Brain Stimulation 18 (2025) 1855–1870

Contents lists available at ScienceDirect

Brain Stimulation

journal homepage: www.journals.elsevier.com/brain-stimulation

Hierarchical Bayesian estimation of motor-evoked potential recruitment curves yields accurate and robust estimates

Vishweshwar Tyagi [a,*], Lynda M. Murray [g,h], Ahmet S. Asan [d], Christopher Mandigo [b,e], Michael S. Virk [c], Noam Y. Harel [f,g,h], Jason B. Carmel [a,c], James R. McIntosh [a,c,**]

[a] *Neurology, Columbia University, New York, NY, 10032, United States of America*
[b] *Neurological Surgery, Columbia University, New York, NY, 10032, United States of America*
[c] *Neurological Surgery, Weill Cornell Medicine, New York, NY, 10065, United States of America*
[d] *Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, 02142, United States of America*
[e] *New York Presbyterian, The Och Spine Hospital, New York, NY, 10034, United States of America*
[f] *Neurology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, United States of America*
[g] *Rehabilitation and Human Performance, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, United States of America*
[h] *James J. Peters Veterans Affairs Medical Center, Bronx, 10468, NY, United States of America*

ARTICLE INFO

Keywords:
Hierarchical Bayesian
Hypothesis testing
Motor threshold
Transcranial magnetic stimulation
Spinal cord stimulation
Mathematical modeling
Evoked potential

ABSTRACT

Purpose: We aim to develop a robust method to improve the estimation accuracy of motor-evoked potential (MEP) recruitment curves (RCs), including motor threshold, in small-sample settings which typically involve fewer than 40 stimuli.
Methods: We present a hierarchical Bayesian (HB) method to model MEP size as a rectified-logistic function of stimulation intensity. This method is designed to account for small samples, handle outliers without discarding data, quantify estimation uncertainty, and simulate synthetic data that closely matches real observations, useful for optimizing experimental design. We validate its performance on transcranial magnetic stimulation (TMS), epidural spinal cord stimulation (SCS), and synthetic TMS datasets, and provide an open-source library for Python, called hbMEP, for diverse applications.
Results: The rectified-logistic outperformed sigmoidal functions in predictive accuracy on TMS and SCS datasets, as demonstrated through cross-validation. A mixture extension of the HB model improved robustness to outliers by further increasing its predictive accuracy. The HB model reduced threshold estimation error by up to 70% on sparse synthetic TMS data compared to non-hierarchical models. Bayesian estimation with the HB model reduced the required number of participants by at least 23% to detect a shift in threshold with 80% power, compared to frequentist testing. Empirical results on human SCS data further validated its applicability to real data.
Conclusion: By improving accuracy on sparse data, our method minimizes the number of stimuli needed to probe each individual's neuromuscular parameters across multiple muscles simultaneously, thereby reducing session duration and the risk of inadvertent neuromodulation. Our approach provides a more statistically powerful and conclusive framework for inferring changes in threshold, and therefore corticospinal excitability. The hbMEP library streamlines and unifies the analysis of RCs across stimulation modalities and experimental paradigms.

1. Introduction

Dose–response relationships characterized by an S-shape are fundamental across a broad range of disciplines, from pharmacology [1,2] and toxicology [3,4] to psychophysics [5,6] and cognition [7]. In the sensorimotor domain, a prime example is the recruitment curve, which describes how electrical or electromagnetic stimulation intensity affects the size of a motor-evoked potential (MEP)—an electrical signal recorded from multiple muscles (Fig. 1a) in response to stimulation. MEPs are typically quantified by their peak-to-peak (pk-pk) voltage or area under the curve (AUC) (Fig. 1b,c, magenta and green dots).

[*] Corresponding author at: Neurology, Columbia University, New York, NY, 10032, United States of America.
[**] Corresponding author at: Neurology, Columbia University, New York, NY, 10032, United States of America.
E-mail addresses: vt2353@cumc.columbia.edu (V. Tyagi), jrm2263@cumc.columbia.edu (J.R. McIntosh).

https://doi.org/10.1016/j.brs.2025.09.008
Received 6 March 2025; Received in revised form 25 August 2025; Accepted 10 September 2025
Available online 18 September 2025

Recruitment curves are derived using techniques such as transcranial magnetic stimulation (TMS), spinal cord stimulation (SCS), or peripheral nerve stimulation. These techniques are widely used in monitoring and planning clinical interventions or mapping muscle activation [8–14], assessing the extent of injury and tracking recovery [15–17], and evaluating the efficacy of therapeutic interventions, including neuromodulation [18–21].

Accurate estimation of recruitment curves is critically important to assess nervous system state, including corticospinal excitability, and to evaluate therapeutic efficacy [14,22–28]. The recruitment curve exhibits a characteristic sigmoidal or S-shape, with a steep increase above the threshold and a plateau phase at high intensities. Its core properties (see Fig. 1b, c) include 1. offset (background noise floor of the recording), 2. saturation (upper asymptotic or maximal MEP size), 3. S_{50} (intensity to produce 50% of the maximal response above the offset), 4. threshold (intensity to produce minimal consistent response above the offset), and 5. gradient (how MEP size increases with increasing intensity). While each of these properties may have specific neurophysiological interpretations, both the threshold [22,29–31] and S_{50} [26,32,33] have been used to infer changes in corticospinal excitability, although the use of S_{50} is less prominent.

Current approaches [34–37] to model recruitment curves predominantly use sigmoidal functions, typically assumed to be a four-parameter logistic function (Fig. 1b, black curve), and rely on numerical optimization methods applied to non-convex search spaces for estimating the curve parameters. These methods are susceptible to suboptimal solutions with small sample sizes [31,38,39], typically involving fewer than 40 stimuli, require repeated random reinitializations to avoid local minima, and provide only point estimates. However, collecting an adequate number of samples is often infeasible due to constraints concerning experimental time [13], discomfort to participants [40], and the risk of inadvertent neuromodulation when large numbers of stimuli are delivered [41–43].

Moreover, the conventional approach of modeling recruitment curves using sigmoidal functions [22,32,38,44–48] often involves estimating the S_{50} parameter, which is subsequently used to test hypotheses related to shifts in this parameter [26,32,33]. In sigmoidal functions, the slope is directly proportional to S_{50}, and in the logistic-4 function specifically, the maximum gradient or the steepest point occurs at S_{50}. By definition, the estimation of S_{50} is contingent upon observing adequate saturation in data, a condition often unmet due to discomfort experienced by participants at higher stimulation intensities [21,38]. Conversely, the threshold can be estimated accurately independent of saturation, making it a more reliable parameter for testing. However, sigmoidal functions cannot be used to estimate the threshold since these are smooth functions and cannot capture sharp deflections from the offset MEP size [22,38]. Previously, threshold was estimated using a rectified-linear function [13,22,49–51], which proved overly simplistic as it does not capture the curvature in data, leading to biased threshold estimates. In addition, area under the recruitment curve, which characterizes the overall corticospinal output over the intensity range [33,52–59], can also be distorted if curvature is not accurately modeled.

In contrast, Bayesian methodology has shown great potential for improving the statistical modeling process [60], including in neuroscience [61,62]. We introduce a rectified-logistic function (Fig. 1c, black curve) for modeling recruitment curves and integrate it within a hierarchical Bayesian framework that accounts for small sample sizes, handles outliers, and returns a posterior distribution (Fig. 1c, bottom and side panels) over the curve parameters, thereby quantifying estimation uncertainty. We evaluate the predictive performance of rectified-logistic function against commonly used sigmoidal alternatives using cross-validation on empirical TMS and SCS data. Our framework is generative and enables simulation of high-fidelity synthetic data for model comparison and optimization of experimental design. In simulations, we assess its accuracy in estimating threshold on sparse data and compare it with conventional non-hierarchical models. Further, we develop a Bayesian estimation method for detecting shift in threshold and evaluate its statistical power against frequentist null hypothesis testing. Finally, we present a common use case involving human epidural SCS data and provide an open-source library for Python called hbMEP for diverse applications.

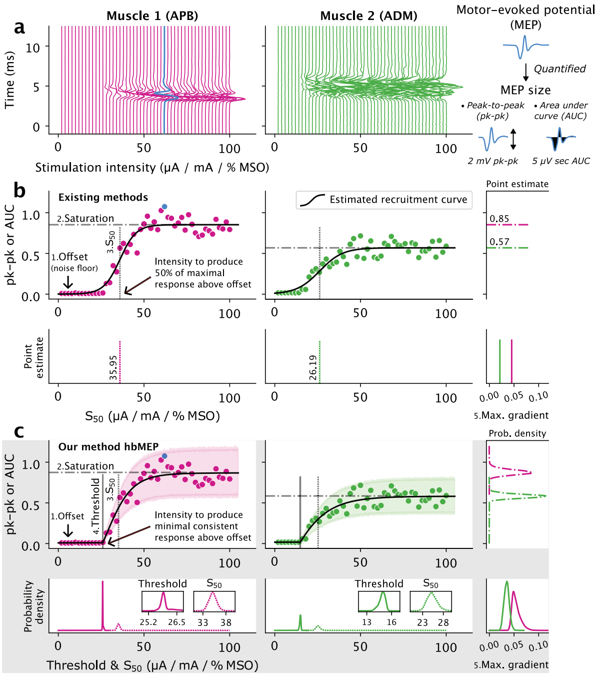


Fig. 1. Hierarchical Bayesian estimation yields posterior distributions over recruitment curve parameters for each participant across multiple muscles simultaneously. (a) Example motor-evoked potentials (MEPs) recorded at different stimulation intensities from a participant's APB (left, magenta) and ADM (middle, green) muscles. The abscissa represents stimulation intensity, specified in units of current such as μA or mA for spinal cord stimulation, or % maximum stimulator output (% MSO) for transcranial magnetic stimulation. Right panel: schematic quantification of MEPs into MEP size using either peak-to-peak (pk-pk) amplitude or area under the curve (AUC). **(b)** Example recruitment curves modeled as a four-parameter logistic function (Boltzmann sigmoid) using least squares minimization. It provides only point estimates for the curve parameters, lacks a threshold estimate, and fails to capture sharp deflection from the offset. Bottom panels: point estimate of S_{50}. Top right: saturation. Bottom right: maximum gradient. **(c)** Example recruitment curves modeled as a five-parameter rectified-logistic function within a hierarchical Bayesian framework. Shading represents the 95% highest density interval (HDI) of the posterior predictive distribution. It accurately estimates the threshold, S_{50}, and saturation. The estimation uncertainty for each parameter is quantified by the width of the 95% HDI. Data from multiple participants and muscles is handled simultaneously. Bottom panels: posterior distribution of the threshold and S_{50}. Inset: zoom over posterior. Top right: saturation. Bottom right: maximum gradient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Methods

In Section 2.1 and Section 2.2, we introduce Bayesian hierarchical models for estimating MEP size recruitment curves. Section 2.1 begins with a review of conventional functions used to model recruitment curves, including the rectified-linear, logistic-4, and logistic-5 functions. We then introduce the rectified-logistic function, which combines aspects of these functions to enable accurate estimation of the threshold parameter. A gamma likelihood model that uses the rectified-logistic

Table 1
Overview of hierarchical and non-hierarchical models, their application to datasets, and the corresponding analyses and figures.

Model	Analysis	Dataset
[a]Standard HB	Parameter recovery (threshold, S_{50})	Synthetic TMS (Fig. 2e–f, Supplementary Fig. S1)
	Statistical power of detecting shift in threshold	Synthetic TMS (Fig. 2g,h, Fig. 6e), Human SCS (Fig. 5c, Supplementary Fig. S2)
	Comparison of predictive performance of different recruitment curve functions	Rat SCS, Human TMS and SCS (Fig. 3)
	Optimizing experimental design	Synthetic TMS (Fig. 6a–d)
[b]Mixture HB	Predictive performance compared to standard HB for robustness to outliers	Rat SCS, Human TMS and SCS (Fig. 4)
[c]Paired comparison (HBe)	Statistical power of detecting shift in threshold	Synthetic TMS (Fig. 2g,h, Fig. 6e), Human SCS (Fig. 5, Supplementary Fig. S2)
[d]Non-hierarchical (nHB, ML, LSM)	Parameter recovery (threshold)	Synthetic TMS (Fig. 2e,f)
	Statistical power of detecting shift in threshold	Synthetic TMS (Fig. 2g,h), Human SCS (Fig. 5c, Supplementary Fig. S2)

[a] Standard hierarchical Bayesian (HB) model.
[b] Mixture extension of the standard HB model.
[c] Hierarchical Bayesian estimation (HBe) model.
[d] Non-hierarchical Bayesian (nHB), maximum likelihood (ML), least squares method (LSM).

function is presented, followed by its integration into a hierarchical Bayesian framework in Section 2.2. This section begins with a standard model and introduces a paired comparison model that uses Bayesian estimation to compare the threshold parameter in participants undergoing repeat experiments. We conclude the section with a mixture extension of the gamma likelihood model to account for outliers in data.

In Sections 2.3–2.5, we describe the application of these models to obtain the results presented in the manuscript. Table 1 provides an overview of the models used and the datasets to which they were applied, including the corresponding analyses and figures. Section 2.3 validates the robustness and efficiency of hierarchical Bayesian methods on simulated data, where we assess the accuracy of the standard model on sparse data for threshold estimation and evaluate the statistical power of the Bayesian estimation approach for detecting shifts in the threshold. In Section 2.4, we validate the predictive performance of the rectified-logistic function through cross-validation on empirical data and compare the results against conventional alternatives. We also validate the mixture extension of the gamma likelihood model using the same cross-validation procedure. In Section 2.5, we present a common use case of the paired comparison model on human SCS data and evaluate its statistical power on bootstrapped data. We conclude the section by describing how the generative capability of the framework can be used to optimize experimental design.

Section 2.6 outlines the statistical methods used and reproducibility of the analyses presented. Sections 2.7 and 2.8 describe the human TMS and SCS, and the rat SCS datasets, respectively. Table 2 provides a concise overview of these datasets. Appendix A1–A3 and B1–B3 show recruitment curves estimated on these datasets using the standard hierarchical Bayesian model and its mixture extension, respectively, with the rectified-logistic function.

We denote $\mathbb{R}$ as the set of real numbers and $\mathbb{R}^+$ as the set of positive reals. $\mathrm{N}(\mu, \sigma)$ or $\mathrm{Normal}(\mu, \sigma)$ denotes the normal distribution (location-scale), and $\mathrm{Gamma}(\alpha, \beta)$ the gamma distribution (shape-rate). $\mathrm{TN}(\mu, \sigma)$ and $\mathrm{HN}(\sigma)$ are the truncated-normal (location-scale, with left truncation at zero) and half-normal (scale) distributions, respectively. $\mathrm{Bernoulli}(p)$ is the Bernoulli distribution with success probability p, and $\mathrm{Uniform}(a, b)$ is the uniform distribution on $[a, b]$. For convenience, the terms statistical power and true positive rate are used interchangeably.

2.1. Modeling MEP size

The different choices for modeling recruitment curves include a three-parameter rectified-linear function [13,22,49–51] (Eq. (1), Fig. 3a) and a four-parameter logistic-4 function [22,38,45–48] (Eq. (2), Fig. 1b, 3b). Additionally, a five-parameter logistic-5 [44] (Eq. (3), Fig. 3c) is a more generalized version of logistic-4 and contains an extra parameter υ to control near which asymptote (lower L, or upper $L + H$) the maximum growth or the inflection point occurs. Unlike logistic-4, the logistic-5 function is not necessarily symmetrical about its inflection point.

$$\forall a, b, L > 0 \quad x \mapsto L + \max\{0, b(x-a)\} \tag{1}$$

$$\forall a, b, L, H > 0 \quad x \mapsto L + \frac{H}{1 + e^{-b(x-a)}} \tag{2}$$

$$\forall a, b, L, H, \upsilon > 0 \quad x \mapsto L + \frac{H}{\left\{1 + (2^{\upsilon} - 1)\, e^{-b(x-a)}\right\}^{\frac{1}{\upsilon}}} \tag{3}$$

In the rectified-linear function, parameter a is the threshold, and in the logistic functions, it is the S_{50}. L represents the offset MEP size, $(L + H)$ defines the saturation, and b is the growth rate. The logistic functions do not have a parameter for the threshold since they are smooth functions, and the rectified-linear function does not have a parameter for the S_{50} since it does not saturate.

Likelihood model

We introduce a five-parameter rectified-logistic function (Eq. (4), Fig. 1c, 3d) that can estimate both the threshold and S_{50}. Supplementary Fig. S3a–f shows the effect of varying its different parameters. Parameters b, L, H have similar interpretation as in the logistic-4 function, and a is the threshold. Similar to the logistic-5, there is an additional parameter ℓ that controls the location of inflection point, whether near the offset L or saturation $(L + H)$. Eq. (5) gives the S_{50} of the rectified-logistic function. For $a, b, L, H, \ell > 0$, define the rectified-logistic function $F : \mathbb{R} \to \mathbb{R}^+$ as

$$F(x) = L + \max\left\{0, -\ell + \frac{H + \ell}{1 + \left(\frac{H}{\ell}\right) e^{-b(x-a)}}\right\} \tag{4}$$

$$S_{50}(F) = a - \frac{1}{b} \ln\left(\frac{\ell}{H + 2\ell}\right) \tag{5}$$

We use a gamma likelihood model in the shape-rate parametrization (Eq. (6)–(8)) to model the relationship between MEP size (y) and stimulation intensity (x). Specifically, we model the expected MEP size

Table 2
Summary of real datasets used in this study.

Dataset	Number of participants	[a]Dataset size	Recruitment curves (RCs)	[b]Number of samples per RC
[c]Rat SCS	8	7676 samples × 6 muscles	150	51.2 ± 4.5
[d]Human TMS	27	1843 samples × 6 muscles	27	68.3 ± 14.2
[e]Human SCS	13	1150 samples × 4 muscles	26	44.2 ± 22.3

[a] Each sample corresponds to stimulation at a given intensity, and responses are recorded simultaneously from multiple muscles.

[b] Mean ± standard deviation of the number of samples per recruitment curve.

[c] Electrode arrays were placed over the cervical spinal cord and stimulated using 21 distinct spatial configurations. Intensity is measured in μA, AUC response in μV s.

[d] Includes 14 uninjured and 13 spinal cord injury participants. Intensity in % MSO, peak-to-peak amplitude response in mV.

[e] Each participant was stimulated at both midline and lateral cervical spinal locations. Intensity in mA, AUC response in μV s.

as a rectified-logistic function of intensity, since $\mathbb{E}\left(y \mid x, \Omega, c_1, c_2\right) = \mu = F(x \mid \Omega)$. For $c_1, c_2 > 0$ and $\Omega = \{a, b, L, \ell, H\}$

$$y \mid x, \Omega, c_1, c_2 \sim \text{Gamma}\,(\mu\beta, \beta) \tag{6}$$

$$\mu = F(x \mid \Omega) \tag{7}$$

$$\beta = \frac{1}{c_1} + \frac{1}{c_2 \mu} \tag{8}$$

We chose a gamma distribution to capture the long-tailed distribution of MEP size around the recruitment curve. We specify the rate parameter β of the gamma distribution as a linear function of the reciprocal of expected MEP size $\left(\frac{1}{\mu}\right)$ with positive weights $\left(\frac{1}{c_1}, \frac{1}{c_2}\right)$ to capture the heteroskedastic spread that increases with increasing MEP size.

Recruitment curves

More generally, F (Eq. (7)) is called the recruitment curve function in the context of modeling MEP size, which transforms the input stimulation intensity (x), and links it to the expected MEP size $\mathbb{E}(y \mid x)$. F can be replaced by other available choices, including the rectified-linear, logistic-4, or logistic-5.

The parametric definition of threshold discussed thus far differs from the resting motor threshold (RMT) used in TMS studies, which is the minimum stimulation intensity that produces a predefined MEP size of 50 μV in at least 50% of repetitions [63]. In general, RMT detects the threshold after the MEP size is already above the offset noise floor (see Discussion).

2.2. Hierarchical Bayesian model

Standard model

The simplest form of a standard three-stage hierarchical Bayesian model (Eq. (9)–(11)) for modeling MEP size can be described as follows. Let there be $N_P \times N_M$ exchangeable sequences $\left\{ \left(x_i{}^p, y_i{}^{p,m}\right)_{i=1}^{n(p)} \mid p = 1, 2 \ldots N_P,\ m = 1, 2 \ldots N_M \right\}$, where $y_i{}^{p,m} \in \mathbb{R}^+$ represents the MEP size recorded at stimulation intensity $x_i{}^p \in \mathbb{R}^+ \cup \{0\}$ from muscle m of participant p, for a total of N_M muscles of N_P participants. Here $n(p)$ denotes the number of intensities, or stimuli, tested for participant p, which is independent of muscle m, since MEP size $y_i{}^{p,m}$ is recorded simultaneously from all muscles $m = 1, 2 \ldots N_M$ at a given intensity $x_i{}^p$.

The first stage of hierarchy is the participant-level (Eq. (9)). It specifies the likelihood model $P\left(y_i{}^{p,m} \mid x_i{}^p, \theta^{p,m}\right)$ for each of the N_M muscles of N_P participants, and models the MEP size $y_i{}^{p,m}$ as a function of intensity $x_i{}^p$ and participant-level parameters $\theta^{p,m}$. In the second stage (Eq. (10)), the participant-level parameters $\theta^{p,m}$ are assumed to be drawn from a common distribution $P(\theta^{p,m} \mid \gamma)$, which depends on the population-level hyper-parameters γ. In the third stage (Eq. (11)), the population-level hyper-parameters γ are assumed to be unknown and assigned a weakly informative prior $P(\gamma)$, also called the hyperprior.

$$\text{Stage I} \qquad y_i{}^{p,m} \sim P\left(y_i{}^{p,m} \mid x_i{}^p, \theta^{p,m}\right) \tag{9}$$

$$\text{Stage II} \qquad \theta^{p,m} \sim P(\theta^{p,m} \mid \gamma) \tag{10}$$

$$\text{Stage III} \qquad \gamma \sim P(\gamma) \tag{11}$$

Supplementary Fig. S3g,h specifies the standard hierarchical Bayesian model for human TMS data that was evaluated for its accuracy and statistical power (Results 3.1, 3.2, see Methods 2.3 for details). We also compared the rectified-logistic function against other available choices using the same model structure (Results 3.3, see Methods 2.4 for detailed implementations). Appendix A1–A3 show curves estimated using the standard HB model for rat SCS, human TMS, and human SCS data, respectively. The shaded region shows the 95% highest density interval (HDI) of the posterior predictive distribution, which is narrow and covers most of the observed data, indicating the model does a good job of capturing the variability. The vertical dashed line gives the point estimate of the threshold, with its full posterior distribution shown below. The right-side density plot shows the posterior distribution of the saturating MEP size.

Paired comparison

This section presents a hierarchical model that is useful for estimating shift in curve parameters. This is applicable in settings where the same set of participants are tested for multiple experimental conditions (repeat measurements), such as pre- and post-intervention phases, stimulation locations (e.g., midline or lateral), or stimulation parameters (e.g., electrode size, stimulation frequency).

Supplementary Fig. S4 gives the graphical representation of such a model used to summarize differences in the threshold parameter. Here we have the threshold $a^{p,c,m}$ of participant p, at tested condition c and muscle m given by,

$$a^{p,c,m} = \begin{cases} a_{\text{fixed}}^{p,m} & c = 1 \\ a_{\text{fixed}}^{p,m} + a_{\Delta}^{p,c,m} & c > 1 \end{cases} \tag{12}$$

The threshold is broken (Eq. (12)) into a fixed component $(c = 1)$ and a shift component $(\forall c > 1)$ that measures the difference from the fixed component. The shift component is parametrized by condition-level location $\left(\mu_{a_\Delta}^{c,m}\right)$ and scale $\left(\sigma_{a_\Delta}^{c,m}\right)$ hyperparameters. The location hyperparameters $\left(\mu_{a_\Delta}^{c,m}\right)$ summarize the shift of each tested condition $(\forall c > 1)$ from the fixed component $(c = 1)$ for each muscle, and the scale parameters $\left(\sigma_{a_\Delta}^{c,m}\right)$ measure the variability in the estimated shifts.

A priori we assume there is no shift from the fixed component and the location hyperparameters $\left(\mu_{a_\Delta}^{c,m}\right)$ are given a flat prior which is symmetric about zero. Once the model is fit, the 95% HDI of the posterior is used to assess the strength of shift for condition c and muscle m. The same model structure can also be used to summarize differences in other curve parameters, once they are parametrized by location-scale hyperparameters.

Extension to mixture model

The models discussed so far can be extended to handle outliers by replacing the gamma distribution (Eq. (6)) with a two-component mixture of gamma and half-normal distributions. The resultant likelihood model (Eq. (13)–(16)) is given as,

$$y \mid x \sim \left(1 - q_y\right) \cdot \text{Gamma}\,(\mu\beta, \beta) + q_y \cdot \text{HN}\left(\sigma_{\text{outlier}}\right) \tag{13}$$

Table 3
Number of parameters in a standard hierarchical Bayesian model.

Model component	Choice	[a]Number of parameters
Curve function (N_{curve})	Rectified-logistic	5 (a, b, L, H, ℓ)
	Logistic-5	5 (S_{50}, b, L, H, v)
	Logistic-4	4 (S_{50}, b, L, H)
	Rectified-linear	3 (a, b, L)
Likelihood model (N_{like})	Gamma	2 (c_1, c_2)
[b]Mixture extension	Outlier probability	1 (p_{outlier})
Total parameters	[c]Standard HB model (N_{HB})	$N_{\text{curve}} + N_{\text{like}} + N_{\text{parent}}$
	Mixture extension of standard HB	$N_{\text{HB}} + 1$

[a] a threshold, b controls slope, L offset, H distance to saturation, ℓ controls location of inflection point, S_{50} intensity at half-maximum response above offset, v controls asymmetry.

[b] Includes Bernoulli indicators q_y for each observation, which are latent and not counted.

[c] Curve and likelihood parameters scale linearly with the number of recruitment curves, and include a constant number of parent hyperparameters (N_{parent}) determined by the choice of curve and likelihood.

$$q_y \sim \text{Bernoulli}\left(p_{\text{outlier}}\right) \tag{14}$$

$$p_{\text{outlier}} \sim \text{Uniform}\left(0, C_{p_{\text{outlier}}}\right) \tag{15}$$

$$\sigma_{\text{outlier}} \sim \text{HN}\left(L + H\right) \tag{16}$$

where $C_{p_{\text{outlier}}} \in (0, 1)$ is a constant and chosen to be small, usually in the range $(0.01, 0.05)$. Intuitively, this means that we expect roughly $1\%–5\%$ outliers, to be captured by the half-normal distribution. Supplementary Fig. S5 shows the mixture extension of the standard model. Table 3 provides a component-wise breakdown of the number of parameters in the standard HB model and its mixture extension. Appendix B1–B3 show curves estimated using the mixture extension of the standard HB model for rat SCS, human TMS, and human SCS data, respectively.

More generally, the gamma likelihood (Eq. (6)) can be replaced with alternative distributions [39,64–67], for example, a log-normal. Supplementary Methods S1.1 discusses a few such alternatives.

2.3. Robustness & efficiency

Accurate estimation of threshold on sparse data

In Results 3.1 (Fig. 2a–d), we used the standard model (Supplementary Fig. S3g,h) to estimate participant- and population-level parameters from TMS data. We used data from the APB muscle ($N_M = 1$), which was the target muscle for 21 of the total 27 participants ($N_P = 27$, Fig. 2a). The model was conditioned on the estimated participant-level parameters $\left(c_1, c_2, a \ldots H\right)$ to replicate the observed participants (Fig. 2b). It was conditioned on the estimated population-level parameters $\left(\sigma_{c_1}, \sigma_{c_2}, \mu_a \ldots \sigma_H\right)$ to simulate new participants (Fig. 2c).

We used the first two components of principal component analysis (PCA) to visualize the participant-level parameters on Cartesian plane (Fig. 2d). The PCA map was fit on parameters estimated from existing TMS participants (Fig. 2d, pink dots). The map was used to project parameters simulated from the prior predictive distribution (blue dots) and parameters of the new simulated participants (green dots) on the Cartesian plane.

The estimated population-level parameters consisted of 10,000 posterior samples (10 chains, 1000 samples each). A total of 16 participants were simulated conditioned on the estimated population-level parameters—resulting in 10,000 distinct draws, each consisting of 16 participants. The threshold values for these draws were used as ground truth for a comparative analysis in Results 3.1 (Fig. 2e,f).

In Results 3.1 (Fig. 2e,f), we compared the standard hierarchical Bayesian model (HB, Supplementary Fig. S3g,h) against three non-hierarchical models to assess how partial pooling across participants affects the accuracy of estimating the threshold. These included a non-hierarchical Bayesian (nHB) model, implemented equivalently to the HB model using the same priors, except without any pooling; maximum likelihood (ML) model implemented using uniform priors for the participant-level parameters; and the least squares method (LSM), which utilized the SciPy [68] library to minimize the residual sum of squares between the observed data points and the estimated recruitment curve fit. The cost function was minimized using the Nelder–Mead method [69] for which reasonable boundaries were set for each parameter of the curve. The Nelder–Mead method was reinitialized 20 times with different starting points to avoid local minima, and the threshold point estimate was chosen based on the minimum cost function. For the HB, nHB, and ML models, point threshold estimates were calculated using the mean of the threshold posterior. These point estimates were used to compute mean absolute error from the ground truth thresholds (described below). Fig. 2e consisted of 48 equispaced stimulation intensities between 0–100% MSO (in Supplementary Fig. S3g, $n(p) = 48 \ \forall p$). Fig. 2f consisted of the first 8 participants of each draw ($N_P = 8$). Both analyses involved a single repetition per intensity and were repeated for 4000 draws, which were randomly chosen without replacement from the total 10,000 simulated draws.

The mean absolute errors were calculated as follows—let $a_1^d, a_2^d, \ldots, a_{16}^d$ be the true thresholds for the sixteen participants of the dth draw, and let $\hat{a}_1^d, \hat{a}_2^d, \ldots, \hat{a}_n^d$ be the corresponding point estimates of a model for the first $n \in \{1, 2, \ldots, 16\}$ participants. Then, the error for n participants of the dth draw is given by $e_{n,d} = \frac{1}{n} \sum_{p=1}^{n} |a_p^d - \hat{a}_p^d|$. Finally, the error for n participants (Fig. 2e,f) across all 4000 draws is given by $e_n = \frac{1}{4000} \sum_{d=1}^{4000} e_{n,d}$, which is the sample mean of $\{e_{n,1}, e_{n,2} \ldots e_{n,4000}\}$. The error bars (Fig. 2e,f) represent the standard error of e_n given by $\text{SE}_{e_n} = \frac{\sigma_{e_n}}{\sqrt{4000}}$, where $\sigma_{e_n} = \left\{ \sum_{d=1}^{4000} \left(e_{n,d} - e_n\right)^2 / (4000 - 1) \right\}^{\frac{1}{2}}$ is the sample standard deviation.

Bayesian estimation for detecting a shift in threshold

In Results 3.2 (Fig. 2g,h), we simulated a total of 20 participants. The parameters for both pre- and post-intervention phases, except for the post-intervention thresholds, were simulated by conditioning the model (Supplementary Fig. S3g,h) on the estimated population-level parameters. The post-intervention thresholds were obtained by subtracting values from the pre-intervention thresholds, where the values to be subtracted were simulated from a normal distribution $\text{N}(\mu = -5, \sigma = 2.5)$ for a negative shift, and $\text{N}(0, 2.5)$ for no shift. This assumed that the intervention did not alter the distribution of any parameter other than the threshold.

The null hypothesis assumed zero shift from pre- to post-intervention, whereas the alternative hypothesis posited a non-zero shift. Supplementary Fig. S6 (with $2 \leq N_p \leq 20$) specifies the paired comparison model, or the hierarchical Bayesian estimation (HBe) approach, that was compared against the standard hierarchical Bayesian (HB, Supplementary Fig. S3g,h) and non-hierarchical models. For the HBe model, the null hypothesis was rejected if the 95% HDI of the population-level location hyperparameter $\left(\mu_{a_\Delta}\right)$ excluded zero; otherwise, the null hypothesis was not rejected. For the HB, nHB, ML, and LSM models, a two-sided Wilcoxon signed-rank test [70] was conducted on their point threshold estimates. The significance level was set at 5% and the null hypothesis was rejected if the p-value was below 0.05. A t-test was not applicable due to non-normality of estimated pairwise threshold differences as indicated by Shapiro–Wilk test. The analyses (Fig. 2g,h) consisted of a single repetition of 48 equispaced stimulation intensities between 0–100% MSO, and were repeated for 2000 draws, randomly chosen without replacement from the total 10,000 simulated draws.

The true and false positive rates were calculated as follows—let H_0 be the null hypothesis (zero shift), H_1 be the alternative hypothesis (non-zero shift). We define the indicator variable $\mathbf{1}_{n,d} \in \{0,1\}$ which evaluates to 1 if a statistical test (e.g., the 95% HDI test for the HBe model, and the two-sided signed-rank test for other models) rejects H_0 based on the first n participants of the dth draw, and 0 otherwise. We distribute the 2000 draws into 20 blocks of 100 draws each, so that the bth block consists of draws with indices in $\mathrm{B}_b = \{100b-99, 100b-98 \dots 100b\}$. Define $\pi_{n,b} = \frac{1}{100}\sum_{d\in\mathrm{B}_b} \mathbf{1}_{n,d}$, which is the sample mean of the set of binary values $\left\{\mathbf{1}_{n,d} \mid d \in \mathrm{B}_b\right\}$. Additionally, define $\pi_n = \frac{1}{20}\sum_{b=1}^{20} \pi_{n,b}$, which is the sample mean of the set of values $\left\{\pi_{n,1} \dots \pi_{n,20}\right\}$. When the differences come from $\mathrm{N}(-5, 2.5)$, the null hypothesis is false and the true positive rate (Fig. 2g) is given by π_n. When the differences come from $\mathrm{N}(0, 2.5)$, the null hypothesis holds and the false positive rate (Fig. 2h) is given by the same π_n. The error bars (Fig. 2g,h) represent the standard error of π_n given by $\mathrm{SE}_{\pi_n} = \frac{\sigma_{\pi_n}}{\sqrt{20}}$, where $\sigma_{\pi_n} = \left\{\sum_{b=1}^{20}\left(\pi_{n,b} - \pi_n\right)^2 / (20-1)\right\}^{\frac{1}{2}}$ is the sample standard deviation.

2.4. Choice of recruitment curve function

In Results 3.3 (Fig. 3e–g), we used the standard model structure for cross-validation [71]. Supplementary Fig. S7 specifies the standard model for rat epidural SCS, human TMS and human epidural SCS datasets with the rectified-logistic, logistic-5, logistic-4 and rectified-linear functions. For the rat epidural SCS data, 150 recruitment curves were fit simultaneously on six muscles: abductor digiti minimi (ADM), biceps, deltoid, extensor carpi radialis longus (ECR), flexor carpi radialis (FCR), and triceps—for a total of 900 curves. For the human TMS data, 27 curves were fit simultaneously on six muscles: ADM, abductor pollicis brevis (APB), biceps, ECR, FCR, and triceps—for a total of 162 curves. For the human SCS data, 26 curves were fit simultaneously on four muscles: ADM, APB, biceps, and triceps—for a total of 104 curves. An ArviZ [72] implementation of cross-validation [71] was used to compute expected log-pointwise predictive density (ELPD) scores and pairwise differences from the best-ranked model.

In Results 3.4 (Fig. 4c), the standard model with rectified-logistic function and gamma distribution (Supplementary Fig. S7 with rectified-logistic function, Eq. (6)–(8)) was compared to its mixture extension (Eq. (13)–(16)). Supplementary Fig. S8 specifies the mixture model on all datasets.

Additionally, we evaluated how varying levels of saturation in the data affect the accuracy of estimating the threshold and S_{50} parameters (Supplementary Fig. S1). Using the standard model (Supplementary Fig. S3g,h) with the rectified-logistic function (Eq. (4)), we estimated the threshold parameter for the first eight participants ($N_P = 8$) of synthetic TMS data (Methods 2.3). This analysis consisted of a single repetition of 48 equispaced stimulation intensities ranging from 0–70% MSO ($n(p) = 48\ \forall p$), and was repeated 4000 times. The full stimulator output range 0–100% MSO was not included to mimic realistic experimental constraints where higher intensities are often excluded due to participant discomfort. For each participant and draw, the mean absolute error was calculated as described in Methods 2.3. These errors were subsequently grouped into bins based on the saturation levels observed in the data, which we define as follows—let F represent the rectified-logistic function (Eq. (4)), and let $\Omega = \{a, b, L, \ell, H\}$ represent the set of simulated parameters for a given participant and draw. Then, $\frac{F(70|\Omega)}{L+H}$ gives the proportion of saturation observed in this case, which is the ratio of the function evaluated at the highest tested intensity $F(70 \mid \Omega)$ and the actual saturation level $(L + H)$. Supplementary Fig. S1 displays the sample mean of the errors for each bin, and the error bars represent the standard error of the mean. The estimation error for the S_{50} parameter was calculated similarly, except the rectified-logistic function in the standard model was reparametrized to explicitly include S_{50} as a model parameter, to enable its partial pooling across participants. Eq. (17) gives the reparametrized rectified-logistic function, where the parameter a represents the S_{50}. For $a, b, L, \ell, H > 0$, define the rectified-logistic function in the S_{50} parametrization $G : \mathbb{R} \rightarrow \mathbb{R}^+$ as

$$G(x) = L + \max\left\{0, -\ell + \frac{H+\ell}{1 + \left(\frac{H}{H+2\ell}\right) e^{-b(x-a)}}\right\} \tag{17}$$

2.5. Common use case

Paired comparison on human SCS data

In Results 3.5 (Fig. 5), we used the mixture extension of the paired comparison model to estimate the threshold differences between midline and lateral stimulation positions. Supplementary Fig. S9 specifies the model used for this analysis. Here, $c = 1$ and $c = 2$ represent lateral and midline stimulation positions, respectively. Fig. 5b displays the HDI of the location hyperparameter $\left(\mu_{a_\Delta}^{m}\right)$ for each muscle. All muscles of the arm and hand (biceps, triceps, APB, and ADM) recorded simultaneously from all 13 participants were analyzed.

In Fig. 5c and Supplementary Fig. S2, we bootstrapped the data from 13 participants to evaluate the statistical power and family-wise error rate of the HBe model. The results were compared against the standard hierarchical Bayesian (HB, Supplementary Fig. S7 with human SCS hyperpriors) and non-hierarchical models. Due to sparse data, the least squares method was reinitialized 1000 times to avoid local minima. For evaluating the power (Fig. 5c), participants were randomly sampled with replacement. For evaluating the family-wise error rate (Supplementary Fig. S2), the midline and lateral conditions were, with equal probability, either interchanged or kept the same, independently for each participant sampled with replacement. This removed any effect between the two conditions. However, since the muscles were recorded simultaneously, the conditions were either interchanged for all muscles or for none of them, for a given participant. Hence, this shuffling procedure removed the effect in all muscles. Due to this constraint, we could only evaluate the family-wise error rate in the weak sense, i.e., the probability of at least one false positive when the null hypothesis (no difference between the thresholds of the two conditions) is true in all muscles. Since we do not account for outliers in the nHB, ML, and LSM models, the mixture distribution in the HBe model was turned off prior to evaluating it on the bootstrapped data (Supplementary Fig. S9, $q_y = 0$). Both analyses in Fig. 5c and Supplementary Fig. S2 were repeated for 2000 bootstrap draws.

For the standard HB and non-hierarchical models, a two-sided Wilcoxon signed-rank test [70] was conducted on their point threshold estimates, similar to Methods 2.3. A Bonferroni-Holm correction [73] was applied to the p-values of the four tested muscles to control the family-wise error rate at the 5% significance level. For the HBe model, a similar procedure was applied to the posterior probabilities of rejecting the null hypothesis. Consider the posterior of the location hyperparameter $\mu_{a_\Delta}^{m}$ of muscle m, and define p_m^+ and p_m^- as the proportions of its posterior samples that are greater and less than 0, respectively. Define $p_m = \max\left\{p_m^+, p_m^-\right\}$ and let H_m be the null hypothesis for muscle m. Without loss of generality, we can assume $p_1 \geq p_2 \geq p_3 \geq p_4$. With $\alpha = 0.05$, the correction procedure for the HBe model was as follows,

```
$m \leftarrow 1$
while $m \leq 4$ do
    if $p_m \geq 1 - \frac{\alpha}{2\cdot(4-m+1)}$ then
        reject $H_m$
    else
```

```
            break
        end if
        $m \leftarrow m + 1$
    end while
```

Whenever the above procedure exits, we fail to reject the remaining hypotheses. The true positive, false positive, and family-wise error rates were calculated as follows—define the indicator variable $\mathbf{1}^m_{n,d} \in \{0,1\}$ which evaluates to 1 if a corrected testing procedure rejects H_m based on the first n participants of the dth bootstrap draw, and 0 otherwise. We distribute the 2000 bootstrap draws into 20 blocks of 100 draws each, so that the bth block consists of draws with indices in $\mathrm{B}_b = \{100b - 99, 100b - 98 \ldots 100b\}$. Define $\pi^m_{n,b} = \frac{1}{100} \sum_{d \in \mathrm{B}_b} \mathbf{1}^m_{n,d}$ and $\pi^m_n = \frac{1}{20} \sum_{b=1}^{20} \pi^m_{n,b}$. When the conditions are not interchanged, the null hypothesis is false for all muscles and the true positive rate (Fig. 5c) for muscle m is given by π^m_n. When the conditions are, with equal probability, either interchanged or kept the same, independently for each sampled participant, the null hypothesis holds in all muscles and the false positive rate (Supplementary Fig. S2a–d) for muscle m is given by the same π^m_n. The error bars (Fig. 5c, Supplementary Fig. S2a–d) represent the standard error of π^m_n given by $\mathrm{SE}_{\pi^m_n} = \sigma_{\pi^m_n}$ where $\sigma_{\pi^m_n} = \left\{ \sum_{b=1}^{20} \left(\pi^m_{n,b} - \pi^m_n \right)^2 / (20 - 1) \right\}^{\frac{1}{2}}$ is the sample standard deviation.

Furthermore, we define $\pi_{n,b} = \frac{1}{100} \sum_{b \in \mathrm{B}_b} \mathbf{1}_{n,d}$ where $\mathbf{1}_{n,d} = \max\{\mathbf{1}^m_{n,d} \mid m = 1 \ldots 4\} \in \{0,1\}$ is again an indicator variable that evaluates to 1 if a corrected testing procedure rejects at least one of the hypotheses $H_1 \ldots H_4$ based on the first n participants of the dth draw, and 0 otherwise. Then, $\pi_n = \frac{1}{20} \sum_{b=1}^{20} \pi_{n,b}$ gives the overall family-wise error rate (Supplementary Fig. S2e) when the conditions are interchanged. The error bars (Supplementary Fig. S2e) represent the standard error of π_n given by $\mathrm{SE}_{\pi_n} = \sigma_{\pi_n}$ where $\sigma_{\pi_n} = \left\{ \sum_{b=1}^{20} \left(\pi_{n,b} - \pi_n \right)^2 / (20 - 1) \right\}^{\frac{1}{2}}$ is the sample standard deviation.

Optimizing experimental design

In Results 3.6 (Fig. 6a–d), we evaluated the effect of single versus multiple repetitions per stimulation intensity on the accuracy of threshold estimation. For the first eight participants of synthetic TMS data (Methods 2.3), we simulated a total of eight observations per intensity, across a total of 64 equispaced intensities between 0–100% MSO. With repetition counts $r \in \{1, 4, 8\}$ and total number of stimuli (including repetitions) $T \in \{32, 40, 48, 56, 64\}$ ($n(p) = T \; \forall p$), the number of unique intensities tested was given by T/r, which were subsampled from the initial set of 64 equispaced intensities. Supplementary Fig. S3g,h specifies the model used to estimate the thresholds. This analysis was repeated for 4000 draws, and the errors were calculated as described in Methods 2.3.

Additionally, we evaluated the statistical power of the hierarchical Bayesian approach (HBe, Supplementary Fig. S4) to detect threshold shift between pre- and post-intervention phases, depending on the number of repetitions per intensity (Fig. 6e). For a fixed total of 64 stimuli ($T = 64$), and repetition counts $r \in \{1, 4, 8\}$, the number of unique intensities tested was again given by T/r. The parameters were simulated, and true positive rates were calculated as described in Methods 2.3. This analysis was repeated for 2000 draws.

2.6. Statistics & reproducibility

All Bayesian models were implemented in NumPyro [74,75], using the No-U-Turn Sampler (NUTS) [76].

For paired comparison of point threshold estimates across conditions, we used the two-sided Wilcoxon signed-rank test [70]. When multiple muscles were tested simultaneously, the Bonferroni-Holm correction [73] was applied to control the family-wise error rate at 5%. For the hierarchical Bayesian estimation (HBe) model, a procedure similar to Bonferroni-Holm correction was applied to the posterior probabilities of rejecting the null hypothesis (see Methods 2.5).

The code to reproduce the presented analyses is available on GitHub (see Code availability).

2.7. Human TMS and SCS data

All procedures were reviewed and approved by the Institutional Review Board (IRB) of James J. Peters Veterans Affairs Medical Center (JJP VAMC); Weill Cornell Medicine (WCM-IRB, 1806019336); and Columbia University Irving Medical Center (IRB 2, protocol AAAT6563). The study was pre-registered at clinicaltrials.gov (NCT05163639). Written informed consent was obtained prior to study enrollment, and all experimental procedures were conducted in compliance with institutional and governmental regulations guiding ethical principles for participation of human volunteers. The goal of the study consisted of assessing the synergistic [77] and plasticity inducing effects of combining brain (TMS or transcranial electrical stimulation) and spinal cord (transcutaneous or epidural) stimulation in uninjured and SCI participants. TMS and epidural SCS recruitment curve data was extracted from one session per participant in the presented analysis.

Human TMS data

Individuals between the ages of 18 and 80 without neurological injury (uninjured volunteers) and individuals with chronic (> 1 year) cervical SCI, were eligible for recruitment. SCI participants required partially retained motor hand function, scoring 1–4 (out of 5) on manual muscle testing, with detectable TMS-evoked MEPs (greater than 50 μV) of the left or right target muscle. Recruitment curve data from 13 individuals living with SCI and 14 individuals with no neurological deficits available at the time of this study were used for analysis. SCI motor level and impairment severity were determined by clinical examination according to the International Standards for the Neurological Classification of SCI (ISNCSCI). Surface electromyography (EMG) preamplifiers were placed bilaterally over the APB, FDI, ADM, FCR, ECR, biceps brachii short head (which we refer to as biceps), triceps brachii long head (which we refer to as triceps), and tibialis anterior (TA) muscle in a belly-tendon montage, as described previously [78]. EMG signals were bandpass filtered between 15 and 2000 Hz, and sampled at 5000 Hz via an MA400 EMG system (Motion Lab Systems Inc., Louisiana, USA).

TMS was delivered with a MagPro X100 system (MagVenture Inc., Georgia, USA) with 80 mm winged coil (D-B80; MagVenture Inc.) placed over the hand motor cortex (M1) hotspot for optimal response in the target muscle. Electromagnetic stimulation was delivered as a single biphasic sinusoidal (anodic-first; 1.0 ms) pulse. The TMS coil, and an adhesive headpiece donned on the participants were fitted with passive markers detected by a stereotactic neuronavigation system (BrainSight; Rogue Research, Montreal, Canada). The coil was oriented at a 45-degree angle from the medial-sagittal plane so that a posterior–anterior directed electric field perpendicular to the central sulcus was induced in the underlying cortical tissue.

Recruitment curves were assembled via delivery of TMS pulses of varying intensities in pseudorandom order ranging from subthreshold to 200% or more of threshold. Analog-to-digital data acquisition and output systems (National Instruments (NI) USB-6363 and NI USB-6229; Emerson Electric Co., Missouri, USA) were controlled with customized LabVIEW software (Emerson Electric Co., Missouri, USA) in order to integrate electromyographic recordings and synchronize stimulator triggers. In each participant, a total of 61.4 ± 14.0 stimulation pulses between 26.0 ± 13.3% to 81.1 ± 16.1% MSO. In 16 participants (SCI $n = 7$) the total number of pulses were divided into 7–8 repetitions per stimulation intensity. In the remaining 11 participants (SCI $n = 6$), the stimulation protocol was changed so that each stimulation trial had a unique intensity based on the preliminary development of our hbMEP approach (see Fig. 6). MEPs in triceps, biceps, ECR, FCR, APB and ADM contralateral to the site of stimulation were quantified as peak-to-peak in an 83.5 ms window starting at 6.5 ms after the start of the first stimulation pulse. Due to temporal jitter in the recording system for the triceps muscle, the starting point of the window was increased to 10.6 ± 2.1 ms in order to avoid stimulation artifacts.

Human epidural SCS data

Detailed protocols can be found in McIntosh et al. 2023 [13], relevant sections are reproduced here. Participants were adult patients with cervical spondylotic myelopathy and/or multilevel foraminal stenosis requiring surgical intervention. Patients were enrolled from the clinical practices of the spine surgeons participating in the study.

Epidural electrodes were used for stimulation during clinically indicated surgeries, with EMG recordings taken from muscles selected as per standard of care. Recordings were made at a sampling rate between 6 kHz and 10.4 kHz, bandpass filtered between 10 Hz and 2000 Hz. A three-pulse train was used for epidural spinal cord stimulation to reduce the necessary intensity to evoke an MEP.

In 13 participants, stimulation was applied at the most caudal exposed segment at midline and lateral locations to compare recruitment curves. A handheld double-ball tip epidural electrode was positioned at midline, in line with the dorsal root entry zone. Stimulation intensity was incremented from 0 up to 8 mA to assess the activation threshold and estimate the subsequent recruitment curve (minimum 5 MEPs per stimulation intensity). The experiments proceeded by repeating the stimulation intensity ramp and fixed-intensity stimulation at the equivalent lateral site. MEPs were quantified in biceps, triceps, APB and ADM ipsilateral to the side of stimulation with the rectified AUC calculated in a window between 6.5 ms and 75 ms after the start of the first stimulation pulse.

2.8. Rat epidural SCS data

Eight Sprague Dawley rats were used in this study for a terminal physiology experiment. All procedures were conducted in compliance with the guidelines of the Institutional Animal Care and Use Committee at Columbia University in New York, NY, and followed aseptic techniques. Detailed methodology of the protocol used can be found in Mishra et al. 2017 [79] and Pal et al. 2022 [20].

EMG activity was recorded from 8 different muscles: left ECR, FCR, biceps, triceps, ADM, deltoid, biceps femoris, and right biceps. Flexible, braided stainless steel wires were employed for EMG recording. A hole was drilled into the skull between the eyes, and the ground screw electrode was inserted into the hole. EMG and ground electrodes were soldered to the connector and covered with epoxy to ensure insulation. Subsequently, the connector was attached to the recording system.

After placing the EMG electrodes, the animal's head was fixed and the T1 spinous process was clamped to stabilize the spine. The C4 spinal cord was exposed by laminectomy. Custom designed electrode arrays [80] were placed in the dorsal epidural space in the midline of the spinal cord over the cervical enlargement (C5–C8). The arrays consisted of 12 electrodes arranged in a 4 by 3 configuration, with the outer columns aligned to each of the C5–C8 dorsal root entry zones and the central columns aligned to the spinal cord midline. The muscles over the spinal cord were brought back together to prevent temperature loss and reduce dryness around the spinal cord opening. Omnetics connectors (Omnetics Connector Corp.; Minneapolis, USA) for the spinal array and EMG wires were mounted on the skull for stimulation and recording.

Connectors for the spinal cord array and EMG wires were attached to a headstage ZIF-clip (Tucker–Davis Technologies; Florida, USA) via Omnetics connectors. Raw signals were sampled at 10 kHz. ZIF-connectors were used to interface the implanted electrodes with a PZ5 amplifier (Tucker–Davis Technologies) in turn connected to a real-time RZ2 signal processing system (Tucker–Davis Technologies).

A 16-channel IZ2H constant current stimulator (Tucker–Davis Technologies), controlled via custom Matlab (R2022a) scripts, delivered biphasic single-pulse stimulation of 200 μs every 2 s, with intensities linearly increased from 0 to an average of 325±88.6 μA across 51 steps. Stimulation patterns were randomly applied across 21 spatial combinations on the left side and midline of the array, excluding high impedance electrodes from testing, resulting in an average of 18.8 ± 3.2 combinations tested per rat. EMG signals were high-pass filtered using a 20 Hz cutoff 10th order IIR filter. MEPs were quantified in biceps, triceps, ECR, FCR, APB and ADM ipsilateral to the side of stimulation by calculating the AUC within a 1.5 to 10 ms window post-stimulation onset.

3. Results

3.1. Accurate estimation of threshold on sparse data

We introduced a rectified-logistic function (Fig. 1c) for modeling MEP size recruitment curves and integrated it into a standard hierarchical Bayesian model. To enable a comparative analysis of different estimation methods based on their accuracy of recovering curve parameters, the resultant model was used to simulate synthetic data that closely matched real TMS data (Fig. 2a–d). We used the model to estimate both the participant- and population-level parameters from TMS data, which consisted of 27 participants (Fig. 2a). Using the estimated participant-level parameters, the model successfully replicated observations from existing TMS participants (Fig. 2b). Additionally, the model was conditioned on the estimated population-level parameters to simulate new participants (Fig. 2c). A principal component analysis (Fig. 2d) showed a large overlap between the parameters of new simulated participants and those estimated from existing TMS participants, validating the quality of the synthetic data.

We evaluated the standard hierarchical Bayesian (HB) model for its accuracy in recovering the simulated thresholds (Fig. 2c, green line) and compared it against three non-hierarchical models: the conventionally used maximum likelihood (ML) and least squares (LSM) methods, and an equivalent non-hierarchical Bayesian (nHB) model. The HB model demonstrated improved accuracy compared to the non-hierarchical models (Fig. 2e,f). For a single participant, the mean absolute error (e) of the HB model was not different from the nHB model (Fig. 2e, $e_{\text{nHB}} - e_{\text{HB}}$ mean ± sem : -0.02 ± 0.02), but lower than the non-Bayesian models ($e_{\text{ML}} - e_{\text{HB}}$: 2.25 ± 0.13, $e_{\text{LSM}} - e_{\text{HB}}$: 4.74 ± 0.21). As the number of participants increased, the HB model further reduced the error over the non-hierarchical models (for $N = 16$ participants, $e_{\text{nHB}} - e_{\text{HB}} : 2.2 \pm 0.04$, $e_{\text{ML}} - e_{\text{HB}} : 4.36 \pm 0.06$, $e_{\text{LSM}} - e_{\text{HB}} : 6.74 \pm 0.04$). In contrast, the errors of the non-hierarchical models did not change irrespective of the number of participants (N) used for analysis ($e_{N=1} - e_{N=16}$ for nHB : 0.13 ± 0.2, ML : 0.23 ± 0.25, LSM : 0.35 ± 0.18).

The HB model also accounted for small sample sizes, with its advantage most apparent when the total number of stimuli (samples) were low (Fig. 2f, example shown with eight participants of the synthetic TMS data). For instance, with only 16 samples, the error difference between the nHB and HB models ($e_{\text{nHB}} - e_{\text{HB}}$) was 3.29 ± 0.07, with a 46% reduction in the error over the nHB model. In contrast, for a relatively large number of 64 samples, the difference was less pronounced at 1.77 ± 0.05, albeit with a similar reduction at 47%. Notably, the ML model performed worse than the LSM model when the data was very sparse (for 16 samples, $e_{\text{LSM}} - e_{\text{ML}} : -3.13 \pm 0.11$), with its performance approaching that of the nHB model as the number of samples increased (for 64 samples, $e_{\text{nHB}} - e_{\text{ML}} : -1.63 \pm 0.04$).

3.2. Bayesian estimation for detecting a shift in threshold

The flexibility of hierarchical Bayesian estimation allows us to directly model the differences in participant-level thresholds, for example, between pre- and post-intervention phases. These differences can be summarized across multiple participants using a population-level parameter, and its 95% highest density interval (HDI) is used for hypothesis testing [81–85]. To evaluate the statistical power in the context of assessing effectiveness of an intervention, we simulated two hypothetical scenarios: one with a negative shift in the threshold from pre- to post-intervention, and another with a zero shift. A negative shift indicates that the intervention results in a lower threshold, thereby

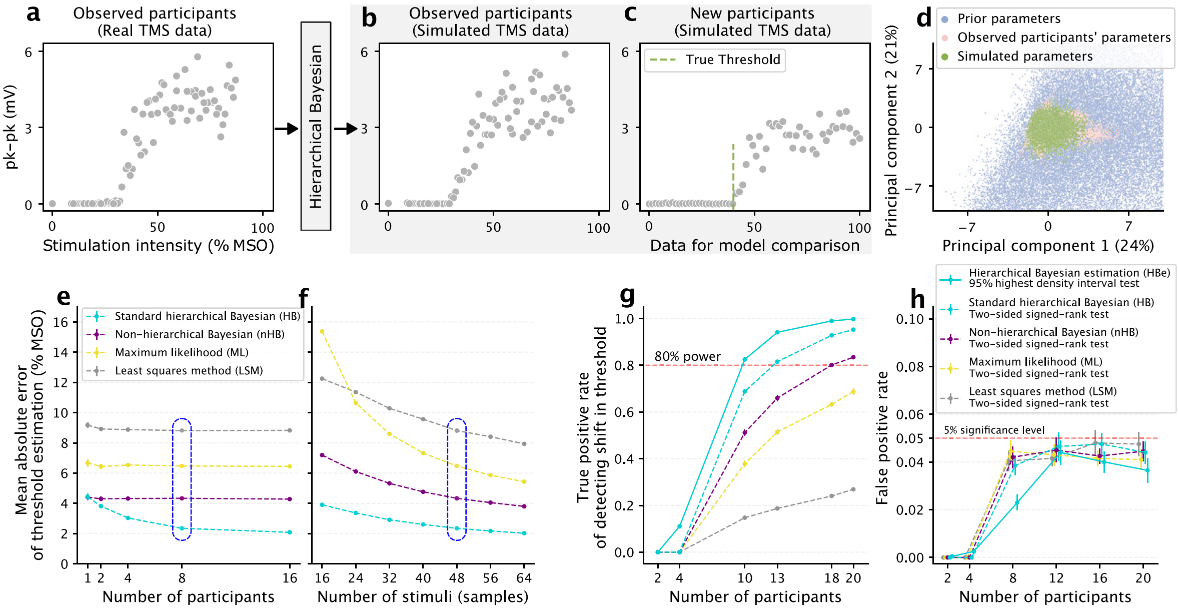


Fig. 2. (a–d) Generative hierarchical Bayesian model simulates high-fidelity synthetic TMS data. (a) Example participant from human TMS data used by the standard hierarchical Bayesian model to estimate participant- and population-level parameters. **(b)** Data simulated from the model conditioned on estimated participant-level parameters. The model can replicate observed participants. **(c)** Data simulated from the model conditioned on estimated population-level parameters for subsequent model comparison. **(d)** Principal component analysis shows a large overlap between the new simulated parameters (green dots) and those estimated from observed TMS data (pink dots). Blue dots represent parameters simulated from the weakly informative prior predictive distribution. **(e–f) Standard hierarchical Bayesian model improves threshold estimation accuracy over non-Bayesian and non-hierarchical models on simulated data. (e)** The standard hierarchical Bayesian (HB) model benefits from partial pooling across participants and uniquely reduces mean absolute error of threshold estimation as the number of participants increase. Error bars represent standard error of the mean. **(f)** For the first eight of the sixteen participants in (e), the error of the HB model remains below the non-hierarchical models at all tested number of stimuli, and its advantage is most pronounced for a low number of stimuli, i.e., on sparse data. **(g–h) Bayesian estimation is more powerful when detecting a shift in the threshold compared to frequentist testing. (g)** Hierarchical Bayesian estimation (HBe) requires fewer participants to achieve 80% power when detecting a shift in the threshold from pre- to post-intervention phase. Here the threshold differences are simulated from a Normal $(\mu = -5, \sigma = 2.5)$ distribution, where the alternative hypothesis (non-zero shift) is true. **(h)** Comparison of false positive rates of Bayesian estimation and frequentist tests against the set significance level of 5% of signed-rank test. The differences are simulated from Normal $(0, 2.5)$, where the null hypothesis (zero shift) is true. Except for (f), all simulations consisted of 48 equispaced stimulation intensities between 0–100% maximum stimulator output (% MSO). Blue rounded rectangles represent the same simulation configurations that are directly comparable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

facilitating muscle activation. Even though our illustrative example is designed to simulate a detection of a negative shift, the framework inherently considers shifts in either direction, and the physiological significance of such a shift would depend on the experimental context. For hierarchical Bayesian estimation (HBe), the decision rule rejects the null hypothesis (zero shift) in favor of the alternative (non-zero shift) if the 95% HDI excludes zero. This was compared against a two-sided Wilcoxon signed-rank test [70] conducted on the point threshold estimates of the standard hierarchical Bayesian (HB) and non-hierarchical models (a t-test was not applicable due to non-normality of estimated pairwise threshold differences as indicated by Shapiro–Wilk test). The significance level was set at 5% and the alternative hypothesis was accepted if the p-value was less than 0.05.

The hierarchical Bayesian methods required fewer participants to achieve 80% power compared to non-hierarchical models, with the HBe model requiring the fewest participants (Fig. 2g). Of the three non-hierarchical models, LSM and ML methods did not reach 80% power in the tested range of participants, while the nHB model required 18 participants to achieve 80% power (at $N = 18$ participants, true positive rate mean ± sem : $80.1 \pm 0.91\%$). In contrast, the standard HB model required only 13 participants ($81.55 \pm 0.82\%$), with a reduction of 28% in the number of participants. The HBe model further reduced the number of participants to 10 ($82.5 \pm 1.09\%$), a reduction of 23% compared to the standard HB model and 44% compared to the nHB model. Additionally, the false positive rates for all methods did not exceed the significance level of 5% (Fig. 2h), validating the observed differences in statistical power.

3.3. Choice of recruitment curve function

The different choices for modeling recruitment curves include a three-parameter rectified-linear function [13,22,49–51] (Fig. 3a) and the most commonly used four-parameter logistic-4 function (Boltzmann sigmoid) [22,38,45–48] (Fig. 1b, 3b). Additionally, a five-parameter logistic-5 function [44] (Fig. 3c) is a more generalized version of logistic-4 that is not necessarily symmetrical about its inflection point.

We evaluated these functions, including the five-parameter rectified-logistic function (Fig. 1c, 3d), for their out-of-sample predictive performance using approximate leave-one-out cross-validation (PSIS-LOO-CV) [71] on empirically obtained human and rat datasets for spinal cord stimulation (SCS) and TMS. Since PSIS-LOO-CV estimates predictive accuracy on held-out data, it inherently accounts for model complexity and favors models that generalize well over those that overfit. The rectified-logistic function demonstrated superior predictive accuracy over the rectified-linear and logistic-4 functions on all datasets (Fig. 3e–g, with respect to logistic-4 on rat SCS $\Delta_\mu \pm \Delta_{\text{sem}}$: 3600.8 ± 118.1, human TMS : 156.3 ± 33.4, human SCS : 60.4 ± 19.5, $\Delta_\mu \geq 3\Delta_{\text{sem}}$ on all datasets). It also outperformed logistic-5 on the largest tested rat SCS dataset (Fig. 3e, 1891.5 ± 94.6, $\Delta_\mu \geq 3\Delta_{\text{sem}}$). For human TMS and SCS datasets, it maintained comparable performance to logistic-5 (Fig. 3f,g human TMS : 39.9 ± 21.5, human SCS : -6.8 ± 14.8).

While logistic functions are standard for estimating S_{50}, the threshold is neither an explicit parameter nor can it be derived from their equations since they are smooth functions. Conversely, the rectified-linear function includes a threshold parameter but exhibits suboptimal

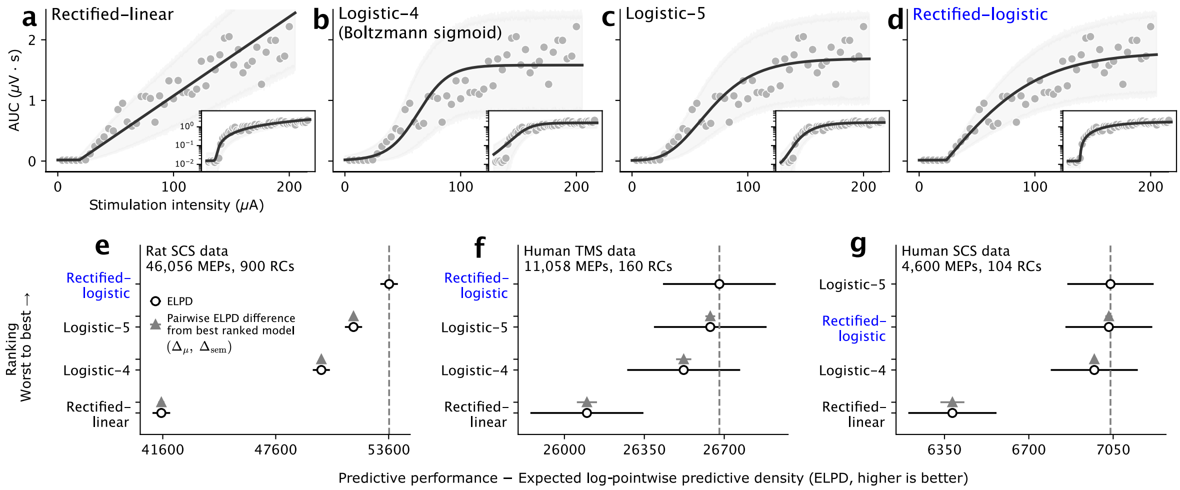


Fig. 3. Rectified-logistic function has at par or superior predictive performance compared to traditional alternatives while having the unique advantage of estimating the threshold along with curvature and saturation. (a) Example recruitment curve fitted to rat epidural SCS data using a three-parameter rectified-linear function. It underestimates the threshold at low intensities due to curvature in data, and subsequently overshoots at higher intensities while failing to capture saturation. Gray dots represent MEP size data, black curve shows the fitted curve, and gray shading represents the 95% HDI of the posterior predictive distribution. Inset: same curve shown on a log scale to highlight responses around threshold. **(b)** Four-parameter logistic-4 function (Boltzmann sigmoid) is symmetric about its inflection point, saturates early, and fails to capture the sharp deflection from offset. **(c)** Five-parameter logistic-5 function shows improved deflection and saturation. **(d)** Five-parameter rectified-logistic function is flexible enough to accurately capture the deflection, curvature, and saturation, resulting in narrower 95% HDI. **(e)** Predictive performance measured with expected log-pointwise predictive density (ELPD) using leave-one-out cross-validation on rat SCS dataset. Black circles represent mean ELPD score, black bars are standard error of mean ELPD, gray triangles are mean pairwise ELPD difference from the best-ranked rectified-logistic model (Δ_μ), and gray bars are standard error of the mean ELPD difference (Δ_{sem}). **(f)** Same as (e), but for human TMS dataset. **(g)** Same as (e), but for human epidural SCS dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

predictive performance (Fig. 3e–g). Furthermore, the estimation of S_{50} requires observing adequate amount of saturation in data, whereas the threshold can be accurately estimated independent of that (Supplementary Fig. S1). Given that observing adequate saturation is often infeasible, this makes the threshold a more reliable parameter for inferring changes in corticospinal excitability. Therefore, the rectified-logistic function addresses the limitations of traditional logistic functions by enabling estimation of threshold while either surpassing or matching their predictive performance on all datasets.

3.4. Robustness to outliers

Inaccuracies in model fitting often arise from sources of variability that occur independently of stimulation intensity, such as fasciculations, movement, or technical anomalies. We define outliers as these rare occurrences, as well as observations that do not conform to our model assumptions. To account for these outliers without manual exclusion, we introduced a mixture extension [60,86] of our gamma likelihood model which assigns a small, learnable probability that an observed sample comes from a broad distribution independent of stimulation intensity.

This adjustment yielded robust estimates that were otherwise biased by outliers (Fig. 4a, overestimated growth rate) and enabled automatic classification of outliers by returning an outlier probability for each observed sample (Fig. 4b, red dots). It further improved the predictive accuracy on all datasets (Fig. 4c, rat SCS : 2261.5 ± 141, human TMS : 1384.9 ± 104.1, human SCS : 721 ± 84, $\Delta_\mu \geq 3\Delta_{\text{sem}}$ on all datasets).

3.5. Paired comparison on human SCS data

To validate the applicability of our method to real data, we conducted a secondary analysis of epidural SCS data from 13 participants who underwent clinically indicated cervical spine surgery, which resulted in more effective muscle activation with lateral stimulation over the dorsal entry zone compared to midline stimulation [13]. Utilizing the mixture extension of our hierarchical Bayesian estimation (HBe) approach, we modeled the differences between the thresholds of midline and lateral stimulation for the arm and hand muscles (Fig. 5a).

Due to the nature of these experiments, multiple muscles are recorded simultaneously, and we would like to know for each individual muscle whether or not a statistically significant effect is present. This requires controlling for the family-wise error rate. Fig. 5b presents the model summary for all participants. The 98.75% HDIs, corrected for multiple comparisons at 5% significance level, show that the summarized threshold differences are entirely to the right of zero for three of the four tested muscles. This indicates strong evidence that lateral stimulation resulted in significantly lower thresholds and facilitated activation of the arm and hand muscles.

To evaluate the statistical power and family-wise error rate of the HBe model on empirically obtained human SCS data, we bootstrapped the participants and compared the results against a Bonferroni-Holm corrected [73] Wilcoxon signed-rank test [70] conducted on the point threshold estimates of the standard hierarchical Bayesian (HB) and non-hierarchical models. Consistent with our simulations (Fig. 2g), hierarchical Bayesian methods required fewer participants to achieve 80% power compared to non-hierarchical models (Fig. 5c). For the biceps muscle, the nHB model required 18 participants to achieve 80% power (at $N = 18$ participants, true positive rate mean ± sem : $83.65 \pm 3.82\%$, at $N = 17$: $79.25 \pm 3.52\%$, samples not shown), while the standard HB and HBe models required only 12 (at $N = 12$, HB : $80.8 \pm 3.22\%$ and HBe : $86.25 \pm 3.66\%$), with a reduction of 33% in the number of participants. Notably, the ML model performed worse than the LSM model for three of the four muscles, potentially due to the sparse nature of the data, which included only 16.73 ± 8.27 (mean ± sd) unique stimulation intensities per recruitment curve, consistent with our simulations (Fig. 2f). Additionally, the false positive (Supplementary Fig. S2a–d) and family-wise error rates (Supplementary Fig. S2e) for all methods did not exceed the significance level of 5%, validating the observed differences in statistical power.

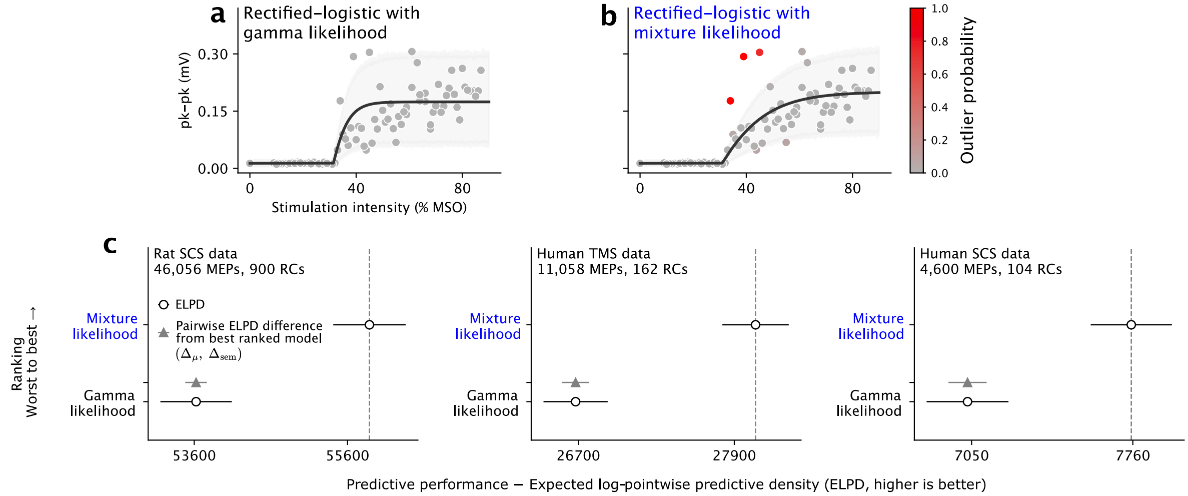


Fig. 4. Mixture model accounts for outliers and further improves predictive performance on all datasets. (a) Example recruitment curve fitted to human TMS data using rectified-logistic function within a gamma likelihood model. It overestimates the growth rate and saturates early due to presence of outliers. **(b)** Mixture extension of the gamma likelihood model is robust to outliers, resulting in narrower 95% HDI of the posterior predictive distribution. It returns an outlier probability for each observed sample which enables automatic outlier classification. Dots are colored by outlier probabilities. **(c)** Predictive performance measured with expected log-pointwise predictive density (ELPD) using leave-one-out cross-validation. Gray triangles represent the mean pairwise ELPD difference from the best-ranked mixture model (Δ_μ), and gray bars are standard error of the mean ELPD difference (Δ_{sem}). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.6. Optimizing experimental design

In this section, we illustrate how the generative capability of our model can be used to prototype and optimize experimental designs. It is common practice to test multiple repetitions or trials of the same stimulation intensity when constructing recruitment curves. The number of repetitions used is highly variable, typically ranging between 4 and 20 [22,28,44,47,50,51,87–92], although extending from as low as 1 [14] to 40 [43]. Since the number of stimuli in a session are constrained by the session duration, increasing the number of repetitions reduces the number of unique stimulation intensities that can be tested, which consequently reduces the sampling resolution. For instance, when administering a total of 64 stimuli in a session, one could test 64 unique intensities (single repetition each), 16 unique intensities (four repetitions each), or 8 unique intensities (eight repetitions each), and so on. We hypothesized that this reduction in resolution would adversely affect the accuracy of threshold estimates derived during post-hoc or offline analyses.

Our generative framework enables us to evaluate such strategies. To demonstrate this, we used data from the first eight participants of the synthetic TMS dataset (Results 3.1), and generated up to eight observations per stimulation intensity, for a total of 64 intensities equispaced between 0 to 100% MSO. Using the standard hierarchical Bayesian model (HB), we estimated the thresholds with one, four, and eight observations per intensity (Fig. 6a–c). The results indicated that conducting experiments without repetition — testing each intensity only once — was the most efficient approach for accurate threshold estimation (Fig. 6d). Notably, the largest improvement in accuracy was observed when the total number of stimuli were low. For instance, distributing 32 stimuli evenly across the full intensity range reduced the mean absolute error from 5.36 ± 0.03% (mean ± sem) with eight repetitions of 4 equispaced intensities to 2.91 ± 0.03% with a single repetition of 32 equispaced intensities.

We also examined how the statistical power of the hierarchical Bayesian estimation model (HBe) to detect a threshold shift varies with the different repetitions per intensity. For a total number of 64 stimuli, we simulated a negative threshold shift from pre- to post-intervention, similar to Results 3.2 (Fig. 2g), and used the model to detect the shift with one, four, and eight repetitions per intensity. The model achieved 80% power with fewer participants when fewer repetitions were used (Fig. 6e). With eight repetitions of 8 equispaced intensities, the model required 12 participants to achieve 80% power (at $N = 12$ and eight repetitions, true positive rate mean ± sem : 82.25 ± 0.84%), while with four repetitions of 16 equispaced intensities, it required only 10 participants ($85 \pm 0.64\%$), reducing the number of participants by 17%. Further reducing to a single repetition of 64 equispaced intensities decreased the required sample size to nine participants ($82.7 \pm 0.77\%$), a reduction of 25% compared to eight repetitions.

4. Discussion

Our method introduces a rectified-logistic function that provides a consistent parametric definition of motor threshold, and integrates it within a hierarchical Bayesian framework for estimating MEP size recruitment curves. This framework improved threshold estimation accuracy on sparse data and required fewer participants to achieve comparable statistical power than non-hierarchical methods. We validated these advantages on empirical TMS and SCS datasets, and synthetic TMS data. Our open-source library for Python, hbMEP, makes our approach broadly accessible to researchers seeking accurate and robust estimation of recruitment curves across multiple stimulation modalities.

We demonstrated that hierarchical Bayesian methods, including Bayesian estimation (HBe) and the standard hierarchical Bayesian model (HB), require substantially fewer participants to detect threshold shifts compared to non-hierarchical models. The HBe approach views the mean threshold shift across participants as a model parameter and yields a complete distribution of its credible values. This approach offers a more satisfactory and informative conclusion compared to frequentist tests, which only provide a dichotomous outcome of significant or not significant based on a *p*-value. However, it requires a more nuanced understanding of Bayesian inference, including the incorporation of hypothesis-specific modifications to the model structure. Conversely, the standard HB model, while slightly less powerful, is easier to use and generalizable across datasets and hypotheses because it decouples the estimation of recruitment curves from statistical inference. This approach strikes a balance between hierarchical Bayesian methods for accurate estimation of curve parameters and frequentist methods for significance testing. For these reasons, the hbMEP library implements

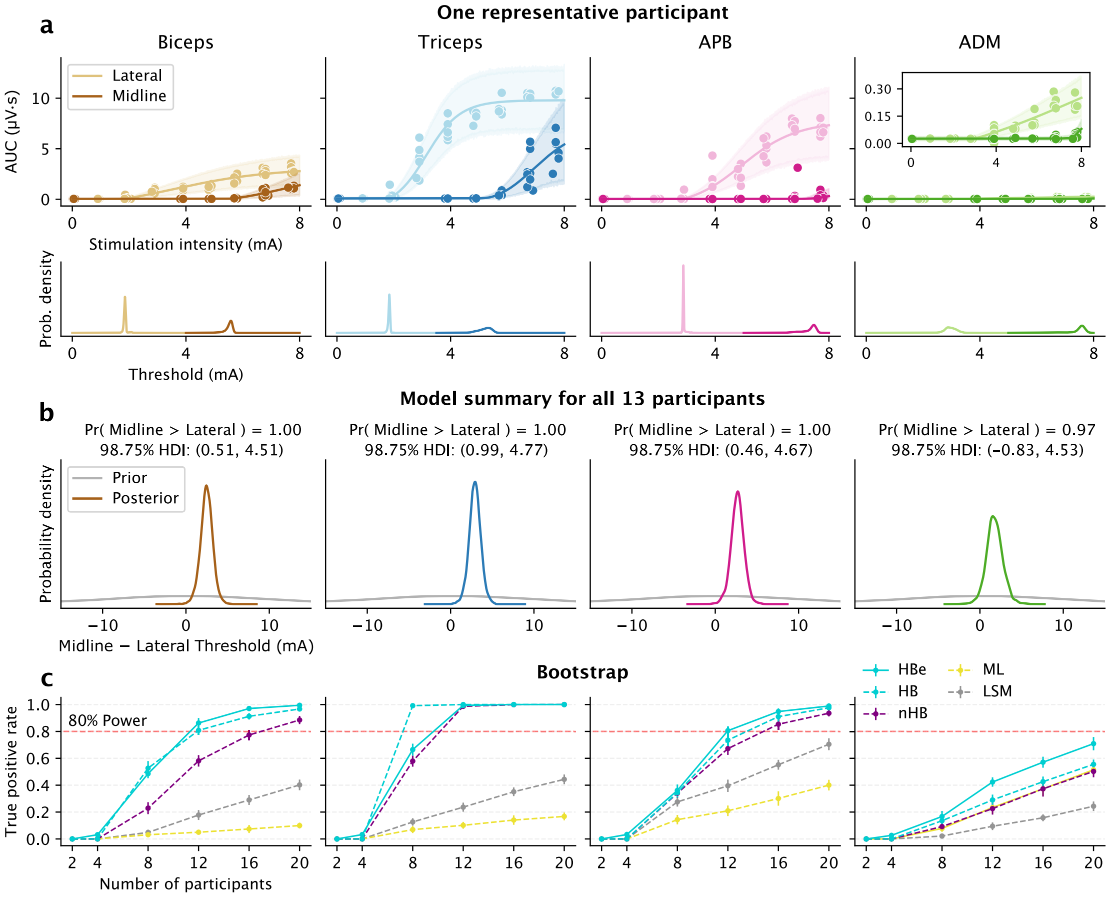


Fig. 5. Comparison of midline versus lateral stimulation thresholds on human epidural SCS data. (a) Example participant showing lateral (light color) and midline (dark color) stimulation. Inset: zoom to show presence of threshold, despite small MEP size. Bottom panels: posterior distribution of the threshold. **(b)** Posterior distribution of the shift between midline and lateral thresholds summarized across all participants ($N = 13$). A priori, the model assumes no shift, indicated by a flat prior (gray distribution) centered at zero. The 98.75% HDIs for the biceps, triceps and APB muscles are entirely to the right of zero. This, together with the Bayesian probabilities, indicates strong evidence that lateral stimulation resulted in significantly lower thresholds for the arm and hand muscles. **(c)** Validation of hierarchical Bayesian methods on bootstrapped human SCS data. The standard hierarchical Bayesian (HB) and hierarchical Bayesian estimation (HBe) methods require fewer participants to achieve 80% power compared to non-hierarchical models when detecting a shift between midline and lateral stimulation thresholds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and adopts the standard HB model as its default, making estimation of recruitment curves and inference accessible to a broad research community.

We demonstrated how MEP size recruitment curves, whether derived from TMS or epidural SCS, can be accurately modeled using the same rectified-logistic function. Using a recruitment curve function that does not appropriately represent the data can produce systematic errors in parameter estimates, thereby compromising subsequent analyses. Through cross-validation on TMS and SCS data, we determined that the rectified-linear function and the most frequently used logistic-4 function are suboptimal for modeling recruitment curves. This is due to the strict assumptions these functions make—rectified-linear assumes linear growth post-threshold, and logistic-4 assumes symmetry about its inflection point. We found that the more flexible logistic-5 function performs significantly better, but is limited to estimating the S_{50} and not the threshold. However, we illustrated how the estimation of S_{50} depends on observing adequate saturation in data, a condition often unmet, whereas the threshold can be accurately estimated independent of it. To shift the focus on analyzing a more reliable threshold parameter, we introduced a rectified-logistic function that matches or exceeds the predictive performance of the logistic-5 function and includes an explicit parameter for the threshold.

The rectified-logistic function provides a consistent parametric representation of threshold as a deflection of MEP size from the estimated offset. However, this differs in approach from the traditional resting motor threshold (RMT) used in TMS studies, defined as the minimum stimulation intensity to produce a predefined MEP size of 50 µV in at least 50% of repetitions [63]. RMT is typically estimated in real time and independently of the recruitment curve, with measurements repeated iteratively across muscles. In general, this method detects the threshold after the MEP size is already above the offset noise floor. For instance, Appendix B2 17.1, 17.2, 17.6 show a TMS participant with SCI with consistent and reliable MEPs above offset, yet below 0.05 mV (50 µV), for three of six recorded muscles, including the target APB muscle. In such cases, the RMT would fail to detect a threshold despite the presence of consistent MEPs. This limitation becomes especially relevant in clinical contexts, where some weakened muscles may exhibit reliable but subthreshold MEPs by the 50 µV criterion.

In contrast, recruitment curves are recorded across multiple muscles simultaneously and provide a more robust characterization of corticospinal output than MEPs recorded at a single intensity [22,31,38]. They utilize data from multiple trials across various stimulation intensities and, in addition to threshold, yield other curve parameters such as slope and saturation, which can subsequently be estimated offline. Furthermore, the 50 µV RMT convention, while suitable for resting-state human TMS, may not generalize across species, clinical conditions, stimulation modalities, or MEP metrics such as AUC. Nevertheless, given its widespread use, we deemed it important to maintain backward

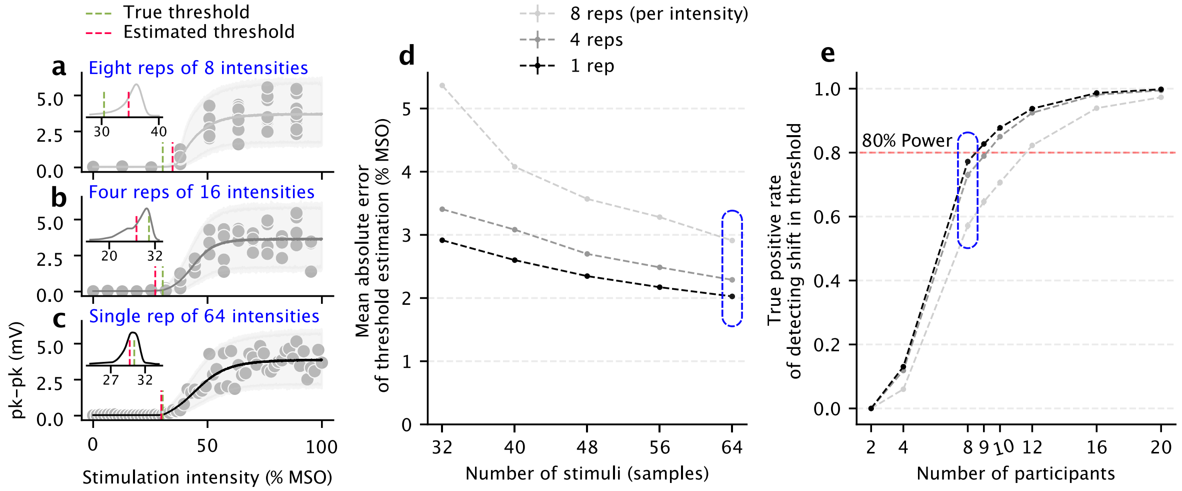


Fig. 6. Efficient data sampling is performed with fewer repetitions per intensity. Example fits on a simulated participant with a total of 64 stimuli and **(a)** eight repetitions, **(b)** four repetitions, and **(c)** a single repetition per intensity. Inset: zoom over threshold posterior with the ground truth (green line) and estimated threshold (red line). **(d)** For eight participants of synthetic TMS data, sampling with a single repetition produces the lowest mean absolute error for threshold estimation, regardless of the total number of stimuli. **(e)** For a total of 64 stimuli, sampling with fewer repetitions requires fewer participants to achieve 80% power when detecting a shift in the threshold from pre- to post-intervention phase. Here the threshold differences are simulated from Normal $(\mu = -5, \sigma = 2.5)$ distribution, where the alternative hypothesis (non-zero shift) is true. Blue rounded rectangles and blue highlighted text represent similar simulation configurations that are directly comparable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

compatibility. Our library enables recovery of the RMT as a post-processing step by intersecting the estimated recruitment curve with a predefined MEP size, provided that this predefined size lies between the estimated offset and saturation. For compatibility, our library also enables the estimation of recruitment curves using logistic-5, logistic-4, and rectified-linear functions within the standard HB model.

By integrating a mixture model, our approach circumvents the need for manual artifact rejection prior to estimating recruitment curves. It isolates observations unlikely to result from stimulation, such as electrical artifacts or unrelated muscle activity, preventing them from biasing parameter estimates. While the primary motivation is to account for such artifacts, the mixture component may also absorb legitimate physiological responses that deviate from the model assumptions. Importantly, the mixture component is constrained to account for only a small fraction of the data, ensuring the primary model remains grounded in stimulation-evoked responses. However, this approach does not account for all sources of variability in MEP measurements and may require extension in specific contexts. For instance, when applying epidural [93] or transcutaneous [78] SCS, posterior roots are thought to be preferentially activated. However, if sufficient stimulation intensity is applied, efferent fibers may also be activated, contributing to the early portion of the MEP. Without a careful choice of the time window for calculating the MEP size, the measured recruitment curve may inadvertently represent a mixture of efferent and afferent recruitment curves, which would not be accounted for by any S-shaped curve.

The broad applicability and scalability of hbMEP comes at the expense of biophysical fidelity. The model does not attempt to explain the biological origins of motor unit activation [12,94–96]. It also does not attempt to decompose MEP variability into distinct mechanistic sources [39,65–67]. Such models often involve convolutions of distributions without closed-form solution, making them difficult to implement in standard probabilistic programming libraries. While our focus here is on robustness and general usability, we view mechanistic decomposition as complementary and future work will explore whether better variance structures can be integrated within the hierarchical Bayesian framework.

In conclusion, the proposed hierarchical Bayesian method substantially reduces the experimental burden associated with quantifying S-shaped relationships. By improving parameter estimation accuracy offline on sparse data, it minimizes the number of stimuli needed to probe each individual's neuromuscular parameters, thereby shortening session duration and reducing the risk of inadvertent neuromodulation, while simultaneously increasing the number of muscles across which these insights are obtained. As a consequence, researchers will gain from increased experimental throughput and enhanced statistical reliability. Future work will explore adaptations for clinical use, as well as integrating closed-loop designs [39,97] to optimize sampling in real time by extending our Bayesian approach, further expanding its scope and utility.

CRediT authorship contribution statement

Vishweshwar Tyagi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Lynda M. Murray:** Writing – review & editing, Data curation. **Ahmet S. Asan:** Writing – review & editing, Data curation. **Christopher Mandigo:** Writing – review & editing, Data curation. **Michael S. Virk:** Writing – review & editing, Funding acquisition, Data curation. **Noam Y. Harel:** Writing – review & editing, Resources, Investigation, Funding acquisition, Data curation. **Jason B. Carmel:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation. **James R. McIntosh:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Code availability

The hbMEP Python package was used to perform all analyses in the paper. The code, together with tutorials, is available at https://github.com/hbmep/hbmep Code to reproduce the presented analyses is available at https://github.com/hbmep/hbmep-paper.

Funding

This research was supported by the National Institutes of Health (NIH: R01NS124224, R01NS115470, 1R03NS141040); the Congressionally Directed Medical Research Programs (HT9425-24-1-0218); and the Travis Roy Foundation Boston, MA (Investigator Initiated).

Declaration of competing interest

Jason B. Carmel is a Founder and stock holder in BackStop Neural and a scientific advisor and stockholder in SharperSense. He has received honoraria from Pacira, Motric Bio, and Restorative Therapeutics. Noam Y. Harel is a consultant for RubiconMD. Michael S. Virk has been a consultant and has received honorarium from Depuy Synthes and BrainLab Inc; he is on the Medical Advisory Board and owns stock with OnPoint Surgical. The other authors have nothing to disclose.

Acknowledgments

We thank A. Gelman and J. M. Mikhaeil at Columbia University and S. Fara at Medtronic for their helpful feedback and comments on the manuscript.

Human TMS data: We thank J. A. Goldsmith, Y. Wu, K. S. Holbrook, M. Liu, T. Magno, G. Famodimu, G. A. Mendez, J. M. Robbins, L. E. Kinne, and A. Villaroel-Sanchez at the James J. Peters VAMC for helping obtain human TMS data.

Human epidural SCS data: We thank neurologists A. Mendiratta (Weill Cornell Medicine), S. C. Karceski, and M. Bell (The Och Spine Hospital At New York Presbyterian Hospital) and intraoperative monitoring technologists E. Thuet (The Och Spine Hospital At New York Presbyterian Hospital), E. Shelkov (Weill Cornell Medicine) and their teams including O. Modik and N. Patel (Weill Cornell Medicine). We also thank M. Vulapalli, C. Mykolajtchuk, and M. Michael (Weill Cornell Medicine) for help in administrative matters.

Computing resources: We thank Columbia University's Neurology High Performance Computing (HPC) Infrastructure and the Department of Systems Biology IT (NIH: S10OD032433, S10OD021764) for providing the computational resources used for simulations.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.brs.2025.09.008.

Data availability

The datasets used in this study are publicly available [98].

References

[1] Tallarida RJ, Jacob LS. The Dose—Response Relation in Pharmacology. New York, NY: Springer US; 1979, http://dx.doi.org/10.1007/978-1-4684-6265-4.

[2] Lin D, Shkedy Z, Yekutieli D, Amaratunga D, Bijnens L, editors. Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R. Berlin, Heidelberg: Springer; 2012, http://dx.doi.org/10.1007/978-3-642-24007-2.

[3] Cox C. Threshold Dose-Response Models in Toxicology. Biometrics 1987;43(3):511–23. http://dx.doi.org/10.2307/2531991, Publisher: [Wiley, International Biometric Society].

[4] Gelman A, Chew GL, Shnaidman M. Bayesian Analysis of Serial Dilution Assays. Biometrics 2004;60(2):407–17. http://dx.doi.org/10.1111/j.0006-341X.2004.00185.x.

[5] Treutwein B. Adaptive psychophysical procedures. Vis Res 1995;35(17):2503–22. http://dx.doi.org/10.1016/0042-6989(95)00016-X.

[6] Serrano-Pedraza I, Vancleef K, Herbert W, Goodship N, Woodhouse M, Read JCA. Efficient estimation of stereo thresholds: What slope should be assumed for the psychometric function? PLoS One 2020;15(1):e0226822. http://dx.doi.org/10.1371/journal.pone.0226822, Publisher: Public Library of Science.

[7] Capuano AW, Wilson RS, Leurgans SE, Dawson JD, Bennett DA, Hedeker D. Sigmoidal mixed models for longitudinal data. Stat Methods Med Res 2018;27(3):863–75. http://dx.doi.org/10.1177/0962280216645632.

[8] MacDonald DB, Skinner S, Shils J, Yingling C. Intraoperative motor evoked potential monitoring – A position statement by the American Society of Neurophysiological Monitoring. Clin Neurophysiol 2013;124(12):2291–316. http://dx.doi.org/10.1016/j.clinph.2013.07.025.

[9] Picht T, Schmidt S, Brandt S, Frey D, Hannula H, Neuvonen T, Karhu J, Vajkoczy P, Suess O. Preoperative functional mapping for rolandic brain tumor surgery: comparison of navigated transcranial magnetic stimulation to direct cortical stimulation. Neurosurgery 2011;69(3):581–8; discussion 588. http://dx.doi.org/10.1227/NEU.0b013e3182181b89.

[10] Sayenko DG, Atkinson DA, Dy CJ, Gurley KM, Smith VL, Angeli C, Harkema SJ, Edgerton VR, Gerasimenko YP. Spinal segment-specific transcutaneous stimulation differentially shapes activation pattern among motor pools in humans. J Appl Physiol 2015;118(11):1364–74. http://dx.doi.org/10.1152/japplphysiol.01128.2014.

[11] Hofstoetter US, Perret I, Bayart A, Lackner P, Binder H, Freundl B, Minassian K. Spinal motor mapping by epidural stimulation of lumbosacral posterior roots in humans. IScience 2021;24(1):101930. http://dx.doi.org/10.1016/j.isci.2020.101930.

[12] Greiner N, Barra B, Schiavone G, Lorach H, James N, Conti S, Kaeser M, Fallegger F, Borgognon S, Lacour S, Bloch J, Courtine G, Capogrosso M. Recruitment of upper-limb motoneurons with epidural electrical stimulation of the cervical spinal cord. Nat Commun 2021;12(1):435. http://dx.doi.org/10.1038/s41467-020-20703-1.

[13] McIntosh JR, Joiner EF, Goldberg JL, Murray LM, Yasin B, Mendiratta A, Karceski SC, Thuet E, Modik O, Shelkov E, Lombardi JM, Sardar ZM, Lehman RA, Mandigo C, Riew KD, Harel NY, Virk MS, Carmel JB. Intraoperative electrical stimulation of the human dorsal spinal cord reveals a map of arm and hand muscle responses. J Neurophysiol 2023;129(1):66–82. http://dx.doi.org/10.1152/jn.00235.2022.

[14] Koponen LM, Martinez M, Wood E, Murphy DLK, Goetz SM, Appelbaum LG, Peterchev AV. Transcranial magnetic stimulation input–output curve slope differences suggest variation in recruitment across muscle representations in primary motor cortex. Front Hum Neurosci 2024;18. http://dx.doi.org/10.3389/fnhum.2024.1310320.

[15] Chen R, Cros D, Curra A, Di Lazzaro V, Lefaucheur J-P, Magistris MR, Mills K, Rösler KM, Triggs WJ, Ugawa Y, Ziemann U. The clinical diagnostic utility of transcranial magnetic stimulation: Report of an IFCN committee. Clin Neurophysiol 2008;119(3):504–32. http://dx.doi.org/10.1016/j.clinph.2007.10.014.

[16] Millet GY, Martin V, Martin A, Vergès S. Electrical stimulation for testing neuromuscular function: from sport to pathology. Eur J Appl Physiol 2011;111(10):2489–500. http://dx.doi.org/10.1007/s00421-011-1996-y.

[17] Balbinot G, Li G, Kalsi-Ryan S, Abel R, Maier D, Kalke Y-B, Weidner N, Rupp R, Schubert M, Curt A, Zariffa J. Segmental motor recovery after cervical spinal cord injury relates to density and integrity of corticospinal tract projections. Nat Commun 2023;14(1):723. http://dx.doi.org/10.1038/s41467-023-36390-7.

[18] Stefan K, Kunesch E, Cohen LG, Benecke R, Classen J. Induction of plasticity in the human motor cortex by paired associative stimulation. Brain 2000;123(3):572–84. http://dx.doi.org/10.1093/brain/123.3.572.

[19] Bunday KL, Perez MA. Motor Recovery after Spinal Cord Injury Enhanced by Strengthening Corticospinal Synaptic Transmission. Curr Biology 2012;22(24):2355–61. http://dx.doi.org/10.1016/j.cub.2012.10.046, URL https://www.sciencedirect.com/science/article/pii/S0960982212012675.

[20] Pal A, Park H, Ramamurthy A, Asan AS, Bethea T, Johnkutty M, Carmel JB. Spinal cord associative plasticity improves forelimb sensorimotor function after cervical injury. Brain 2022;awac235. http://dx.doi.org/10.1093/brain/awac235.

[21] Arora T, Desai N, Kirshblum S, Chen R. Utility of transcranial magnetic stimulation in the assessment of spinal cord injury: Current status and future directions. Front Rehabil Sci 2022;3. http://dx.doi.org/10.3389/fresc.2022.1005111, Publisher: Frontiers.

[22] Devanne H, Lavoie BA, Capaday C. Input-output properties and gain changes in the human corticospinal pathway. Exp Brain Res 1997;114(2):329–38. http://dx.doi.org/10.1007/PL00005641.

[23] Ridding MC, Rothwell JC. Stimulus/response curves as a method of measuring motor cortical excitability in man. Electroencephalogr Clin Neurophysiol 1997;105(5):340–4. http://dx.doi.org/10.1016/s0924-980x(97)00041-6.

[24] Boroojerdi B, Battaglia F, Muellbacher W, Cohen LG. Mechanisms influencing stimulus-response properties of the human corticospinal system. Clin Neurophysiol 2001;112(5):931–7. http://dx.doi.org/10.1016/S1388-2457(01)00523-5, URL https://www.sciencedirect.com/science/article/pii/S1388245701005235.

[25] Rosenkranz K, Kacar A, Rothwell JC. Differential Modulation of Motor Cortical Plasticity and Excitability in Early and Late Phases of Human Motor Learning. J Neurosci 2007;27(44):12058–66. http://dx.doi.org/10.1523/JNEUROSCI.2663-07.2007.

[26] Houdayer E, Degardin A, Cassim F, Bocquillon P, Derambure P, Devanne H. The effects of low- and high-frequency repetitive TMS on the input/output properties of the human corticospinal pathway. Exp Brain Res 2008;187(2):207–17. http://dx.doi.org/10.1007/s00221-008-1294-z.

[27] Möller C, Arai N, Lücke J, Ziemann U. Hysteresis effects on the input–output curve of motor evoked potentials. Clin Neurophysiol 2009;120(5):1003–8. http://dx.doi.org/10.1016/j.clinph.2009.03.001, URL https://www.sciencedirect.com/science/article/pii/S1388245709002399.

[28] Nardone R, Höller Y, Thomschewski A, Bathke AC, Ellis AR, Golaszewski SM, Brigo F, Trinka E. Assessment of corticospinal excitability after traumatic spinal cord injury using MEP recruitment curves: a preliminary TMS study. Spinal Cord 2015;53(7):534–8. http://dx.doi.org/10.1038/sc.2015.12.

[29] Farzan F. Single-Pulse Transcranial Magnetic Stimulation (TMS) Protocols and Outcome Measures. In: Rotenberg A, Horvath JC, Pascual-Leone A, editors. Transcranial magnetic stimulation. New York, NY: Springer; 2014, p. 69–115. http://dx.doi.org/10.1007/978-1-4939-0879-0_5.

[30] Schicktanz N, Schwegler K, Fastenrath M, Spalek K, Milnik A, Papassotiropoulos A, Nyffeler T, de Quervain DJ-F. Motor threshold predicts working memory performance in healthy humans. Ann Clin Transl Neurol 2014;1(1):69–73. http://dx.doi.org/10.1002/acn3.22.

[31] Bestmann S. On the use of the motor threshold as a dependent variable in TMS research. Brain Stimul: Basic Transl Clin Res in Neuromodulation 2024;17(4):780–1. http://dx.doi.org/10.1016/j.brs.2024.06.009, URL https://www.brainstimjrnl.com/article/S1935-861X(24)00115-3/fulltext. Publisher: Elsevier.

[32] Cho HJ, Panyakaew P, Thirugnanasambandam N, Wu T, Hallett M. Dynamic Modulation of Corticospinal Excitability and Short-Latency Afferent Inhibition during Onset and Maintenance Phase of Selective Finger Movement. Clin Neurophysiol : Off J the International Fed Clin Neurophysiol 2016;127(6):2343–9. http://dx.doi.org/10.1016/j.clinph.2016.02.020.

[33] Kraus D, Naros G, Guggenberger R, Leão MT, Ziemann U, Gharabaghi A. Recruitment of Additional Corticospinal Pathways in the Human Brain with State-Dependent Paired Associative Stimulation. J Neurosci 2018;38(6):1396–407. http://dx.doi.org/10.1523/JNEUROSCI.2893-17.2017, Publisher: Society for Neuroscience Section: Research Articles.

[34] van de Ruit M, Pearson T, Grey MJ. Novel tools for rapid online data acquisition of the TMS stimulus-response curve. Brain Stimul 2019;12(1):192–4. http://dx.doi.org/10.1016/j.brs.2018.09.015.

[35] Ratnadurai Giridharan S, Gupta D, Pal A, Mishra AM, Hill NJ, Carmel JB. Motometrics: A Toolbox for Annotation and Efficient Analysis of Motor Evoked Potentials. Front Neuroinformat. 2019;13. http://dx.doi.org/10.3389/fninf.2019.00008.

[36] Skelly M, Salameh A, McCabe J, Pundik S. MEP-ART: A system for real-time feedback and analysis of transcranial magnetic stimulation motor evoked potentials. Brain Stimul: Basic Transl Clin Res in Neuromodulation 2020;13(6):1614–6. http://dx.doi.org/10.1016/j.brs.2020.09.012.

[37] Hassan U, Pillen S, Zrenner C, Bergmann TO. The Brain Electrophysiological recording & STimulation (BEST) toolbox. Brain Stimul 2022;15(1):109–15. http://dx.doi.org/10.1016/j.brs.2021.11.017.

[38] Kukke SN, Paine RW, Chao C, de Campos AC, Hallett M. Efficient and reliable characterization of the corticospinal system using transcranial magnetic stimulation. J Clin Neurophysiol 2014;31(3):246–52. http://dx.doi.org/10.1097/WNP.0000000000000057.

[39] Alavi SMM, Goetz SM, Peterchev AV. Optimal Estimation of Neural Recruitment Curves Using Fisher Information: Application to Transcranial Magnetic Stimulation. IEEE Trans Neural Syst Rehabil Eng 2019;27(6):1320–30. http://dx.doi.org/10.1109/TNSRE.2019.2914475, Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

[40] Manson GA, Calvert JS, Ling J, Tychhon B, Ali A, Sayenko DG. The relationship between maximum tolerance and motor activation during transcutaneous spinal stimulation is unaffected by the carrier frequency or vibration. Physiol Rep 2020;8(5):e14397. http://dx.doi.org/10.14814/phy2.14397.

[41] Antal A, Nitsche MA, Kincses TZ, Lampe C, Paulus W. No correlation between oving phosphene and motor thresholds: a transcranial magnetic stimulation study. NeuroReport 2004;15(2):297. http://dx.doi.org/10.1097/00001756-200402090-00017.

[42] Hassanzahraee M, Zoghi M, Jaberzadeh S. Longer Transcranial Magnetic Stimulation Intertrial Interval Increases Size, Reduces Variability, and Improves the Reliability of Motor Evoked Potentials. Brain Connect 2019;9(10):770–6. http://dx.doi.org/10.1089/brain.2019.0714.

[43] Sharma P, Rampersaud H, Shah PK. Repeated epidural stimulation modulates cervical spinal cord excitability in healthy adult rats. Exp Brain Res 2024;243(1):22. http://dx.doi.org/10.1007/s00221-024-06965-x.

[44] Pitcher JB, Ogston KM, Miles TS. Age and sex differences in human motor cortex input–output characteristics. J Physiol 2003;546(2):605–13. http://dx.doi.org/10.1113/jphysiol.2002.029454.

[45] Klimstra M, Zehr EP. A sigmoid function is the best fit for the ascending limb of the Hoffmann reflex recruitment curve. Exp Brain Res 2008;186(1):93–105. http://dx.doi.org/10.1007/s00221-007-1207-6.

[46] Smith AC, Rymer WZ, Knikou M. Locomotor training modifies soleus monosynaptic motoneuron responses in human spinal cord injury. Exp Brain Res 2015;233(1):89–103. http://dx.doi.org/10.1007/s00221-014-4094-7.

[47] Murray LM, Knikou M. Transspinal stimulation increases motoneuron output of multiple segments in human spinal cord injury. In: Nógrádi A, editor. PLoS One 2019;14(3):e0213696. http://dx.doi.org/10.1371/journal.pone.0213696.

[48] de Freitas RM, Sasaki A, Sayenko DG, Masugi Y, Nomura T, Nakazawa K, Milosevic M. Selectivity and excitability of upper-limb muscle activation during cervical transcutaneous spinal cord stimulation in humans. J Appl Physiol 2021;131(2):746–59. http://dx.doi.org/10.1152/japplphysiol.00132.2021.

[49] Willer JC, Miserocchi G, Gautier H. Hypoxia and monosynaptic reflexes in humans. J Appl Physiol 1987;63(2):639–45. http://dx.doi.org/10.1152/jappl.1987.63.2.639.

[50] Shkorbatova P, Lyakhovetskii V, Pavlova N, Popov A, Bazhenova E, Kalinina D, Gorskii O, Musienko P. Mapping of the Spinal Sensorimotor Network by Transvertebral and Transcutaneous Spinal Cord Stimulation. Front Syst Neurosci 2020;14:555593. http://dx.doi.org/10.3389/fnsys.2020.555593.

[51] Malone IG, Kelly MN, Nosacka RL, Nash MA, Yue S, Xue W, Otto KJ, Dale EA. Closed-Loop, Cervical, Epidural Stimulation Elicits Respiratory Neuroplasticity after Spinal Cord Injury in Freely Behaving Rats. ENeuro 2022;9(1). http://dx.doi.org/10.1523/ENEURO.0426-21.2021, ENEURO.0426–21.2021.

[52] Carson RG, Nelson BD, Buick AR, Carroll TJ, Kennedy NC, Cann RM. Characterizing Changes in the Excitability of Corticospinal Projections to Proximal Muscles of the Upper Limb. Brain Stimul 2013;6(5):760–8. http://dx.doi.org/10.1016/j.brs.2013.01.016.

[53] Potter-Baker KA, Varnerin NM, Cunningham DA, Roelle SM, Sankarasubramanian V, Bonnett CE, Machado AG, Conforto AB, Sakaie K, Plow EB. Influence of Corticospinal Tracts from Higher Order Motor Cortices on Recruitment Curve Properties in Stroke. Front Neurosci 2016;10. http://dx.doi.org/10.3389/fnins.2016.00079, Publisher: Frontiers.

[54] Iyer PC, Madhavan S. Characterization of stimulus response curves obtained with transcranial magnetic stimulation from bilateral tibialis anterior muscles post stroke. Neurosci Lett 2019;713:134530. http://dx.doi.org/10.1016/j.neulet.2019.134530.

[55] Kemlin C, Moulton E, Leder S, Houot M, Meunier S, Rosso C, Lamy J-C. Redundancy Among Parameters Describing the Input-Output Relation of Motor Evoked Potentials in Healthy Subjects and Stroke Patients. Front Neurol 2019;10:535. http://dx.doi.org/10.3389/fneur.2019.00535.

[56] Filipović SR, Kačar A, Milanović S, Ljubisavljević MR. Neurophysiological Predictors of Response to Medication in Parkinson's Disease. Front Neurol 2021;12. http://dx.doi.org/10.3389/fneur.2021.763911, Publisher: Frontiers.

[57] Eginyan G, Zhou X, Williams AMM, Lam T. Effects of motor stimulation of the tibial nerve on corticospinal excitability of abductor hallucis and pelvic floor muscles. Front Rehabil Sci 2023;3:1089223. http://dx.doi.org/10.3389/fresc.2022.1089223.

[58] Murphy HM, Fetter CM, Snow NJ, Chaves AR, Downer MB, Ploughman M. Lower corticospinal excitability and greater fatigue among people with multiple sclerosis experiencing pain. Mult Scler J - Exp Transl Clin 2023;9(1):20552173221143398. http://dx.doi.org/10.1177/20552173221143398.

[59] Bryson N, Lombardi L, Hawthorn R, Fei J, Keesey R, Peiffer JD, Seáñez I. Enhanced selectivity of transcutaneous spinal cord stimulation by multielectrode configuration. BioRxiv: Prepr Serv Biology 2023. http://dx.doi.org/10.1101/2023.03.30.534835, 2023.03.30.534835.

[60] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. New York: Chapman and Hall/CRC; 1995, http://dx.doi.org/10.1201/9780429258411.

[61] Cronin B, Stevenson IH, Sur M, Körding KP. Hierarchical Bayesian Modeling and Markov Chain Monte Carlo Sampling for Tuning-Curve Analysis. J Neurophysiol 2010;103(1):591–602. http://dx.doi.org/10.1152/jn.00379.2009, Publisher: American Physiological Society.

[62] Coventry BS, Bartlett EL. Practical Bayesian Inference in Neuroscience: Or How i Learned to Stop Worrying and Embrace the Distribution. ENeuro 2024;11(7). http://dx.doi.org/10.1523/ENEURO.0484-23.2024, Publisher: Society for Neuroscience Section: Research Article: Methods/New Tools.

[63] Rossini PM, Burke D, Chen R, Cohen LG, Daskalakis Z, Di Iorio R, Di Lazzaro V, Ferreri F, Fitzgerald PB, George MS, Hallett M, Lefaucheur JP, Langguth B, Matsumoto H, Miniussi C, Nitsche MA, Pascual-Leone A, Paulus W, Rossi S, Rothwell JC, Siebner HR, Ugawa Y, Walsh V, Ziemann U. Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.c.n. Committee. Clin Neurophysiol: Off J the International Fed Clin Neurophysiol 2015;126(6):1071–107. http://dx.doi.org/10.1016/j.clinph.2015.02.001.

[64] Nielsen JF. Logarithmic Distribution of Amplitudes of Compound Muscle Action Potentials Evoked by Transcranial Magnetic Stimulation. J Clin Neurophysiol 1996;13(5):423. http://dx.doi.org/10.1097/00004691-199609000-00005.

[65] Goetz SM, Luber B, Lisanby SH, Peterchev AV. A Novel Model Incorporating Two Variability Sources for Describing Motor Evoked Potentials. Brain Stimul 2014;7(4):541–52. http://dx.doi.org/10.1016/j.brs.2014.03.002.

[66] Goetz SM, Alavi SMM, Deng Z-D, Peterchev AV. Statistical Model of Motor Evoked Potentials. IEEE Trans Neural Syst Rehabil Eng : A Publ the IEEE Eng Med Biology Soc 2019;27(8):1539–45. http://dx.doi.org/10.1109/TNSRE.2019.2926543.

[67] Ma K, Liu S, Qin M, Goetz SM. Extraction of three mechanistically different variability and noise sources in the trial-to-trial variability of brain stimulation. IEEE Trans Neural Syst Rehabil Eng 2024. http://dx.doi.org/10.1109/TNSRE.2024.3522681, 1–1.

[68] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 10 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020;17:261–72. http://dx.doi.org/10.1038/s41592-019-0686-2.

[69] Nelder JA, Mead R. A Simplex Method for Function Minimization. Comput J 1965;7(4):308–13. http://dx.doi.org/10.1093/comjnl/7.4.308.

[70] Wilcoxon F. Individual Comparisons by Ranking Methods. Biom Bull 1945;1(6):80–3. http://dx.doi.org/10.2307/3001968.

[71] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 2017;27(5):1413–32. http://dx.doi.org/10.1007/s11222-016-9696-4.

[72] Kumar R, Carroll C, Hartikainen A, Martin O. ArviZ a unified library for exploratory analysis of Bayesian models in Python. J Open Source Softw 2019;4(33):1143. http://dx.doi.org/10.21105/joss.01143.

[73] Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat 1979;6(2):65–70.

[74] Phan D, Pradhan N, Jankowiak M. Composable effects for flexible and accelerated probabilistic programming in NumPyro. 2019, http://dx.doi.org/10.48550/arXiv.1912.11554, ArXiv.

[75] Bingham E, Chen JP, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip PA, Horsfall P, Goodman ND. Pyro: Deep universal probabilistic programming. J Mach Learn Res 2019;20:28:1–6. http://dx.doi.org/10.48550/arXiv.1810.09538, URL http://jmlr.org/papers/v20/18-403.html.

[76] Hoffman MD, Gelman A. The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 2011, http://dx.doi.org/10.48550/arXiv.1111.4246, arXiv.

[77] McIntosh JR, Joiner EF, Goldberg JL, Greenwald P, Dionne AC, Murray LM, Thuet E, Modik O, Shelkov E, Lombardi JM, Sardar ZM, Lehman RA, Chan AK, Riew KD, Harel NY, Virk MS, Mandigo C, Carmel JB. Timing-dependent synergies between motor cortex and posterior spinal stimulation in humans. J Physiol 2024;602(12):2961–83. http://dx.doi.org/10.1113/JP286183.

[78] Wu YK, Levine JM, Wecht JR, Maher MT, LiMonta JM, Saeed S, Santiago TM, Bailey E, Kastuar S, Guber KS, Yung L, Weir JP, Carmel JB, Harel NY. Posteroanterior cervical transcutaneous spinal stimulation targets ventral and dorsal nerve roots. Clin Neurophysiol 2020;131(2):451–60. http://dx.doi.org/10.1016/j.clinph.2019.11.056.

[79] Mishra AM, Pal A, Gupta D, Carmel JB. Paired motor cortex and cervical epidural electrical stimulation timed to converge in the spinal cord promotes lasting increases in motor responses. J Physiol 2017;595(22):6953–68. http://dx.doi.org/10.1113/JP274663.

[80] Garcia-Sandoval A, Pal A, Mishra AM, Sherman S, Parikh AR, Joshi-Imre A, Arreaga-Salas D, Gutierrez-Heredia G, Duran-Martinez AC, Nathan J, Hosseini SM, Carmel JB, Voit W. Chronic softening spinal cord stimulation arrays. J Neural Eng 2018;15(4):045002. http://dx.doi.org/10.1088/1741-2552/aab90d.

[81] Gelman A, Tuerlinckx F. Type s error rates for classical and Bayesian single and multiple comparison procedures. Comput Statist 2000;15(3):373–90. http://dx.doi.org/10.1007/s001800000040.

[82] Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. 2009, http://dx.doi.org/10.1080/19345747.2011.618213, arXiv.

[83] Kruschke JK. Bayesian estimation supersedes the t test.. J Exp Psychol [Gen] 2013;142(2):573–603. http://dx.doi.org/10.1037/a0029146.

[84] Kruschke J. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. second ed.. Boston: Academic Press; 2014, http://dx.doi.org/10.1016/B978-0-12-405888-0.00001-5.

[85] Vincent BT. Hierarchical Bayesian estimation and hypothesis testing for delay discounting tasks. Behav Res Methods 2016;48(4):1608–20. http://dx.doi.org/10.3758/s13428-015-0672-2.

[86] Ratcliff R, Tuerlinckx F. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. Psychon Bull Rev 2002;9(3):438–81. http://dx.doi.org/10.3758/BF03196302.

[87] Carroll TJ, Riek S, Carson RG. Reliability of the input–output properties of the cortico-spinal pathway obtained from transcranial magnetic and electrical stimulation. J Neurosci Methods 2001;112(2):193–202. http://dx.doi.org/10.1016/S0165-0270(01)00468-X, URL https://www.sciencedirect.com/science/article/pii/S016502700100468X.

[88] Gangitano M, Valero-Cabré A, Tormos JM, Mottaghy FM, Romero JR, Pascual-Leone A. Modulation of input-output curves by low and high frequency repetitive transcranial magnetic stimulation of the motor cortex. Clin Neurophysiol: Off J the International Fed Clin Neurophysiol 2002;113(8):1249–57. http://dx.doi.org/10.1016/s1388-2457(02)00109-8.

[89] Lotze M, Braun C, Birbaumer N, Anders S, Cohen LG. Motor learning elicited by voluntary drive. Brain: A J Neurol 2003;126(Pt 4):866–72. http://dx.doi.org/10.1093/brain/awg079.

[90] Nitsche MA, Seeber A, Frommann K, Klein CC, Rochford C, Nitsche MS, Fricke K, Liebetanz D, Lang N, Antal A, Paulus W, Tergau F. Modulating parameters of excitability during and after transcranial direct current stimulation of the human motor cortex. J Physiol 2005;568(1):291–303. http://dx.doi.org/10.1113/jphysiol.2005.092429.

[91] Goodwill AM, Pearce AJ, Kidgell DJ. Corticomotor plasticity following unilateral strength training. Muscle & Nerve 2012;46(3):384–93. http://dx.doi.org/10.1002/mus.23316.

[92] Murray LM, Tahayori B, Knikou M. Transspinal Direct Current Stimulation Produces Persistent Plasticity in Human Motor Pathways. Sci Rep 2018;8(1):717. http://dx.doi.org/10.1038/s41598-017-18872-z.

[93] Sharma P, Shah PK. In vivo electrophysiological mechanisms underlying cervical epidural stimulation in adult rats. J Physiol 2021;599(12):3121–50. http://dx.doi.org/10.1113/JP281146.

[94] Moezzi B, Schaworonkow N, Plogmacher L, Goldsworthy MR, Hordacre B, McDonnell MD, Iannella N, Ridding MC, Triesch J. Simulation of electromyographic recordings following transcranial magnetic stimulation. J Neurophysiol 2018;120(5):2532–41. http://dx.doi.org/10.1152/jn.00626.2017.

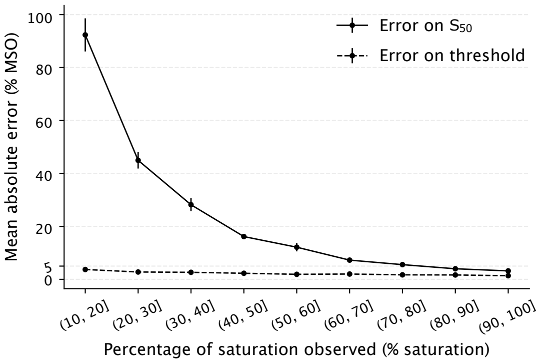
[95] Wilson MT, Moezzi B, Rogasch NC. Modeling motor-evoked potentials from neural field simulations of transcranial magnetic stimulation. Clin Neurophysiol 2021;132(2):412–28. http://dx.doi.org/10.1016/j.clinph.2020.10.032.

[96] Balaguer J-M, Prat-Ortega G, Verma N, Yadav P, Sorensen E, Freitas Rd, Ensel S, Borda L, Donadio S, Liang L, Ho J, Damiani A, Grigsby E, Fields DP, Gonzalez-Martinez JA, Gerszten PC, Fisher LE, Weber DJ, Pirondini E, Capogrosso M. Supraspinal control of motoneurons after paralysis enabled by spinal cord stimulation. 2023, http://dx.doi.org/10.1101/2023.11.29.23298779, medRxiv. Pages: 2023.11.29.23298779.

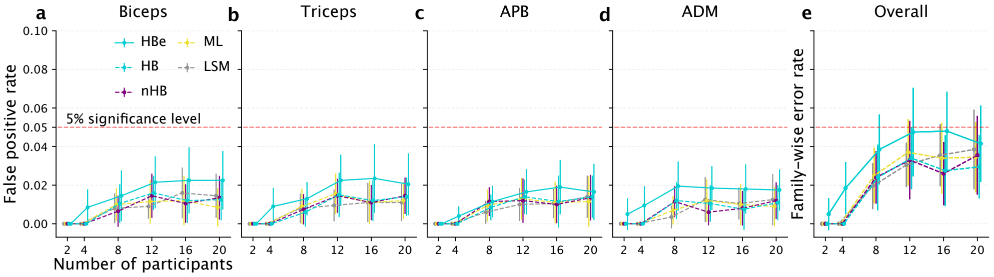
[97] Wang B, Peterchev AV, Goetz SM. Three novel methods for determining motor threshold with transcranial magnetic stimulation outperform conventional procedures. J Neural Eng 2023. http://dx.doi.org/10.1088/1741-2552/acf1cc.

[98] Tyagi V, Murray L, Asan A, Mandigo C, Virk M, Harel N, Carmel J, McIntosh JR. Hierarchical Bayesian estimation of motor-evoked potential recruitment curves yields accurate and robust estimates. 2025, http://dx.doi.org/10.5281/zenodo.15865734, Zenodo.

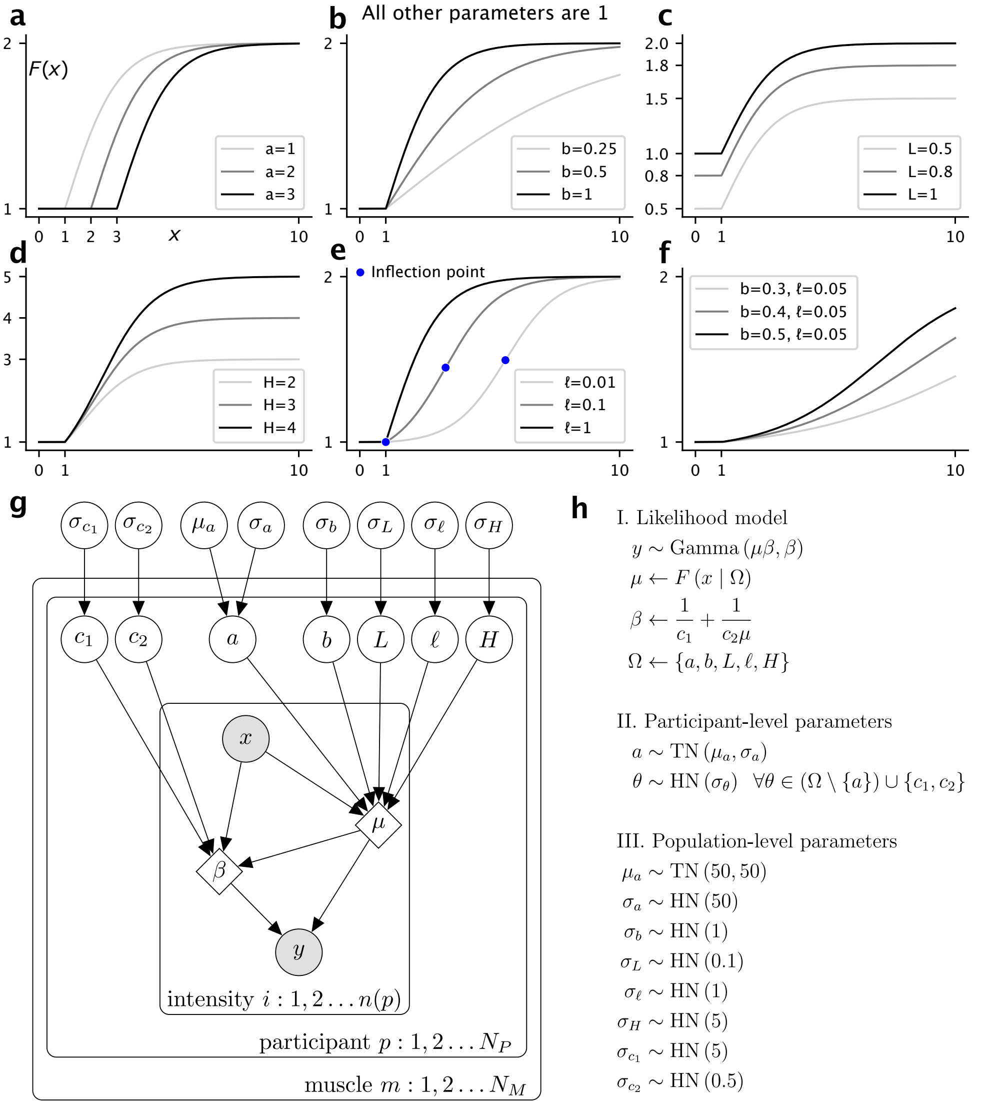
Supplementary information



Supplementary Fig. S1. Accurate estimation of threshold is independent of saturation observed in data and requires less experimental burden on participants, as opposed to S_{50}. For the first eight participants of synthetic TMS data, the mean absolute error for estimating the threshold remains consistently low irrespective of the amount of saturation observed in data. In contrast, partial saturation can result in highly inaccurate and unreliable estimates for the S_{50} parameter. Abscissa bin $(a, b]$ represents the maximum percentage of the actual saturation which is observed in data that is greater than a and less than or equal to b. Error bars represent standard error of the mean absolute error.



Supplementary Fig. S2. Comparison of false positive and family-wise error rates against the set significance level of 5% on bootstrapped human epidural SCS data. False positive rate for the **(a)** biceps, **(b)** triceps, **(c)** APB, and **(d)** ADM muscles. The midline and lateral stimulation conditions were, with equal probability, either interchanged or kept the same for each sampled participant, ensuring that the null hypothesis of no difference between the thresholds of the two conditions holds in all muscles. **(e)** The correction procedure for each method accounts for multiple comparisons across the four tested muscles and the family-wise error rate does not exceed the 5% level.



I. Likelihood model

$y \sim \text{Gamma}\,(\mu\beta, \beta)$

$\mu \leftarrow F\,(x \mid \Omega)$

$\beta \leftarrow \dfrac{1}{c_1} + \dfrac{1}{c_2\mu}$

$\Omega \leftarrow \{a, b, L, \ell, H\}$

II. Participant-level parameters

$a \sim \text{TN}\,(\mu_a, \sigma_a)$

$\theta \sim \text{HN}\,(\sigma_\theta) \quad \forall\theta \in (\Omega \setminus \{a\}) \cup \{c_1, c_2\}$

III. Population-level parameters

$\mu_a \sim \text{TN}\,(50, 50)$

$\sigma_a \sim \text{HN}\,(50)$

$\sigma_b \sim \text{HN}\,(1)$

$\sigma_L \sim \text{HN}\,(0.1)$

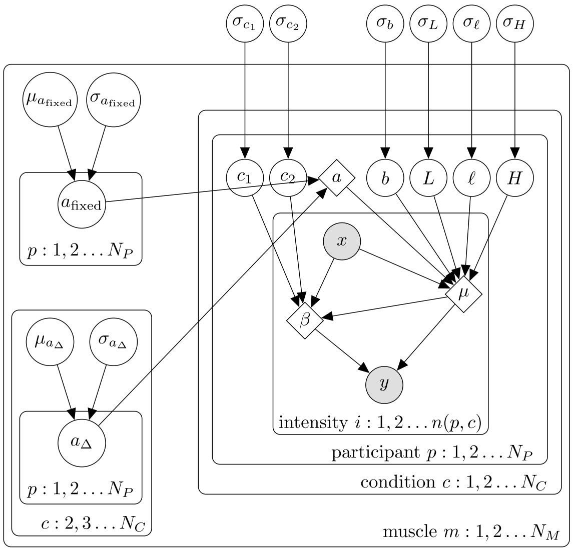
$\sigma_\ell \sim \text{HN}\,(1)$

$\sigma_H \sim \text{HN}\,(5)$

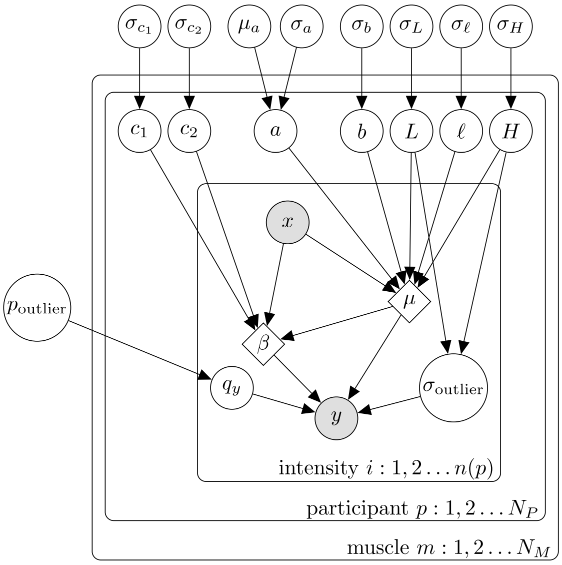
$\sigma_{c_1} \sim \text{HN}\,(5)$

$\sigma_{c_2} \sim \text{HN}\,(0.5)$

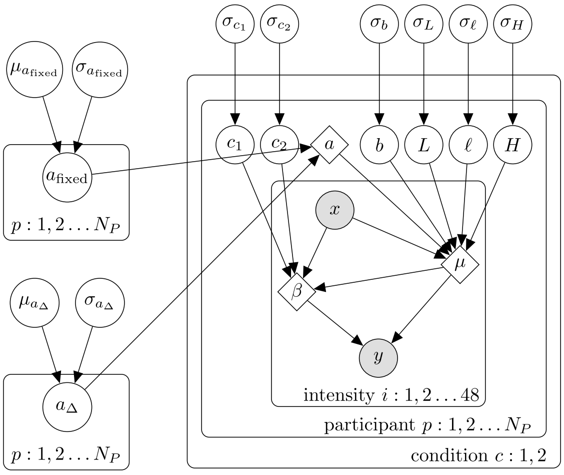
Supplementary Fig. S3. (a–f) Effect of varying parameters of the rectified-logistic function. (a) a shifts the threshold. **(b)** b changes the growth rate. **(c)** L changes the offset MEP size. **(d)** H controls the distance between offset L and saturation $(L + H)$. **(e)** ℓ affects the location of inflection point or point of maximum gradient, whether near offset L or saturation $(L + H)$. **(f)** Varying b, ℓ simultaneously. **(g–h) Standard hierarchical Bayesian model for TMS data. (g)** Graphical model. The model yields parameter estimates for each participant across multiple muscles simultaneously. Circular nodes represent random variables. Filled circular nodes represent observed data. Diamonds represent deterministic variables. Arrows represent that the child node is informed by the distribution of its parent node. Plates denote re-instantiation of nodes. **(h)** Bayesian model specification with participant- and population-level parameters and hyperpriors for TMS data. Here F is the rectified-logistic function.



Supplementary Fig. S4. Paired comparison model. This models the differences in the threshold parameter (a) of participants under multiple experimental conditions, and μ_{a_Δ} summarizes these differences across all participants.



Supplementary Fig. S5. Mixture extension of the standard hierarchical Bayesian model.



$$
\begin{aligned}
y &\sim \text{Gamma}\left(\mu\beta, \beta\right) \\
\mu &\leftarrow F\left(x \mid \Omega\right) \\
\beta &\leftarrow \frac{1}{c_1} + \frac{1}{c_2\mu} \\
\Omega &\leftarrow \{a, b, L, \ell, H\}
\end{aligned}
$$

$$
\begin{aligned}
a^p_{\text{fixed}} &\sim \text{TN}\left(\mu_{a_{\text{fixed}}}, \sigma_{a_{\text{fixed}}}\right) \\
a^p_\Delta &\sim \text{N}\left(\mu_{a_\Delta}, \sigma_{a_\Delta}\right) \\
a^{p,c} &\leftarrow \begin{cases} a^p_{\text{fixed}} & c = 1 \\ a^p_{\text{fixed}} + a^p_\Delta & c = 2 \end{cases} \\
\theta &\sim \text{HN}\left(\sigma_\theta\right) \quad \forall \theta \in \left(\Omega \setminus \{a\}\right) \cup \{c_1, c_2\}
\end{aligned}
$$

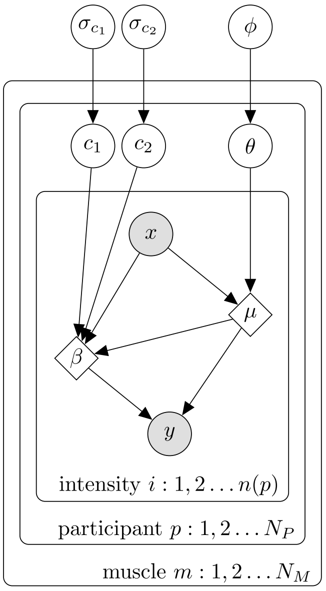
$$
\begin{aligned}
\mu_{a_{\text{fixed}}} &\sim \text{TN}\left(50, 50\right) \\
\sigma_{a_{\text{fixed}}} &\sim \text{HN}\left(50\right)
\end{aligned}
$$

$$
\begin{aligned}
\mu_{a_\Delta} &\sim \text{N}\left(0, 50\right) \\
\sigma_{a_\Delta} &\sim \text{HN}\left(50\right)
\end{aligned}
$$

$$
\begin{aligned}
\sigma_b &\sim \text{HN}\left(1\right) \\
\sigma_L &\sim \text{HN}\left(0.1\right) \\
\sigma_\ell &\sim \text{HN}\left(1\right) \\
\sigma_H &\sim \text{HN}\left(5\right)
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{c_1} &\sim \text{HN}\left(5\right) \\
\sigma_{c_2} &\sim \text{HN}\left(0.5\right)
\end{aligned}
$$

Supplementary Fig. S6. Paired comparison model for detecting a shift in the threshold between pre- and post-intervention. Here, $c = 1$ and $c = 2$ represent pre- and post-intervention conditions, respectively. A priori the model assumes no shift, as indicated by a flat prior on μ_{a_Δ}, which is symmetric about zero. F is the rectified-logistic function.


σ_{c_1}
σ_{c_2}
ϕ
c_1
c_2
θ
x
μ
β
y
intensity $i : 1, 2 \ldots n(p)$
participant $p : 1, 2 \ldots N_P$
muscle $m : 1, 2 \ldots N_M$

$$y \sim \text{Gamma}\left(\mu\beta, \beta\right)$$
$$\mu \leftarrow F\left(x \mid \theta\right)$$
$$\beta \leftarrow \frac{1}{c_1} + \frac{1}{c_2\mu}$$

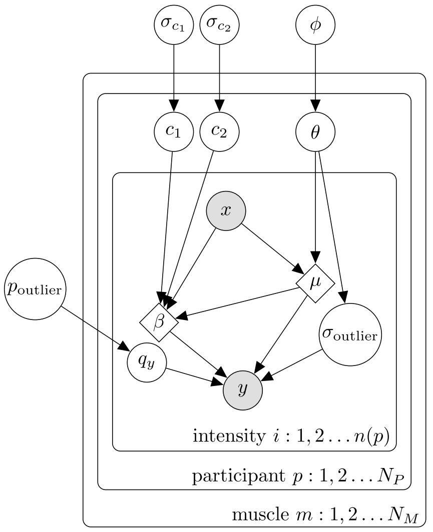
$$a \sim \text{TN}\left(\mu_a, \sigma_a\right)$$
$$b \sim \text{HN}\left(\sigma_b\right)$$

$$v \sim \text{HN}\left(\sigma_v\right)$$
$$L \sim \text{HN}\left(\sigma_L\right)$$
$$\ell \sim \text{HN}\left(\sigma_\ell\right)$$
$$H \sim \text{HN}\left(\sigma_H\right)$$

$$c_1 \sim \text{HN}\left(\sigma_{c_1}\right)$$
$$c_2 \sim \text{HN}\left(\sigma_{c_2}\right)$$

Parameter	Distribution	Rat SCS	Human TMS	Human SCS
μ_a	TN	$(150, 50)$	$(50, 50)$	$(5, 10)$
σ_a	HN	150	50	10
σ_b	HN	1	1	5
σ_v	HN	1	1	1
σ_L	HN	0.1	0.1	0.1
σ_ℓ	HN	1	1	1
σ_H	HN	5	5	10
σ_{c_1}	HN	5	5	5
σ_{c_2}	HN	0.5	0.5	0.5
N_M		6	6	4
N_P		150	27	26

Supplementary Fig. S7. Standard hierarchical Bayesian model for cross-validation of recruitment curve functions on TMS and SCS data. We have $F \in \{F_1, F_2, F_3, F_4\}$, where $F_1 \ldots F_4$ are the rectified-logistic, logistic-5, logistic-4 and rectified-linear functions, respectively. We have $\theta = \theta_f$ and $\phi = \phi_f$ when $F = F_f$ for $f \in \{1, 2, 3, 4\}$. Here, $\theta_1 = \{a, b, L, \ell, H\}$, $\theta_2 = \{a, b, v, L, H\}$, $\theta_3 = \{a, b, L, H\}$, $\theta_4 = \{a, b, L\}$, and $\phi_1 = \{\mu_a, \sigma_a, \sigma_b, \sigma_L, \sigma_\ell, \sigma_H\}$, $\phi_2 = \{\mu_a, \sigma_a, \sigma_b, \sigma_v, \sigma_L, \sigma_H\}$, $\phi_3 = \{\mu_a, \sigma_a, \sigma_b, \sigma_L, \sigma_H\}$, $\phi_4 = \{\mu_a, \sigma_a, \sigma_b, \sigma_L\}$. $\theta_f \cup \{c_1, c_2\}$ are the sets of participant-level parameters and $\phi_f \cup \{\sigma_{c_1}, \sigma_{c_2}\}$ are the corresponding sets of hyperparameters for each function F_f. The table gives the respective hyperprior distributions.


σ_{c_1}
σ_{c_2}
ϕ
c_1
c_2
θ
x
p_{outlier}
μ
β
σ_{outlier}
q_y
y
intensity $i : 1, 2 \ldots n(p)$
participant $p : 1, 2 \ldots N_P$
muscle $m : 1, 2 \ldots N_M$

$$\begin{aligned}
y &\sim (1 - q_y) \cdot \text{Gamma}\left(\mu\beta, \beta\right) + q_y \cdot \text{HN}\left(\sigma_{\text{outlier}}\right) \\
\mu &\leftarrow F\left(x \mid \theta\right) \\
\beta &\leftarrow \frac{1}{c_1} + \frac{1}{c_2\mu}
\end{aligned}$$

$$\begin{aligned}
q_y &\sim \text{Bernoulli}\left(p_{\text{outlier}}\right) \\
p_{\text{outlier}} &\sim \text{Uniform}\left(0, 0.01\right) \\
\sigma_{\text{outlier}} &\sim \text{HN}\left(L + H\right)
\end{aligned}$$

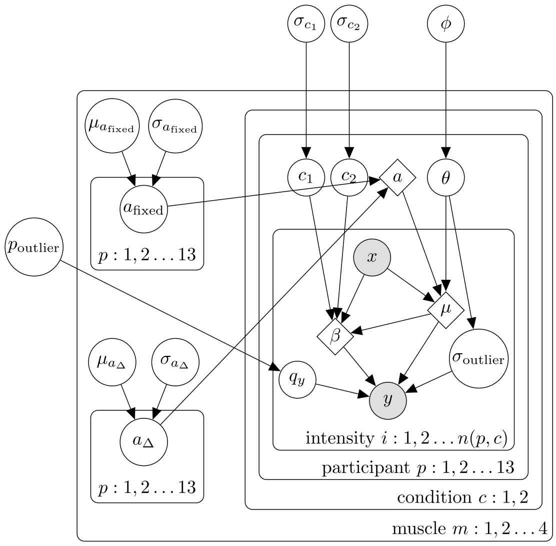
$$\begin{aligned}
a &\sim \text{TN}\left(\mu_a, \sigma_a\right) \\
b &\sim \text{HN}\left(\sigma_b\right)
\end{aligned}$$

$$\begin{aligned}
L &\sim \text{HN}\left(\sigma_L\right) \\
\ell &\sim \text{HN}\left(\sigma_\ell\right) \\
H &\sim \text{HN}\left(\sigma_H\right)
\end{aligned}$$

$$\begin{aligned}
c_1 &\sim \text{HN}\left(\sigma_{c_1}\right) \\
c_2 &\sim \text{HN}\left(\sigma_{c_2}\right)
\end{aligned}$$

Parameter	Distribution	Rat SCS	Human TMS	Human SCS
μ_a	TN	$(150, 50)$	$(50, 50)$	$(5, 10)$
σ_a	HN	150	50	10
σ_b	HN	1	1	5
σ_L	HN	0.1	0.1	0.1
σ_ℓ	HN	1	1	1
σ_H	HN	5	5	10
σ_{c_1}	HN	5	5	5
σ_{c_2}	HN	0.5	0.5	0.5
N_M		6	6	4
N_P		150	27	26

Supplementary Fig. S8. Mixture extension of the standard model for cross-validation on TMS and SCS data. Here F is the rectified-logistic function, $\theta = \{a, b, L, \ell, H\}$, and $\phi = \{\mu_a, \sigma_a, \sigma_b, \sigma_L, \sigma_\ell, \sigma_H\}$. The table gives the hyperpriors.



$$\begin{aligned}
y &\sim (1 - q_y) \cdot \text{Gamma}\left(\mu\beta, \beta\right) + q_y \cdot \text{HN}\left(\sigma_\text{outlier}\right) \\
\mu &\leftarrow \text{F}\left(x \mid a, \theta\right), \quad \beta \leftarrow \frac{1}{c_1} + \frac{1}{c_2 \mu} \\
\theta &\leftarrow \{b, L, \ell, H\}, \ \phi = \{\sigma_b, \sigma_L, \sigma_\ell, \sigma_H\}
\end{aligned}$$

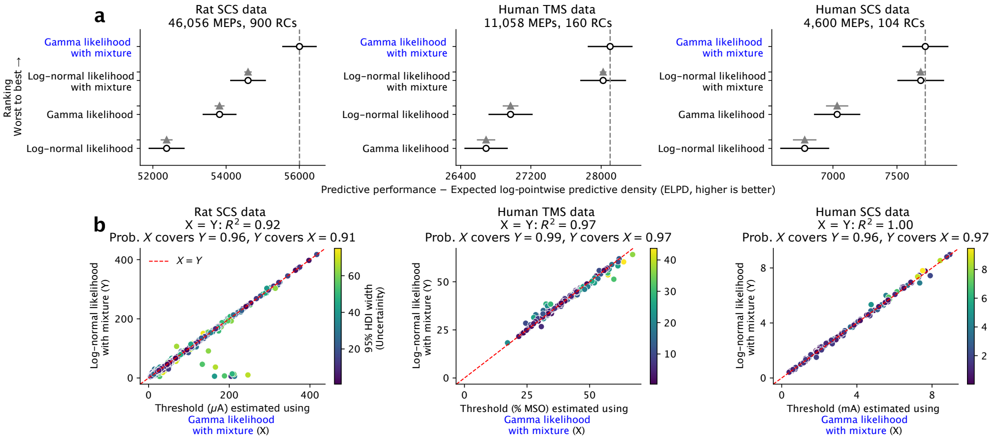
$$\begin{aligned}
q_y &\sim \text{Bernoulli}\left(p_\text{outlier}\right) \\
p_\text{outlier} &\sim \text{Uniform}\left(0, 0.01\right) \\
\sigma_\text{outlier} &\sim \text{HN}\left(L + H\right)
\end{aligned}$$

$$\begin{aligned}
a_\text{fixed}^{p,m} &\sim \text{TN}\left(\mu_{a_\text{fixed}}^m, \sigma_{a_\text{fixed}}^m\right) & \mu_{a_\text{fixed}} &\sim \text{TN}\left(5, 10\right) \\
a_\Delta^{p,m} &\sim \text{N}\left(\mu_{a_\Delta}^m, \sigma_{a_\Delta}^m\right) & \sigma_{a_\text{fixed}} &\sim \text{HN}\left(10\right) \\
a^{p,c,m} &\leftarrow \begin{cases} a_\text{fixed}^{p,m} & c = 1 \\ a_\text{fixed}^{p,m} + a_\Delta^{p,m} & c = 2 \end{cases} & \mu_{a_\Delta} &\sim \text{N}\left(0, 10\right) \\
& & \sigma_{a_\Delta} &\sim \text{HN}\left(10\right)
\end{aligned}$$

$$\begin{aligned}
b &\sim \text{HN}\left(\sigma_b\right) & \sigma_b &\sim \text{HN}\left(5\right) \\
L &\sim \text{HN}\left(\sigma_L\right) & \sigma_L &\sim \text{HN}\left(0.1\right) \\
\ell &\sim \text{HN}\left(\sigma_\ell\right) & \sigma_\ell &\sim \text{HN}\left(1\right) \\
H &\sim \text{HN}\left(\sigma_H\right) & \sigma_H &\sim \text{HN}\left(10\right)
\end{aligned}$$

$$\begin{aligned}
c_1 &\sim \text{HN}\left(\sigma_{c_1}\right) & \sigma_{c_1} &\sim \text{HN}\left(5\right) \\
c_2 &\sim \text{HN}\left(\sigma_{c_2}\right) & \sigma_{c_2} &\sim \text{HN}\left(0.5\right)
\end{aligned}$$

Supplementary Fig. S9. Paired comparison model for comparing the midline and lateral stimulation thresholds on human SCS data. Here, $c = 1$ and $c = 2$ represent lateral and midline stimulation, respectively. A priori the model assumes no difference between the midline and lateral thresholds, as indicated by a flat prior on μ_{a_Δ}, which is symmetric about zero. F is the rectified-logistic function.



Supplementary Fig. S10. Comparison of predictive performance and threshold estimation of standard hierarchical Bayesian (HB) models using gamma (two likelihood parameters) and log-normal (one likelihood parameter) likelihoods, with and without mixture extension. (a) Predictive performance measured with expected log-pointwise predictive density (ELPD) using leave-one-out cross-validation. Gray triangles represent the mean pairwise ELPD difference from the best-ranked mixture extension of gamma HB model (Δ_μ), and gray bars are standard error of the mean ELPD difference (Δ_{sem}). **(b)** Point threshold estimates from the two mixture models are consistent, with a high coefficient of determination (R^2) along the identity line ($R^2 > 0.9$ on rat SCS, and $R^2 \simeq 1$ on human datasets). Additionally, the 95% highest density interval (HDI) of threshold posterior of each model contains, with high probability, the corresponding point threshold estimate of the other, indicating high mutual overlap. Here, the coverage probability that X covers Y denotes the proportion of cases where the 95% HDI of model X contains the corresponding point estimate of model Y. Dots are colored by the maximum width of 95% HDIs across the two models, lighter shades indicate wider HDIs and greater uncertainty in threshold estimates. In a few rat SCS cases, the log-normal model fails to fit (Appendix D1 123.4, 123.6, 125.4, 125.6), and the corresponding threshold estimates deviate substantially from those of the gamma model, which fits these cases well (see corresponding entries of Appendix B1).

S1 Methods

S1.1 Alternative specification

An alternative specification replaces the gamma likelihood (Eq. 6–8) with a log-normal (location-scale) (Nielsen, 1996; Goetz et al., 2014, 2019; Alavi et al., 2019; Ma et al., 2024) (Eq. 18, 19) to estimate the median MEP size as a recruitment curve function of stimulation intensity, since median $(y \mid x, \Omega, \sigma) = e^{\ln \mu} = \mu = F(x \mid \Omega)$. In reference to Table 3, this introduces one likelihood parameter (σ) for the scale of the log-normal, as opposed to two parameters introduced by the gamma model. Equation 20 gives its mixture extension, analogous to that of the gamma model (Eq. 13–16).

For $\sigma > 0$

$$y \mid x, \Omega, \sigma \sim \text{Lognormal}\left(\ln \mu, \sigma\right) \tag{18}$$

$$\mu = F\left(x \mid \Omega\right) \tag{19}$$

$$y \mid x \sim \left(1 - q_y\right) \cdot \text{Lognormal}\left(\ln \mu, \sigma\right) + q_y \cdot \text{HN}\left(\sigma_{\text{outlier}}\right) \tag{20}$$

Appendix C1–C3 show curves estimated using the standard HB model with log-normal likelihood for rat SCS, human TMS, and human SCS data, respectively. In general, the 95% HDIs of the posterior predictive distributions are wider than those of the gamma model (Appendix A1–A3), even in regions with ample data (e.g., C1 2.3 compared to A1 2.3). This is attributable to the model assumption that log-transformed MEP size measurements are homoscedastic around the log-transformed recruitment curve $\ln \mu = \ln F(x \mid \Omega)$, since $\ln y \sim \text{Normal}(\ln \mu, \sigma)$, where the variance σ^2 is constant and independent of the estimated MEP size. While log transformation is often used to stabilize variance in measurements, it does not fully regularize the variance in MEP sizes, which continue to exhibit skewness and increased spread in the transition region between offset and saturation (Goetz et al., 2014).

Appendix D1–D3 show curves estimated using the log-normal mixture HB model for the same datasets. Supplementary Fig. S10a compares the predictive performance of these models. On the largest tested rat SCS dataset, the gamma model, both with and without the mixture extension, significantly outperforms the log-normal counterpart. On human TMS and SCS data, their performance is comparable, making the log-normal model a viable alternative, particularly for TMS data, which tends to exhibit greater skewness. Accordingly, the log-normal is included as an alternative model in the hbMEP library.

Supplementary Fig. S10b shows the point threshold estimates of the two mixture models are consistent ($R^2 > .9$ on rat SCS data, and $R^2 \simeq 1$ on human TMS and SCS datasets). Additionally, the 95% HDI of threshold posterior of each model contains, with high probability, the corresponding point threshold estimate of the other, indicating high mutual coverage. However, in a few rat SCS cases, the log-normal model fails to fit the data (see Appendix D1, F1 123.4, 123.6, 125.4, 125.6). The corresponding threshold estimates deviate substantially from those of the gamma model, which fits these cases well (see corresponding entries of Appendix B1, E1).

An extension of the log-normal model is a triple-variability model (Ma et al., 2024) (Eq. 21–25) which uses the logistic-4 function with zero offset ($L = 0$). In addition to multiplicative noise which characterizes the log-normal, this model further decomposes the mechanistic sources of noise in TMS MEP recordings by introducing an additive Gaussian noise term on the input intensity and an additive offset term for the background noise, which is drawn from a generalized extreme value (GEV, location-scale-shape) distribution.

Let G be the logistic-4 function, then for $\sigma_x, \sigma_{\text{add}} > 0$ and $\mu_{\text{add}}, k_{\text{add}} \in \mathbb{R}$

$$y = \tilde{y} + v_{\text{add}} \tag{21}$$

$$v_{\text{add}} \sim \text{GEV}(\mu_{\text{add}}, \sigma_{\text{add}}, k_{\text{add}}) \tag{22}$$

$$\tilde{y} \mid \tilde{x}, \sigma \sim \text{Lognormal}(\ln \tilde{\mu}, \sigma) \tag{23}$$

$$\tilde{\mu} = G(\tilde{x} \mid L = 0, \Omega), \quad \Omega = \{a, b, H\} \tag{24}$$

$$\tilde{x} \sim \text{Normal}(x, \sigma_x) \tag{25}$$

The triple-variability model introduces five likelihood parameters $(\sigma, \sigma_x, \mu_{\text{add}}, \sigma_{\text{add}}, k_{\text{add}})$ and includes perturbed stimulation intensities $\tilde{x}$, which are latent variables and not counted as model parameters. It involves sampling from a convolution of independent GEV and log-normal distributions (Eq. 21), which does not have a closed-form and is therefore not supported by standard probabilistic programming libraries (Bingham et al., 2019; Phan et al., 2019). Additionally, the support of GEV distribution depends on its parameters, which poses challenges to gradient-based Markov Chain Monte Carlo samplers such as No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2011). In the current framework (Ma et al., 2024), model parameters are estimated using numerical optimization. Specifically, GEV parameters for the offset are estimated separately from baseline-only trials collected at 0% MSO and subsequently held fixed during estimation of the logistic-4 recruitment curve using particle swarm optimization.

Appendix A4 and B4 show curves estimated using the gamma HB model and its mixture extension, respectively, on data simulated from the triple-variability model. We simulated responses for 10 participants using fixed model parameters estimated from real data (Ma et al., 2024). The first column shows data generated with 25 stimuli, evenly spaced between 30% MSO and 100% MSO, with 20 repetitions per stimulus, for a total of 500 samples per participant (Ma et al., 2024). The subsequent columns show datasets with 64, 32, and 16 stimuli, each with a single repetition. In all cases, both the gamma HB model and its mixture extension do a good job of capturing the variability in data, as indicated by the shaded 95% HDI of the posterior predictive distribution, which covers most of the observed data.