SI Appendix

A microstructural neural network biomarker for dystonia diagnosis identified by a DystoniaNet deep learning platform

Davide Valeriani[a,b,c] and Kristina Simonyan[a,b,c1]

[a]Department of Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear Infirmary, Boston, MA 02114; [b]Department of Otolaryngology – Head and Neck Surgery, Harvard Medical School, Boston, MA 02114; and [c]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114

SI Materials and Methods

Study participants and data collection

The data were collected by the senior author on this study over the period from 2004 to 2020 at three different sites, including Massachusetts Eye and Ear/Mass General Brigham (MEE/MGB), Icahn School of Medicine at Mount Sinai (ISMMS), and Intramural National Institute of Neurological Disorders of Stroke, National Institutes of Health (NIH). Out of 392 patients with dystonia, MRI data of 325 patients were acquired specifically for this study, and data of 67 patients were obtained from our other studies, which used similar protocols for clinical evaluation and brain imaging data acquisition (Tables S1 and S2). Out of 220 healthy controls, MRI data of 69 controls were acquired specifically for this study; data of 43 controls were obtained from other studies conducted by us, and data of 108 controls were obtained through the publicly available Information eXtraction from Images (IXI) dataset (brain-development.org/ixi-dataset/) collected at Hammersmith Hospital, the UK, matching them by their age, sex, and scanning parameters with patient cohorts. The study was approved by the institutional review boards of the Icahn School of Medicine at Mount Sinai and Mass General Brigham Research Program. All subjects gave written informed consent for study participation. Those subjects whose data were obtained from other studies conducted by our laboratory gave written informed consent for their data sharing between different study protocols.

www.pnas.org/cgi/doi/10.1073/pnas.2009165117

The diagnosis of isolated dystonia was confirmed by at least two clinicians and the senior author on this study using a recommended multi-disciplinary approach, which was based on a detailed case history and neurological and laryngological evaluations, as applicable (1-3). Only patients with confirmed diagnosis of isolated focal dystonia were included in the study. That is, those with unclear diagnosis or those whose diagnosis was not agreed upon by all examiners were not included.

None of the patients had any other neurological disorders (except for co-occurring tremor in 32.4% of patients), psychiatric, or laryngeal problems. Those patients who received botulinum toxin injections for the management of their dystonia symptoms were included at the end of their treatment cycle, at least three months after the last injections, when fully symptomatic. Healthy controls were healthy individuals without any neurological, psychiatric, or laryngeal problems. None of patients or controls were on any medications affecting the central nervous system, and none had any surgery to the body region affected by dystonia.

Subjects were assigned to the three groups as follows (Table 1):

(I) the training set of 160 patients with laryngeal dystonia and 160 healthy controls, which was used to train all machine-learning models;

(II) the first independent test set of 60 patients with laryngeal dystonia and 60 healthy controls, which was used to evaluate the performance of all machine-learning algorithms and further optimize DystoniaNet;

(III) the second independent test set of 59 patients with laryngeal dystonia, 54 patients with blepharospasm, and 59 patients with cervical dystonia, which was used to evaluate the generalizability of the best-performing biomarker and its optimized algorithmic platform.

To facilitate the cross-validation of our findings, harmonization within training and testing data sets was achieved by using cohorts of patients and healthy controls as clinically homogeneous as possible while controlling for several variables, including the age, sex, scanner magnetic field strength, scanner vendor, head coil, acquisition sequence, and data collection site. Clinical homogeneity was achieved by including only patients with confirmed diagnosis of isolated focal dystonia and excluding those with unclear diagnosis and any other neurological (except for co-occurring tremor in 32.4% of patients), psychiatric, or laryngeal disorders. Healthy controls were healthy volunteers who expressed interest in research study participation and were enrolled from the general population or the publicly available databases according to the same study inclusion/exclusion criteria.

The age and sex of subjects were tightly balanced between the patient and control groups in the training set (all $p \geq 0.37$), on which all machine-learning pipelines were designed and trained. Subjects in the first and second independent test sets were randomized into their respective groups. The possible differences in sex or age distributions in the first and second independent test sets did not affect the performance of machine-learning algorithms because their internal parameters were fixed before the testing of these data sets was performed.

To balance the effects of the scanner vendors, hardware, and data collection sites in our large dataset to the best of our abilities, we used MRI data from all three scanner vendors and different hardware (i.e., head coils) across all acquisition sites in both training and testing sets (Table S2). The parameters of the whole-brain T1-weighted sequence protocol (MPRAGE, SPGR, T1-weighted FLAIR) were kept as stable as possible across all data sets, both acquired by us and obtained through the public databases (Table S3). In all subjects, head movements during scanning were minimized by tightly cushioning and restricting the head inside the coil. All images were manually inspected to ensure their quality and the absence of gross radiological abnormalities or image artifacts. Raw whole-brain MR images in all patients and controls were qualitatively examined for the absence of artifacts, including field-of-view clipping anatomy, wrapping artifacts, ringing, striping, blurring, ghosting, radio frequency noise, and signal inhomogeneity.

Based on these stringent inclusion/exclusion criteria and quality control procedures, the final study cohort of 504 subjects (392 patients and 112 controls) was selected from our larger database of 695 subjects that includes 513 patients with dystonia and 182 healthy subjects. Similarly, 108 healthy subjects of the IXI database used in the training and first independent test sets were selected from the available cohort of 592 subjects to match our subjects by their age, sex, and scanning parameters.

To further validate the performance of optimized DystoniaNet with respect to its specificity, we assembled a third independent test set of 1,480 healthy controls, whose raw structural MRIs were obtained from the Human Connectome Project (HCP; N = 1,112; 506 females/606 males; age 28.8 ± 3.7 years), the IXI dataset collected at the Hammersmith and Guy's Hospitals (N = 349; 177 females/172 males; age 48.5 ± 16.3 years), and other similar studies in our laboratory (N = 19; 1 female/18 males; age 39.7 ± 6.9 years). None of these healthy controls were included in either training or testing sets of the main study, thus representing an entirely independent cohort. MRIs were collected using a T1-weighted MPRAGE sequence using 3.0 Tesla or 1.5 Tesla scanners and 8- or 32-channel head coils, similar to the data sets in the main study (Tables S1 and S2).

Table S1. Scanner hardware and parameters used at different acquisition sites

	NIH	ISMMS		MEE/MGB						IXI	HCP
Number of subjects	104	281		138						457	1112
Scanner	GE	Philips	Siemens	Siemens		GE				Philips	Siemens
Strength	3T	3T	3T	1.5T	3T	1.5T		3T		3T / 1.5T	3T
Sequence	MPRAGE	MPRAGE	MP2RAGE	MPRAGE	MPRAGE	SPGR	T1w-FLAIR	SPGR	MPRAGE	MPRAGE	MPRAGE
Head coil (channels)	1 / 8	8	32	8 / 12 / 20	12 / 20 / 32	8	32	8	8	8	32
TE (ms)	3.0	3.4	2.0	4.8 / 4.1 / 3.4	1.6 / 2.1 / 2.0	6.5	26	3.2	2.1	4.6	2.14
TI (ms)	725 / 450	900	633-1000	1100 /1000 / 1100	1200 / 1000 / 1000	450	809	450	900	n/a	1000
FA (degree)	12 / 10	8	8	15 / 8 / 15	7 / 8 / 8	13	160	12	8	8	8
FOV (mm)	240x240	210x210	240x240	250x250	256x256	260x260	240x240	240x240	240x240	240x240	224x224
Slice thickness (mm)	1.2 / 1.3	1.0	1.0	1.0 / 1.3 / 1.0	1.0/0.8/1.0	1.0	5.0	0.9	1.0	1.2	0.7

T, Tesla; TE, echo time; TI, inversion time; FA, flip angle; FOV, field of view; SPGR, spoiled gradient-recalled echo; MPRAGE, magnetization prepared-rapid gradient echo; MP2RAGE, magnetization-prepared 2 rapid gradient echoes; T1w-FLAIR, T1-weighted fluid-attenuated inversion recovery; NIH, National Institute of Health; ISMMS, Icahn School of Medicine at Mount Sinai; MEE/MGB, Massachusetts Eye and Ear/Mass General Brigham; IXI, Information eXtraction from Images; HCP, Human Connectome Project; n/a, not available. All sequences were acquired in a 3-dimentional (3D) mode.

Table S2. MRI acquisition parameters and data acquisition sites used in each dataset

	Training set 160 patients/ 160 controls	**1st independent test set** 60 patients/ 60 controls	**2nd independent test set** 172 patients	**3rd independent test set** 1480 controls
Scanner vendor				
Philips	50%	62%	20%	24%
Siemens	31%	23%	40%	75%
GE	19%	15%	40%	1%
Scanner strength				
3.0 Tesla	100%	100%	61%	79%
1.5 Tesla	-	-	39%	21%
Head coil				
Single-channel	2%	2%	8%	-
8-channel	66%	75%	52%	24%
12-channel	-	-	13%	-
20-channel	1%	-	9%	-
32-channel	31%	23%	18%	76%
Scanning site				
NIH	19%	15%	10%	1%
ISMMS	55%	42%	25%	1%
MEE/MGB	6%	6%	65%	-
IXI	20%	37%	-	23%
HCP	-	-	-	75%

NIH, National Institute of Health; ISMMS, Icahn School of Medicine at Mount Sinai; MEE/MGB, Massachusetts Eye and Ear/Mass General Brigham; IXI, information extraction from images; HCP, human connectome project.

Model training and performance evaluations

Deep learning DystoniaNet pipeline. Details on DystoniaNet architecture are given in the Results section of the paper. The total number of floating-point operations (FLOPs) required to generate the output was 0.1 megaFLOPs, showing that DystoniaNet was more energy efficient (4) than other pipelines by up to six orders of magnitudes (5). As a comparison, popular AlexNet requires 727 megaFLOPs, hence being more than three orders of magnitude less energy efficient than DystoniaNet, even though it classifies 2D rather than 3D images, as in the case of DystoniaNet. The latter was trained for up to 200 epochs using the gradient-based stochastic optimizer Adam with a global learning rate of 10^{-4} without decay. Early stopping was employed to speed up the training of the network by monitoring the validation loss in the last 40 epochs. DystoniaNet was implemented in Python 3.6.0 using Keras library v2.2.4 with a

Tensorflow v1.14 backend. We trained the model on a Tesla K80 GPU on a Deep Learning Amazon Machine Image (AMI) v24.0, run on the Amazon Web Services EC2 p2.xlarge instance.

Shallow machine-learning pipelines. To select features for the shallow machine-learning pipelines, we conducted meta-analysis of neuroimaging literature in laryngeal dystonia following the PRISMA guidelines (6). That is, on May 31, 2019, we searched the PubMed online library (https://pubmed.ncbi.nlm.nih.gov) for original research articles with the following search query: "((laryngeal AND dystonia) OR (spasmodic AND dysphonia)) AND ((functional AND MRI) OR (speech AND production AND MRI) OR (resting AND state) OR (fMRI) OR (positron AND emission AND tomography) OR (brain AND activity) OR (brain AND activation))) NOT (Review[Publication Type])". This search resulted in 46 papers with no duplicates. We reviewed all 46 papers, and those meeting all of the following criteria were included in meta-analysis:

(1) assessed structural and/or functional data in patients with laryngeal dystonia vs. healthy controls;
(2) included at least 6 subjects per group;
(3) reported *x,y,z* coordinates of abnormal brain regions in standard Talairach-Tournoux or MNI space;
(4) were original peer-reviewed research articles written in English.

Fourteen papers met these inclusion criteria and were selected for meta-analysis; four papers reported both structural and functional abnormalities and were considered as separate studies from meta-analytical perspective (Table S3). From each study, the number of subjects and the coordinates of functionally or structurally abnormal brain regions were manually extracted, resulting in a total of 1,084 subjects (both patients and controls) and 221 clusters of structural and functional abnormalities. Coordinates originally reported in the MNI standard space were converted into the Talairach-Tournoux standard space using publicly available GingerALE 3.0.2 software.

Table S3. Summary of studies included in the meta-analysis

Study	N	Measure	x	y	z	Standard space
Waugh *et al.* (11)	48	GMV	-6.97	-24.69	6.64	TT
		GMV	7	-22	8	TT
		GMV	-5	-20	8	TT
Kirke *et al.* (12)	60	GMV	-45	9	24	TT
		GMV	-30	-9	-5	TT

Study	N	Measure	x	y	z	Standard space
		GMV	30	5	1	TT
		GMV	33	-12	0	TT
		GMV	-16	-7	0	TT
		WMV	-36	7	19	TT
Simonyan and Ludlow (13)	80	GMV	-39	-18	46	MNI
		GMV	51	17	28	MNI
		GMV	-60	-39	24	MNI
		GMV	-18	3	-12	MNI
		GMV	-34	-42	-60	MNI
		CT	-41	-10	34	TT
		CT	51	-10	36	TT
		CT	-57	-17	34	TT
		CT	31	-30	55	TT
		CT	-35	9	55	TT
		CT	-49	33	1	TT
		CT	-58	-43	-4	TT
		CT	43	-43	6	TT
		CT	-34	-10	-8	TT
		CT	-40	-43	38	TT
		CT	-55	-37	44	TT
		CT	50	-38	23	TT
		CT	48	-43	24	TT
Kostic *et al.* (14)	46	CSA	-42	-32	35	TT
		CSA	-63	-22	-1	TT
		CSA	-19	8	52	TT
		CSA	39	-11	54	TT
		CSA	54	-18	44	TT
		CSA	-57	-38	36	TT
		CSA	-16	-52	62	TT
		CSA	-26	-61	6	TT
		CSA	-58	-5	10	TT
		CSA	-55	5	2	TT
		CSA	52	7	5	TT
		CSA	47	-61	35	TT
Termsarasab *et al.* (15)	114	GMV	-50	-18	22	TT
		GMV	-21	27	36	TT
		GMV	64	-27	-5	TT
		GMV	26	-14	12	TT
		GMV	-50	-11	40	TT

Study	N	Measure	x	y	z	Standard space
		GMV	-6	-61	-12	TT
		GMV	-10	-28	33	TT
		GMV	39	-28	25	TT
		GMV	38	-7	13	TT
		GMV	-39	-11	17	TT
		GMV	-44	0	21	TT
		GMV	-54	-39	25	TT
		GMV	18	3	-1	TT
Ramdhani *et al.* (16)	48	GMV	-36	-14	60	MNI
		GMV	-51	-50	38	MNI
		GMV	-18	-54	27	MNI
		GMV	-18	-80	-24	MNI
		GMV	45	-24	32	MNI
		GMV	48	-17	10	MNI
		GMV	49	-27	36	MNI
		GMV	-45	-9	9	MNI
		GMV	40	5	2	MNI
		GMV	-50	-33	-14	MNI
		GMV	-47	-45	37	MNI
		GMV	49	-47	-12	MNI
		WMV	-27	42	-3	MNI
		WMV	33	42	-2	MNI
		WMV	10	5	1	MNI
		WMV	-15	37	1	MNI
		WMV	-9	-51	-30	MNI
Bianchi *et al.* (17)	32	GMV	-29	-68	50	MNI
		GMV	31	-9	64	MNI
		WMV	8	-62	27	MNI
Haslinger *et al.* (18)	24	fMRI	-32	-34	62	MNI
		fMRI	-36	-34	54	MNI
		fMRI	64	-12	24	MNI
		fMRI	64	-6	18	MNI
		fMRI	4	6	36	MNI
		fMRI	6	20	40	MNI
		fMRI	14	0	54	MNI
		fMRI	22	22	62	MNI
		fMRI	22	12	62	MNI
		fMRI	-24	30	54	MNI
		fMRI	-8	12	64	MNI
		fMRI	54	20	10	MNI

Study	**N**	**Measure**	**x**	**y**	**z**	**Standard space**
		fMRI	-14	30	38	MNI
		fMRI	-12	60	26	MNI
		fMRI	26	-70	52	MNI
		fMRI	28	-82	30	MNI
		fMRI	-38	-40	-8	MNI
		fMRI	-22	-6	-28	MNI
		fMRI	-28	-70	-50	MNI
		fMRI	-24	-28	60	MNI
		fMRI	38	-42	56	MNI
		fMRI	46	-42	52	MNI
		fMRI	54	-42	40	MNI
		fMRI	-50	-40	36	MNI
		fMRI	-44	-44	50	MNI
		fMRI	4	-40	52	MNI
		fMRI	52	18	-20	MNI
		fMRI	66	-38	4	MNI
		fMRI	54	-62	-20	MNI
		fMRI	18	-74	32	MNI
Simonyan and Ludlow (13)	30	fMRI	-43	-33	54	TT
		fMRI	27	-41	34	TT
		fMRI	-47	-13	32	TT
		fMRI	43	-7	30	TT
		fMRI	-48	3	23	TT
		fMRI	60	2	15	TT
		fMRI	-59	-3	14	TT
		fMRI	-32	-10	-10	TT
		fMRI	-55	-37	12	TT
		fMRI	-53	-15	-6	TT
		fMRI	59	-19	-18	TT
		fMRI	1	-25	-4	TT
		fMRI	-45	-62	-34	TT
		fMRI	27	-53	-38	TT
Kiyuna *et al.* (19)	12	fMRI	-51	-42	17	TT
		fMRI	-55	16	14	TT
		fMRI	36	12	9	TT
		fMRI	38	37	4	TT
		fMRI	22	-64	-27	TT
		fMRI	-32	-62	-27	TT
		fMRI	-53	-16	30	TT
		fMRI	-26	0	-2	TT

Study	N	Measure	x	y	z	Standard space
		fMRI	24	4	0	TT
Battistella *et al.* (20)	113	fMRI	-22	-32	63	TT
		fMRI	-60	-26	23	TT
		fMRI	-30	-6	7	TT
		fMRI	2	-4	57	TT
		fMRI	62	-18	15	TT
		fMRI	-26	-42	37	TT
Kiyuna *et al.* (21)	28	fMRI	-20	-10	26	MNI
		fMRI	-50	-6	-44	MNI
		fMRI	48	52	-8	MNI
		fMRI	44	-12	28	MNI
		fMRI	28	-62	56	MNI
		fMRI	44	-20	58	MNI
		fMRI	54	-54	46	MNI
		fMRI	-50	28	22	MNI
		fMRI	0	-86	-10	MNI
		fMRI	18	-92	16	MNI
		fMRI	-48	-4	-16	MNI
		fMRI	-26	-28	14	MNI
		fMRI	-26	-16	64	MNI
		fMRI	40	8	0	MNI
		fMRI	24	-52	-38	MNI
		fMRI	10	-54	-32	MNI
		fMRI	54	-14	50	MNI
		fMRI	-44	-16	54	MNI
		fMRI	26	18	4	MNI
		fMRI	-30	-38	52	MNI
		fMRI	-10	-48	-22	MNI
		fMRI	14	8	70	MNI
		fMRI	-20	-90	-28	MNI
		fMRI	24	-80	-20	MNI
		fMRI	-64	-28	18	MNI
		fMRI	-16	-70	-16	MNI
Putzel *et al.* (22)	87	fMRI	-22	-32	63	TT
		fMRI	2	-4	57	TT
		fMRI	-60	-26	23	TT
		fMRI	62	-18	15	TT
		fMRI	-30	-6	7	TT
Termsarasab *et al.* (15)	114	fMRI	-33	15	38	TT
		fMRI	-37	19	30	TT

Study	N	Measure	x	y	z	Standard space
		fMRI	-30	20	31	TT
		fMRI	-23	8	46	TT
		fMRI	-51	-25	0	TT
		fMRI	-42	-36	8	TT
		fMRI	-40	-35	4	TT
		fMRI	27	-31	44	TT
		fMRI	57	-17	28	TT
		fMRI	1	-23	64	TT
		fMRI	-5	-33	56	TT
		fMRI	1	21	40	TT
		fMRI	-1	25	30	TT
		fMRI	-23	-77	-36	TT
		fMRI	11	-70	-15	TT
		fMRI	15	-67	-14	TT
		fMRI	57	-21	12	TT
		fMRI	-37	-21	54	TT
		fMRI	-37	-25	42	TT
		fMRI	-31	-21	48	TT
		fMRI	-32	6	29	TT
		fMRI	-7	-49	26	TT
		fMRI	37	-12	17	TT
		fMRI	31	19	32	TT
Kirke *et al.* (12)	60	fMRI	-40	-14	34	TT
		fMRI	44	-13	29	TT
		fMRI	-49	-9	44	TT
		fMRI	-49	1	6	TT
		fMRI	-57	-27	6	TT
		fMRI	-25	-7	0	TT
		fMRI	-15	-21	-4	TT
		fMRI	15	-19	-2	TT
		fMRI	41	11	4	TT
		fMRI	53	-7	12	TT
		fMRI	5	-3	58	TT
		fMRI	45	-53	22	TT
		fMRI	25	21	46	TT
Battistella and Simonyan (23)	75	fMRI	-30	-14	-1	TT
		fMRI	-22	-16	55	TT
		fMRI	-36	-16	55	TT
		fMRI	-58	-26	21	TT
		fMRI	24	-24	63	TT

Study	N	Measure	x	y	z	Standard space
de Lima Xavier and Simonyan (24)	81	fMRI	54	-11	38	TT
		fMRI	40	10	-8	TT
		fMRI	51	-50	20	TT
		fMRI	-12	10	-4	TT
		fMRI	47	-11	34	TT
		fMRI	37	-22	10	TT
		fMRI	-51	-50	20	TT
		fMRI	51	-64	20	TT
		fMRI	37	-15	45	TT
		fMRI	44	-25	31	TT
		fMRI	54	-53	34	TT
		fMRI	-51	-46	17	TT
		fMRI	33	-67	48	TT
		fMRI	23	20	52	TT
		fMRI	54	-11	38	TT
		fMRI	47	-11	34	TT
Bianchi *et al.* (17)	32	fMRI	44	-47	55	MNI

GMV, gray matter volume; CT, cortical thickness; CSA, cortical surface area; fMRI, functional magnetic resonance imaging; TT, Talairach-Tournoux standard space; MNI, Montreal Neurological Institute standard space.

Meta-analysis was performed using GingerALE software, which calculated the activation likelihood estimation (ALE) at each voxel (7) and identified significant locations of common brain abnormalities across published studies. GingerALE uses a random-effects algorithm to determine the agreement between groups and reported clusters, incorporates variable uncertainty based on the study cohort, and limits the effect of a single experiment (8). The ALE significance threshold was set at family-wise error (FWE)-corrected $p \leq 0.05$ by simulating 500 random datasets with the same characteristics as the input dataset in terms of the number of clusters, groups, and subjects and applying cluster-forming $p \leq 0.01$.

The final ALE map of significant clusters was used as a binary mask to extract the average gray matter volume and cortical thickness in each cluster in each subject of the training and first independent test sets. For this, individual brain images in 440 subjects comprising the training and first independent test sets were skull-stripped and segmented into gray matter, white matter, and cerebrospinal fluid tissues using standard SPM tissue probability maps of the CAT12 toolbox of SPM12 software. Gray matter probability maps were normalized to the AFNI standard Talairach-Tournoux space using a diffeomorphic nonlinear registration (DARTEL) and smoothed using a 6-mm Gaussian kernel full-width at half-maximum. The final image quality was assessed by visual inspection of the quality check module of the CAT12 toolbox. Cortical thickness was estimated using the standard pipeline of FreeSurfer software. All brain images were visually inspected for accuracy of cortical boundaries; manual corrections of reconstructed surfaces were made, as needed, including the correction of erroneous skull stripping by adjusting watershed parameters, manual editing of skull tissue, and an addition of control points to normalize intensity for white matter surface reconstruction. Cortical thickness measures were calculated based on the shortest distance between gray matter and white matter, as well as gray matter and cerebrospinal fluid boundaries at each vertex on the tesselate surface. The cortical thickness maps were smoothed using a 6-mm Gaussian kernel full-width at half-maximum. Each participant was, therefore, represented by 12 features: six extracted from gray matter volume and six extracted from cortical thickness.

The shallow pipelines used three different classifiers: linear discriminant analysis (LDA), support vector machine (SVM), and one-layer artificial neural network (ANN). All classifiers were implemented in Python 3.6.0 using the Scikit-learn v0.21.0 library. The *predict_proba* method of each classifier instance was used to estimate the probability of each patient to have dystonia (i.e., value between 0 and 1) from the selected features. LDA was chosen for the absence of hyperparameters to tune. LDA models the class conditional distribution of the input features makes probabilistic predictions using Bayes rule for each class and selects the class, which maximizes this conditional probability. LDA assumes each class has a normal distribution with the same covariance matrix, hence applying a linear decision boundary between classes.

SVM was selected for its ability to identify nonlinear associations between features and labels, without making assumptions about the distribution of the dataset, as opposed to LDA. SVM learns a multi-dimensional representation (kernel) of the features where the classes could be separated by a hyperplane. Depending on the kernel, SVM can efficiently perform both linear and nonlinear classification. In our case, we chose a radial basis function kernel, so that SVM could perform nonlinear transformations of the features that predict the diagnosis of the subject. Moreover, through the regularization parameter C, SVM could assign higher or lower penalties to misclassified data points of the training set. We chose C = 1 as we sought a compromise between underfitting and overfitting. Both LDA and SVM showed a previously promising performance on resting-state fMRI data in dystonia patients (9, 10).

The one-layer ANN was chosen for its comparability with deep neural networks. The ANN is a multilayer perceptron classifier composed of an input layer, a hidden layer of six neurons, and an output layer. Each neuron used a rectified linear unit (ReLU) activation function to define the output of that neuron given a set of features. The ANN was trained using the stochastic gradient-based optimizer Adam to optimize the weights associated with the connections between neurons. We used a regularization term of 0.0001 to control for overfitting. The input features were normalized by subtracting the mean and dividing by the standard deviation on the training set.

Impact of MRI scanner, acquisition sequence, and data collection site on the accuracy of DystoniaNet

To determine whether the performance of DystoniaNet depends on the MRI scanner, acquisition protocols, or data collection site, we stratified our subjects from the most heterogeneous second independent test set based on the MR scanner magnetic field strength [3.0 Tesla (N=105), 1.5 Tesla (N=67)], scanner vendor [GE (N=69), Siemens (N=68), Philips (N=35)], head coil [number of channels: 1 (N=13), 8 (N=90), 12 (N=22), 20 (N=15), 32 (N=32)], T1-weighted acquisition sequence [MPRAGE/MP2RAGE (N=123), SPGR (N=41), FLAIR (N=8)], and data collection site [MEE/MGB (N=112), ISMMS (N=43), NIH (N=17)]. We reassessed the performance of the final optimized DystoniaNet in these stratified cohorts for its accuracy of dystonia diagnosis and the referral rate using the same pipeline as described above (Fig. 1).

Impact of shallow machine-learning and DystoniaNet algorithmic complexity on diagnostic performance

We used the visualized components (clusters of layers I, II, III) of DystoniaNet-identified biomarker to extract the corresponding white matter and gray matter (cortical thickness and gray matter volume) values as input features for LDA, SVM and ANN classifiers. Shallow machine-learning pipelines were trained using the training set as described above and tested using the first independent test set. By keeping the input features of shallow learning pipelines as comparable as possible to those of DystoniaNet, this analysis assessed the impact of algorithmic complexity of these pipelines on their overall diagnostic potential.

SI Results

Performance of DystoniaNet does not depend on MRI scanner, acquisition sequence or site

The diagnostic accuracy of DystoniaNet was consistently high across the MR scanner magnetic field strength and vendors. Specifically, DystoniaNet achieved 98.0% accuracy in diagnosing dystonia from 3.0 Tesla data with a 3.8% referral rate, and 100% accuracy from 1.5 Tesla images with a 3.0% referral rate (Fig. S1*A*). Based on the scanner vendor, DystoniaNet achieved 100% accuracy and referred two patients (2.9%) using MRI data acquired on the GE scanners, 96.9% accuracy with a referral four patients (5.9%) using MRI data acquired on the Siemens scanners, and 100% accuracy with no referrals using MRI data acquired on the Philips scanners data (Fig. S1*B*).

Similarly, we did not find any dependency of the DystoniaNet performance on the number of channels of the head coil used for data acquisition. DystoniaNet achieved 100% accuracy using the head coils with 1, 8, and 20 channels, 95.2% accuracy with 12 channels, and 96.8% accuracy with 32 channels (Fig. S1*C*). Referral rate was slightly higher for 20-channel coil (13.3%) compared to 12-channel coil (4.5%), 32-channel coil (3.1%), 8-channel coil (2.2%).

The performance of DystoniaNet was similarly independent of the T1-weighted sequence acquisition protocol. It achieved 98.3% accuracy with MPRAGE sequence, referring five patients (4.1%), 100% accuracy with SPGR sequence, referring one patient (2.4%), and 100% accuracy with the T1-weighted FLAIR sequence, with no referred patients (Fig. S1*D*). MPRAGE data represented 72% of images in our independent test set, hence showing a proper generalization of the model to a larger dataset.

Finally, DystoniaNet showed consistently high accuracy when tested on the datasets collected at three different sites. The diagnostic accuracy was 99.1% with 3.6% referral rate for data collected at MEE/MGB, 97.6% with 2.3% referral rate for data collected at ISMMS, and 100% with 5.9% referral rate for data collected at NIH (Fig. S1*E*).

Overall, the DystoniaNet-based biomarker performance did not depend on a particular MRI scanner magnetic field strength, vendor, head coil, T1-weighted acquisition protocol, or data collection site. Particularly interesting is its exceptional performance (100% accuracy with 3% referral rate and no misclassified patients) on 1.5 Tesla data, given that the DystoniaNet model was trained on 3.0 Tesla data only. These results suggest that the microstructural neural network biomarker identified by DystoniaNet may also be used with data acquired at lower resolution. This is particularly encouraging for the clinical translation of DystoniaNet, given that clinical MR scanners usually have a limited magnetic field strength.

Moreover, the consistently high performance of DystoniaNet across the head coils excludes the possibility that the differences in channel distribution between patients and healthy controls in the training set have acted as confounding factors in our analysis.

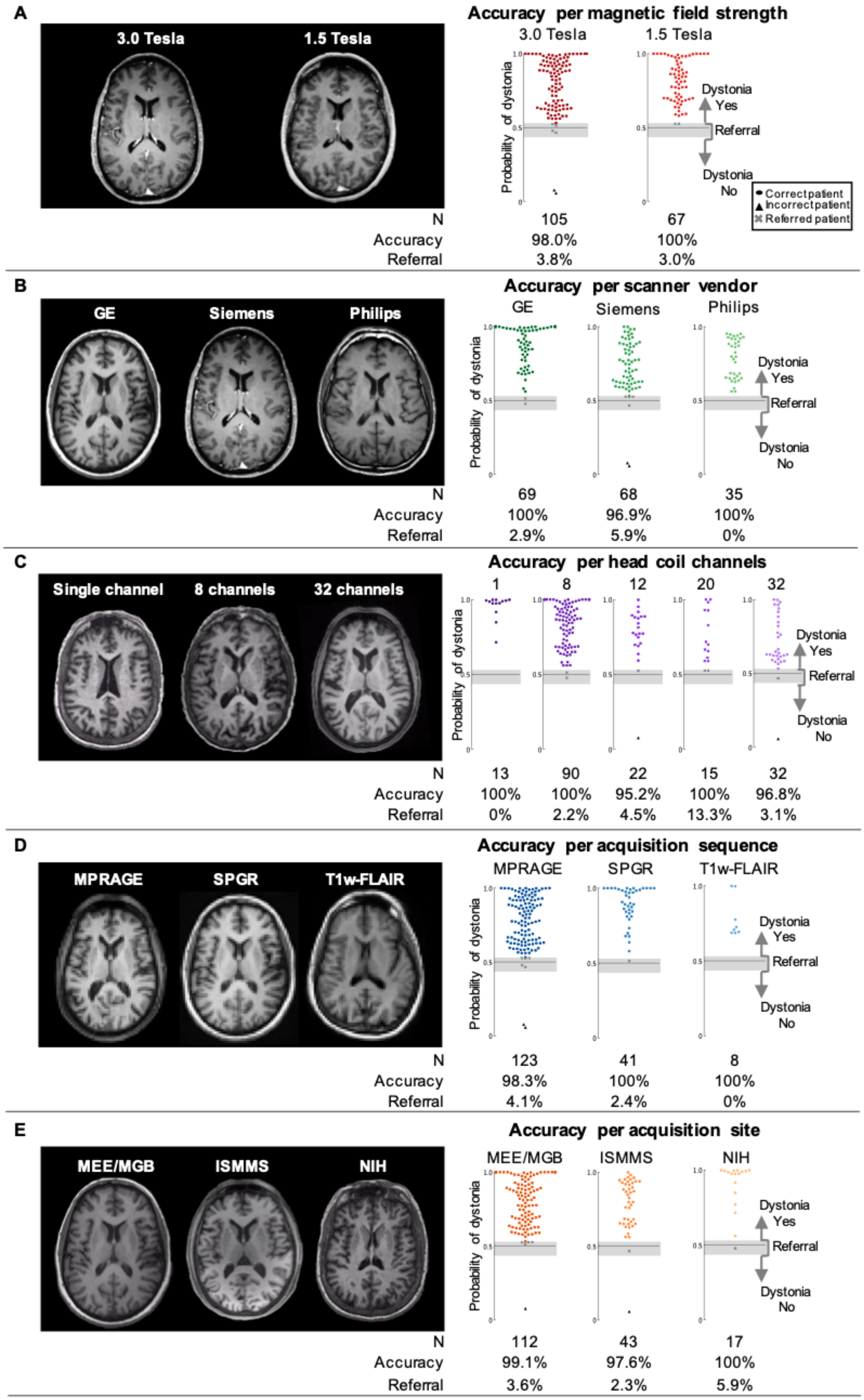

Fig. S1. Performance of DystoniaNet based on MR scanner magnetic field strength, vendor, acquisition sequence, head coil, and data collection site. (*A*) Sample axial MR images acquired with 3.0 Tesla and 1.5 Tesla scanners (*left*) and diagnostic performance of DystoniaNet for MR images in the second independent test set grouped by scanner's magnetic field strength (*right*). (*B*) Sample axial MR images acquired with MRI scanners from different vendors (*left*) and the diagnostic performance of DystoniaNet with MR images in the second independent test set grouped by the scanner vendor (*right*). (*C*) Sample axial MR images acquired using the head coils with a different number of channels (*left*) and diagnostic performance of DystoniaNet for MR images in the second independent test set grouped by the number of channels of head coil (*right*). (*D*) Sample axial MR images acquired with different T1-weighted acquisition sequences (*left*) and diagnostic performance of the DystoniaNet for MR images in the second independent test set grouped

by the MRI acquisition sequence (*right*). (*E*) Sample axial MR images acquired at different sites (*left*) and diagnostic performance of DystoniaNet for MR images in the second independent test set grouped by a site of data collection (*right*). In (*A-E*), correctly classified patients are represented by circles; misclassified patients are represented by triangles; patients referred to further examinations are represented by crosses. Colored symbols represent correct dystonia diagnosis; black symbols represent incorrect diagnosis. The *y* axis represents the probability of dystonia as assessed by DystoniaNet; the gray line represents the decision boundary; the gray shading represents the area of uncertainty where DystoniaNet refers the subject for further evaluation. The corresponding sample size, accuracy (computed excluding referrals), and referral rate values are given.

Shallow machine-learning algorithms lack the diagnostic power of DystoniaNet

To examine the impact of algorithmic complexity of shallow learning pipelines vs. DystoniaNet, their performance was re-assessed using the visualized components (clusters of layers I, II, III) of DystoniaNet-identified biomarker as input features for LDA, SVM and ANN (Fig. S2*A*). Using the data of the first independent test set, their AUCs were as follows: 72.0% for LDA, 77.7% for SVM, and 58.8% for ANN (Fig. S2*B* and *D*). The sensitivity and specificity were 46.7% and 75.0% for LDA, 58.3% and 83.3% for SVM, and 50.0% and 60.0% for ANN, respectively (Fig. S2*C*). The positive predictive value (PPV) and negative predictive value (NPV) were 65.1% and 58.4% for LDA, 77.8% and 66.7% for SVM, and 55.6% and 54.5% for ANN, respectively. The McNemar's tests showed that the diagnostic performance of shallow pipelines was similar when using clusters identified by DystoniaNet and clusters identified by meta-analysis (LDA $p = 0.10$; SVM $p = 0.54$; ANN $p = 0.17$).

The overall underperformance of shallow learning pipelines, even when using DystoniaNet-identified input features, may be explained by the fundamentally low complexity of their internal models compared to the deep-learning architecture of DystoniaNet. In addition, the shallow learning architecture lacked the dense layer of 40 filters as part of feature extraction (Fig. 1*A*), which was present in DystoniaNet and non-linearly combined the features extracted from informative clusters for increased diagnostic power of the deep learning platform. It is also worth noting that DystoniaNet-identified input features for shallow learning pipelines were derived only from the visualized components of biomarker in the first three layers of DystoniaNet because the significant clusters in the fourth layer could not be visualized due to the low spatial resolution (Fig. 2). Hence, the diagnostic information present in the fourth layer of DystoniaNet was likely critically missing during training and testing of these shallow machine-learning pipelines. Taken together, compared to LDA, SVM and ANN, DystoniaNet demonstrated improved discriminative accuracy based on the higher complexity of its model *and* a superior combination of feature extraction and classification components. These results signify the importance of DystoniaNet as a combined framework for both biomarker detection and diagnosis of dystonia.

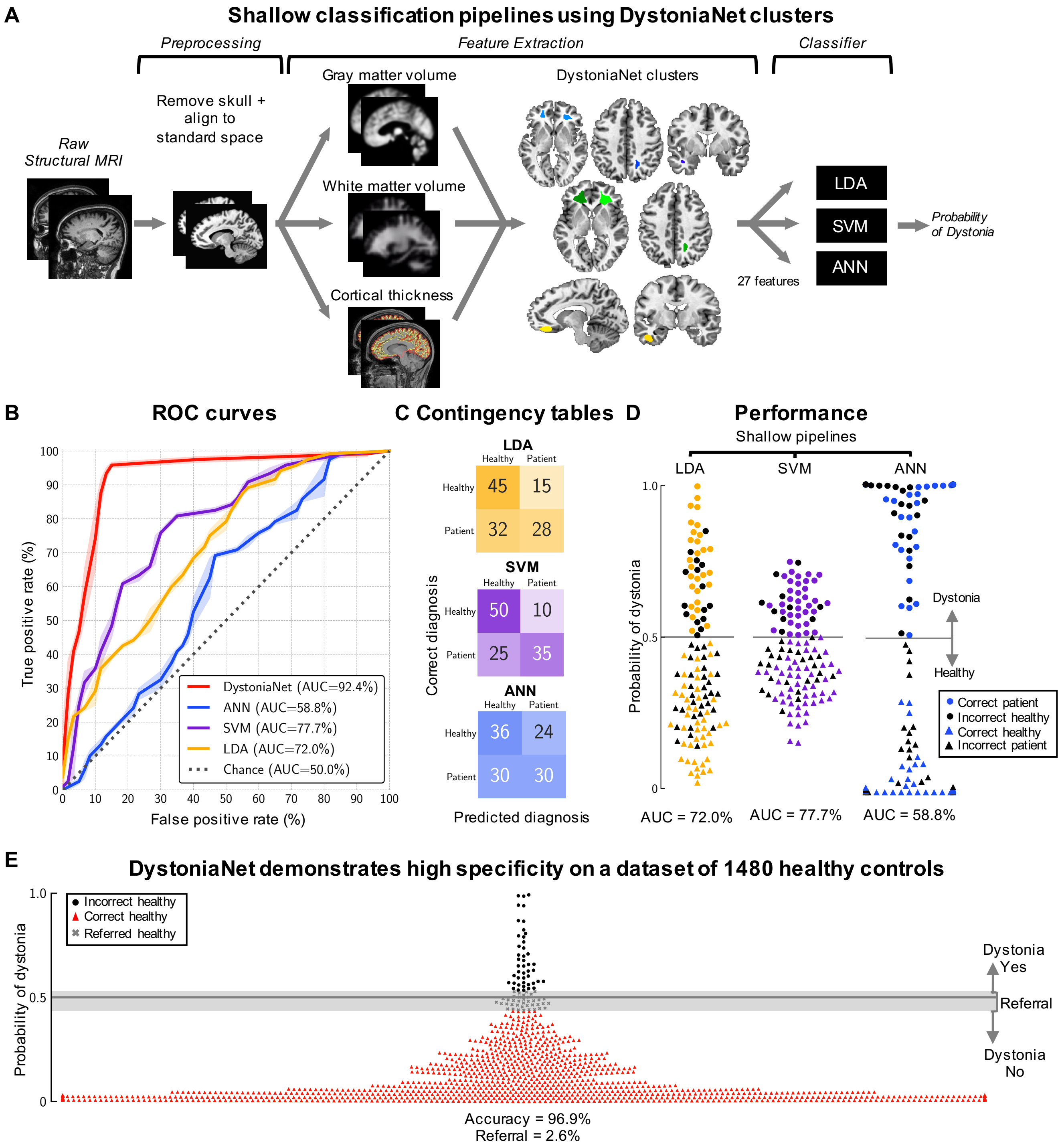

Fig. S2. (*A-D*) Performance of shallow machine-learning pipelines using the visualized components of DystoniaNet-identified biomarker. (*A*) Diagram of the shallow machine-learning pipelines extracting features from cortical thickness, gray matter and white matter volumes based on the visualized clusters of a biomarker identified by DystoniaNet. The 27 features per subject were used as input for three different classifiers: linear discriminant analysis (LDA), support vector machine (SVM), and single-layer artificial neural network (ANN). (*B*) Receiving operating characteristic (ROC) curves for each shallow machine-learning pipeline using DystoniaNet-identified clusters in the first independent test set of 60 patients with laryngeal dystonia and 60 healthy controls. The area under the ROC curve (AUC) values for each pipeline are shown in the legend. The dotted line represents the performance of a random classifier. (*C*) The corresponding contingency tables report the number of healthy controls and patients that are correctly and incorrectly classified by each shallow machine-learning pipeline. (*D*) The diagnostic performance of each shallow pipeline in the first independent test set. Each symbol represents a subject. Subjects classified as patients are represented by circles; subjects classified as healthy controls are represented by triangles. Colored symbols represent correct diagnosis; black symbols represent misclassifications. The *y* axis represents the probability of dystonia as assessed by each pipeline; the gray line represents the decision boundary. The corresponding AUC values are given for each pipeline. **(*E*) Performance of DystoniaNet in an additional third independent set of 1480 healthy controls.** Each symbol represents a subject. Subjects correctly classified as healthy controls are represented by red triangles; misclassified subjects are represented by black circles; referral are represented by gray crosses. The *y* axis represents the probability of dystonia; the gray line represents the decision boundary; the gray shading shows the area of diagnostic uncertainty (referral). The corresponding accuracy and referral rate are reported. Data are visualized using the Matplotlib library (25).

DystoniaNet delivers high specificity on a large independent dataset of healthy controls

When tested on the third independent test set of 1,480 healthy individuals, optimized DystoniaNet achieved 96.9% accuracy in correctly classifying healthy controls, with 2.6% referral rate (Fig. S2*E*). These results provide additional support for the overall robust and accurate performance of DystoniaNet based on the identified microstructural neural network biomarker.

References

1. C. L. Ludlow *et al.*, Consensus-Based Attributes for Identifying Patients With Spasmodic Dysphonia and Other Voice Disorders. *JAMA Otolaryngol Head Neck Surg* (2018).
2. C. L. Ludlow *et al.*, Research priorities in spasmodic dysphonia. *Otolaryngol Head Neck Surg* **139**, 495-505 (2008).
3. B. Balint, K. P. Bhatia, Dystonia: an update on phenomenology, classification, pathogenesis and treatment. *Curr Opin Neurol* **27**, 468-476 (2014).
4. R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI. *arXiv preprint arXiv:1907.10597* (2019).
5. S. Albanie (2017) Memory consumption and FLOP count estimates for convnets. (https://github.com/albanie/convnet-burden).

6. D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* **6**, e1000097 (2009).

7. S. B. Eickhoff *et al.*, Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum Brain Mapp* **30**, 2907-2926 (2009).

8. S. B. Eickhoff, D. Bzdok, A. R. Laird, F. Kurth, P. T. Fox, Activation likelihood estimation meta-analysis revisited. *Neuroimage* **59**, 2349-2361 (2012).

9. G. Battistella, S. Fuertinger, L. Fleysher, L. J. Ozelius, K. Simonyan, Cortical sensorimotor alterations classify clinical phenotype and putative genotype of spasmodic dysphonia. *Eur J Neurol* **23**, 15-17-1527 (2016).

10. Z. Li *et al.*, Alterations of resting-state fMRI measurements in individuals with cervical dystonia. *Hum Brain Mapp* **38**, 4098-4108 (2017).

11. J. L. Waugh *et al.*, Thalamic Volume Is Reduced in Cervical and Laryngeal Dystonias. *PLoS One* **11**, e0155302 (2016).

12. D. N. Kirke *et al.*, Neural correlates of dystonic tremor: a multimodal study of voice tremor in spasmodic dysphonia. *Brain Imaging Behav* **11**, 166-175 (2017).

13. K. Simonyan, C. L. Ludlow, Abnormal structure-function relationship in spasmodic dysphonia. *Cereb Cortex* **22**, 417-425 (2012).

14. V. S. Kostic *et al.*, Brain structural changes in spasmodic dysphonia: A multimodal magnetic resonance imaging study. *Parkinsonism Relat Disord* **25**, 78-84 (2016).

15. P. Termsarasab *et al.*, Neural correlates of abnormal sensory discrimination in laryngeal dystonia. *Neuroimage Clin* **10**, 18-26 (2016).

16. R. A. Ramdhani *et al.*, What's special about task in dystonia? A voxel-based morphometry and diffusion weighted imaging study. *Mov Disord* **29**, 1141-1150 (2014).

17. S. Bianchi, S. Fuertinger, H. Huddleston, S. J. Frucht, K. Simonyan, Functional and structural neural bases of task specificity in isolated focal dystonia. *Mov Disord* **34**, 555-563 (2019).

18. B. Haslinger *et al.*, "Silent event-related" fMRI reveals reduced sensorimotor activation in laryngeal dystonia. *Neurology* **65**, 1562-1569 (2005).

19. A. Kiyuna *et al.*, Brain activity related to phonation in young patients with adductor spasmodic dysphonia. *Auris Nasus Larynx* **41**, 278-284 (2014).

20. G. Battistella, S. Fuertinger, L. Fleysher, L. J. Ozelius, K. Simonyan, Cortical sensorimotor alterations classify clinical phenotype and putative genotype of spasmodic dysphonia. *European Journal of Neurology* **23**, 1517-1527 (2016).

21. A. Kiyuna *et al.*, Brain Activity in Patients With Adductor Spasmodic Dysphonia Detected by Functional Magnetic Resonance Imaging. *J Voice* **31**, 379 e371-379 e311 (2017).

22. G. G. Putzel *et al.*, Polygenic Risk of Spasmodic Dysphonia is Associated With Vulnerable Sensorimotor Connectivity. *Cereb Cortex* **28**, 158-166 (2018).

23. G. Battistella, K. Simonyan, Top-down alteration of functional connectivity within the sensorimotor network in focal dystonia. *Neurology* **92**, e1843-e1851 (2019).

24. L. de Lima Xavier, K. Simonyan, The extrinsic risk and its association with neural alterations in spasmodic dysphonia. *Parkinsonism Relat Disord* (2019).

25. J. D. Hunter, Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90-95 (2007).