An Annotated Multi-Site and Multi-Contrast Magnetic Resonance Imaging Dataset for the study of the Human Tongue Musculature

Fernanda L. Ribeiro[1*], Xiangyun Zhu[1,2], Xincheng Ye[1], Sicong Tu[3,4], Shyuan T. Ngo[5,6], Robert D. Henderson[7,5], Frederik J. Steyn[5,8], Matthew C. Kiernan[3,9], Markus Barth[1,10,11,12], Steffen Bollmann[1,10], Thomas B. Shaw[1,11,5*]

Affiliations:

[1] School of Electrical Engineering and Computer Science, The University of Queensland; Queensland, Australia

[2] Griffith School of Medicine and Dentistry, Queensland, Australia

[3] Neuroscience Research Australia, Sydney, NSW, Australia.

[4] Brain and Mind Centre, Faculty of Medicine and Health, University of Sydney, NSW

[5] Royal Brisbane and Women's Hospital, Herston; Queensland, Australia

[6] Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St Lucia; Queensland, Australia

[7] University of Queensland Centre for Clinical Research, Queensland, Australia

[8] School of Biomedical Sciences, The University of Queensland, St Lucia; Queensland, Australia

[9] Scientia Professor of Neuroscience, The University of NSW, and Department of Neurology, Southeastern Sydney Local Health District, Sydney, NSW, Australia.

[10] Queensland Digital Health Centre (QDHeC), The University of Queensland, Herston; Queensland, Australia

[11] Centre for Advanced Imaging, Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St Lucia; Queensland, Australia

[12] ARC Training Centre for Innovation in Biomedical Imaging and Technology (CIBIT), Centre for Advanced Imaging, The University of Queensland, St Lucia; Queensland, Australia

* **Corresponding authors:** Fernanda L. Ribeiro (fernanda.ribeiro@uq.edu.au), Thomas B. Shaw (t.shaw@uq.edu.au)

Abstract

This dataset provides the first fully annotated, openly available MRI-based imaging dataset for investigations of tongue musculature, including multi-contrast and multi-site MRI data from non-disease participants. The present dataset includes 47 participants collated from three studies: BeLong (four participants; T2-weighted images), EATT4MND (19 participants; T2-weighted images), and BMC (24 participants; T1-weighted images). We provide automatically generated and manually corrected segmentation of five key tongue muscles: the superior longitudinal, combined transverse/vertical, genioglossus, and inferior longitudinal muscles. Other phenotypic measures, including age, sex, weight, height, and tongue muscle volume, are also available for use. This dataset will benefit researchers across domains interested in the structure and function of the tongue in health and disease. For instance, researchers can use this data to train new machine learning models for tongue segmentation, which can be leveraged for segmentation and tracking of different tongue muscles engaged in speech formation in health and disease. Altogether, this dataset provides the means to the scientific community for investigation of the intricate tongue musculature and its role in physiological processes and speech production in health and disease.

Keywords

Tongue segmentation, tongue anatomy, atlas, MRI

Background and Summary

The human tongue is involved in many physiological processes[1,2] and speech production[3]. The tongue plays a crucial role in the manipulation and recognition of food[1] (e.g., food texture[4]), tasting[2] and thermosensation[5,6], breathing[7], and speech[3,8]. This diverse set of functionalities characterises the tongue as both a motor and a sensory organ[2]. Moreover, to subserve these processes, different tongue muscles may compress and/or elongate, but the overall tissue volume is constant, i.e., the tongue is a muscular hydrostat[7]. The tongue is also implicated in neurodegenerative diseases[9,10], developmental speech pathologies[11], and sleep disorders[12], proving to be a potential marker for a wide range of diseases that implicate tongue function during speech, food processing, and breathing. For example, studies have found diffuse T1-weighted hyperintensity of the tongue musculature in Amyotrophic Lateral Sclerosis (ALS) patients[13,14] and reduced tongue volume in ALS patients with bulbar palsy[15]. Despite the potential of the measures of the tongue (e.g., morphometry and volume) as a biomarker, there exists no comprehensive annotated and publicly available MRI dataset to help inform understanding of tongue anatomy in health and in disease. We therefore introduce this critical resource that will enable the identification of new biomarkers and interventions.

The tongue is comprised of anatomically distinguishable and interconnected intrinsic and extrinsic muscles[16,17]. The intrinsic muscles of the tongue both originate and insert within the tongue itself. There are four pairs of these intrinsic muscles: the superior longitudinal, inferior longitudinal, transverse, and vertical muscles. In contrast, the extrinsic muscles originate from structures outside the tongue, including the genioglossus, hyoglossus, styloglossus, and palatoglossus muscles. Each of these muscles can move the tongue in a particular direction[7], but coordinated contractions of multiple muscles work together to enable movements like protrusion, retraction, and elevation of the tongue, as well as changing its shape and position during activities such as chewing and swallowing[18].

Magnetic resonance imaging (MRI) is an important non-invasive technique that allows for the imaging and identification of the intricate tongue musculature[19]. Accordingly, in our recent study (Shaw et al., 2024; *in prep*), we compiled a detailed guideline for the identification and segmentation of the superior longitudinal, transverse/vertical combined, genioglossus and inferior longitudinal muscles of the tongue. Although

detailed manual annotation was performed using T2-weighted (T2w) images—mostly from patients with motor neuron disease (MND)—where the contrast differences were clear across these different muscles, we developed a semi-automated segmentation pipeline for tongue segmentation of participants across different studies and MRI contrasts (T1-weighted (T1w) and T2-weighted images). Here we provide a subset of that data, i.e., the data from non-disease controls, to aid future research into the structure and function of the tongue in health and disease.

This is the first openly available fully annotated MRI-based imaging data for tongue segmentation. This dataset includes structural MRI data (T2w and/or T1w) from non-disease controls, with accompanying template space and demographic information, from three studies/scanners. Moreover, we provide automatically generated and manually corrected tongue segmentation labels, to aid segmentation model training in prospective studies. Finally, we also distribute an atlas generated with manually corrected segmentations to provide a more accurate model of tongue muscle location and size. This dataset will benefit researchers across domains interested in the structure and function of the tongue in health and disease. This data could aid machine learning model training for tongue segmentation, which can be leveraged for segmentation and tracking of different tongue muscle engagement in speech formation using high-speed real-time MRI (e.g., CINE)[20,21]. Similarly, these data can be used to inform longitudinal studies of disease progression, in which physiological functions of the tongue are implicated, e.g., bulbar onset ALS[22]. In sum, this new imaging/segmentation dataset of the human tongue provides the means to the scientific community for investigation of the intricate tongue musculature and its role in physiological processes and speech production in health and disease. Since the tongue is typically included in standard MRI scans of the brain and head/neck, the annotated data and atlas we provide here can serve as a helpful resource for analyzing both existing and future MRI datasets.

Methods

Participants

The present dataset includes information from 47 non-diseases “healthy” participants (20 females, 25-80 years old) collated from three studies: The Biomarkers of Long surviving MND (BeLong; 4 participants), Exploring Appetite Targets and Therapies for

Motor Neurone Disease (EATT4MND;19 participants), and the Brain and Mind Centre Motor Neuron Disease neuroimaging database (BMC; 24 participants). Table 1 shows basic demographic information across studies.

Ethics

All studies were approved by their relevant Human Research Ethics Committees. Specifically, BeLong was approved by the University of Queensland HREC (2021/HE000975), EATT4MND was approved by the University of Queensland HREC and Royal Brisbane and Women's Hospital (RBWH) HRECs (HREC/17/QRBW/616), and Uniting Care Health Human Research Ethics Committee (#1801), and the BMC dataset was approved by the University of Sydney HREC (2021/283). All participants provided written and informed consent.

Table 1. Demographic information.

Study	*Number of individuals (number of females)*	*Age range (years)*	*Weight/Height availability*
BELong	4 (2)	25-74	Available
EATT4MND	19 (5)	26-73	Available
BMC	24 (13)	37-80	Not available

Image acquisition

BeLong. Data from the BeLong study were collected between 2020-2023 and include 4 non-neurodegenerative healthy control (HC) participants. Imaging was performed at the University of Queensland, Centre for Advanced Imaging using a 3T Siemens Prisma (PrismaFit, Siemens Healthineers, Erlangen, Germany) using a 64-channel head and neck coil.

Participants were imaged using a 3D SPACE T2w sequence for spinal cord imaging covering the tongue with an isotropic resolution of 0.8mm^3. This sequence shows high contrast of the tongue and surrounding tissue and was acquired with the following parameters: TR=1500ms, TE=120ms, TA=4m:02s, FA=120°, matrix size=256x320, and number of slices=64.

EATT4MND. Participant information from the EATT4MND study has been described elsewhere[23]. Briefly, 24 HCs were imaged at the Herston Imaging Research Facility, of which 19 participants are included in our data repository. Data were collected using a 3T Siemens Prisma (Siemens Healthcare, Erlangen, Germany). T2-weighted (T2w) scans were obtained from a 3D SPACE 1mm^3 isotropic sequence with the following parameters: TR = 5000ms, TE = 386ms, TI = 1800ms, TA = 5m:52s, matrix size =256x256 and number of slices=176[23].

BMC. Whole-brain imaging was performed using a 3T MRI scanner (GE MR750, DV29; 32-channel Nova head coil) at the Brain and Mind Centre (BMC), The University of Sydney, Australia. Imaging data from 25 HCs were collected, 24 participants are included in our repository. Coronal T1-weighted images were acquired using an 1mm^3 isotropic MPRAGE sequence (parameters: TE = 2.3ms, TR = 6.2ms; TI = 500ms, FA = 12°; matrix size = 256x256; number of slices = 204, TA = 5m 31s).

Semi-automated tongue segmentation

Stage 1 – Active learning. To generate annotated data for segmentation model training, we initially manually annotated the BeLong dataset T2w images from healthy controls and MND patients. Specifically, three tongue volumes from three scans were manually annotated by XZ (Medical Principal House Officer with three years medical experience). These initially labelled data have been used in conjunction with MONAI Label[24] within the Slicer application[25] available on Neurodesk[26] to interactively annotate the BeLong study data and iteratively train a model for T2w MRI tongue muscle segmentation. MONAI Label is based on active learning, which is a strategy that starts off by training a segmentation model on limited annotated data, which is then used for selecting a new sample from a pool of unlabelled data that may be labelled to improve model's performance in the next iteration of model training. In this context, sample selection is based on the model's uncertainty[27]. Using this framework, we trained a DynUNet model to segment five key tongue muscles: genioglossus, transverse and vertical muscles combined, superior longitudinal, and inferior longitudinal (Figure 1a), using default data augmentation strategies, including random flip and intensity scaling. In this iterative process, XZ, FLR (six years medical imaging experience), and TBS (ten years medical imaging experience) corrected and labelled 9 additional scans with assistance from XY.

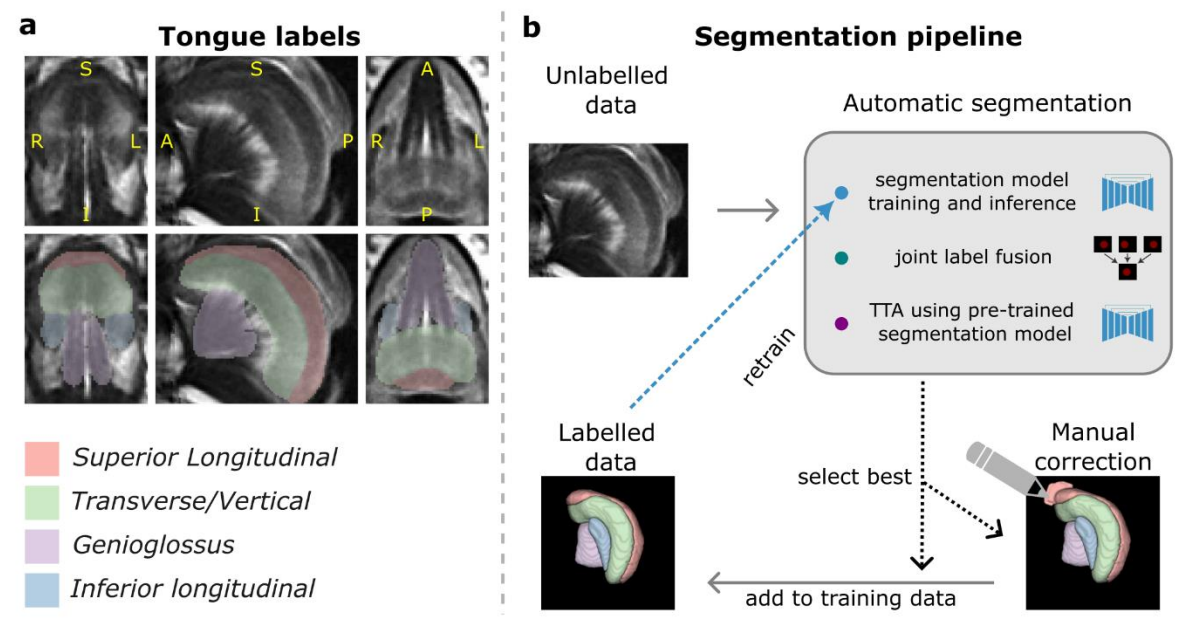


Figure 1. Tongue segmentation method. **a:** Atlas generated with Joint Label Fusion showing the labelled muscles and overlaid on the BeLong dataset template using T2w data. **b:** The segmentation pipeline: we gathered unlabelled data from each study and applied one of the three semi- or automatic segmentation methods to obtain rough segmentations. We then selected the best outputs by visual inspection, and either manually corrected the labels and/or added these data to our training for further segmentation model retraining. We then repeated this procedure until all data were labelled.

Stage 2 – Segmentation model training and inference on new datasets. After annotating and training on data from 12 individuals interactively and iteratively using MONAI Label, we used our own training implementation where data augmentation strategies were adjusted, and a new model was trained and used to predict tongue muscle segmentation on new unlabelled data. These initial predictions were further refined with manual correction (by FLR and TBS) and were leveraged with different approaches to speed up data annotation, including Joint Label Fusion (JLF)[28] and test-time adaptation (TTA)[29,30]. We performed a few iterations of model training and prediction on unlabelled data whenever we had more refined segmentations available. This involved manually inspecting segmentations generated across all strategies and selecting the ones to be added to the training set in the next model training iteration (Figure 1b).

Joint Label Fusion: JLF[28] is a multi-atlas segmentation method that allows for combining a representative set of manually labelled datasets (or atlases) through data warping and weighted voting (label fusion). In detail, the technique involves non-linearly registering the atlases to the input image, then assigning a segmentation label

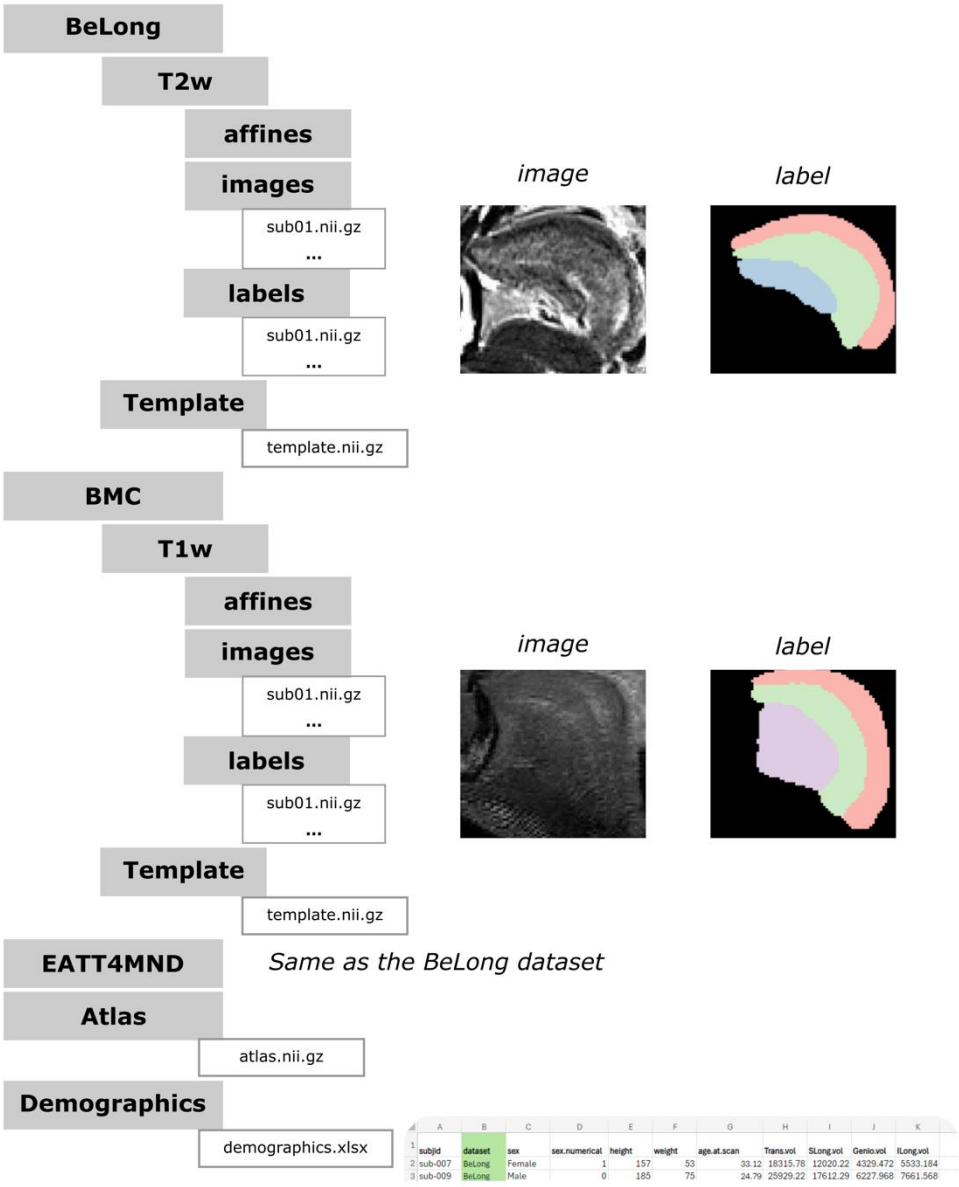
to each voxel based on intensity similarities. Here, we used the 12 manually labelled data from Stage 1, which included both HC and patients' data. We used the ANTs implementation of JLF[31,32].The JLF atlas was used to estimate segmentation in unlabelled data through image registration or as a proxy (suboptimal segmentation) for segmentation model adaptation.

Test-time adaptation: TTA refers to a strategy for improving deep learning model generalisability to new data that follows a different distribution than the original training data[29,30]. For example, a model pre-trained on T2w MRI data from the BeLong study performs poorly on data from different collection sites (EATT and Sydney) or of different contrast weightings (i.e. T1w). Here, we used TTA to adapt our previously pre-trained model using a model adapter (a smaller convolutional neural network prepended to the segmentation model), aiming to improve the predicted segmentations for T1w data and data from different studies. Note that only the parameters of the model adapter were trained while the segmentation model's parameters were fixed to retain the segmentation knowledge learned from the BeLong dataset. Specifically, to guide model adaptation in a supervised fashion, we leverage a proxy (or suboptimal) segmentation that consists of the JLF atlas registered to each individual's space. We performed instance-wise adaptation, i.e., the model adapter's parameters were adapted (or trained) for each individual separately.

Stage 3 – Final manual correction. Each of these approaches were used to grow the pool of annotated and manually corrected data for a new iteration of segmentation model training. We iteratively "bootstrapped" the best segmentation out of the three methodologies (using a pre-trained model, TTA, and JLF) to manually correct the segmentation if required (Figure 1b). The new labelled data were then used to train a new segmentation model for the following iteration. We repeated this process until all data (HC and patient data) were segmented accurately, i.e., following segmentation landmarks according to our established method (Shaw et al., 2024; *in prep*). Our approach was intentionally multi-modal, as it was necessary to leverage the strengths of each method, given the variety of data (patient and control, different scanners, different MR contrasts). Finally, for the final release of HC data, FLR and TBS inspected, manually corrected, and smoothed all segmentations using a 1mm gaussian smoothing kernel across all labels independently.

Data records

This dataset is deposited in the Open Science Framework (OSF), a free and open platform to support open research. The data can be accessed through this link: https://osf.io/wt9fc/. The files are organized per study, as shown in Figure 2.



	A	B	C	D	E	F	G	H	I	J	K
1	subjid	dataset	sex	sex.numerical	height	weight	age.at.scan	Trans.vol	SLong.vol	Genio.vol	ILong.vol
2	sub-007	BeLong	Female	1	157	53	33.12	18315.78	12020.22	4329.472	5533.184
3	sub-009	BeLong	Male	0	185	75	24.79	25929.22	17612.29	6227.968	7661.568

Figure 2. Summary of the dataset. The dataset is structured into distinct studies. For each study we provide either T1w or T2w images with corresponding segmentations, and a study template that can be used for new imaging data registration. We also provide an atlas generated with Joint Label Fusion and demographic information.

Imaging data: Under each study directory are folders for cropped T1w or T2w data. Within those folders, anatomical images are found in the images folder and corresponding segmentations are found in the labels folder. Note that this data has

been registered to study templates generated as described below. Finally, to protect the identity of the participants, we constrained the field of view of the anatomical data to the mouth by cropping out data from elsewhere. The same transformation was applied to the labels.

Also under each study directory is a folder for a template that can be used for registration of new imaging data, if required. These templates were generated using data from both HC and patients using *antsMultivariateTemplateConstruction2.sh* from ANTs[31]. Parameters were: 3 iterations using an affine template, 8 iterations using a refined SyN template at 0.25 gradient step, and 6 more iterations using refined SyN template at 0.15 gradient step.

JLF atlas: We provide a JLF atlas generated with manually corrected segmentation from the BMC study, which can be found under the Atlas folder. The JLF atlas can be used to estimate segmentation in new unlabelled data through image registration, for example, using ANTs[31] or the manual registration tool from ITK-SNAP[33]. Note that the provided atlas was generated with the final segmentations from the BMC study, i.e., it is not the same as the one described in the “Semi-automated tongue segmentation” section. The provided atlas was generated as described previously and with minimal manual correction of small mis-segmentations.

Demographic information: This dataset also includes demographic information (age, weight, height, and sex) and is available under the Demographics folder.

Technical validation

The quality of anatomical data has been evaluated by estimating signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR). SNR was calculated by first defining a noise region outside the brain and skull and dividing the mean intensity within the whole tongue musculature by the standard deviation of the noise region. CNR was calculated by subtracting the mean intensity value of the first tissue class (Transverse/Vertical) by the second tissue class (Superior Longitudinal), then dividing the result by the standard deviation of the noise region. All data processing was conducted using Nibabel[34] in Python. Both measures are available in the demographic information spreadsheet. We performed a 3x1 ANOVA to test differences in SNR and CNR (independently) across datasets. A highly significant difference in the SNR was observed across datasets, $F(2, 43) = 8.979$, $p < 0.001$. Two-sample t-test revealed a

significant difference in SNR between the EATT and Sydney datasets, with EATT showing higher SNR (64.36 vs. 23.08, $p < 0.01$). As for CNR, a near-significant difference across datasets was observed, $F(2, 43) = 2.904$, $p = 0.0656$, indicating a potential variation in contrast-to-noise ratio between datasets (mean CNR BeLong: 0.82, Sydney: 1.80, EATT: 4.23). We followed these with pairwise t-tests, which revealed a non-significant trend towards higher CNR in the EATT dataset compared to Sydney ($p = 0.054$). No follow up tests were conducted for BeLong due to sample size (three with successful calculation).

We also performed ANOVAs to evaluate differences in age and muscle volumes across the three datasets. We found a significant difference in age of participants across dataset, $F(2, 43) = 3.449$, $p < 0.05$. The post-hoc pairwise t-tests (with Bonferroni adjustment) revealed no significant differences between the specific datasets, though the small sample size and high variability in BeLong is likely driving this result. Finally, we did not find significant differences in muscle volume across datasets (Figure 3a). With aggregated data from all datasets (Figure 3b), we determined the correlation (Pearson's r) between pairs of muscle's volumes (Figure 3c), revealing different muscle volumes are generally well correlated with the exception of the genioglossus being less correlated with inferior longitudinal and transverse/vertical muscles.

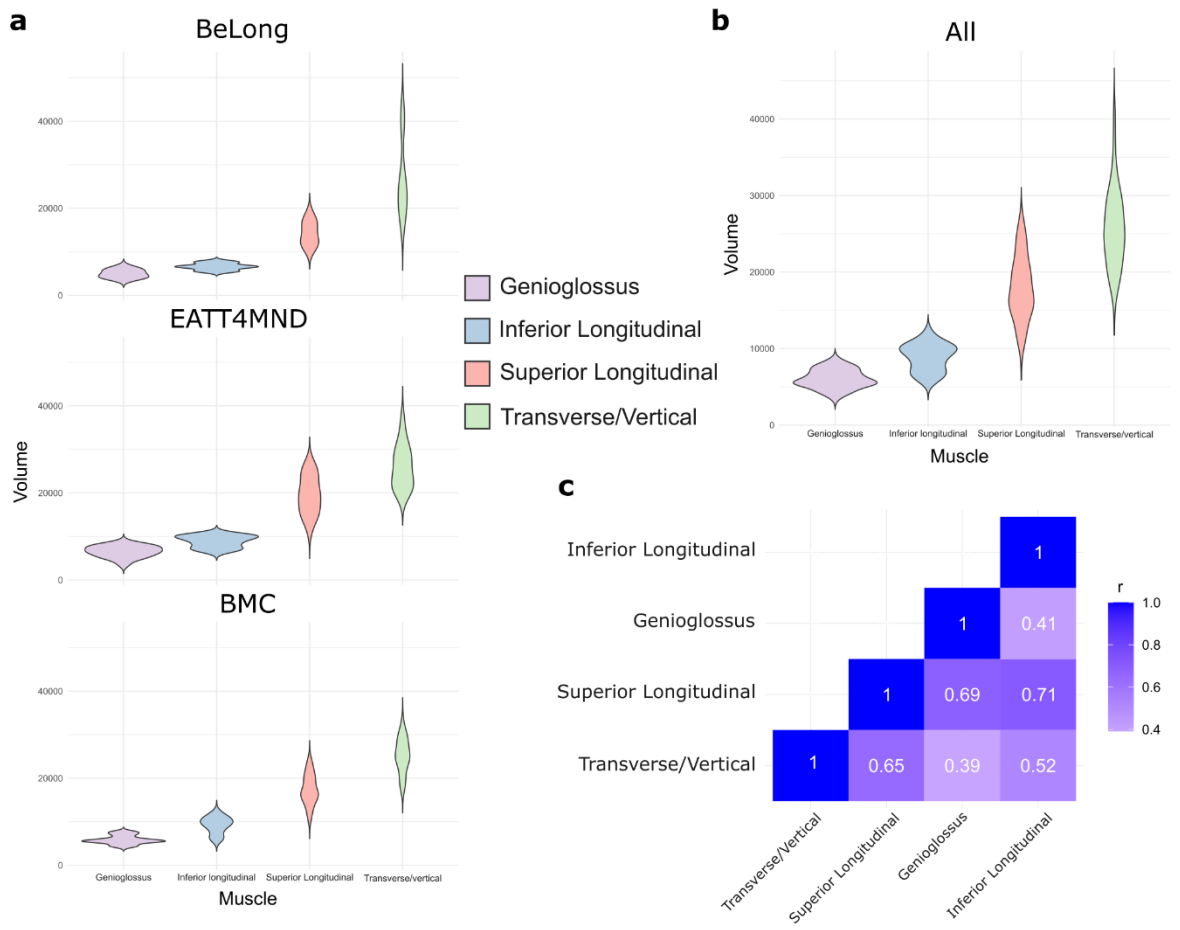


Figure 3. Muscle volume distribution and correlation. a, Distributions of tongue muscles' volumes across datasets. **b,** Distributions of tongue muscles' volumes across all aggregated data. **c,** Pair-wise correlation among inferior longitudinal, genioglossus, superior longitudinal, and transverse/vertical muscles volumes using all data.

Code availability

All accompanying Python, R, and Bash source code is available on GitHub (https://github.com/thomshaw92/TongueSegMND).

Acknowledgement

The authors acknowledge funding by a Motor Neurone Disease Research Australia (MNDRA) Postdoctoral Research Fellowship (PDF2112), NHMRC Ideas grant APP202987, FightMND Collaborative Initiative Grant, Lenity Australia, and an ARC Linkage grant (LP200301393). The data collection for EATT4MND was supported by funding from the Wesley Medical Research grant (#2017-07) and the University of Queensland, Faculty of Medicine. STN acknowledges funding from the Scott Sullivan

MND Research Fellowship (MND and Me Foundation, RBWH, and the University of Queensland).

We thank all participants for their contributions. We also thank the radiology and professional staff at the Herston Imaging Research Facility, National Imaging Facility, UQ Centre for Clinical Research, and the UQ Research Computing Centre. We also thank Dr Diana Lucia for their support in collecting data. We acknowledge support from Aiman Al Najjar, Nicole Atcheson, Sarah Daniel, and the Centre for Advanced Imaging.

Author contribution

Fernanda L. Ribeiro: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – Original Draft, Review & Editing, Supervision, Project administration, Visualization. **Xiangyun Zhu:** Methodology, Validation, Data curation, Writing – Review & Editing. **Xincheng Ye:** Methodology, Data curation, Writing – Review & Editing. **Sicong Tu:** Methodology, Resources, Data curation, Writing – Review & Editing, Funding acquisition. **Shyuan Ngo:** Resources, Writing – Review & Editing, Funding acquisition. **Robert Henderson:** Resources, Writing – Review & Editing, Funding acquisition. **Frederik Steyn:** Resources, Writing – Review & Editing, Funding acquisition. **Matthew Kiernan:** Funding acquisition. **Markus Barth:** Resources, Supervision, Writing – Review & Editing, Funding acquisition. **Steffen Bollmann:** Methodology, Writing – Review & Editing, Funding acquisition. **Thomas B. Shaw:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – Original Draft, Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

Competing interests

The authors have nothing to declare.

References

1. Chen, J. Food oral processing—A review. *Food Hydrocolloids* **23**, 1–25 (2009).

2. Doyle, M. E., Premathilake, H. U., Yao, Q., Mazucanti, C. H. & Egan, J. M. Physiology of the tongue with emphasis on taste transduction. *Physiological Reviews* **103**, 1193–1246 (2023).
3. Ekström, A. G. & Edlund, J. Evolution of the human tongue and emergence of speech biomechanics. *Front. Psychol.* **14**, 1150778 (2023).
4. Okada, A., Honma, M., Nomura, S. & Yamada, Y. Oral behavior from food intake until terminal swallow. *Physiology & Behavior* **90**, 172–179 (2007).
5. Lemon, C. H. Tasting temperature: neural and behavioral responses to thermal stimulation of oral mucosa. *Current Opinion in Physiology* **20**, 16–22 (2021).
6. Manrique, S. & Zald, D. Individual differences in oral thermosensation. *Physiology & Behavior* **88**, 417–424 (2006).
7. Cheng, S., Butler, J. E., Gandevia, S. C. & Bilston, L. E. Movement of the tongue during normal breathing in awake healthy humans. *The Journal of Physiology* **586**, 4283–4294 (2008).
8. Hiiemae, K. M. & Palmer, J. B. Tongue movement in feeding and speech. *Critical Reviews in Oral Biology & Medicine* **14**, 413–429 (2003).
9. Northall, A. *et al.* An Automated Tongue Tracker for Quantifying Bulbar Function in ALS. *Front. Neurol.* **13**, 838191 (2022).
10. Hensiek, N. *et al.* Sonographic and 3T-MRI-based evaluation of the tongue in ALS. *NeuroImage: Clinical* **26**, 102233 (2020).
11. Potter, N. L., Nievergelt, Y. & VanDam, M. Tongue Strength in Children With and Without Speech Sound Disorders. *Am J Speech Lang Pathol* **28**, 612–622 (2019).
12. Kim, A. M. *et al.* Tongue Fat and its Relationship to Obstructive Sleep Apnea. *Sleep* **37**, 1639–1648 (2014).

13. Fox, M. D. & Cohen, A. B. “Bright tongue sign” in ALS. *Neurology* **79**, 1520–1520 (2012).

14. Saxena, S., Tiwari, S., Khera, P. & Midha, N. Bright Tongue Sign in Amyotrophic Lateral Sclerosis. *Neurol India* **70**, 824 (2022).

15. Vernikouskaya, I., Müller, H.-P., Ludolph, A. C., Kassubek, J. & Rasche, V. AI-assisted automatic MRI-based tongue volume evaluation in motor neuron disease (MND). *Int J CARS* **19**, 1579–1587 (2024).

16. Abd-El-Malek, S. Observations on the morphology of the human tongue. *J Anat* **73 (Pt 2)**, (1939).

17. Mu, L. & Sanders, I. Human tongue neuroanatomy: Nerve supply and motor endplates. *Clinical Anatomy* **23**, 777–791 (2010).

18. Kayalioglu, M., Shcherbatyy, V., Seifi, A. & Liu, Z.-J. Roles of intrinsic and extrinsic tongue muscles in feeding: Electromyographic study in pigs. *Archives of Oral Biology* **52**, 786–796 (2007).

19. Stone, M. *et al.* Structure and variability in human tongue muscle anatomy. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **6**, 499–507 (2018).

20. Iltis, P. W. *et al.* High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. *Quantitative Imaging in Medicine and Surgery* **5**, (2015).

21. Narayanan, S., Nayak, K., Lee, S., Sethy, A. & Byrd, D. An approach to real-time magnetic resonance imaging for speech production. *J. Acoust. Soc. Am.* **115**, 1771–1776 (2004).

22. Kühnlein, P. *et al.* Diagnosis and treatment of bulbar symptoms in amyotrophic lateral sclerosis. *Nat Rev Neurol* **4**, 366–374 (2008).

23. Chang, J. *et al.* Lower hypothalamic volume with lower body mass index is associated with shorter survival in patients with amyotrophic lateral sclerosis. *Euro J of Neurology* **30**, 57–68 (2023).

24. Diaz-Pinto, A. *et al.* MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images. Preprint at http://arxiv.org/abs/2203.12362 (2022).

25. Kikinis, R., Pieper, S. D. & Vosburgh, K. G. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. in *Intraoperative Imaging and Image-Guided Therapy* (ed. Jolesz, F. A.) 277–289 (Springer New York, New York, NY, 2014). doi:10.1007/978-1-4614-7657-3_19.

26. Renton, A. I. *et al.* Neurodesk: an accessible, flexible and portable data analysis environment for reproducible neuroimaging. *Nat Methods* (2024) doi:10.1038/s41592-023-02145-x.

27. Wang, G. *et al.* Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019).

28. Hongzhi Wang *et al.* Multi-Atlas Segmentation with Joint Label Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 611–623 (2013).

29. Guo, J., Zhang, W., Sinclair, M., Rueckert, D. & Chen, C. Pay Attention to the Atlas: Atlas-Guided Test-Time Adaptation Method for Robust 3D Medical Image Segmentation. Preprint at http://arxiv.org/abs/2307.00676 (2023).

30. Karani, N., Erdil, E., Chaitanya, K. & Konukoglu, E. Test-Time Adaptable Neural Networks for Robust Medical Image Segmentation. *Medical Image Analysis* **68**, 101907 (2021).

31. Avants, B., Tustison, N. J. & Song, G. Advanced Normalization Tools: V1.0. *The Insight Journal* (2009) doi:10.54294/uvnhin.

32. Avants, B., Epstein, C., Grossman, M. & Gee, J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis* **12**, 26–41 (2008).

33. Yushkevich, P. A. *et al.* User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage* **31**, 1116–1128 (2006).

34. Brett, M. *et al.* nipy/nibabel: 5.2.1. Zenodo https://doi.org/10.5281/ZENODO.10714563 (2024).