Archival Report

Biological Psychiatry: GOS

Segmentation and Volume Estimation of the Habenula Using Deep Learning in Patients With Depression

Yusuke Kyuragi, Naoya Oishi, Momoko Hatakoshi, Jinichi Hirano, Takamasa Noda, Yujiro Yoshihara, Yuri Ito, Hiroyuki Igarashi, Jun Miyata, Kento Takahashi, Kei Kamiya, Junya Matsumoto, Tomohisa Okada, Yasutaka Fushimi, Kazuyuki Nakagome, Masaru Mimura, Toshiya Murai, and Taro Suwa

ABSTRACT

BACKGROUND: The habenula is involved in the pathophysiology of depression. However, its small structure limits the accuracy of segmentation methods, and the findings regarding its volume have been inconsistent. This study aimed to create a highly accurate habenula segmentation model using deep learning, test its generalizability to clinical magnetic resonance imaging, and examine differences between healthy participants and patients with depression.

METHODS: This multicenter study included 382 participants (patients with depression: N = 234, women 47.0%; healthy participants: N = 148, women 37.8%). A 3-dimensional residual U-Net was used to create a habenula segmentation model on 3T magnetic resonance images. The reproducibility and generalizability of the predictive model were tested on various validation cohorts. Thereafter, differences between the habenula volume of healthy participants and that of patients with depression were examined.

RESULTS: A Dice coefficient of 86.6% was achieved in the derivation cohort. The test-retest dataset showed a mean absolute percentage error of 6.66, indicating sufficiently high reproducibility. A Dice coefficient of >80% was achieved for datasets with different imaging conditions, such as magnetic field strengths, spatial resolutions, and imaging sequences, by adjusting the threshold. A significant negative correlation with age was observed in the general population, and this correlation was more pronounced in patients with depression ($p < 10^{-7}$, $r = -0.59$). Habenula volume decreased with depression severity in women even when the effects of age and scanner were excluded ($p = .019$, $\eta^2 = 0.099$).

CONCLUSIONS: Habenula volume could be a pathophysiologically relevant factor and diagnostic and therapeutic marker for depression, particularly in women.

https://doi.org/10.1016/j.bpsgos.2024.100314

The morphology of the habenula, a small cerebral structure comprising the epithalamus and pineal gland, is conserved across species (1). With extensive input from the frontal lobe and the limbic system as well as output to the monoaminergic nuclei in the brainstem, the habenula is involved in regulating aversion, reward, motor output, cognitive functions, sleep, circadian rhythm, and pain functions through the regulation of the dopaminergic and serotonergic nervous systems (2). The habenula has been hypothesized to play an essential role in the pathophysiology of major depression owing to its roles in cognitive, emotional, and autonomic functions. Several studies, mainly studies using animal disease models, have suggested the involvement of the habenula in the pathophysiology of major depression (3–5).

A postmortem study of the human brain reported that habenula volume was reduced in patients with depression (6). Because this structure is visible on magnetic resonance (MR) images of the living human brain, habenula volume can potentially be used as a diagnostic or predictive marker of depression. However, the relatively small size of the habenula poses a challenge when investigating its structure in the living human brain. Previous studies of the habenula have reported inconsistent findings. For example, a recent meta-analysis reported a discrepancy in the left-right differences in habenula volume found in previous studies (7).

Earlier MR imaging (MRI) studies were primarily based on manual segmentation (8–10). However, manual segmentation is a laborious task, and it is associated with disadvantages in terms of validity and reliability owing to inevitable inter- and intrarater biases. Thus, automated or semiautomated segmentation methods have been used in recent studies (11–15). Nevertheless, these automated methods have the limitation of time-consuming or potential missegmentation, especially in images with enlarged ventricular systems because of registration-based algorithms (16). Recently, a deep learning–based method has emerged that can instantly segment the habenula in native space. The application using 7T MR images accomplished highly accurate habenula segmentation (17).

ISSN: 2667-1743

The application using clinically available 3T MR images is expected to be developed.

Previous findings of habenula volumetry in patients with depression have been inconsistent, with MRI studies in humans reporting decreased (8,18), unchanged (11,19), or increased (20) volume of the habenula compared with that of healthy participants. These discrepancies may be attributed to small sample sizes or differences in segmentation methods or study populations. Thus, to obtain more conclusive findings regarding habenula volume in patients with depression, a reliable segmentation model should be applied to an MRI sample of a sufficiently large size, and the whole sample, as well as the sample stratified according to major confounding variables, should be evaluated.

This multicenter, collaborative study aimed to develop a new segmentation model of the habenula in T1-weighted 3T MR images using deep learning to achieve the highest reliability and validity compared with previous studies. Generalization performance along with reproducibility was verified using images with different spatial resolutions, magnetic field strengths (both 3T and 7T), and imaging parameters as an external validation dataset obtained in various MR scanners. The habenula volumes of patients with depression and healthy participants were calculated. Group comparisons and associations with symptoms were examined subsequently. Sex difference was considered important among the variables and was investigated extensively because many studies suggest substantial sex differences in terms of neurobiology, epidemiology, and therapeutic response in patients with depression (21,22).

METHODS AND MATERIALS

Participants and Clinical Assessments

A total of 382 participants (234 patients with depression and 148 healthy participants) were recruited from the Kyoto University Hospital, Keio University Hospital, and the National Center Hospital, National Center of Neurology and Psychiatry. The Mini-International Neuropsychiatric Interview was conducted during the initial visit to confirm the diagnosis of a depressive episode (23). Psychological evaluation was performed using the 17-item Hamilton Depression Rating Scale (HDRS) to assess the severity of the depressive symptoms (24). The severity of depression was categorized based on HDRS scores according to the following criteria: mild depression was a score between 8 and 16; moderate depression was between 17 and 23; severe depression was more than 24 (25). See details in Supplemental Methods.

This study was approved by the Committee on Medical Ethics of the Kyoto University, Keio University, and National Center Hospital, National Center of Neurology and Psychiatry. The study adhered to the principles of the Declaration of Helsinki. All participants provided written informed consent.

MRI Acquisition

Patients with depression and healthy participants underwent structural MRI on the same day as the psychological evaluation. MR images were acquired using five 3T and one 7T scanner, labeled as scanners 1, 2, 3, 4, 5, and 7T, respectively. Scanners 1, 2, and 3 acquired images using 3-dimensional (3D) magnetization-prepared rapid acquisition gradient-echo (MPRAGE) with a spatial resolution of 0.8 × 0.8 × 0.8 mm. Scanners 4 and 5 acquired images using MPRAGE with a spatial resolution of 0.9 × 0.9 × 1 mm. The 7T scanner acquired images using MP2RAGE (research prototype sequence) (26) with a spatial resolution of 0.7 × 0.7 × 0.7 mm. For the test-retest dataset, 46 healthy participants who underwent imaging with scanner 1 underwent a second MRI with the same scanner (6 or 16 weeks after the first imaging). Furthermore, 40 healthy participants who underwent imaging with scanners 1 and 2 underwent scanning with the 7T scanner with different sequences and spatial resolution from the traveling subject dataset. See details in Supplemental Methods.

Overview of the Analysis

An overview of this analysis is presented in Figure 1. A precise predictive model of the habenula structure was constructed using the data acquired from healthy participants using scanner 1. The habenula volume of healthy individuals and patients with depression were evaluated using the model, and the generalization of the model was verified. See details in Supplemental Methods.

Constructing the Habenula Prediction Model

A predictive model constructed using deep learning was trained using the annotated habenula of the healthy participants as supervisory data. The model was constructed using TensorFlow (version 2.4.0), with a 3D U-Net as the basic structure (27) modified by the architecture that we reported (28). A schema of the architecture of the model is provided in Figure 2A, and its details are provided in Figure S1. Generalization of the model was evaluated using 5-fold cross-validation in the healthy participants included in the derivation cohort. Prediction models were created with the same epoch and learning rate decay using all training data, and

Table 1. Demographic and Clinical Characteristics of the Participants

	Participants With Depression, N = 234	Healthy Participants, N = 148	p Value
Age, Years	45.3 (16.3)	42.2 (12.9)	.051[a]
Female Sex	110 (47.0%)	56 (37.8%)	.078[b]
HDRS Score	16.5 (6.8)	0.44 (1.08)	<.001
Mild Depression	134 (57.3%)	–	–
Moderate Depression	63 (26.9%)	–	–
Severe Depression	37 (15.8%)	–	–
Scanner 1	68 (29.1%)	51 (34.5%)	.19[b]
Scanner 2	12 (5.13%)	6 (4.05%)	–
Scanner 3	74 (31.6%)	35 (23.6%)	–
Scanner 4	51 (21.8%)	43 (29.1%)	–
Scanner 5	29 (12.4%)	13 (8.78%)	–
TIV, cm^3	1486.8 (142.8)	1481.2 (143.5)	.71[a]

The values in the cell are mean (SD) or n (%).
HDRS, Hamilton Depression Rating Scale; TIV, total intracranial volume.
[a] t test.
[b] χ^2 test.

automatic habenula segmentation of the test dataset was performed. See details of annotation of the habenula and constructing the segmentation model in Supplemental Methods.

Volume Assessment With the Validation Cohort

The reliability of the prediction model was verified using various validation datasets. The position of the habenula predicted on the images acquired using scanners 1, 2, 3, 4, and 5 was validated using the intersection over union and Dice coefficient on small samples of annotations. The predictive accuracy for the same sequence but different acquisition parameters, including voxel size images acquired using scanners 4 and 5 (Figure 1), was validated at several thresholds. For the test-retest dataset, the habenula volumes calculated from the images obtained during the 2 imaging sessions (Figure 1) were validated for agreement and error. Correlation coefficients and mean absolute percentage errors were calculated, and the distribution of the error was evaluated using Bland-Altman analysis to examine these consistencies. In addition, the consistency of the left-right difference in volume was evaluated using each dataset (see Supplemental Methods).

Habenula Volume in Healthy Individuals

Habenula volume differences between the left-right side and sex were assessed in healthy participants. Only images acquired using scanners 1, 2, and 3 (n = 92), which had identical imaging parameters and spatial resolution (0.8 mm), were used

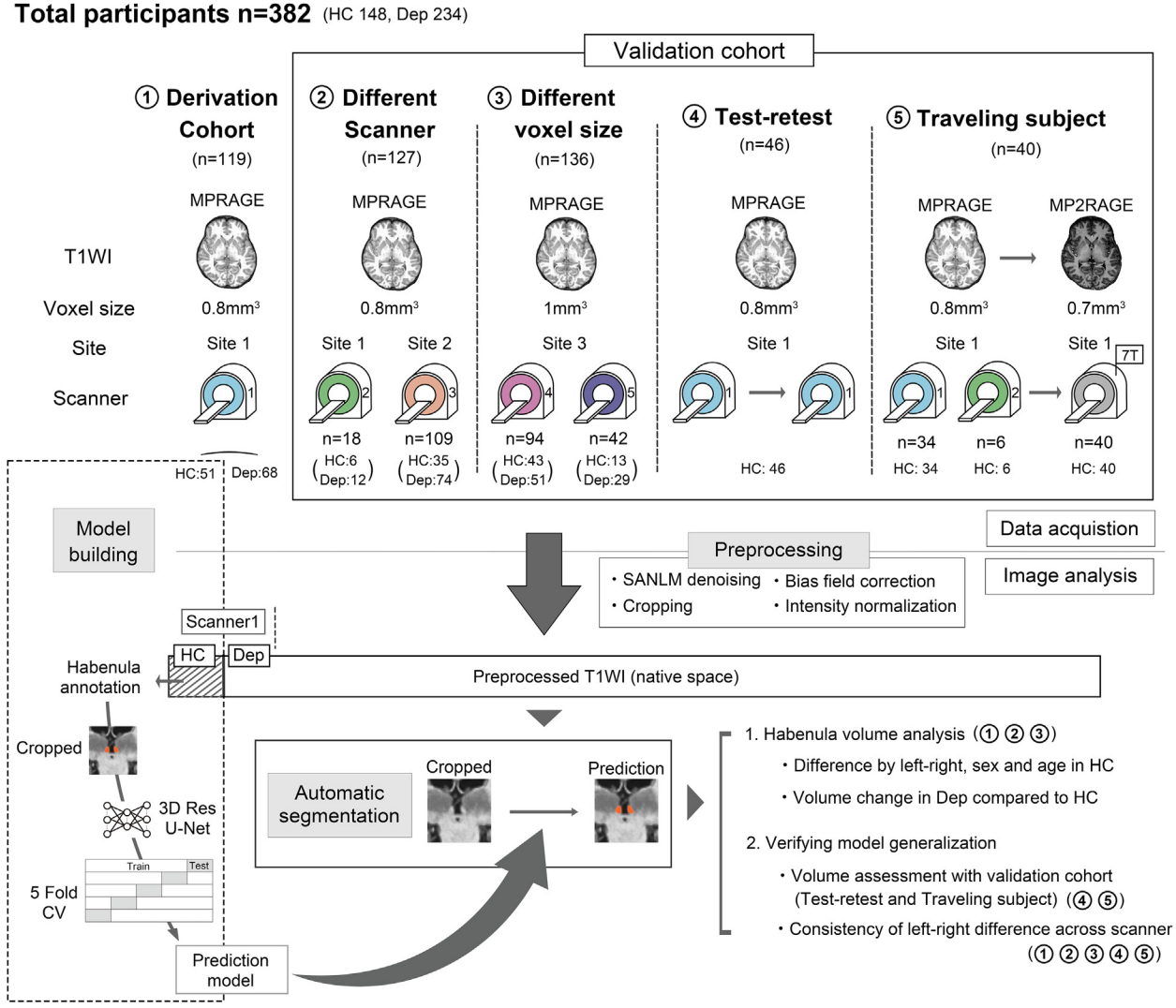


Figure 1. Overview of data acquisition and analysis. Structural images of the brain were obtained using five 3T magnetic resonance imaging scanners at 3 facilities. Using magnetic resonance images from healthy control participants (HC) acquired using scanner 1 as derivation cohort, a prediction model was built using deep learning, and the habenula was predicted on other images from a validation cohort. A habenula volume analysis was conducted, and the generalizability of the model was verified using the predicted habenula volume. 3D Res U-Net, 3-dimensional residual U-Net; CV, cross-validation; Dep, patient with depression; MPRAGE, magnetization-prepared rapid acquisition gradient-echo; SANLM, spatial-adaptive nonlocal means; T1WI, T1-weighted image.

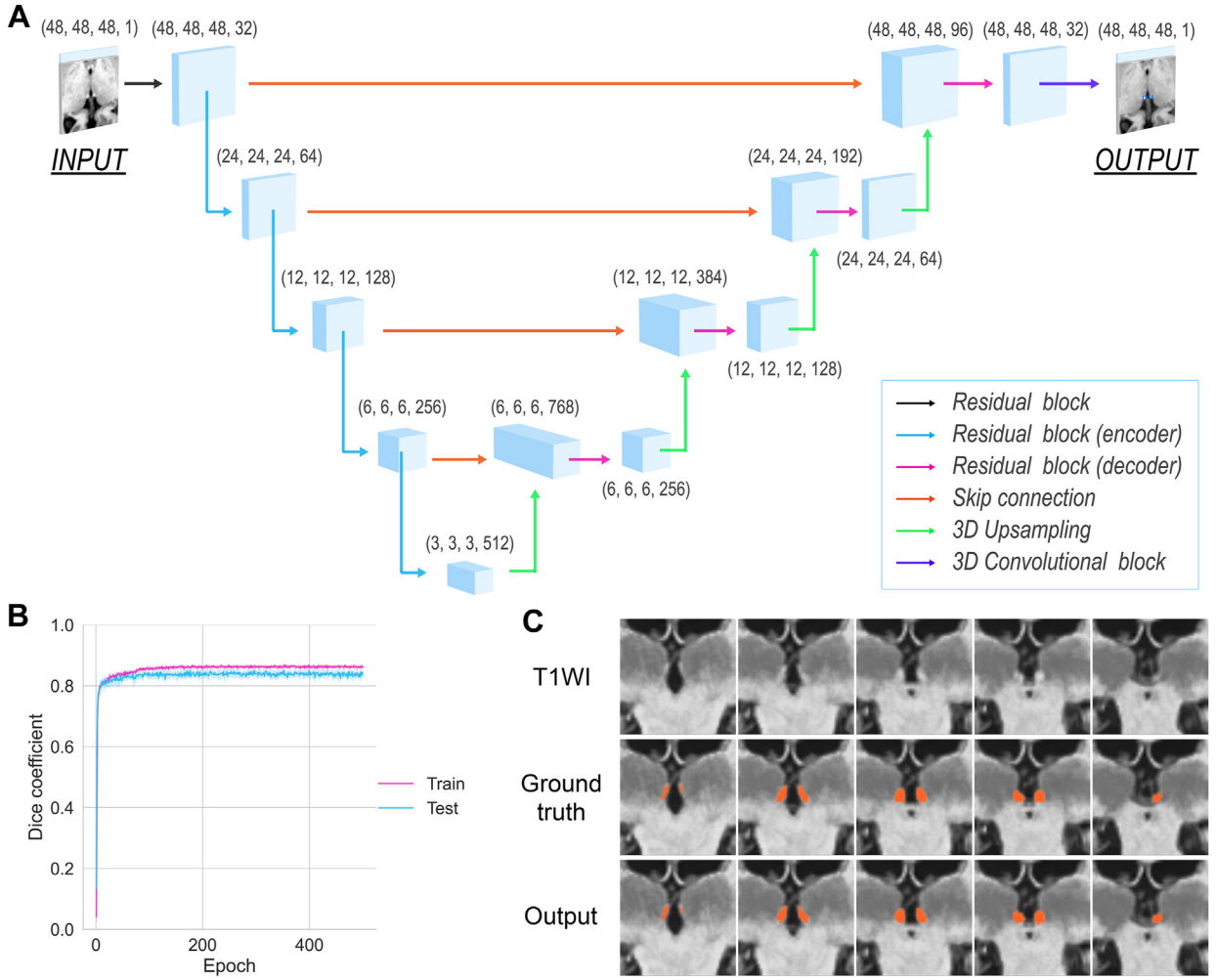


Figure 2. Habenula segmentation using deep learning. **(A)** Three-dimensional (3D) residual U-Net model architecture, **(B)** training and testing Dice coefficient curves of the model, **(C)** example of prediction in 1 participant. T1WI, T1-weighted image.

to minimize the effects of voxel size and partial volume (see Supplemental Methods).

Habenula Analysis in Patients With Depression

Habenula volume in the patients with depression was calculated with images acquired using scanners 1, 2, and 3 (n = 154). The habenula volumes of patients with depression were compared with those of healthy participants of the same sex. A correlation analysis was performed between volume and age in both groups to examine the correlation with age. Partial correlation coefficients between habenula volume and age were calculated separately for the groups and sex with the scanner and total intracranial volume (TIV) as covariates. Volume change was compared in the patients with mild, moderate, and severe depression. A partial correlation coefficient between HDRS scores and volume was calculated using age, sex, scanner, and TIV as covariates for each severity level to investigate the association between depressive symptoms and habenula volume. Additionally, habenula volume differences were examined according to severity and sex. See details in Supplemental Methods.

The same tests were performed as described above for patients with depression, with the habenula volume divided by TIV as a correction for intracranial volume.

Statistical Analysis

Clinical and neuropsychological data were analyzed. Statistical significance was set at a p value $<.05$. All tests were 2-sided. See details in Supplemental Methods.

RESULTS

Clinical and Neuropsychological Data

The behavioral data are presented in Table 1. Age and the female sex ratio were higher in the depression group, although not significantly (p = .051 and 0.078, respectively). More than one-half of the patients had mild depression (mild, 57.3%; moderate, 26.9%; severe, 15.8%), and the mean HDRS score

was 16.5 in the depressed group, which was significantly higher than that in the control group ($p < .001$). No significant differences were observed between healthy participants and the depression group in the proportion of scanner ($p = .19$) and the average of TIV ($p = .71$). The data of 1 patient with severe depression were not used because the HDRS could not be administered owing to catatonia (Table S5).

Precision of the Prediction Model

Figure 2B presents the learning curve of the 5-fold cross-validation in the derivation cohort. The reduction in loss of function reached a plateau early in the learning phase, and no under- or overlearning was observed during the learning process. The average Dice coefficient and standard deviation were 87.5 ± 0.19% and 86.6 ± 0.63% in the training and test phases, respectively. The location and shape of the habenula in the predicted images were confirmed to be visually appropriate, and regions other than the habenula were not predicted. Figure 2C presents a representative predicted image indicating that the segmentation of the habenula is similar to human evaluation.

Volume Assessment With the Validation Cohort

Table 2 presents the validation results of the accuracy of the predictions using images with the same spatial resolution as the data used to create the prediction model. The accuracy of each scanner was almost identical to that of the model. The average Dice coefficient was 86.14 ± 3.52%, and the intersection of union was 75.9 ± 5.35% across scanners (see Supplemental Results).

The threshold for maximum agreement with the supervised data differed from that of the derivation cohort for datasets with different voxel sizes. The accuracy of predictions using datasets with different voxel sizes was slightly lower than that for datasets with the same spatial resolution, indicating that the accuracy was maximum at a threshold of 0.9999 (Dice coefficient: scanner 4, 82.4%; scanner 5, 82.2%) (Table S1).

Figure 3A presents the results for the prediction accuracy in the test-retest dataset. Table S2 presents the predicted volume and accuracy. The correlation coefficient between each volume in the test-retest dataset was relatively high ($r = 0.788$, $p < 10^{-10}$) and was higher for the volume with TIV correction ($r = 0.856$, $p < 10^{-13}$). The mean absolute percentage error was 6.47% and 6.66% for the volume with and without TIV correction, respectively. No obvious systematic error was observed in the Bland-Altman analysis.

The validation results of the prediction accuracy of the traveling subject dataset are presented in Figure 3B. The details of the prediction accuracy with and without TIV correction, respectively, are presented in Tables S3 and S4. The mean absolute percentage error was lowest for the volume without TIV correction at a threshold of 0.999 and the volume with TIV correction at a threshold of 0.9999 (8.42% and 8.06%, respectively). The correlation coefficient between the volume without TIV correction at these thresholds was significantly positive ($r = 0.651$, $p < .0001$) and was higher for the volume with TIV correction ($r = 0.734$, $p < .0001$). The error distribution exhibited no specific trends.

The left-right differences in habenula volume were consistent in each dataset (see Figure S2 and Supplemental Results).

Habenula Volume in Healthy Individuals

Figure 4A presents the results of the habenula volume analysis in the healthy participants. The habenula volume on the left side was significantly higher than that on the right side (left, 36.89 ± 5.42 mm^3; right, 34.83 ± 5.24 mm^3; $p = .00022$, $t_{91} = -3.9$, Cohen's $d = 0.39$). No significant difference in volume was observed between men and women (men, 70.94 ± 10.44 mm^3; women, 72.17 ± 8.71 mm^3, $p = .54$, $t_{90} = 0.60$, $d = 0.13$). Comparison of the left and right sides according to sex revealed that the habenula volume on the left side was significantly larger than that on the right side in women (left, 37.28 ± 4.93 mm^3; right, 34.89 ± 5.16 mm^3; $p = .00075$, $t_{57} = -3.6$, $d = 0.47$); however, no difference was observed between the 2 sides in men (left, 36.22 ± 6.18 mm^3; right, 34.73 ± 5.45 mm^3; $p = .10$, $t_{33} = -1.7$, $d = 0.26$).

Figure 4B presents the analysis of the TIV-corrected habenula volume in healthy participants. The left-right ($p = .00019$, $t_{91} = -3.9$, $d = 0.34$) and sex ($p < 10^{-5}$, $t_{90} = 4.9$, $d = 1.05$) differences were significant, and the ipsilateral sex difference was more pronounced when comparing the left and right sides and sex (left: $p = .00016$, $t_{90} = -3.9$, $d = 0.85$; right: $p < 10^{-5}$, $t_{90} = -4.7$, $d = 1.02$). The left-right differences within the same sex were similar to those without TIV correction (men: $p = .11$, $t_{33} = -1.6$, $d = 0.24$; women: $p = .00069$, $t_{57} = -3.6$, $d = 0.45$).

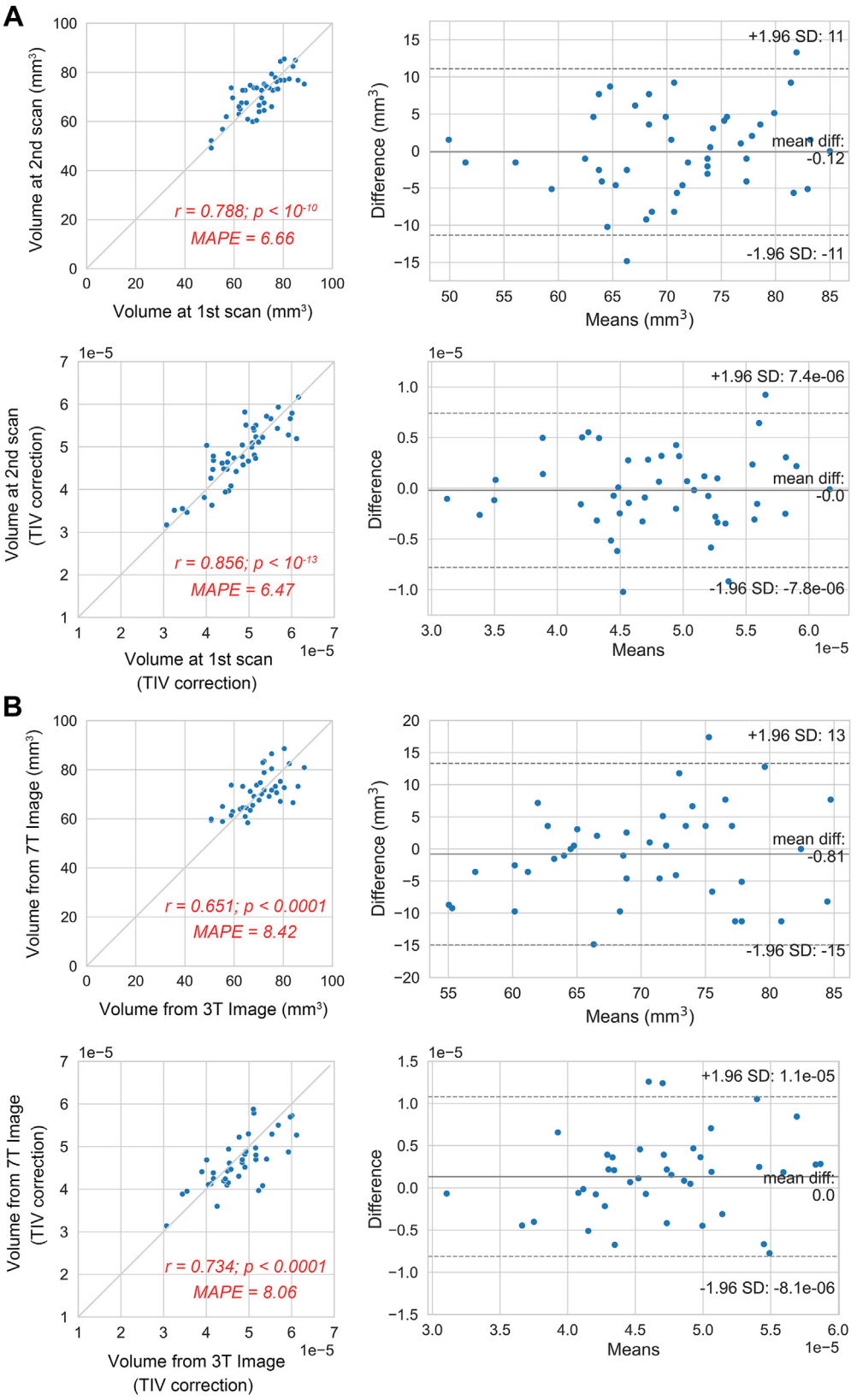
Habenula Volume in Patients With Depression

Figure 5A shows that no significant differences were observed between the groups of the same sex (men: $p = .62$, $t_{104} = 0.50$, $d = 0.10$; women: $p = .10$, $t_{138} = 1.6$, $d = 0.28$). Analysis of covariance revealed no significant difference in the main effect between groups ($p = .20$, $F_{1,240} = 1.6$, $\eta^2 = 0.0068$). Figure 5B shows the negative correlation between habenula volume and

Table 2. Validation Score of the Prediction Model

	Group	Intersection Over Union, Mean (SD) %	Dice Coefficient, Mean (SD) %
Scanner 1 (0.8 mm)	Depression, $n = 10$	81.2% (3.94%)	89.6% (2.39%)
Scanner 2 (0.8 mm)	Depression, $n = 10$	73.3% (7.28%)	84.4% (4.96%)
	Healthy, $n = 6$	76.7% (4.53%)	86.8% (2.93%)
Scanner 3 (0.8 mm)	Depression, $n = 10$	73.6% (5.32%)	84.7% (3.53%)
	Healthy, $n = 10$	74.5% (5.70%)	85.2% (3.79%)

T1-weighted images obtained using scanner 1 were used to construct a prediction model. The images obtained using scanners 2 and 3 had the same parameters as the magnetic resonance imaging sequence obtained using scanner 1 (derivation cohort).


A
Volume at 2nd scan (mm³)
Volume at 1st scan (mm³)
r = 0.788; p < 10-10
MAPE = 6.66
Difference (mm³)
Means (mm³)
+1.96 SD: 11
mean diff: -0.12
-1.96 SD: -11
Volume at 2nd scan (TIV correction)
Volume at 1st scan (TIV correction)
1e−5
r = 0.856; p < 10-13
MAPE = 6.47
Difference
Means
+1.96 SD: 7.4e-06
mean diff: -0.0
-1.96 SD: -7.8e-06
B
Volume from 7T Image (mm³)
Volume from 3T Image (mm³)
r = 0.651; p < 0.0001
MAPE = 8.42
Difference (mm³)
Means (mm³)
+1.96 SD: 13
mean diff: -0.81
-1.96 SD: -15
Volume from 7T Image (TIV correction)
Volume from 3T Image (TIV correction)
r = 0.734; p < 0.0001
MAPE = 8.06
Difference
Means
+1.96 SD: 1.1e-05
mean diff: 0.0
-1.96 SD: -8.1e-06

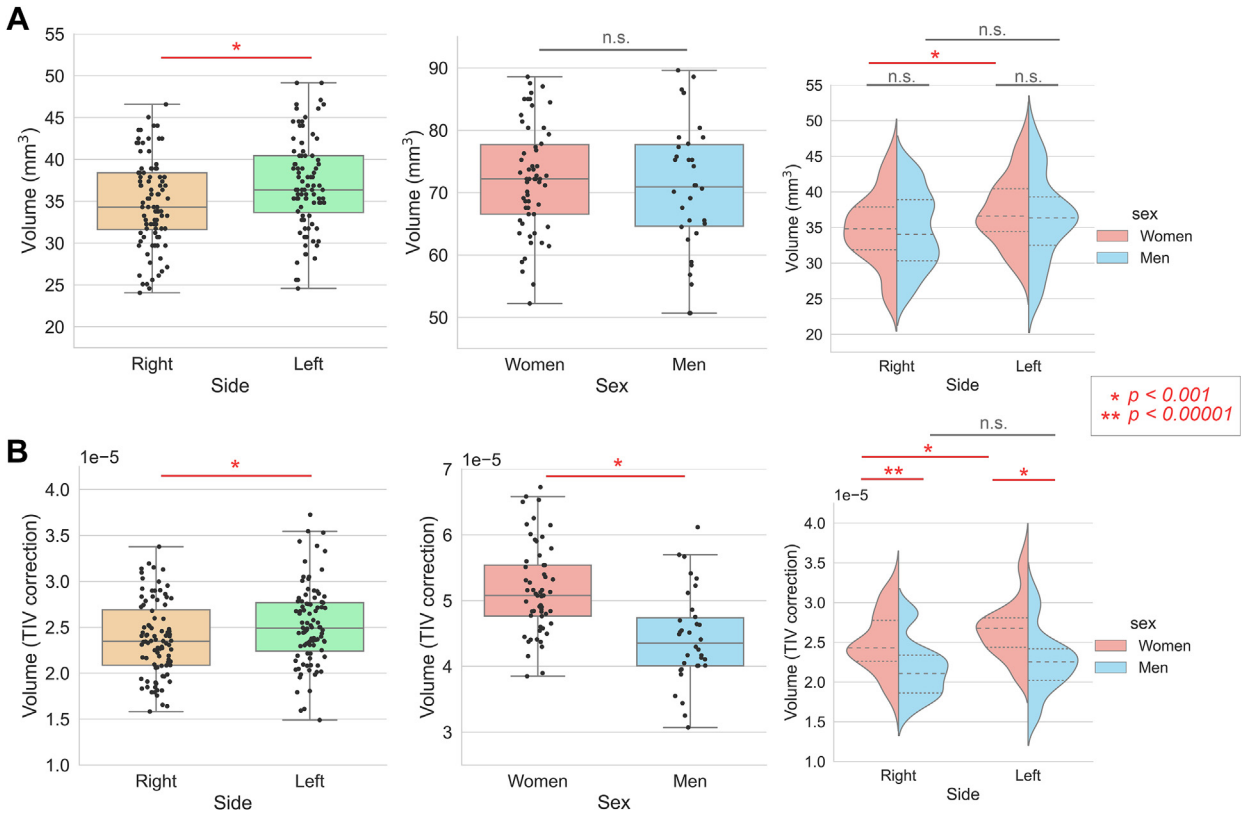


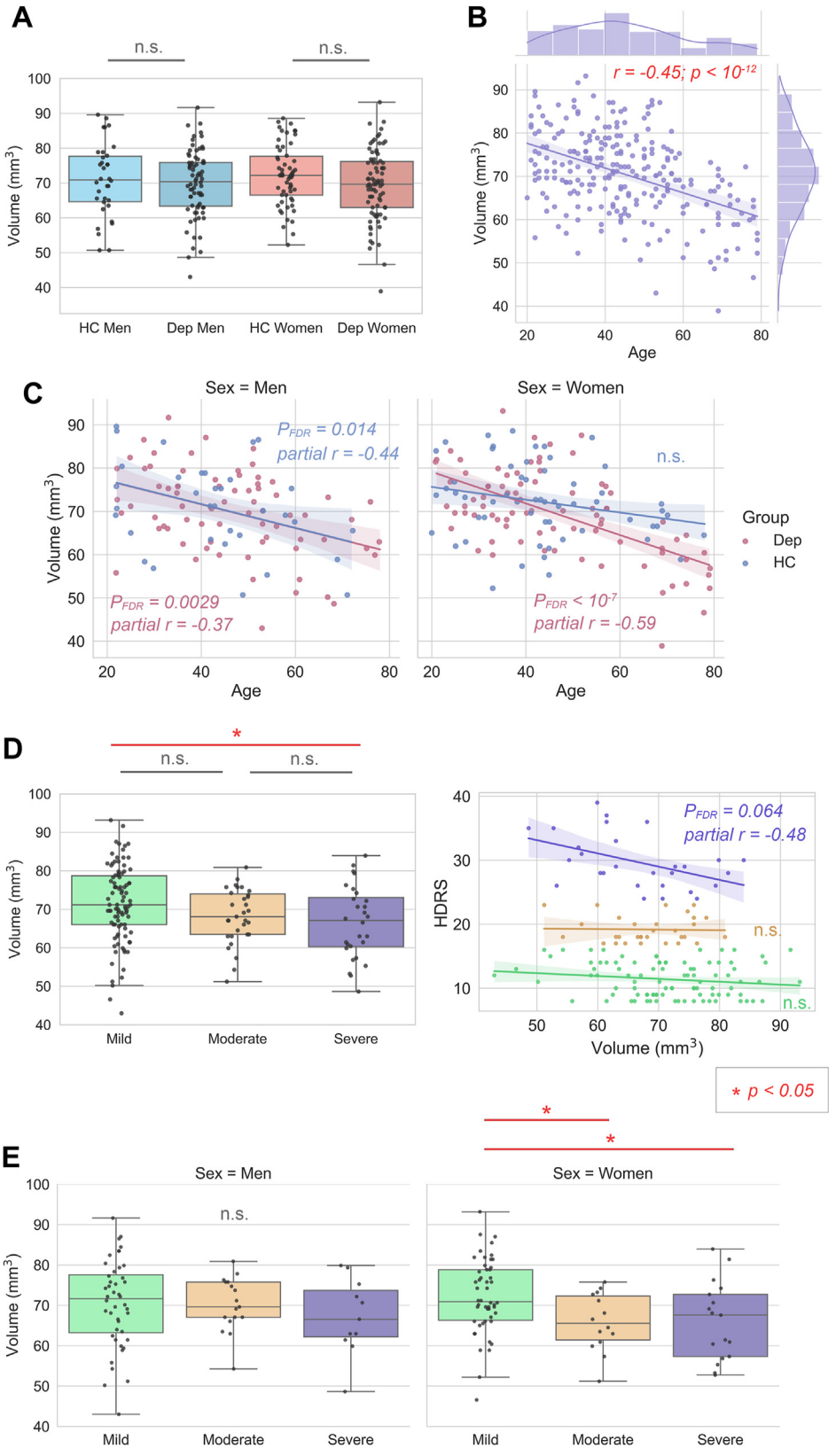
Figure 4. Habenula volume in the healthy control group. **(A)** Habenula volume according to the side (left), sex (middle), and left-right and sex (right). **(B)** Habenula volume divided by the total intracranial volume (TIV) according to the side (left), sex (middle), and left-right and sex (right). n.s., not significant.

age in the combined group of healthy individuals and patients with depression ($r = -0.45$, $p < 10^{-12}$). Figure 5C shows that the regression line between age and volume had a different slope in the groups for women and that age and volume were more strongly correlated in patients with depression than in healthy participants in the partial correlation analysis, especially for women (healthy men: $p = .011$, false discovery rate [FDR]–corrected p [p_{FDR}] = .014, partial $r = -0.44$; healthy women: $p = .071$, $p_{FDR} = .071$, partial $r = -0.24$; men with depression: $p = .0015$, $p_{FDR} = .0029$, partial $r = -0.37$; women with depression: $p < 10^{-8}$, $p_{FDR} < 10^{-7}$, partial $r = -0.59$).

The left panel of Figure 5D shows a significant difference in volume, with the volume decreasing by severity in analysis of variance ($p = .029$, $F_{2,150} = 3.6$, $\eta^2 = 0.046$). Post hoc analysis revealed a significant difference between mild and severe depression (mild to severe: $p = .021$, $t_{120} = 2.3$, $d = 0.50$; mild to moderate: $p = .099$, $t_{123} = 1.66$, $d = 0.34$; moderate to severe: $p = .42$, $t_{57} = 0.80$, $d = 0.21$). Analysis of covariance showed no significant difference in the main effect in terms of severity ($p = .32$, $F_{2,145} = 1.2$, $\eta^2 = 0.016$). Partial correlation analysis between HDRS scores and the habenula volume in each group revealed a correlation, which was not statistically significant after correction for multiple comparisons in patients with severe depression (mild: $p = .17$, $p_{FDR} = .25$, partial $r = -0.15$; moderate: $p = .62$, $p_{FDR} = .62$, partial $r = 0.10$; severe: $p = .021$, $p_{FDR} = .064$, partial $r = -0.48$) (Figure 5D, right panel).

Figure 5E presents the volume according to the severity of illness by sex. A significant difference according to severity was observed in women in an analysis of variance (men: $p = .62$, $F_{2,69} = 0.48$, $\eta^2 = 0.014$; women: $p = .011$, $F_{2,78} = 4.8$, $\eta^2 = 0.11$). Post hoc analysis revealed significant differences between women with mild and moderate depression and between women with mild and severe depression (mild-severe: $p = .019$, $t_{65} = 2.4$, $d = 0.67$; mild-moderate: $p = .019$, $t_{62} = 2.4$, $d = 0.73$; moderate-severe: $p = .96$, $t_{29} = -0.049$, $d = 0.018$). Analysis of covariance with age and scanner as covariates also revealed significant differences between the severity groups in women ($p = .019$, $F_{2,75} = 4.1$, $\eta^2 = 0.099$).

Figure 3. Volume assessment with validation dataset. **(A)** The test-retest dataset. The figures on the left show the correlation with the habenula volume calculated from the first and second scans in the same healthy participants. The figures on the right show the Bland-Altman plots. The predicted habenula volumes were used in the upper figure, and those corrected by total intracranial volume (TIV) in the lower figure. **(B)** Traveling subject dataset. The figures on the left show the correlation with the habenula volume calculated from the images acquired using the 3T and 7T scanners in the same healthy participants. The figures on the right show the Bland-Altman plots. The predicted habenula volumes were used in the upper figure, and those corrected by TIV in the lower figure. MAPE, mean absolute percentage error.


A
n.s.
n.s.
Volume (mm3)
HC Men
Dep Men
HC Women
Dep Women
B
r = -0.45; p < 10-12
Volume (mm3)
Age
C
Sex = Men
Sex = Women
PFDR = 0.014
partial r = -0.44
n.s.
PFDR = 0.0029
partial r = -0.37
PFDR < 10-7
partial r = -0.59
Volume (mm3)
Age
Group
Dep
HC
D
*
n.s.
n.s.
Volume (mm3)
Mild
Moderate
Severe
PFDR = 0.064
partial r = -0.48
n.s.
n.s.
HDRS
Volume (mm3)
* p < 0.05
E
*
*
Sex = Men
Sex = Women
n.s.
Volume (mm3)
Mild
Moderate
Severe
Mild
Moderate
Severe

An additional analysis performed wherein the habenula volume was divided by TIV yielded similar results (Figure S3). Partial correlation analysis between HDRS scores and the habenula volume divided by TIV revealed a significant correlation with severe depression (mild: $p = .21$, $p_{FDR} = .31$, partial $r = -0.13$; moderate: $p = .75$, $p_{FDR} = .75$, partial $r = 0.064$; severe: $p = .0094$, $p_{FDR} = .028$, partial $r = -0.52$).

DISCUSSION

A segmentation methodology for the habenula with sufficiently high accuracy was developed in this study. The deep learning model developed in this study was generalizable to independent datasets with different imaging parameters or magnetic field strengths. Volume estimation of the habenula was found to be highly reproducible across test-retest datasets. This multicenter study revealed sex differences in the volume of the habenula in both healthy participants and patients with depression using a large number of images. Habenula volume was found to be inversely associated with the severity of depression in women but not in men.

A 3D residual U-Net, which is an advanced form of deep learning, was used in this study. U-Net, which is based on a convolutional neural network and has an encoder-decoder structure, is a highly accurate model for segmentation tasks (29). Only one study has applied U-Net to human habenula segmentation and achieved satisfactory accuracy, with a Dice coefficient of 85.2% (17). The current study achieved a similar accuracy with a Dice coefficient of 86.6% using 3T MR images. This may be attributed to the differences in the applied U-Net architecture. A previous study used a 2D attention U-Net with the 4 deepest layers, whereas the method proposed in the current study was based on 3D residual U-Net, as reported previously (28). Compared with 2D U-Net, 3D U-Net can learn more complex parameters and improve the continuity between slices with proper training, thereby improving the segmentation accuracy (27). Moreover, the current study also addressed the problem of gradient vanishing by introducing a residual block into the model (30) and constructing a complex model with the 5 deepest layers. These architectural features may have contributed to the satisfactory accuracy of the results.

Another strength of this study is that validation of the model was performed by applying it to multiple external datasets. Differences in MRI scanners can affect the volumes calculated from brain images (31). The model was satisfactorily applicable to 2 external datasets acquired using a different MRI scanner but with the same parameters (3T MRI, 0.8 × 0.8 × 0.8 mm) (Table 2). The application of the original model to images acquired using different parameters, i.e. the dataset with T1-weighted images with different voxel sizes (3T MRI, 0.9 × 0.9 × 1 mm), as well as those with different magnetic field strengths and different voxel sizes (7T MRI, MP2RAGE, 0.7 × 0.7 × 0.7 mm), resulted in smaller volume calculations and lower accuracy. However, with a minor adjustment to the output threshold, the proposed model was found to be applicable to these datasets (Tables S1, S3, S4).

Several methodologies of semiautomated or automated segmentation of the habenula using machine learning have been developed. The method of voting procedure using multiple atlases was proposed to estimate the volume robustly and reliably (15). Recently, the method using deep learning made the habenula automatically segmented in T1 map in 7T MR images with high accuracy even in the native space (17). It may be favorable to compare these prominent methods with our method in the future. Furthermore, the microstructure of the habenula could be visualized using quantitative susceptibility mapping sequences (32). The model that was created in this study could have better performance, potentially by including these additional parameters as the feature vectors of the model.

The mean habenula volume in healthy individuals was 36.89 mm^3 (left) and 34.83 mm^3 (right) in the current study, which were largely equivalent to the result of a postmortem brain study ($n = 38$ [22 men and 16 women]; age mean ± SD = 49.97 ± 18.54; men: left 35.55 ± 11.5 mm^3, right 33.08 ± 10.7 mm^3; women: left 39.4 ± 12.6 mm^3, right 35.65 ± 11.4 mm^3) (33). In contrast, previous MRI studies have reported estimates that vary from 10 to 30 mm^3 (8,9,12,17). The stria medullaris and fasciculus retroflexus, which were visualized with high signal intensity on T1-weighted images, were carefully excluded during manual annotation of the habenula in the current study, and the habenula was segmented by assessing its morphology in 3 directions. Potential manual tracing bias was avoided across a large number of datasets via the use of fully automated segmentation in this study. These procedures may have led to the estimation being closer to that of the postmortem study. However, because the habenula was defined using the T1-contrast of MRI in the current study, underestimation of the medial portion of the habenula (12,34), partial volume effects, and spillover effects may have been involved. This issue should be addressed by applying the proposed model to the ex vivo imaging of postmortem brains.

The mean habenula volume of patients with depression was smaller than that in healthy individuals. However, this is not a statistically significant difference. This may be due to the expected possible volume difference, if any. The postmortem study showed small average group differences (left, 4.14 mm^3; right, 7.74 mm^3) (6), which is difficult to address by MRI with a spatial resolution of 0.8 mm^3. Another possibility to explain nonsignificant statistical differences is a substantial diversity of habenula volume in healthy participants, which may be associated with the psychological and neurobiological heterogeneity of healthy participants.

A novel finding of the current study is the age-associated decline in habenula volume in patients with depression, which was found to be more prominent in women. Habenula

Figure 5. Habenula volume change in patients with depression (Dep) compared with healthy control participants (HC). Habenula volume **(A)** according to group and sex, **(B)** correlation with age in both groups, **(C)** correlation according to group and sex, and **(D)** difference according to the severity of depression (left). Association of habenula volume with the 17-item Hamilton Depression Rating Scale (HDRS) scores (**D**, right), and **(E)** the difference in habenula volume according to the severity of depression by sex. Considering the brain size effect, total intracranial volume (TIV) was covaried in the model in which analysis of covariance was conducted. n.s., not significant.

volume was shown to decrease with increasing symptom severity, especially in women. The habenula may play a role in the pathophysiology of depression, especially severe depression, as is suggested by a case report that deep brain stimulation of this region dramatically improved severe treatment-resistant depression in a female patient (35). The regions of the brain involved in mood regulation (such as the limbic system and prefrontal cortex), which may be important in the pathophysiology of depression, are thought to develop under the influence of sex hormones, mainly during puberty (36,37). Moreover, the lateral habenula expresses high levels of estrogen receptor-1, which suggests that its activity is regulated by sex hormones (38,39). A recent study reported that excitatory input from the lateral hypothalamic area to the lateral habenula by neurons expressing estrogen receptor-1 caused aversion (40). These findings suggest that the lateral habenula may be an important region in the neural basis of sex differences in the stress response. The incidence of mood disorders, including depression, begins to vary according to sex after puberty and is significantly higher in women (41). Thus, the finding that the habenula is smaller in older female patients with more severe depression may be attributed to the stress-associated accumulation of microdamage to this small brain structure across development and aging. Future longitudinal studies should address this issue.

This study has some limitations. First, because this was a cross-sectional study, it is unclear whether the findings were associated with the risk of developing depression or any changes that occurred at the onset of depression. Longitudinal studies must be conducted to clarify these associations. Second, although not statistically significant, the distributions of the clinical characteristic variables, including differences in age and sex between the healthy and depressed groups, and variations in the severity of depression was unbalanced. This may have affected the detection power. Third, the sample size became slightly smaller when the participants were classified into subtypes, such as disease severity or sex. In the future, it will be desirable to validate the results of the current study using a larger dataset. Fourth, it should be noted that our training data showed left-right habenula volume differences (Supplemental Results). This might have affected our findings of habenula volume asymmetry, although the data augmentation of the flip was processed. Thus, our results of habenula volume asymmetry should be further reconfirmed in independent samples in future studies.

Conclusions

In conclusion, this study developed a deep learning–based habenula segmentation model and confirmed its generalizability and reproducibility. Habenula volume was calculated with higher prediction accuracy by applying the proposed model to a dataset obtained via a multicenter collaborative study. Moreover, the variations with sex in healthy and depressed groups were clarified. The proposed model can be applied to create regions of interest for resting-state functional and diffusion MRI in the future, which will enable the evaluation of connectivity patterns in many cases and further advance our understanding of the pathophysiology of depression.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by KAKENHI Grant-in-Aid for Scientific Research C (Grant No. 21K07593 [to NO]) and KAKENHI Grant-in-Aid for Scientific Research B (Grant No. 21H02849 [to TM]) from the Japan Society for the Promotion of Science. Part of this study was supported by a Grant-in-Aid for the Strategic International Brain Science Research Promotion Program (Brain/MINDS Beyond) (Grant No. JP19dm0307102 [to TM]) from the Japan Agency for Medical Research and Development. The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

We thank all the participants in this study and all research collaborators, including Makiko Nakamoto, Sayuri Murakami, and Manami Tanaka (a clinical psychologist working at Kyoto University Hospital) for their invaluable work in data collection and analysis. We thank Dr. Tobias Kober (Siemens Healthineers International AG, Switzerland) for providing the MP2RAGE research prototype sequence used for the 7T scan. This study was conducted using the MRI scanner and related facilities of Institute for the Future of Human Society, Kyoto University.

TO received a research grant from Siemens Healthcare KK, Japan. All other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry, Graduate School of Medicine, Kyoto University, Kyoto, Japan (YK, NO, MH, YY, YI, HI, JMi, KT, TM, TS); Department of Neuropsychiatry, Keio University School of Medicine, Tokyo, Japan (JH, KK, MM); Department of Psychiatry, National Center Hospital, National Center of Neurology and Psychiatry, Tokyo, Japan (TN); Department of Psychiatry, Aichi Medical University, Aichi, Japan (JMi); Department of Pathology of Mental Diseases, National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan (JMa); Human Brain Research Center, Graduate School of Medicine, Kyoto University, Kyoto, Japan (TO); Department of Diagnostic Imaging and Nuclear Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan (YF); and National Center of Neurology and Psychiatry, Tokyo, Japan (KN).

Address correspondence to Naoya Oishi, M.D., Ph.D., at noishi@kuhp.kyoto-u.ac.jp, or Taro Suwa, M.D., Ph.D., at tarosuwa@kuhp.kyoto-u.ac.jp.

Received Dec 15, 2023; revised Mar 6, 2024; accepted Mar 27, 2024.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.bpsgos.2024.100314.

REFERENCES

1. Namboodiri VM, Rodriguez-Romaguera J, Stuber GD (2016): The habenula. Curr Biol 26:R873–R877.
2. Hu H, Cui Y, Yang Y (2020): Circuits and functions of the lateral habenula in health and in disease. Nat Rev Neurosci 21:277–295.
3. Li B, Piriz J, Mirrione M, Chung C, Proulx CD, Schulz D, *et al.* (2011): Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. Nature 470:535–539.
4. Li K, Zhou T, Liao L, Yang Z, Wong C, Henn F, *et al.* (2013): βCaMKII in lateral habenula mediates core symptoms of depression. Science 341:1016–1020.
5. Yang Y, Cui Y, Sang K, Dong Y, Ni Z, Ma S, Hu H (2018): Ketamine blocks bursting in the lateral habenula to rapidly relieve depression. Nature 554:317–322.
6. Ranft K, Dobrowolny H, Krell D, Bielau H, Bogerts B, Bernstein HG (2010): Evidence for structural abnormalities of the human habenular complex in affective disorders but not in schizophrenia. Psychol Med 40:557–567.
7. Abuduaini Y, Pu Y, Thompson PM, Kong XZ (2023): Significant heterogeneity in structural asymmetry of the habenula in the human brain: A systematic review and meta-analysis. Hum Brain Mapp 44:4165–4182.
8. Savitz JB, Nugent AC, Bogers W, Roiser JP, Bain EE, Neumeister A, *et al.* (2011): Habenula volume in bipolar disorder and major depressive disorder: A high-resolution magnetic resonance imaging study. Biol Psychiatry 69:336–343.

9. Lawson RP, Drevets WC, Roiser JP (2013): Defining the habenula in human neuroimaging studies. Neuroimage 64:722–727.
10. Bocchetta M, Gordon E, Marshall CR, Slattery CF, Cardoso MJ, Cash DM, *et al.* (2016): The habenula: An under-recognised area of importance in frontotemporal dementia? J Neurol Neurosurg Psychiatry 87:910–912.
11. Schmidt FM, Schindler S, Adamidis M, Strauß M, Tränkner A, Trampel R, *et al.* (2017): Habenula volume increases with disease severity in unmedicated major depressive disorder as revealed by 7T MRI. Eur Arch Psychiatry Clin Neurosci 267:107–115.
12. Kim JW, Naidich TP, Ely BA, Yacoub E, De Martino F, Fowkes ME, *et al.* (2016): Human habenula segmentation using myelin content. Neuroimage 130:145–156.
13. Schafer M, Kim JW, Joseph J, Xu J, Frangou S, Doucet GE (2018): Imaging habenula volume in schizophrenia and bipolar disorder. Front Psychiatry 9:456.
14. Su JH, Thomas FT, Kasoff WS, Tourdias T, Choi EY, Rutt BK, Saranathan M (2019): Thalamus Optimized Multi Atlas Segmentation (THOMAS): Fast, fully automated segmentation of thalamic nuclei from structural MRI. Neuroimage 194:272–282.
15. Germann J, Gouveia FV, Martinez RCR, Zanetti MV, de Souza Duran FL, Chaim-Avancini TM, *et al.* (2020): Fully automated habenula segmentation provides robust and reliable volume estimation across large magnetic resonance imaging datasets, suggesting intriguing developmental trajectories in psychiatric disease. Biol Psychiatry Cogn Neurosci Neuroimaging 5:923–929.
16. Shao M, Han S, Carass A, Li X, Blitz AM, Shin J, *et al.* (2019): Brain ventricle parcellation using a deep neural network: Application to patients with ventriculomegaly. NeuroImage Clin 23:101871.
17. Lim SH, Yoon J, Kim YJ, Kang CK, Cho SE, Kim KG, Kang SG (2021): Reproducibility of automated habenula segmentation via deep learning in major depressive disorder and normal controls with 7 Tesla MRI. Sci Rep 11:13445.
18. Cho SE, Park CA, Na KS, Chung CH, Ma HJ, Kang CK, Kang SG (2021): Left–right asymmetric and smaller right habenula volume in major depressive disorder on high-resolution 7-T magnetic resonance imaging. PLoS One 16:e0255459.
19. Luan SX, Zhang L, Wang R, Zhao H, Liu C (2019): A resting-state study of volumetric and functional connectivity of the habenular nucleus in treatment-resistant depression patients. Brain Behav 9:e01229.
20. Liu WH, Valton V, Wang LZ, Zhu YH, Roiser JP (2017): Association between habenula dysfunction and motivational symptoms in unmedicated major depressive disorder. Soc Cogn Affect Neurosci 12:1520–1533.
21. Bangasser DA, Cuarenta A (2021): Sex differences in anxiety and depression: Circuits and mechanisms. Nat Rev Neurosci 22:674–684.
22. Eid RS, Gobinath AR, Galea LAM (2019): Sex differences in depression: Insights from clinical and preclinical studies. Prog Neurobiol 176:86–102.
23. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, *et al.* (1998): The mini-international neuropsychiatric interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 59(suppl 20):22–33.. quiz 34.
24. Williams JB (1988): A structured interview guide for the Hamilton depression rating scale. Arch Gen Psychiatry 45:742–747.
25. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K (2013): Severity classification on the Hamilton depression rating scale. J Affect Disord 150:384–388.
26. Marques JP, Kober T, Krueger G, Van Der Zwaag W, Van De Moortele PF, Gruetter R (2010): MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage 49:1271–1281.
27. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016): 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer, 424–432.
28. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T, *et al.* (2020): Deep learning-derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion. Stroke 51:1484–1492.
29. Ronneberger O, Fischer P, Brox T (2015): U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer, 234–241.
30. He K, Zhang X, Ren S, Sun J (2016): Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. New York: IEEE, 770–778.
31. Takao H, Hayashi N, Ohtomo K (2011): Effect of scanner in longitudinal studies of brain volume changes. J Magn Reson Imaging 34:438–444.
32. Milotta G, Green I, Roiser JP, Callaghan MF (2023): In vivo multi-parameter mapping of the habenula using MRI. Sci Rep 13:3754.
33. Ahumada-Galleguillos P, Lemus CG, Díaz E, Osorio-Reich M, Härtel S, Concha ML (2017): Directional asymmetry in the volume of the human habenula. Brain Struct Funct 222:1087–1092.
34. Kim JW, Naidich TP, Joseph J, Nair D, Glasser MF, O'halloran R, *et al.* (2018): Reproducibility of myelin content-based human habenula segmentation at 3 Tesla. Hum Brain Mapp 39:3058–3071.
35. Sartorius A, Kiening KL, Kirsch P, von Gall CC, Haberkorn U, Unterberg AW, *et al.* (2010): Remission of major depression under deep brain stimulation of the lateral habenula in a therapy-refractory patient. Biol Psychiatry 67:e9–e11.
36. Andersen SL, Teicher MH (2008): Stress, sensitive periods and maturational events in adolescent depression. Trends Neurosci 31:183–191.
37. Juraska JM, Sisk CL, Doncarlos LL (2013): Sexual differentiation of the adolescent rodent brain: Hormonal influences and developmental mechanisms. Horm Behav 64:203–210.
38. Shughrue PJ, Lane MV, Merchenthaler I (1997): Comparative distribution of estrogen receptor-alpha and -beta mRNA in the Rat central nervous system. J Comp Neurol 388:507–525.
39. Zhang L, Hernández VS, Swinny JD, Verma AK, Giesecke T, Emery AC, *et al.* (2018): A GABAergic cell type in the lateral habenula links hypothalamic homeostatic and midbrain motivation circuits with sex steroid signaling. Transl Psychiatry 8:50.
40. Calvigioni D, Fuzik J, Le Merre P, Slashcheva M, Jung F, Ortiz C, *et al.* (2023): Esr1+ hypothalamic-habenula neurons shape aversive states. Nat Neurosci 26:1245–1255.
41. Epperson CN, Kim DR, Bale TL (2014): Estradiol modulation of monoamine metabolism: One possible mechanism underlying sex differences in risk for depression and dementia. JAMA Psychiatry 71:869–870.