A  Participant Details

We did not genotype any participants. We did not screen for any type of cognitive impairment. Healthy control participants in general were matched on age as a group, with no statistically significant differences in age.

| ID  | Sex | Age at Visit 1 | Dom. Hand |
|-----|-----|----------------|-----------|
| C07 | F   | 55.67          | L         |
| C09 | F   | 55.38          | R         |
| C14 | F   | 56.49          | R         |
| C15 | M   | 52.7           | R         |
| C16 | M   | 57.55          | R         |
| C17 | M   | 57.08          | R         |
| C19 | M   | 63.44          | L         |
| C20 | F   | 58.8           | R         |
| C22 | F   | 59.51          | R         |
| C23 | M   | 49.51          | R         |
| C26 | F   | 54.66          | L         |
| C28 | F   | 57.11          | R         |
| C31 | M   | 56.35          | L         |
| C32 | M   | 56.35          | R         |
| C34 | F   | 57.93          | R         |
| C35 | F   | 57.39          | R         |
| C36 | F   | 63.63          | R         |
| C39 | F   | 60.59          | R         |
| C40 | F   | 53.48          | R         |
| C41 | M   | 57.28          | R         |
| C42 | F   | 58.38          | R         |
| C43 | F   | 65.83          | R         |

Table A1: Healthy control participant demographics. Dom. Hand = dominant hand, as measured by the Edinburg Inventory.
Figure A1: Visit timing for each participant in the study. Each dot represents a MRI session visit. Note that for many participants, the first two visits occurred in quick succession and appear overlapping on the plot.
### Table A2: ALS Participants

| Participant | Age (years) | ALSFRS-R | Number of Visits |
|-------------|-------------|----------|-----------------|
|             | First Visit | Last Visit | First Visit | Last Visit |             |
| A04         | 58.7        | 63.1      | 44           | 43          | 10          |
| A06         | 53.4        | 54.2      | 43           | 38          | 3           |
| A08         | 64.9        | 65.1      | 29           | 29          | 3           |
| A11         | 57.8        | 58.4      | 42           | 39          | 4           |
| A14         | 55.1        | 57.2      | 42           | 35          | 7           |
| A18         | 61.3        | 63.0      | 42           | 35          | 7           |
| A19         | 67.8        | 69.2      | 43           | 31          | 7           |
| A21         | 66.8        | 67.2      | 40           | 30          | 4           |
| A23         | 56.7        | 57.0      | 39           | 31          | 3           |
| A25         | 54.6        | 54.8      | 40           | 37          | 3           |
| A26         | 58.3        | 59.6      | 46           | 27          | 6           |
| A30         | 63.2        | 64.1      | 35           | 24          | 4           |
| A31         | 65.8        | 66.6      | 26           | 25          | 4           |
| A32         | 47.1        | 48.2      | 36           | 30          | 5           |
| A33         | 55.3        | 56.2      | 46           | 44          | 5           |
| A34         | 56.0        | 56.2      | 38           | 31          | 3           |

B Processing and Analysis Details

For each research visit for each subject, MRI data were processed with a blend of SPM (Statistical Parametric Mapping, Version 12, Release 7219, University College London) \([\text{Friston et al.} 2007]\), FSL (Functional Magnetic Resonance Imaging of the Brain Software Library, Version 6, Oxford University) \([\text{Jenkinson et al.} 2012]\), ANTs (2.3.1, University of Pennsylvania) \([\text{Avants et al.} 2011]\), FreeSurfer (Version 6.0, Harvard and Mass General Hospital) \([\text{Fischl} 2012]\), and finally the Human Connectome Project Workbench (Version 1.2.3).

The $T_1$-weighted image from each subject at each visit was first skull-stripped using FSL’s brain extraction tool with robust brain center estimation and a fractional intensity threshold of 0.4. Non-uniformity correction was applied to the original $T_1$-weighted images using ANTs’ N4 algorithm (variant of the N3 algorithm, nonparametric nonuniform normalization). For each individual, the bias-field corrected and skull stripped $T_1$-weighted images of all sessions were then fed into ANTs’ `antsMultivariateTemplateConstruction2` pipeline to create a participant template that is spatially unbiased to the orientation of the input images. This pipeline first generates an unbiased starting point by averaging all inputs, then aligns the center of
mass of all inputs to that of the initial input. All input images are then rigidly registered (i.e., 6 degrees of freedom) to the starting point image and averaged to create a new template, which then becomes the new starting point for subsequent iterations. Within each iteration, there are four levels at which the registration was estimated. At each level, different parameters are set for the shrink factor and smoothing factor. We set parameters for this script as follows: number of iterations of template creation = 2, number of step levels within each iteration = 4; number of iterations per step level = $10 \times 10 \times 10 \times 5$; shrink factor = $8 \times 4 \times 2 \times 1$; smoothing factor = $4 \times 2 \times 1 \times 0$. The participant’s template image was then processed through FreeSurfer, including edits for brain mask, to result in a model of the pial surface and a corresponding spherical surface for the participant.

BOLD time-series were projected to the unique cortical surface of each individual participant. Timeseries data were slice-time corrected and realigned (2 passes). The lower resolution $T_1$-weighted image was co-registered to the mean realigned BOLD image. The high-resolution $T_1$-weighted image was then co-registered to the resulting co-registered low-resolution image.

To reduce computational load, both the left and right pial surface models were resampled to 10,000 vertices per hemisphere. This was accomplished using the Workbench command `-surface-resample`, which leverages the registration between a sphere consisting of 10,000 vertices (created using the Workbench command `-surface-create-sphere`) and the participant’s spherical surface generated using FreeSurfer. The BOLD data was then projected to the left and right hemisphere resampled surfaces using the Work-
Figure B3: Modeled hemodynamic response (HRF) for the right hand clench task and its temporal derivative (dHRF).

bench command `-metric-resample`. Finally, the FreeSurfer labeling of four sensorimotor areas (i.e., the paracentral gyrus, postcentral gyrus, precentral gyrus, and caudal middle frontal gyrus [Verstraete et al. 2011]; all were taken from the Desikan-Killiany atlas [Desikan et al. 2006]) were resampled to 10,000 vertices using the Workbench command `-label-resample`. These labels were combined to produce a participant-specific motor mask to limit the location of statistical estimation.

Before model fitting, we identified and removed noisy volumes based on data-driven leverage scrubbing using the fMRIscrub R package [Mejia et al. 2017] (version 0.1.2), which identifies volumes that differ substantially from the multivariate distribution of images. We employed a threshold of 4 times the median leverage for scrubbing. We also excluded any sessions where more than 25% of volumes were scrubbed. This resulted in exclusion of one visit from one ALS participant, one visit each from three HC participants, and two visits from one HC participant.

In models 2 and 3, the function $f(\cdot)$ is a natural cubic spline, which allows for a non-linear relationship between hand motor disability and activation size. Spline knots were placed at the 33rd and 67th quantiles. Natural splines have boundary conditions that enforce a linear fit beyond the boundary knots, which avoids the extreme boundary fits often observed in standard polynomial regression. A basis was generated in R using the `ns` function from the splines package version 4.0.3.

This final model form in Eqn. 3 was determined by a series of likelihood ratio tests. We compared three models for each predictor: one assuming a linear fit, one allowing a non-linear spline fit, and one excluding the predictor. Each test was based on left-hemispheric (contra-lateral) activation at an effect size of $\gamma = 0\%$, which provided the most robust areas of activation. We found statistically significant evidence for a non-linear relationship between hand disability and size of activation and a
linear relationship between other disability and size of activation. We also considered an alternative model with days since onset as a predictor (spline or linear fit) in place of the disease burden measures. These models were substantially worse in terms of predictive accuracy and Akaike information criterion (AIC) (Akaike 1998).
C Longitudinal spatial Bayesian task fMRI analysis

The model is fit within each hemisphere on the brain separately. The triangular mesh representing the participant-specific cortical surface, after resampling and masking as described above, contained approximately 1,500 vertices per hemisphere. The exact size and shape varied across participants due to differences in cortical anatomy. Fig. C4 shows the surface meshes for one participant with ALS.

In our longitudinal spatial Bayesian modeling framework, the amplitudes and areas of activation are estimated for each visit, but model parameters including the residual variance and spatial properties of the task activation fields (e.g. correlation range, variance) are estimated using data from multiple visits to improve estimation efficiency. The same set of tasks must be performed across visits, though the stimulus timing can vary over visits.

Consider a single subject and hemisphere. Let $j = 1, \ldots, J$ index visits and $k = 1, \ldots, K$ index task stimuli. In our models, $K = 2$ (the canonical HRF and its first derivative), and the number of visits per participant varied between $J = 3$ to $J = 10$. Let $T_j$ be the number of volumes in visit $j$ after scrubbing, and let $V$ be the number of surface vertices within the mask. Note that the surfaces are required to be spatially co-registered across visits within a subject, but not across subjects, as the model is fit separately for each subject. Let $y_{j}(v)$ ($T_j \times 1$) be the processed and scrubbed fMRI data at vertex $v$. Let $x_{jk}$ ($T_j \times 1$) represent the expected BOLD response to task $k$ (excluding scrubbed volumes). In the classical GLM, we would fit a separate linear model at each location $v = 1, \ldots, V$, namely

$$y_{j}(v) = \sum_{k=1}^{K} x_{jk} \beta_{jk}(v) + \epsilon_{j}(v), \quad \epsilon_{j}(v) \sim N(0, \sigma^{2}I_{T_j}),$$  

where $\beta_{jk}(v)$ is the activation amplitude associated with task $k$. In Eqn. (1) the residuals are assumed to be temporally independent, which can be achieved by prewhitening.

To illustrate the construction of our longitudinal spatial Bayesian GLM, we first combine across vertices within a single session, describe the incorporation of spatial priors on the task amplitudes to yield a spatial Bayesian model, then generalize to the longitudinal case. Denote

$$y_{j} = \begin{bmatrix} y_{j}(1) \\ \vdots \\ y_{j}(V) \end{bmatrix}, \quad X_{jk} = I \otimes x_{jk} = \begin{bmatrix} x_{jk} \\ \vdots \\ x_{jk} \end{bmatrix}, \quad \beta_{jk} = \begin{bmatrix} \beta_{jk}(1) \\ \vdots \\ \beta_{jk}(V) \end{bmatrix}, \quad \epsilon_{j} = \begin{bmatrix} \epsilon_{j}(1) \\ \vdots \\ \epsilon_{j}(V) \end{bmatrix},$$  

where $\otimes$ denotes the Kronecker product. Then we can write the single-session model as

$$y_{j} = X_{jk} \beta_{jk} + \epsilon_{j}, \quad \epsilon_{j} \sim N(0, \sigma^{2}I).$$  

7
Assuming spatial process priors on the $\beta_{jk}$, $k = 1, \ldots, K$, along with hyperpriors on their parameters, yields a spatial Bayesian model. Mejia et al. (2020) proposed employing a class of flexible Gaussian Markov random field (GMRF) priors that are appropriate for high-dimensional data in a triangular mesh format, known as stochastic partial differential equation (SPDE) priors (Lindgren et al. 2011). Specifically, SPDE priors are zero-mean multivariate Normal priors with a sparse precision (inverse covariance) structure. The precision matrix has non-zero entries along the diagonal and in cells corresponding to neighboring locations in the triangular mesh. We provide more details on the precision structure in the specification of the longitudinal model below.

Now combining over sessions, denote

$$
\begin{align*}
\mathbf{y} & = \begin{bmatrix} y_1 \\ \vdots \\ y_J \end{bmatrix}, \\
\mathbf{X}_k & = \begin{bmatrix} X_{1k} \\ \vdots \\ X_{Jk} \end{bmatrix}, \\
\beta_k & = \begin{bmatrix} \beta_{1k} \\ \vdots \\ \beta_{Jk} \end{bmatrix}, \\
\epsilon & = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_J \end{bmatrix}. 
\end{align*}
$$

The longitudinal spatial Bayesian model is given by

$$
(y|\beta_1, \ldots, \beta_K) = \sum_{k=1}^{K} \mathbf{X}_k \beta_k + \epsilon
$$

where

$$
\begin{align*}
\epsilon | \sigma^2 & \sim N(0, \sigma^2 \mathbf{I}) \\
\beta_{jk} | \kappa_k, \tau_k & \sim N(0, Q_k^{-1}) \text{ for } j = 1, \ldots, J, k = 1, \ldots, K
\end{align*}
$$

and

$$
\theta \sim \pi(\theta),
$$

where $\theta = (\kappa_1, \tau_1, \ldots, \kappa_K, \tau_K, \sigma^2)$ are all of the hyperparameters and $\pi(\theta)$ is their joint prior density. We adopt the default priors in R-INLA for each hyperparameter, namely independent log-normal priors on the spatial hyperparameters $\kappa_k$ and $\tau_k$ and a gamma prior on the inverse residual variance (Lindgren et al. 2015). Note that the spatial hyperparameters are allowed to vary across tasks, allowing for differences in the spatial properties of different tasks, but are common across visits, which improves estimation efficiency. The form of the spatial precision with parameters $\kappa$ and $\tau$ is

$$
Q = \tau (\kappa^4 \mathbf{C} + 2\kappa^2 \mathbf{G} + \mathbf{G} \mathbf{C}^{-1} \mathbf{G}),
$$

where $\mathbf{C}$ is a diagonal matrix and $\mathbf{G}$ is a sparse symmetric matrix with non-zero entries in cells corresponding to neighboring vertices in the triangular mesh (Lindgren et al. 2015). The parameter $\kappa$ controls the spatial dependence of the field, while $\tau$ controls its variance.

This model can be estimated using the BayesfMRI R package, which uses R-INLA (Lindgren et al. 2015) to compute the necessary posterior quantities for each latent field $\beta_{jk}$, as described in detail in Mejia et al. (2020). Given the posterior mean and precision of each latent field, we can then identify areas of activation based on the joint posterior distribution using an excursions set approach (Bolin and Lindgren 2015, Mejia et al. 2020). This avoids massive multiple comparisons and results in much greater power to detect true activations by leveraging spatial dependencies and avoiding multiplicity correction.
Areas of activation can also be identified through BayesfMRI, which uses the *excursions* package (Bolin and Lindgren 2018) to identify areas exceeding a specified effect size $\gamma$ (e.g. 1% signal change) at a given significance level $\alpha$. For more information on the model estimation and computation of excursions sets, see Mejia et al. (2020).

**Figure C4:** Triangular mesh for the resampled pial surface of each hemisphere for one individual with ALS. The model is fit within the motor cortex, which is shaded in blue. For this individual, the motor cortex includes 1,455 resampled vertices in the left hemisphere and 1,524 in the right hemisphere.
D Computation Time

All computations were performed in R version 4.0.3 (R Core Team 2020) using the BayesfMRI package on a Mac Pro computer with a 2.7 GHz 24-core Intel Xeon W processor with 512 GB of memory. Depending on the number of visits being simultaneously estimated, model estimation per participant and hemisphere took 10 to 30 minutes and required approximately 10 to 25 GB of RAM. Identifying areas of activation took an additional 1-5 minutes per session and effect size. Computation times for all participants are shown in Fig. D5.

Figure D5: Computation time in minutes for model estimation and identifying areas of activation, per participant. Times represent the sum across both hemispheres per participant. Time to identify activations reflects the average time per visit, averaged over visits, at a given effect size. Computation times for both model estimation and identifying areas of activation grew approximately linearly with the number of visits per participant.
Figure E6: Bayesian GLM and classical GLM estimates of activation amplitude and areas of activation in one HC participant. The Bayesian GLM tends to produce estimates that are smoother and areas of activation that are larger and more contiguous.

E Longitudinal Reliability of Spatial Bayesian GLM

We first compared the results of the Bayesian and classical approaches visually for one example HC participant in Fig. E6. The left panel shows estimates of activation amplitude produced from each approach. The Bayesian GLM produced amplitudes of activation that were noticeably smoother than those produced by the classical GLM. This is due to the implicit smoothing in the model estimation for the spatial Bayesian GLM, which accounts for spatial dependencies between neighboring vertices. The degree of smoothing is determined in an optimal fashion and avoids smoothing of noise along with the signal as in data smoothing (Lindquist and Mejia 2015).

The right panel of Fig. E6 shows areas of activation produced from both approaches. Note that the Bayesian GLM with effect size $\gamma = 0\%$ is analogous to the classical GLM with FWER correction, since both provide similar guarantees around false positive control, and setting $\gamma = 0\%$ is comparable to the null hypothesis of no activation. Yet the Bayesian GLM produced much larger areas of activation at $\gamma = 0\%$ compared to the classical GLM with FWER correction. This is due to the power gained in the Bayesian GLM by leveraging spatial dependencies and avoiding the need for multiplicity correction. FDR correction produced larger areas of activation than FWER correction, but in this participant they were still smaller than those produced with the Bayesian GLM at $\gamma = 0\%$, and do not provide similar guarantees around false positive control.
To quantitatively assess the quality of the areas of activation produced by the Bayesian GLM and classical GLM using different methods and effect sizes, we analyzed the longitudinal stability of the size of those areas in HC participants. Since we do not expect much change in HC participants over the duration of the study, smaller variation in the size of activation over time was considered better.

To quantitatively assess the quality of the areas of activation produced by the Bayesian GLM and classical GLM using different methods and effect sizes, we analyzed the longitudinal stability of the size of those areas in HC participants. Since we do not expect much change in HC participants over the duration of the study, smaller variation in the size of activation over time was considered better. We compared the Bayesian GLM at the three effect sizes ($\gamma = 0\%, 1\%, 2\%$) and the classical GLM using FWER and FDR correction. Since some of these methods tend to result in larger areas of activations, they will tend to have larger variance (since variance is not unit-less), so it is important to consider the size of activation when comparing the variance.

Fig. E7 displays two plots of longitudinal variation in size of activation within HC participants. Both plots illustrate that the Bayesian GLM results in lower variation in size of activation across visits compared with the classical GLM, considering size of activation. In Fig. E7(a), we plot the standard deviation (SD) across visits versus the mean across visits. The line from a linear model relating the SD to the mean for each method is also shown. We observe that, considering mean size of activation, the Bayesian GLM results in lower variation in size of activation across visits compared with the classical GLM. For example, Bayesian GLM with an effect size of $\gamma = 0\%$ and classical GLM with FDR correction often result in similar mean sizes of activation, but the Bayesian GLM has lower variance within HC participants over time. Similarly, the Bayesian GLM with $\gamma = 1\%$ and classical GLM with FWER correction often produce activations of similar size, but the Bayesian GLM has lower within-participant variance.

In Fig. E7(b) we explicitly account for differences in the size of activation through the coefficient of variation (CV), a unit-less measure of variability equal to the standard deviation divided by the mean. For each method and effect size, boxplots display the longitudinal CV within each HC participant. Lower within-participant CV indicates more reliable estimates. This plot shows that the Bayesian GLM with $\gamma = 0\%$ produces highly reliable areas of activation in HC participants. The colored diamonds display the CV between HC participants, based on the mean across visits for each participant. Methods that produce higher between-participant CV better preserve differences between participants. The Bayesian GLM with $\gamma = 1\%$ and $\gamma = 2\%$ perform the strongest in this regard, since they have higher between-participant CV relative to the within-subject CV.
Figure E7: Longitudinal reliability of size of activations in HC participants. (a) Standard deviation (SD) of size of activations versus mean size of activations across visits. Each point represents a HC participant and a method. Lines represent a linear regression fit for each method. The Bayesian GLM results in lower variation in size of activation across visits compared with the classical GLM, considering the mean size of activation. For example, when Bayesian GLM with $\gamma = 0\%$ and FDR correction result in similar mean sizes of activation, the Bayesian GLM has lower within-participant variance. Similarly, when Bayesian GLM with $\gamma = 1\%$ and FWER correction produce activations of similar size, the Bayesian GLM has lower within-participant variance. (b) Coefficient of variation (CV) in activation size. For each method and effect size, boxplots display the longitudinal CV within each HC participant; the colored diamonds display the CV between HC participants, based on the mean across visits for each participant. Methods with lower within-participant CV produce more reliable estimates, and those with higher between-participant CV better respect differences between participants. Therefore, the best case scenario is low within-participant CV (boxplot) and high between-participant CV (diamond). The best methods in this respect are the Bayesian GLM at $\gamma = 0\%$ and $\gamma = 1\%$. 
Additional Results Figures

One individual with ALS (A04) had a very slow disease trajectory (see Fig. 1) and had many more visits spanning a much longer duration, compared with other ALS participants (Fig. A1). To avoid undue influence of this unusual individual on the random intercept models given in Eqn. 3, participant A04 was excluded from model fitting. Here, we present the results of the models for each hemisphere and effect size including this individual. Fig. F8 shows coefficient curves that are very similar to those seen in the main text. This illustrates that the relationships observed between ALS disability and size of activation are robust to the inclusion or exclusion of this individual.

Additionally, we excluded visits from several subjects who had visits spanning more than a two-year window in order to avoid their undue influence on model fit. Figs. F9 and F10 shows coefficient curves that are very similar to those seen in the main text. This illustrates that the relationships observed between ALS disability and size of activation are robust to the inclusion or exclusion of these visits.
Figure F8: Coefficient curves for the size of the area of activation in response to right hand clenching, based on the mixed effects model given in Eqn. 2, including subject A04. The curves are very similar to those observed in the main text Fig. 4. This illustrates that the relationships observed between ALS disability and size of activation are robust to the inclusion or exclusion of this subject.
**Figure F9:** Coefficient curves for the size of the area of activation in response to right hand clenching, based on the mixed effects model given in Eqn. (8), including all visits from all subjects. The curves are very similar to those observed in the main text Fig. 4. This illustrates that the relationships observed between ALS disability and size of activation are robust to the inclusion or exclusion of these visits.
Figure F10: Coefficient curves for the size of the area of activation in response to right hand clenching in fast and slow progressors, based on the mixed effects model given in Eqn. 3, including all visits from all subjects. The curves are very similar to those observed in the main text Fig. 5. This illustrates that the relationships observed between ALS disability and size of activation by progression rate are robust to the inclusion or exclusion of these visits.
Figure F11: Coefficient curves for the size of *ipsilateral* activation in response to right hand clenching, based on the mixed effects model given in Eqn. (3), by progression rate. Here we focus on the relationship between activation size and Hand Motor Disability, with Other Disability fixed at zero. The curves are very similar to those observed in the main text Fig. 5.

(a) Contralateral (Left Hemisphere) Activation

(b) Ipsilateral (Right Hemisphere) Activation

Figure F12: Coefficient curves for the size of activation in response to right hand clenching, based on the mixed effects model given in Eqn. (3), by progression rate. Here we focus on the relationship between activation size and Other Disability, with Hand Motor Disability fixed at zero. Fast progressors show larger size of activation at low levels of disability and decline faster as a function of disability, compared with moderate progressors.