ell(s_j) I_{G\ell}(g)}{1 + \sum_{\ell \neq i} I_{G\ell}(g)},$$

and

$$\widehat{\Pr}\left(Y_i(s_j) = 0 \mid G_i = g, \mathcal{D}_{-i}\right) = 1 - \widehat{\Pr}\left(Y_i(s_j) = 1 \mid G_i = g, \mathcal{D}_{-i}\right).$$

Finally, the estimate of G_i is given by

$$\widehat{g}_i = \arg\max_g \left(q_g \prod_{j=1}^{M} \widehat{\Pr}\left(y_i(s_j) \mid G_i = g, \mathcal{D}_{-i}\right) \right).$$

APPENDIX D: SUPPLEMENTARY FIGURES

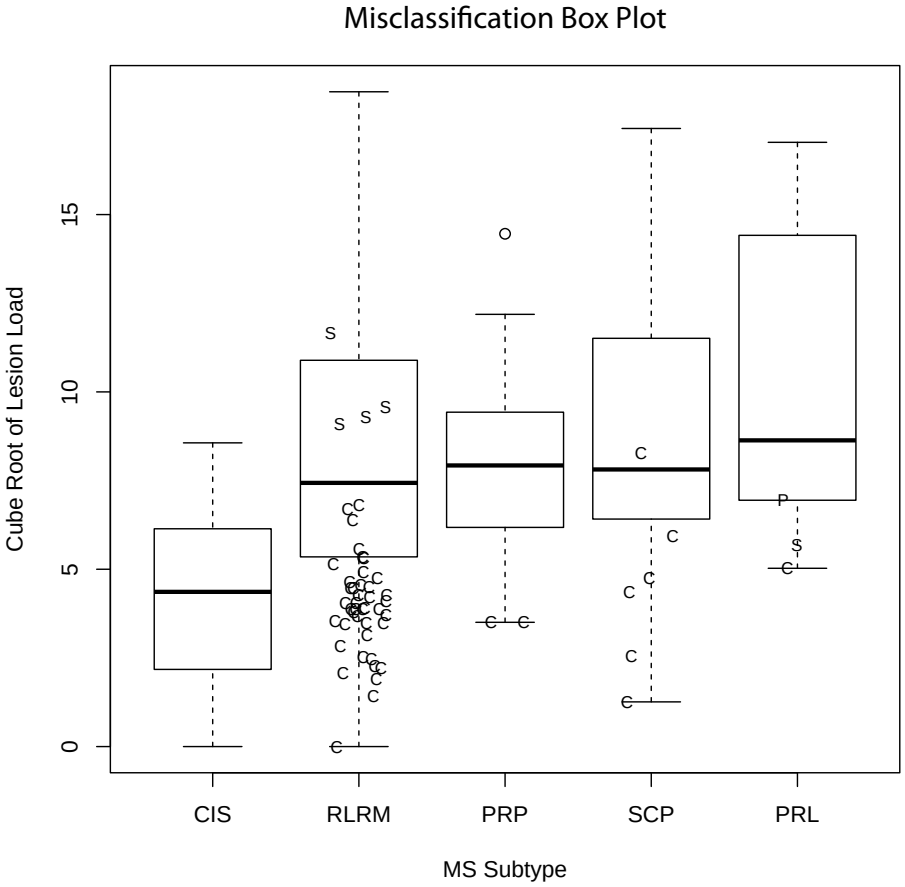

FIG S1. *Box plot of lesion load, measured in cube root of the number of lesion voxels. The letters on the box plot are those misclassified subjects: C – -misclassified as CIS; R — misclassified as RLRM; P — misclassified as PRP; S — misclassified as SCP; L — misclassified as PRL. From this figure it is evident that those misclassified as CIS tend to have relatively small lesion load.*

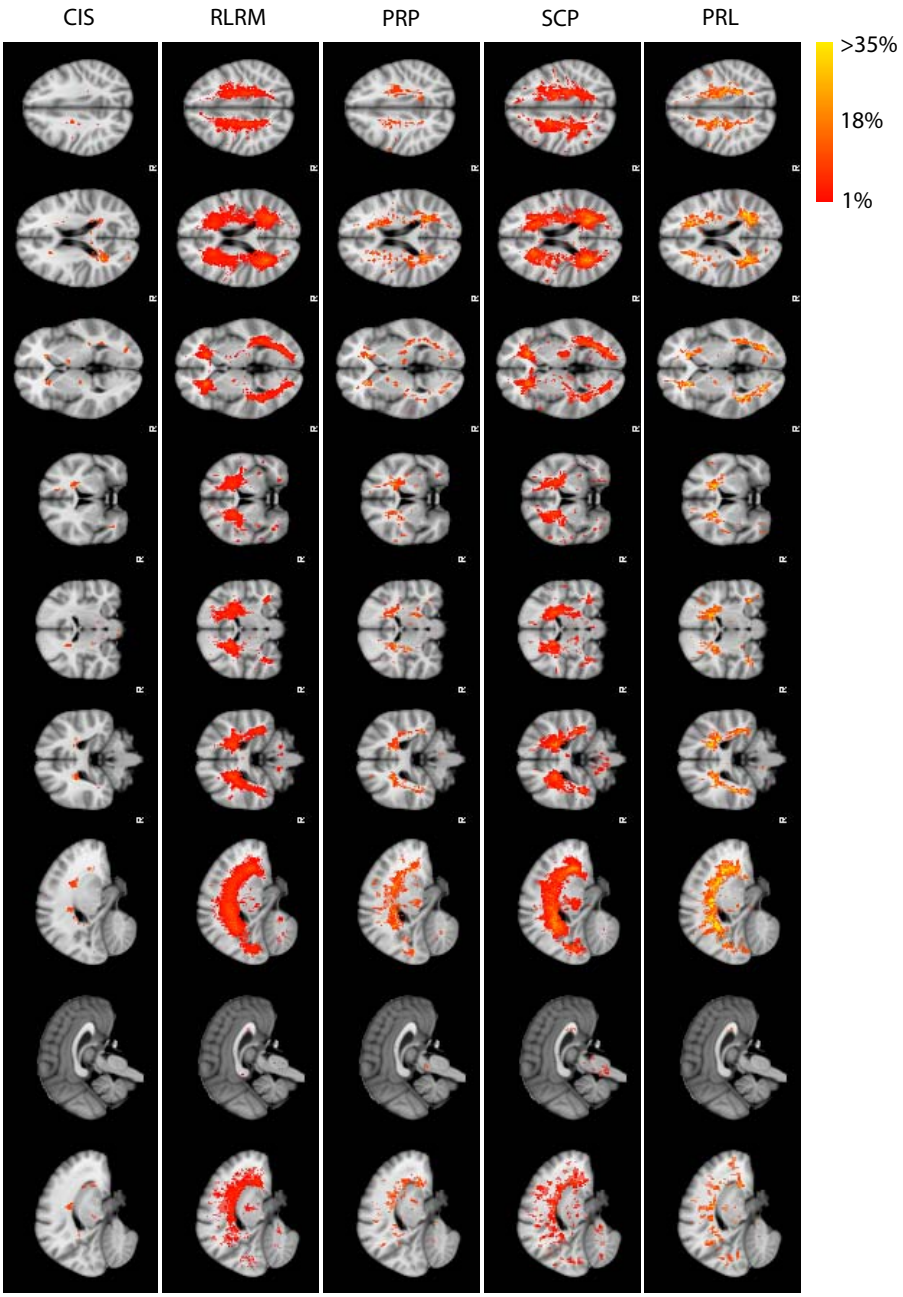

FIG S2. *Empirical lesion rates for each subtype, CIS, RLRM, PRP, SCP, and PRL, from left to right. The intensity shown is the proportion of subjects with a lesion at each voxel. Color scale is set from 1% to 35% (rates below 1% are not shown, rates of 35% or greater have maximal yellow color). Extent of colored regions is less informative about absolute rate than a reflection of sample sizes in each group.*

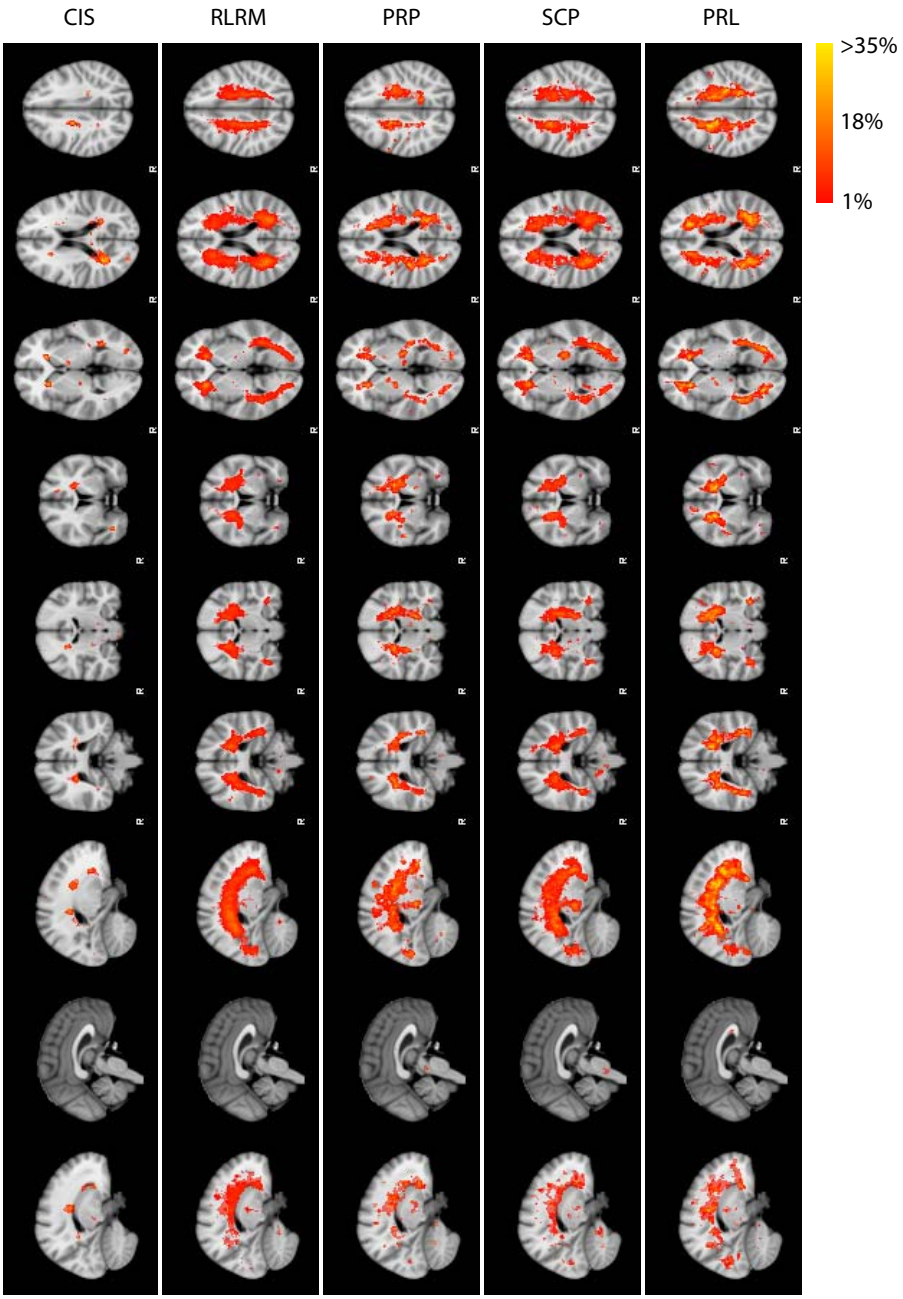

FIG S3. *Model estimates of lesions incidence for each subtype, CIS, RLRM, PRP, SCP, and PRL, from left to right. Intensity at each voxel is the prediction of lesion incidence using our Bayesian spatial model, which implicitly learns the smoothness of the incidence maps and regularizes the rates accordingly. Color scale is set from 1% to 35% (rates below 1% are not shown, rates of 35% or greater have maximal yellow color).*

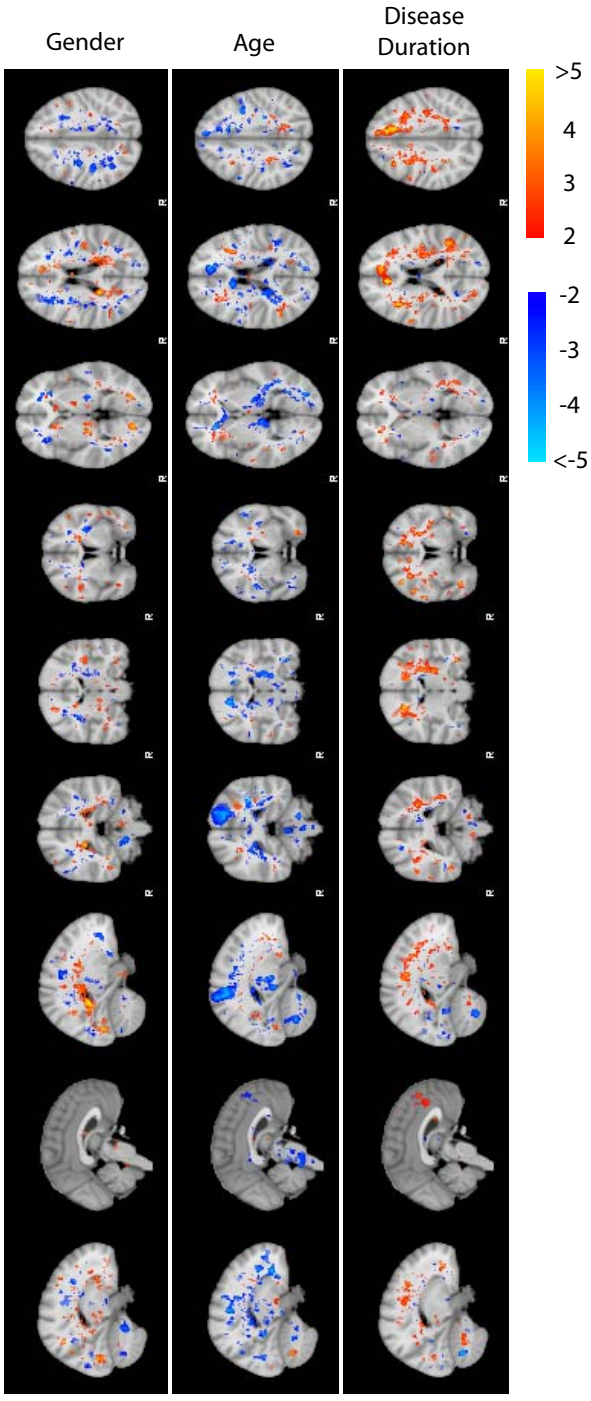

FIG S4. *Standardized (posterior mean divided by posterior standard deviation) spatial maps for gender, age and disease duration from our Bayesian spatial model. Color scale is set from 2 to 5 for positive values (values below 2 are not shown, values of 5 or greater have maximal yellow color), and from -5 to -2 for negative values (values above -2 are not shown, values of -5 or smaller have the lightest blue color).*

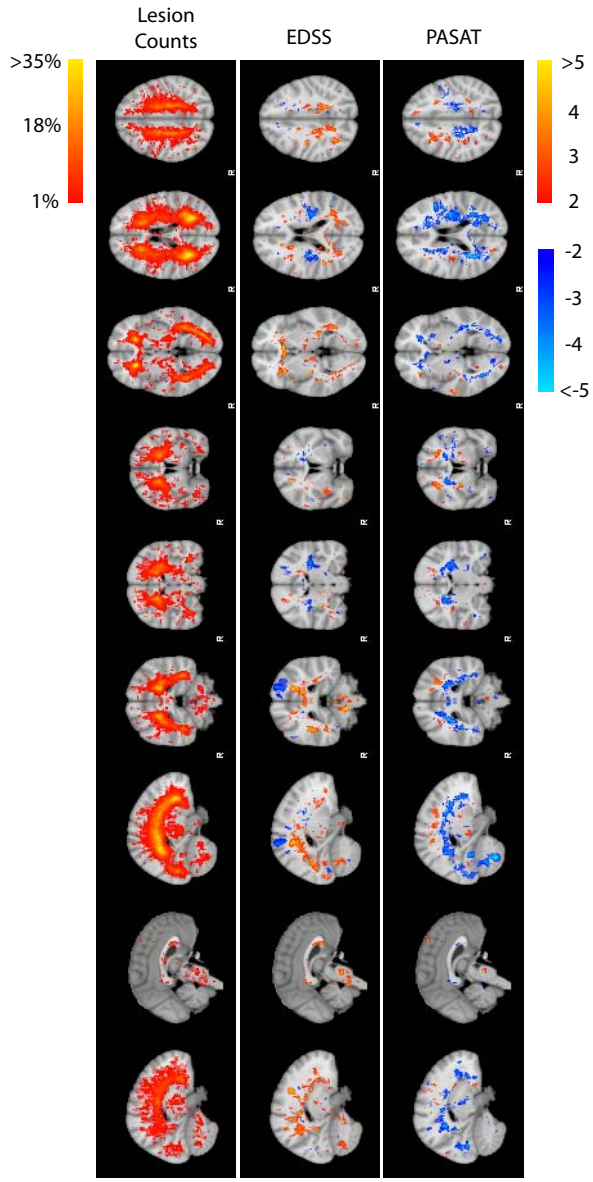

FIG S5. *The left column is the empirical lesion counts. Color scale is set from 1% to 35%. The middle and right columns are standardized (posterior mean divided by posterior standard deviation) spatial maps for EDSS and PASAT respectively from our Bayesian spatial model. Color scale is set from 2 to 5 for positive values (values below 2 are not shown, values of 5 or greater have maximal yellow color), and from -5 to -2 for negative values (values above -2 are not shown, values of -5 or smaller have the lightest blue color).*

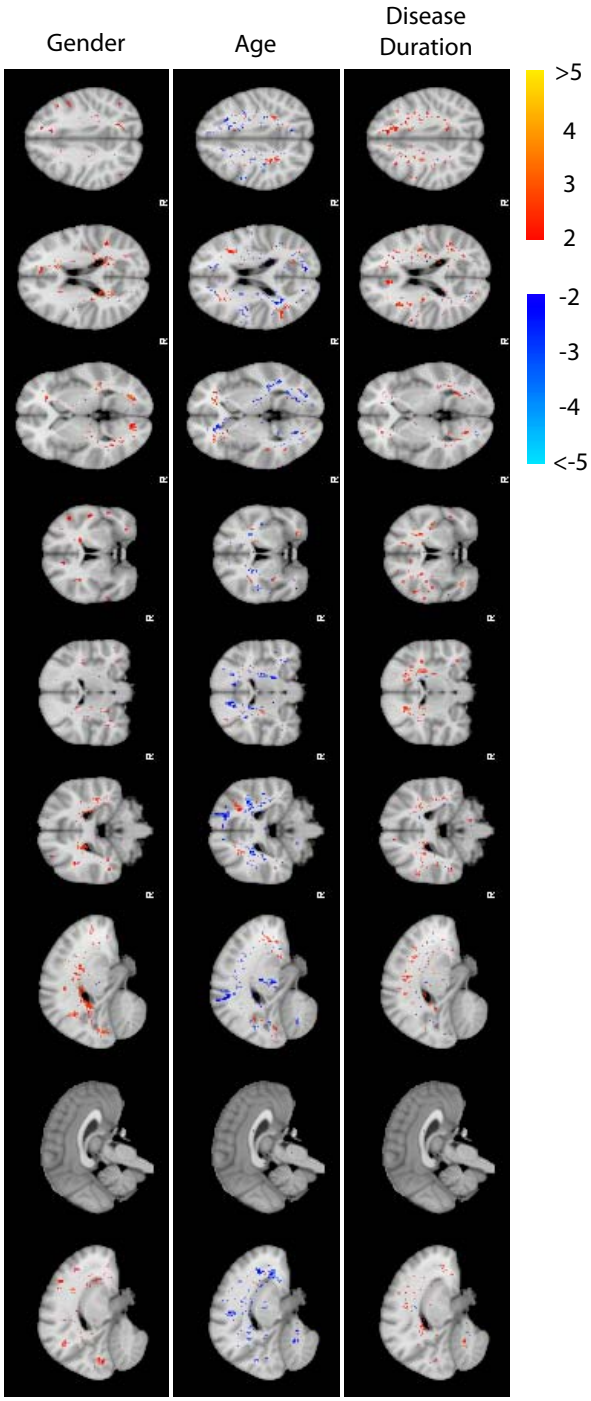

FIG S6. *Standardized (mean divided by standard deviation) spatial maps for gender, age and disease duration from the Firth logistic regression. Color scale is set from 2 to 5 for positive values (values below 2 are not shown, values of 5 or greater have maximal yellow color), and from -5 to -2 for negative values (values above -2 are not shown, values of -5 or smaller have the lightest blue color).*

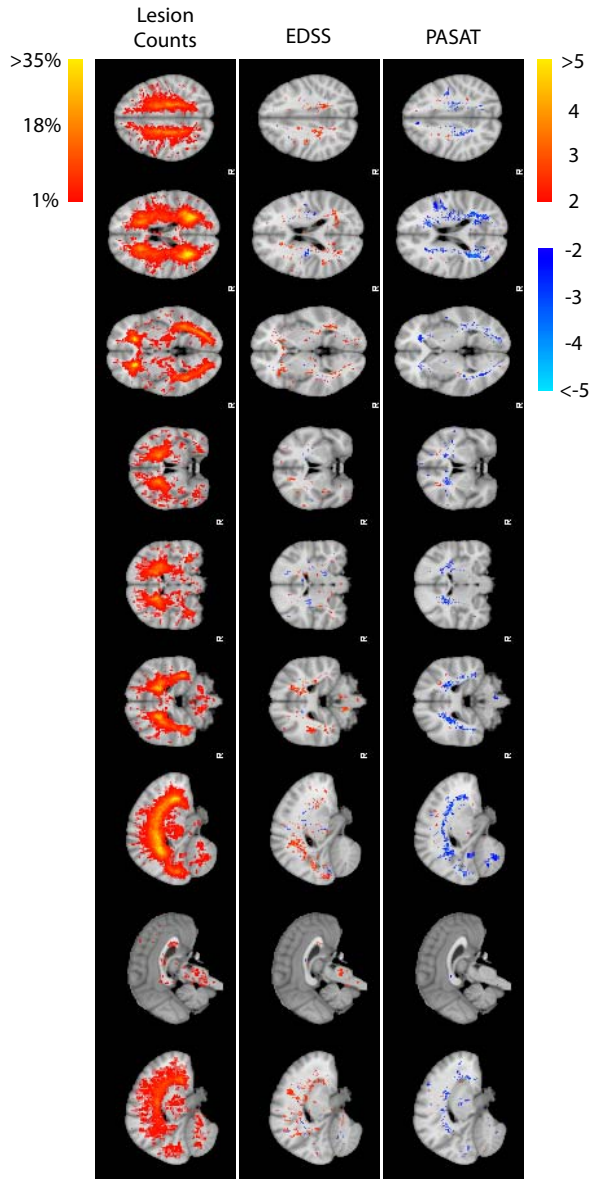

FIG S7. *The left column is the empirical lesion counts. Color scale is set from 1% to 35%. The middle and right columns are standardized (mean divided by standard deviation) spatial maps for EDSS and PASAT from the Firth logistic regression. Color scale is set from 2 to 5 for positive values (values below 2 are not shown, values of 5 or greater have maximal yellow color), and from -5 to -2 for negative values (values above -2 are not shown, values of -5 or smaller have the lightest blue color).*

REFERENCES

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian Statisitcs* **4** 147–167.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5** 121–125.