

1 **Deep learning algorithms for automatic segmentation of acute cerebral infarcts on**  
2 **diffusion-weighted images: Effects of training data sample size, transfer learning, and**  
3 **data features**

4 Yoon-Gon Noh, MSc,<sup>1</sup> Wi-Sun Ryu, MD, PhD,<sup>1, 2</sup> Dawid Schellingerhout, MBChB,<sup>4</sup>  
5 Jonghyeok Park, MSc,<sup>1</sup> Jinyong Chung, PhD,<sup>2, 3</sup> Sang-Wuk Jeong, MD, PhD,<sup>2, 3</sup> Dong-Seok  
6 Gwak, MD, PhD,<sup>2, 3</sup> Beom Joon Kim, MD, PhD,<sup>5</sup> Joon-Tae Kim, MD, PhD,<sup>6</sup> Keun-Sik Hong,  
7 MD, PhD,<sup>7</sup> Kyung Bok Lee, MD, PhD,<sup>8</sup> Tai Hwan Park, MD, PhD,<sup>9</sup> Sang-Soon Park, MD,<sup>9</sup>  
8 Jong-Moo Park, MD, PhD,<sup>10</sup> Kyusik Kang, MD, PhD,<sup>11</sup> Yong-Jin Cho, MD, PhD,<sup>7</sup> Hong-  
9 Kyun Park, MD, MSc,<sup>7</sup> Byung-Chul Lee, MD, PhD,<sup>12</sup> Kyung-Ho Yu, MD, PhD,<sup>12</sup> Mi Sun Oh,  
10 MD, PhD,<sup>12</sup> Soo Joo Lee, MD, PhD,<sup>13</sup> Jae Guk Kim, MD, MSc,<sup>13</sup> Jae-Kwan Cha, MD,  
11 PhD,<sup>14</sup> Dae-Hyun Kim, MD, PhD,<sup>14</sup> Jun Lee, MD, PhD,<sup>15</sup> Man Seok Park, MD,<sup>6</sup> Dongmin  
12 Kim, PhD,<sup>1</sup> Oh Young Bang, MD, PhD,<sup>16</sup> Eung Yeop Kim, MD, PhD,<sup>17</sup> Chul-Ho Sohn, MD,  
13 PhD,<sup>18</sup> Hosung Kim, PhD,<sup>19</sup> Hee-Joon Bae, MD, PhD,<sup>5</sup> Dong-Eog Kim, MD, PhD<sup>2, 3</sup>

14

15 1. Artificial Intelligence Research Center, JLK Inc., Seoul, South Korea

16 2. National Priority Research Center for Stroke and Department of Neurology, Dongguk  
17 University Ilsan Hospital, Goyang, South Korea

18 3. Bioimaging Data Curation Center, South Korea

19 4. Department of Neuroradiology and Imaging Physics, The University of Texas M.D.  
20 Anderson Cancer Center, Houston, USA

21 5. Department of Neurology, Seoul National University Bundang Hospital, Seongnam, South  
22 Korea

23 6. Department of Neurology, Chonnam National University Hospital, Gwangju, South Korea

24 7. Department of Neurology, Inje University Ilsan Paik Hospital, Goyang, South Korea

25 8. Department of Neurology, Soonchunhyang University Hospital, Seoul, South Korea

26 9. Department of Neurology, Seoul Medical Center, Seoul, South Korea

27 10. Department of Neurology, Uijeongbu Eulji Medical Center, Uijeongbu, South Korea

28 11. Department of Neurology, Nowon Eulji Medical Center, Eulji University School of

29 Medicine, Seoul, South Korea

30 12. Department of Neurology, Hallym University Sacred Heart Hospital, Anyang, South  
31 Korea

32 13. Department of Neurology, Eulji University Hospital, Daejeon, South Korea

33 14. Department of Neurology, Dong-A University Hospital, Busan, South Korea

34 15. Department of Neurology, Yeungnam University Hospital, Daegu, South Korea

35 16. Department of Neurology, Samsung Medical Center, Sungkyunkwan University School  
36 of Medicine, Seoul, South Korea.

37 17. Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of  
38 Medicine, Seoul, South Korea.

39 18. Department of Radiology, College of Medicine, Seoul National University, Seoul, South  
40 Korea

41 19. USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC,  
42 University of Southern California, Los Angeles, California, USA

43

44 **Correspondence to**

45 Wi-Sun Ryu, MD, PhD, Artificial Intelligence Research Center, JLK Inc., 5, Teheran-ro 33-  
46 gil, Gangnam-gu, Seoul, South Korea. E-mail: [wisunryu@jlkgroup.com](mailto:wisunryu@jlkgroup.com)

47 Dong-Eog Kim, MD, PhD, Department of Neurology, Dongguk University Ilsan Hospital, 27,  
48 Dongguk-ro, Ilsandong-gu, Goyang, South Korea. E-mail: [kdongeog@duih.org](mailto:kdongeog@duih.org)

49

50 **Disclosure:** Yoon-Gon Noh, Wi-Sun Ryu, and Jonghyeok Park are employees of JLK Inc. Oh  
51 Young Bang, Hee-Joon Bae, and Dong-Eog Kim are stockholders of JLK inc.

52

53 **Acknowledgements:** The authors appreciate the contributions of all members of the  
54 Comprehensive Registry Collaboration for Stroke in Korea to this study. This study was  
55 supported by the National Priority Research Center Program Grant (NRF-

56 2021R1A6A1A03038865), the Basic Science Research Program Grant (NRF-  
57 2020R1A2C3008295), the Multiminsty Grant for Medical Device Development  
58 (KMDF\_PR\_20200901\_0098), and the Bioimaging Data Curation Center Program Grant  
59 (2022M3H9A2083956) of the National Research Foundation, funded by the Korean  
60 government.

## 61 **Abstract**

62 **Background:** Deep learning-based artificial intelligence techniques have been developed for  
63 automatic segmentation of diffusion-weighted magnetic resonance imaging (DWI) lesions,  
64 but currently mostly using single-site training data with modest sample sizes.

65 **Objective:** To explore the effects of 1) various sample sizes of multi-site vs. single-site  
66 training data, 2) domain adaptation, the utilization of target domain data to overcome the  
67 domain shift problem, where a model that performs well in the source domain proceeds to  
68 perform poorly in the target domain, and 3) data sources and features on the performance and  
69 generalizability of deep learning algorithms for the segmentation of infarct on DW images.

70 **Methods:** In this nationwide multicenter study, 10,820 DWI datasets from 10 hospitals  
71 (Internal dataset) were used for the training-and-validation (Training-and-validation dataset  
72 with six progressively larger subsamples: n=217, 433, 866, 1,732, 4,330, and 8,661 sets,  
73 yielding six algorithms) and internal test (Internal test dataset: 2,159 sets without overlapping  
74 sample) of 3D U-net algorithms for automatic DWI lesion segmentation. In addition, 476  
75 DW images from one of the 10 hospitals (Single-site dataset) were used for training-and-  
76 validation (n=382) and internal test (n=94) of another algorithm. Then, 2,777 DW images  
77 from a different hospital (External dataset) and two ancillary test datasets (I, n=50 from three  
78 different hospitals; II, n=250 from Ischemic Stroke Lesion Segmentation Challenge 2022)  
79 were used for external validation of the seven algorithms, testing each algorithm performance  
80 vs. manual segmentation gold standard using DICE scores as a figure of merit. Additional  
81 tests of the six algorithms were performed after stratification by infarct volume, infarct  
82 location, and stroke onset-to-imaging time. Domain Adaptation was performed to fine-tune  
83 the algorithms with subsamples (50, 100, 200, 500, and 1000) of the 2,777 External dataset,  
84 and its effect was tested using a) 1,777 DW images (from the External dataset, without  
85 overlapping sample) and b) 2,159 DW images from the Internal test dataset.

86 **Results:** Mean age of the 8,661 patients in the Training-and-validation dataset was 67.9 years  
87 (standard deviation 12.9), and 58.9% (n = 4,431) were male. As the subsample size of the  
88 multi-site dataset was increased from 217 to 1,732, algorithm performance increased sharply,  
89 with DSC scores rising from 0.58 to 0.65. When the sample size was further increased to  
90 4,330 and 8,661, DSC increased only slightly (to 0.68 and 0.70, respectively). Similar results  
91 were seen in external tests. Although a deep learning algorithm that was developed using the

92 Single-site dataset achieved DSC of 0.70 (standard deviation 0.23) in internal test, it showed  
93 substantially lower performance in the three external tests, with DSC values of 0.50, 0.51,  
94 and 0.33, respectively (all  $p < 0.001$ ). Stratification of the Internal test dataset and the  
95 External dataset into small ( $< 1.7$  ml;  $n = 994$  and  $1,046$ , respectively), medium ( $1.7-14.0$  ml;  
96  $n = 587$  and  $904$ , respectively), and large ( $> 14.0$ ;  $n = 446$  and  $825$ , respectively) infarct size  
97 groups, showed the best performance (DSCs up to  $\sim 0.8$ ) in the large infarct group, lower (up  
98 to  $\sim 0.7$ ) in the medium infarct group, and the lowest (up to  $\sim 0.6$ ) in the small infarct group.  
99 Deep learning algorithms performed relatively poorly on brainstem infarcts or hyperacute ( $<$   
100  $3h$ ) infarcts. Domain adaptation, the use of a small subsample of external data to re-train the  
101 algorithm, was successful at improving algorithm performance. The algorithm trained with  
102 the 217 DW images from the Internal dataset and fine-tuned with an additional 50 DW  
103 images from the External dataset, had equivalent performance to the algorithm trained using a  
104 four-fold higher number ( $n=866$ ) of DW images using the Internal dataset only (without  
105 domain adaptation).

106 **Conclusion:** This study using the largest DWI data to date demonstrates that: a) multi-site  
107 data with  $\sim 1,000$  DW images are required for developing a reliable infarct segmentation  
108 algorithm, b) domain adaptation could contribute to generalizability of the algorithm, and c)  
109 further investigation is required to improve the performance for segmentation of small or  
110 brainstem infarcts or hyperacute infarcts.

111

## 112 **Introduction**

113 Diffusion weighted imaging (DWI) has been a critical imaging technique for the diagnosis  
114 and treatment of acute ischemic stroke because it is highly sensitive in detecting acute  
115 cerebral infarcts.<sup>1</sup> DWI lesion volume<sup>2</sup> and pattern<sup>3</sup> can predict post-stroke functional  
116 outcomes and future cerebrovascular events. Moreover, DWI can guide acute recanalization  
117 therapy<sup>4,5</sup> by triaging patients based on their infarct volumes.

118 There is a real clinical need for automated segmentation of DW images. Since human  
119 segmentation of the infarct core demands time-consuming clinical expertise, multiple deep  
120 learning-based artificial intelligence techniques have been developed for automatic  
121 segmentation of DWI lesions.<sup>6-9</sup> However, such techniques are critically dependent on the  
122 quantity and quality of the training-and-validation data (training data) used to build the  
123 algorithms, and most studies to date have utilized single-site training data with only modest  
124 sample sizes (Supplementary Table 1). Only a few studies have externally tested their deep  
125 learning algorithms, reporting -as expected that dice similarity coefficients (DSCs) were  
126 much higher for internal data than for external data.<sup>10,11</sup>

127 Large-scale, multi-site training data are needed to avoid the two well-known machine  
128 learning failures: a) the failure of generalization problem that prevents a deep learning model  
129 from learning patterns that generalize to unseen data, and b) the domain shift problem where  
130 a model that performs well in the source domain proceeds to perform poorly in the target  
131 domain.<sup>12</sup> However, collecting extensive imaging data from multiple centers is challenging.  
132 Labeling and annotating data are very labor-intensive processes that require thorough  
133 knowledge of neuroimaging. Specifically, regarding deep learning algorithms for DWI lesion  
134 segmentation, the training data sample size that minimizes the generalization problem and  
135 domain shift problem is not known yet.

136 To overcome the domain shift problem, domain adaptation, which fine-tunes the pre-trained  
137 model using source domain data by adjusting its parameters using additional training data  
138 from the target domain, has been successfully applied in various fields.<sup>13</sup> However, studies  
139 exploring the effect of domain adaptation on the performance of deep learning algorithms for  
140 DWI infarct segmentation have not been reported yet. Clearly, the sample sizes of both initial  
141 training data and of the effects of target domain data both would be important variables to  
142 consider in such a study.

143 In this nationwide multi-center study (Figure 1), 10,820 patients' DW images (collected  
144 consecutively from 10 university hospitals) were used to develop deep learning-based infarct  
145 segmentation algorithms. These algorithms were tested using three external datasets (n =  
146 2,777, 50, and 250). We examined effects of 1) various sample sizes of multi-site vs. single-  
147 site training data, 2) domain adaptation, and 3) data sources and features on algorithm  
148 performance.

149

150

## 151 **Methods**

### 152 **Training cohort**

153 **Multi-site data** This study included brain DW images from the Korean nationwide image-  
154 based stroke database project.<sup>14-16</sup> From May 2011 to February 2014, we consecutively  
155 enrolled 12,013 patients with ischemic stroke or transient ischemic attack who were admitted  
156 to 10 stroke centers within 7 days of symptom onset. We excluded the following patients:  
157 contraindication to MRI (n = 258), poor quality or unavailability of DW images (n = 904),  
158 and MRI registration error (n = 31), leaving 10,820 patients for 'Internal dataset' (Figure 1).  
159 This Internal dataset was further split 80/20 into a 'Training-and-validation dataset' (n=8,661)  
160 and 'Internal test dataset' (n = 2,159).

161 **Single-site data** To investigate segmentation performance of a deep learning model that was  
162 trained using data from a single site, we chose one of the 10 hospitals to prepare 'Single-site  
163 dataset' (Figure 1) with 476 DW images, which is comparable to the amounts of training data  
164 in previous studies.<sup>17, 18</sup>

165

### 166 **External test cohorts**

167 Three datasets (Figure 1) were used for external validation of deep learning algorithms. First,  
168 a consecutive series of 2,777 DW images ('External dataset') were collected from patients  
169 who were admitted with acute ischemic stroke or transient ischemic from a university  
170 hospital during the same period as the training cohort. Second, 'Ancillary test dataset I' was  
171 prepared using DW images of 50 patients with ischemic stroke due to atrial fibrillation from

172 three different university hospitals between 2011 and 2014.<sup>19</sup> Third, ‘Ancillary test dataset II’  
173 (n = 250) were Ischemic Stroke Lesion Segmentation Challenge (ISLES) 2022 data.<sup>20</sup>

174 Institutional Review Boards of all participating centers approved this study. All patients or  
175 their legally authorized representatives provided written informed consent for study  
176 participation.

177

## 178 **DW Images and ischemic lesion segmentation**

179 Brain MR images for Training-and-validation dataset and internal test dataset were obtained  
180 using 1.5 Tesla (n = 6,360) or 3.0 Tesla (n = 2,882) MRI systems; data were missing in 1,259  
181 and 319 patients for Training-and-validation dataset and internal test dataset, respectively  
182 (Supplementary Table 2). DWI protocols were: b-values of 0 and 1,000 s/mm<sup>2</sup>, TR of 2,400–  
183 9,000 ms, TE of 50 – 99 ms, voxel size of 1 × 1 × 3–5 mm<sup>3</sup>, interslice gap of 0–2 mm, and  
184 thickness of 3–7 mm. For the External dataset, the majority of DW images were obtained  
185 using a 1.5 Tesla MRI system (n = 2,724, 98.5%; data were missing in 12 patients). Ischemic  
186 lesions on DW images in the Training-and-validation dataset, Internal test dataset, and  
187 External dataset were segmented by experienced researchers using an in-house software  
188 Image\_QNA under the close guidance by an experienced vascular neurologist, as previously  
189 described.<sup>14, 15</sup> During the semi-automatic segmentation, inter-rater reliability was monitored  
190 as previously described.<sup>21</sup> For the Ancillary test dataset I, an experienced neurologist  
191 manually outlined ischemic lesions. In the Ancillary test dataset II, a hybrid human-algorithm  
192 annotation scheme was applied for lesion segmentation.<sup>20</sup>

193

## 194 **Image preprocessing**

195 To train the infarct segmentation model, brain DW images were preprocessed by (1) skull  
196 stripping using Gaussian blur and Otsu's threshold,<sup>22</sup> (2) N4 correction using the SimpleITK  
197 library, and (3) image signal image normalization as described below.

198 *Skull stripping.* Brain parenchyma has relatively high signal intensities in the DWI compared  
199 with skull, cerebrospinal fluid, and noisy areas. To focus on the brain parenchyma, Otsu  
200 thresholding was used to generate a parenchymal brain mask from the Gaussian blur-  
201 processed image. The brain mask was then superimposed on the original image to remove



202 non-parenchymal structures outside the mask.

203 *N4 correction.* Signal intensity values of MR images are frequently non-uniform because of a  
204 bias field effect. DW images from various participating centers had different levels /  
205 distributions of signal non-uniformity. To reduce the inter-site difference, bias field correction  
206 was performed before model training, which was done using Python version of the N4  
207 correction algorithm in the SimpleITK library. However, because the algorithm was  
208 computationally expensive, the maximum number of corrections was set to be 10 to limit the  
209 execution time for each case.

210 *Image normalization.* DWI signal distribution varies depending on imaging conditions such  
211 as MRI equipment vendor, magnetic field strength, echo time, and repetition time. When the  
212 noise area is removed, the peak point of the signal intensity histogram is primarily occupied  
213 by gray and white matter, with lesion and artifact areas exhibiting a higher signal, resulting in  
214 a bimodal distribution. As a normalization process to make signal intensities of each skull-  
215 stripped DW images distribute within a constant range, all the voxels in each slice was  
216 multiplied by a specific coefficient: a number found to shift the peak value in the signal  
217 intensity histogram to 150, when the peak value was divided by the number. In order to  
218 suppress abnormally high signals, which are typically noticed as artifacts in DWI,<sup>23</sup> the  
219 multiplied values were clipped not to exceed 255.

220

## 221 **Model Development**

222 We employed modified version of 3D U-Net<sup>24</sup> for model training. While the model retained  
223 its U-shaped architecture, the number of CNN layers and the filters for these layers were  
224 modified. In addition, each convolution unit (Conv3D-BatchNormalization-ReLU) was  
225 replaced with pre-activation unit (BatchNormalization-ReLU-Conv3D), which was first  
226 utilized to increase ResNet performance<sup>25</sup> and was expected to be able to boost the  
227 performance of our models.

228 To develop multi-site deep learning models and compare segmentation performances as  
229 training data increased, the Training-and-validation dataset was subsampled by a factor of  
230 2.5/5/10/20/50/100% (217, 433, 866, 1,732, 4,330, and 8,661 DW images, respectively;  
231 Supplementary Fig 1), with an 8:2 training-to-validation set ratio. To develop a single-site  
232 deep learning model, a total of 476 patients' DW images were divided into 382 (for training

233 and validation) and 94 (for internal testing). During deep learning, random augmentation was  
234 performed in real-time to increase the diversity of training datasets and to prevent overfitting:  
235 a slice-wise affine transformation, MRI (bias field) artifact simulation, an axis flip, and a  
236 gamma/contrast change. The implementation code was developed using TorchIO, a medical  
237 imaging library written in Python.<sup>26</sup> Further information is available in Supplementary  
238 Material.

239 In addition to the aforementioned 3D U-Net, we employed vision transformer (Swin  
240 UNETR),<sup>27</sup> another well-known medical image segmentation network, for deep learning  
241 (Supplementary Material).

242

### 243 **Model Evaluation**

244 After training models, segmentation performance was evaluated using the Internal test dataset,  
245 External dataset, and Ancillary test datasets I and II. As a primary evaluation metric, Dice  
246 similarity coefficient (DSC) was calculated as follows:

$$DSC \text{ (Dice similarity coefficient)} = \frac{2|A \cap B|}{|A| + |B|}$$

247 A: manual segmentation (gold standard), B: image segmentation by a deep learning algorithm

248 Additionally, voxel-wise sensitivity and precision were calculated by quantifying the number  
249 of missed lesion voxels or incorrectly predicted positive voxels, as follows:

$$250 \quad \text{Sensitivity} = \frac{TP}{TP+FN}, \text{Precision} = \frac{TP}{TP+FP}$$

251 TP: true positive; FN: false negative; FP: false positive

252

253 We also assessed the performance of infarct segmentation depending on the differences in:

- 254 1- infarct volume, which was categorized as small (< 1.7 mL), medium (1.7 mL – 14  
255 mL), and large (> 14 mL)<sup>11</sup>
- 256 2- imaging acquisition time after symptom onset defined as last-known-well (< 3 hours,  
257 3-24 hours, and > 24 hours)
- 258 3- infarct location (cortex, corona radiata, basal ganglia and internal capsule, thalamus,  
259 midbrain, pons, medulla, cerebellum, and multiple)

- 260 4- MRI vendor
- 261 5- the presence vs. absence of chronic infarct, which was defined as a) 3–15 mm
- 262 ischemic lesions outside the basal ganglia, brainstem, thalamus, internal capsule, or
- 263 cerebral white matter or b) ischemic lesions larger than 15 mm in any areas on fluid-
- 264 attenuated inversion recovery images<sup>28</sup>
- 265 6- and the volume of underlying white matter hyperintensity (WMH), which was
- 266 quantified as previously described<sup>21</sup> and classified into deciles

267

## 268 **Domain adaptation**

269 To investigate whether domain adaptation using target domain data as additional training data

270 after initial deep learning affects segmentation performance of a fine-tuned algorithm, we

271 randomly divided the External dataset to 1,000 images (Additional training-and-validation

272 dataset for domain adaptation) and 1,777 images (Test dataset for domain adaptation) (Figure

273 1 and Supplementary Fig 2). The Additional training-and-validation dataset for domain

274 adaptation and the Test dataset for domain adaptation were split so that there was no

275 overlapping sample between them. The sample size for the fine tuning (i.e., additional

276 training and validation) of the initially trained model was increased from 50 to 100, 200, 500,

277 and 1,000 to assess the effect of domain adaptation-related data sample size on segmentation

278 performance. The subsampled data were split at a ratio of 8:2 for training and validation. We

279 calculated the mean and standard deviation of the DSC for both the Internal test dataset and

280 the Test dataset for domain adaptation. Moreover, to evaluate whether the sample size of

281 initial training dataset affects the model's performance after domain adaptation, initial deep

282 learning was performed with 2.5 / 5 / 10 / 20 / 50 / 100% of the Training-and-validation

283 dataset and then fine-tuned with the Additional training-and-validation dataset for domain

284 adaptation (sample size of 50, 100, or 200).

285

## 286 **Statistical analysis**

287 To compare baseline characteristics of the Training-and-validation dataset, Internal test

288 dataset, and External dataset, we used ANOVA, the Kruskal-Wallis test, and the chi-square

289 test as appropriate for continuous variables and categorical variables, respectively. We used

290 Bland-Altman plots and a linear regression analysis to compare ground truth infarct volume

291 and segmented infarct volume by the model. To test whether DSC increased as the training  
292 sample size increased and to compare infarct volumes segmented by deep learning and  
293 manual segmentation, we used a linear regression analysis. Performance difference between  
294 models was tested using paired t-test. P-values less than 0.05 was considered statistically  
295 significant.

296

297

## 298 **Results**

### 299 **Baseline characteristics of study population**

300 Mean age of patients was 67.9 (standard deviation 12.9) years in the Training-and-validation  
301 dataset (n = 8,661). Males accounted for 58.9% (n = 4,431) (Table 1). Median NIHSS score  
302 was 4 (interquartile range 2–9) and median infarct volume was 1.95 mL. Mean age of  
303 patients was  $68.2 \pm 12.7$  years in the Internal test dataset and  $68.2 \pm 12.4$  years in the External  
304 dataset. Males accounted for 60.4% and 58.0% in the Internal test dataset and the External  
305 dataset, respectively. Compared with the Training-and-validation dataset and the Internal test  
306 dataset, External dataset was characterized by more cardioembolic strokes, shorter time  
307 intervals from last-known-well to imaging acquisition, and larger infarct volumes. Moreover,  
308 MR vendors, magnetic strengths, and imaging parameters were different among the Training-  
309 and-validation dataset, Internal test dataset, and External dataset (Table 1 and Supplementary  
310 Table 2). Estimated background noise and estimated signal-to-noise ratios in the Internal  
311 dataset varied widely among the 10 participating hospitals (Supplementary Fig 3).

312

### 313 **Performance of a deep learning algorithm trained using data of a single-center**

314 To develop a single-center deep learning model, we used 382 DW images from a single  
315 hospital for model training and validation. Mean age was  $68.8 \pm 13.2$  years in the Single site  
316 training-and-validation dataset. Males accounted for 60.8%. Median infarct volume was 1.70  
317 (0.53–11.25) mL (Supplementary Table 3). For the Single site internal test dataset, the 3D U-  
318 net model achieved DSC of  $0.70 \pm 0.23$  with a per-pixel sensitivity of 0.69 and a precision of  
319 0.78 (Supplementary Table 4). However, the single-center model showed substantially lower  
320 performance for the tests using the External dataset and the Ancillary test datasets I and II,

321 with DSC values of 0.50, 0.51, and 0.33, respectively (all  $p < 0.001$ ).

322

323 **Effect of training data sample size on the performance of deep learning algorithm to**  
324 **segment acute infarcts on DW images**

325 As the sample size of the Training-and-validation dataset increased from 217 to 433 and 866,  
326 DSC of the 3D U-net algorithm increased sharply from 0.58 to 0.61 and 0.65 for the Internal  
327 test dataset (Figure 2A). When the sample size was further increased to 1,732, DSC seemed  
328 to increase less steeply, nearly reaching a plateau (0.67). When the sample size was further  
329 increased to 4,330 and 8,661, DSC only slightly increased to 0.68 and 0.70, respectively.  
330 Similar results were seen in the tests using the External dataset (see also Supplementary Fig 4  
331 for the Ancillary test datasets I and II). When the sample size was 433 or greater, DSC values  
332 in External dataset were significantly higher than those in Internal test dataset. In both  
333 Internal test dataset and External dataset, infarct volumes that were segmented and quantified  
334 by the 3D-Unet algorithm (trained with 8,661 DWI data) showed strong correlations with  
335 ground truth infarct volumes (both  $r^2 = 0.96$ ,  $p < 0.001$ ; Supplementary Fig 5), although the  
336 deep learning algorithm tended to underestimate infarct volumes. Voxel-wise detection  
337 sensitivity showed a pattern that was comparable to that shown for DSCs except for fewer  
338 differences between Internal test dataset and External dataset (Figure 2B). Contrary to the  
339 exponential increase in DSC and sensitivity, precision values in both Internal test dataset and  
340 External dataset changed only slightly when training data sample size increased (Figure 2C).

341

342 **Effect of training data sample size on performance of deep learning algorithm to**  
343 **segment acute infarcts on DW images according to infarct volume, infarct location,**  
344 **presence of chronic ischemic lesions, onset-to-imaging time, and MRI vendors**

345 When the Internal test dataset and the External dataset were divided into small ( $< 1.7$  ml,  $n =$   
346  $994$  and  $1,046$ ), medium ( $1.7 - 14.0$  ml,  $n = 587$  and  $904$ ), and large ( $> 14.0$ ,  $n = 446$  and  $825$ )  
347 infarct groups, DSCs for the internal and external testing were the highest (up to  $\sim 0.8$ ) in the  
348 large infarct group, lower (up to  $\sim 0.7$ ) in the medium infarct group, and the lowest (up to  $\sim 0.6$ )  
349 in the small infarct group (Figure 2D-F). This finding is consistent with generally higher  
350 performances of our deep learning models in the tests using the External dataset as opposed  
351 to the Internal test dataset, given that the mean infarct volume in the former was about two

352 times bigger than in the latter.

353 With regards to lesion locations (Figure 3), DSCs were generally higher for supratentorial  
354 lesions (~0.65 or higher) than for infratentorial lesions (~0.6 or lower), except for cerebellar  
355 lesions (in the tests using the Internal test dataset and the External dataset) and thalamus (in  
356 the test using the External dataset) with DSCs being about 0.7.

357 When data were divided based on the presence of chronic ischemic lesions and WMH  
358 volumes, similar model performances were observed across groups (Supplementary Fig 6 and  
359 7).

360 When data were divided based on the time from last-known-well to imaging, DSCs were the  
361 highest (up to ~0.75) in the > 24-hour group, slightly lower (up to ~0.7) in the 3–24-hour  
362 group, and the lowest (up to ~0.55 and ~0.65) in the < 3-hour group (Figure 2G-I). With  
363 respect to MRI vendors, the deep learning model showed better performances for Phillips or  
364 GE images than for Siemens images in both tests using the Internal test dataset and the  
365 External dataset (Supplementary Table 5).

366 In tests of the 3D-Unet model trained with 8,661 DW images, DSCs for the Internal test  
367 dataset varied, ranging from 0.45 to 0.78, depending on the participating center and training  
368 data sample size, especially the latter (Supplementary Table 6). When we employed the Swin  
369 UNETR for training with the same data, the performance of the deep learning model was  
370 generally lower than that using the 3D-Unet (Supplementary Table 7).

371

### 372 **Improvement of the external test performance of deep learning algorithms via domain** 373 **adaptation**

374 Domain adaptation using subsamples of the External dataset (target domain) enhanced the  
375 model performance in terms of DSC, voxel-wise sensitivity, and precision of lesion  
376 segmentation in testing with Test dataset for domain adaptation (Table 2 and Figure 3). When  
377 the sample size of the Training-and-validation dataset (source domain) was 217, retraining  
378 with 50 cases that were randomly selected from the Additional training-and-validation dataset  
379 for domain adaptation significantly increased DSC from 0.56 to 0.67 ( $p < 0.001$ ; Figure 3) in  
380 testing with the Test dataset for domain adaptation. When the domain adaptation was  
381 performed with 200 cases, DSC was higher (0.71) than that for the 50 cases ( $p < 0.001$ ). A

382 similar pattern of domain adaptation-mediated performance improvement of the deep  
383 learning algorithm was observed when the sample size of the Additional training-and-  
384 validation dataset was 433. However, when the sample size was higher than 433 (i.e., 866 or  
385 higher), there was only slight improvement of infarct segmentation after domain adaptation.  
386 Thus, in terms of the effectiveness of deep learning algorithms, the training data sample size  
387 of 866 without domain adaptation was practically similar to that of 50 with subsequent  
388 domain adaptation. It is notable that domain adaptation with subsamples of target domain  
389 worsened the model performance in internal testing (i.e., testing with the source domain data).  
390 This deterioration could be partly restored by increasing the sample size of the source domain  
391 data for initial deep learning to as high as 8,661.

392

393

## 394 **Discussion**

395 In the study using the largest DWI data to date, we demonstrated that the performance of 3D  
396 U-Net model for the automatic segmentation of acute infarcts improved steeply with training  
397 data volume as sample size was increased from 217 to 866 but reached a plateau as the  
398 training data was further increased to 1,732. When single-center training data was used, the  
399 performance of the deep learning algorithm degraded dramatically in external testing.  
400 Furthermore, we found that domain adaptation utilizing small amount of data from the target  
401 domain improved segmentation accuracy significantly, making the sample size of 866  
402 without domain adaptation equivalent to that of 217 with domain adaptation.

403 The performance of the deep learning-based DWI lesion segmentation algorithm that was  
404 trained on the single-center dataset ( $n = 382$ ) was much inferior in all three external tests than  
405 in the internal test (DSCs of 0.50, 0.51, and 0.33 vs. 0.70, respectively). To develop a more  
406 robust algorithm that generalizes well and performs better on an unseen data, there is a need  
407 for multi-site training data, which better reflects the heterogeneity of the ischemic stroke  
408 phenotype as well as the diversity of MR equipment and protocols in real-world clinical use.  
409 However, it is challenging to obtain, label, and annotate a high volume of multi-center data.  
410 Our findings suggests that multi-site data with a sample size of about 866 ~ 1732 might be  
411 cost-effective in developing a reasonable deep learning algorithm for DWI lesion  
412 segmentation.

413 To enhance the deep learning model's capacity to generalize to new cases, data augmentation  
414 can be used to artificially increase the amount and diversity of training data by generating  
415 modified copies of a dataset using existing data. However, this method carries the biases of  
416 the existing data, such as noise and resolution-related ones, without increasing the variety of  
417 infarct locations and patterns.<sup>29</sup>

418 Utilizing a small data from the target domain could be used to resolve the domain shift issue,  
419 where the model performs poorly on the target data acquired from a different source or  
420 domain (and unseen during training) due to differences in the data distributions.<sup>12, 30, 31</sup> Our  
421 study showed that on the External dataset, the algorithm that was trained with 217 DW  
422 images and was followed by domain adaptation with 50 additional DW images from target  
423 domain performed comparably to the model trained with 866 DW images without subsequent  
424 domain adaptation. As a trade-off due to diversion of the deep learning model on the target  
425 domain, domain adaptation may result in worse performance in the source domain. However,  
426 resilience was observed with little impact on the model's performance in the source domain  
427 when employing a large multi-site data for training. The post-domain-adaptation (n = 200)  
428 DSC drop for the source domain internal test data was 0.10 and 0.03, respectively, in the  
429 models that were pretrained with 866 DW images and 8,661 DW images.

430 Dice coefficients for DWI lesion segmentation were low when infarcts were small or MRI  
431 was performed early (within 3 hours of symptom onset). Given that the External dataset (for  
432 external testing) had approximately 2-fold bigger infarct volumes than the Internal test  
433 dataset, this finding is in line with higher DSCs for the former (vs. the latter) dataset. In  
434 addition, training on multi-site data may have led to the robustness to external testing. Deep  
435 learning algorithms performed poorly on brainstem infarcts, probably due to small number of  
436 cases even in the large training data (n=8,661) and a relatively complex anatomical structures  
437 and variations of the posterior fossa near the brainstem.<sup>32</sup> A strategy for enhancing the  
438 segmentation performance for brainstem infarcts should be developed in future research.

439 This study has strengths, such as the large sample size of multi-site training data and  
440 extensive external test. There are also limitations. First, using apparent diffusion coefficient  
441 images for training may have enhanced the segmentation performance. Second, the  
442 performance of the algorithm may have been improved by using clinical data for training, as  
443 physicians do in clinical practice. Third, caution should be taken when extrapolating our  
444 findings from Korean stroke patients to other ethnic groups, although previous research found



445 no ethnic differences in the pattern of ischemic infarct on DW images.<sup>33</sup>

446 In conclusion, our study demonstrates that domain adaptation or big (n=~1000) multi-site  
447 DWI data are required for a reliable infarct segmentation algorithm with generalizability. In  
448 addition, future research should focus on improving the relatively low segmentation  
449 performance for small or brainstem infarcts or hyperacute infarcts, which has not been  
450 previously described.

451

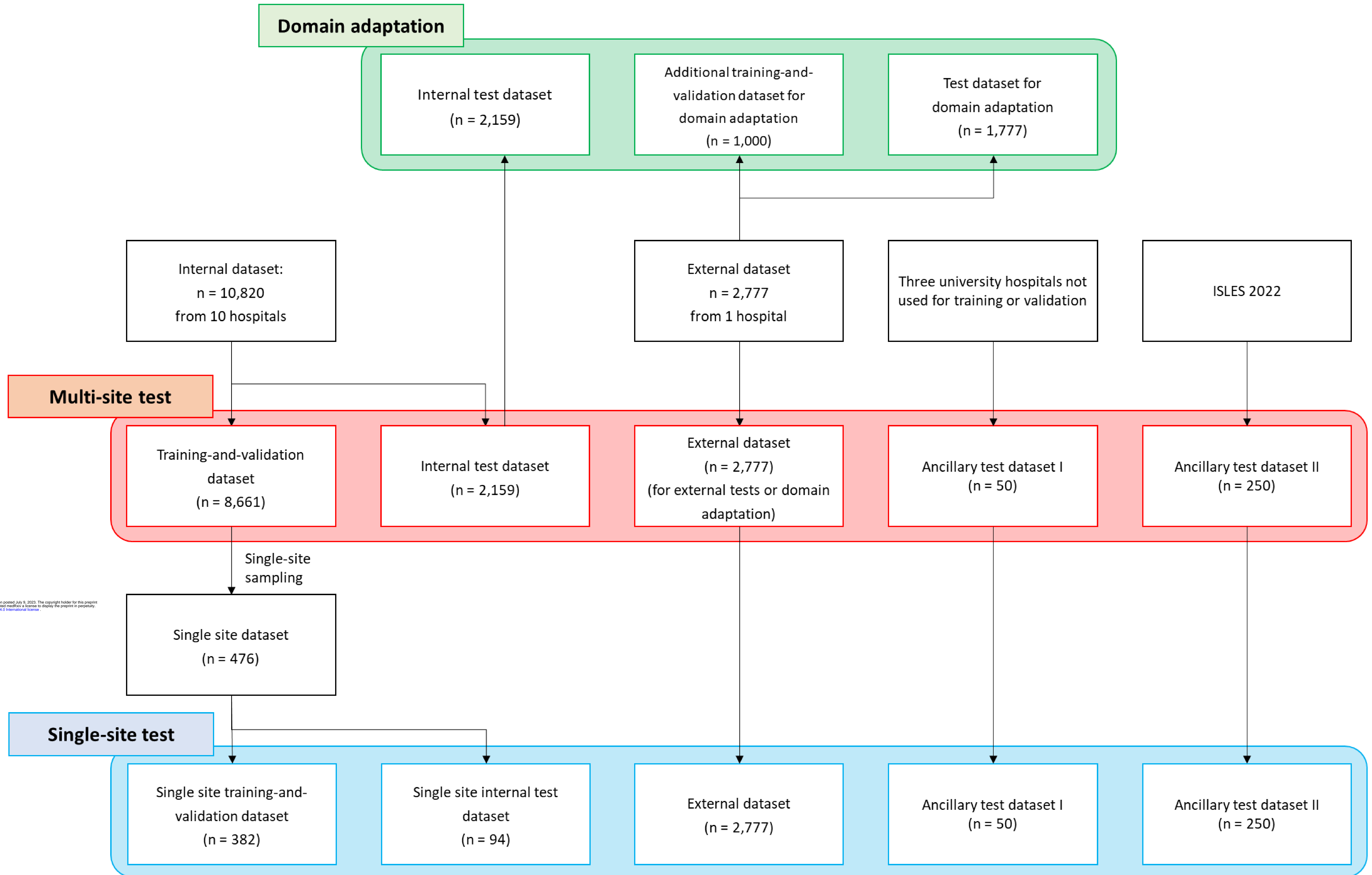
## 452 **References**

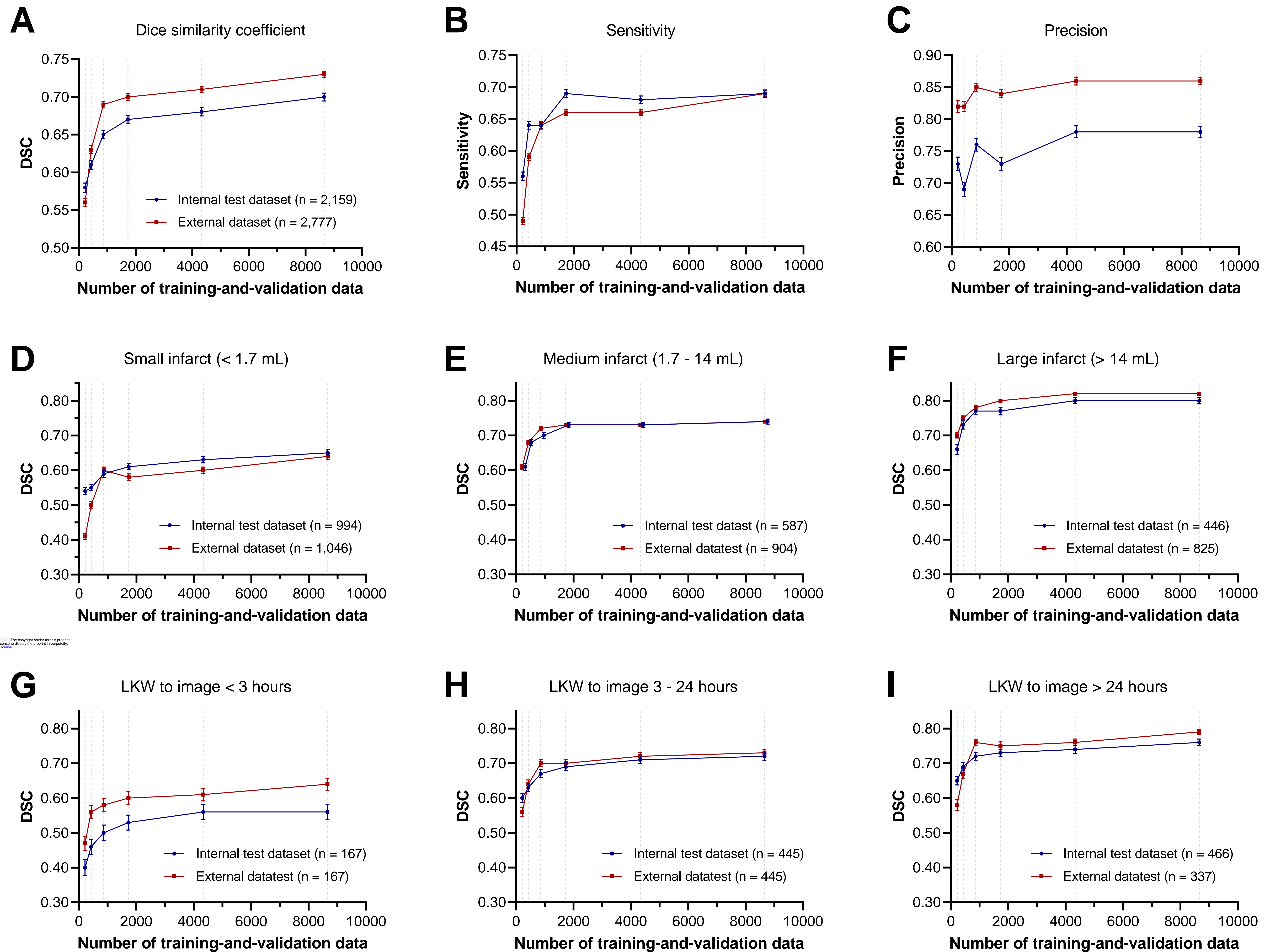
- 453 1. Albers GW. Diffusion-weighted MRI for evaluation of acute stroke. *Neurology*  
454 1998;51:S47-49.
- 455 2. Thijs VN, Lansberg MG, Beaulieu C, Marks MP, Moseley ME, Albers GW. Is early  
456 ischemic lesion volume on diffusion-weighted imaging an independent predictor of stroke  
457 outcome? A multivariable analysis. *Stroke* 2000;31:2597-2602.
- 458 3. Bang OY, Lee PH, Heo KG, Joo US, Yoon SR, Kim SY. Specific DWI lesion patterns  
459 predict prognosis after acute ischaemic stroke within the MCA territory. *J Neurol Neurosurg*  
460 *Psychiatry* 2005;76:1222-1228.
- 461 4. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 Hours after  
462 Stroke with a Mismatch between Deficit and Infarct. *N Engl J Med* 2018;378:11-21.
- 463 5. Albers GW, Marks MP, Kemp S, et al. Thrombectomy for Stroke at 6 to 16 Hours  
464 with Selection by Perfusion Imaging. *N Engl J Med* 2018;378:708-718.
- 465 6. Kim YC, Chung JW, Bang OY, et al. A Deep Learning-Based Automatic Collateral  
466 Assessment in Patients with Acute Ischemic Stroke. *Transl Stroke Res* 2022.
- 467 7. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of Tissue Outcome and  
468 Assessment of Treatment Effect in Acute Ischemic Stroke Using Deep Learning. *Stroke*  
469 2018;49:1394-1401.
- 470 8. Yu Y, Xie Y, Thamm T, et al. Use of Deep Learning to Predict Final Ischemic Stroke  
471 Lesions From Initial Magnetic Resonance Imaging. *JAMA Netw Open* 2020;3:e200772.
- 472 9. Zoetmulder R, Konduri PR, Obdeijn IV, et al. Automated Final Lesion Segmentation  
473 in Posterior Circulation Acute Ischemic Stroke Using Deep Learning. *Diagnostics (Basel)*  
474 2021;11.
- 475 10. Zhang R, Zhao L, Lou W, et al. Automatic segmentation of acute ischemic stroke  
476 from DWI using 3-D fully convolutional DenseNets. *IEEE transactions on medical imaging*  
477 2018;37:2149-2160.
- 478 11. Liu CF, Hsu J, Xu X, et al. Deep learning-based detection and segmentation of  
479 diffusion abnormalities in acute ischemic stroke. *Commun Med (Lond)* 2021;1:61.
- 480 12. Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE*  
481 *Trans Biomed Eng* 2022;69:1173-1185.
- 482 13. Guan H, Liu M. Domain adaptation for medical image analysis: a survey. *IEEE*  
483 *Transactions on Biomedical Engineering* 2021;69:1173-1185.
- 484 14. Ryu WS, Woo SH, Schellingerhout D, et al. Stroke outcomes are worse with larger

- 485 leukoaraiosis volumes. *Brain* 2017;140:158-170.
- 486 15. Ryu WS, Schellingerhout D, Hong KS, et al. Relation of Pre-Stroke Aspirin Use With  
487 Cerebral Infarct Volume and Functional Outcomes. *Ann Neurol* 2021;90:763-776.
- 488 16. Kim DE, Park JH, Schellingerhout D, et al. Mapping the Supratentorial Cerebral  
489 Arterial Territories Using 1160 Large Artery Infarcts. *JAMA Neurol* 2019;76:72-80.
- 490 17. Kim YC, Lee JE, Yu I, et al. Evaluation of Diffusion Lesion Volume Measurements in  
491 Acute Ischemic Stroke Using Encoder-Decoder Convolutional Network. *Stroke*  
492 2019;50:1444-1451.
- 493 18. Woo I, Lee A, Jung SC, et al. Fully automatic segmentation of acute ischemic lesions  
494 on diffusion-weighted imaging using convolutional neural networks: comparison with  
495 conventional algorithms. *Korean Journal of Radiology* 2019;20:1275-1284.
- 496 19. Kim DY, Han S-G, Jeong H-G, et al. Covert Brain Infarction as a Risk Factor for  
497 Stroke Recurrence in Patients With Atrial Fibrillation. *Stroke* 2023;54:87-95.
- 498 20. Hernandez Petzsche MR, de la Rosa E, Hanning U, et al. ISLES 2022: A multi-center  
499 magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data* 2022;9:762.
- 500 21. Ryu WS, Woo SH, Schellingerhout D, et al. Grading and interpretation of white  
501 matter hyperintensities using statistical maps. *Stroke* 2014;45:3567-3575.
- 502 22. Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE*  
503 *Transactions on Systems, Man, and Cybernetics* 1979;9:62-66.
- 504 23. Chilla GS, Tan CH, Xu C, Poh CL. Diffusion weighted magnetic resonance imaging  
505 and its recent trend—a survey. *Quantitative imaging in medicine and surgery* 2015;5:407.
- 506 24. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning  
507 dense volumetric segmentation from sparse annotation. *Medical Image Computing and*  
508 *Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens,*  
509 *Greece, October 17-21, 2016, Proceedings, Part II* 19; 2016: Springer: 424-432.
- 510 25. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks.  
511 *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands,*  
512 *October 11–14, 2016, Proceedings, Part IV* 14; 2016: Springer: 630-645.
- 513 26. Perez-Garcia F, Sparks R, Ourselin S. TorchIO: A Python library for efficient loading,  
514 preprocessing, augmentation and patch-based sampling of medical images in deep learning.  
515 *Comput Methods Programs Biomed* 2021;208:106236.
- 516 27. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin  
517 transformers for semantic segmentation of brain tumors in mri images. *Brainlesion: Glioma,*

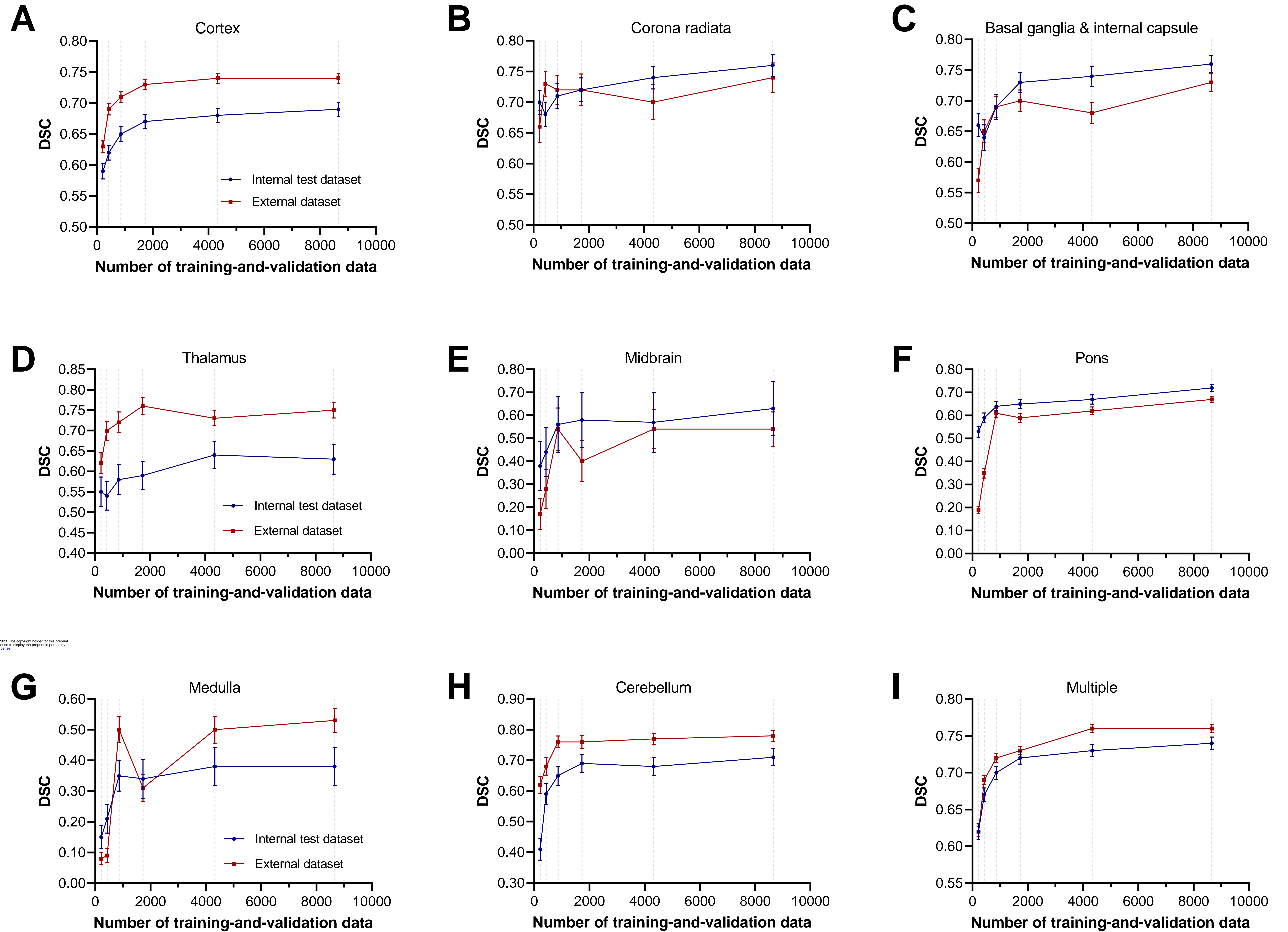
- 518 Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop,  
519 BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021,  
520 Revised Selected Papers, Part I; 2022: Springer: 272-284.
- 521 28. Wardlaw JM, Smith EE, Biessels GJ, et al. Neuroimaging standards for research into  
522 small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol*  
523 2013;12:822-838.
- 524 29. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of  
525 medical image data augmentation techniques for deep learning applications. *J Med Imaging*  
526 *Radiat Oncol* 2021;65:545-563.
- 527 30. Singh T, Saurabh P, Bisen D, Kane L, Pathak M, Sinha GR. Ftl-CoV19: A Transfer  
528 Learning Approach to Detect COVID-19. *Comput Intell Neurosci* 2022;2022:1953992.
- 529 31. Sundaresan V, Zamboni G, Dinsdale NK, Rothwell PM, Griffanti L, Jenkinson M.  
530 Comparison of domain adaptation techniques for white matter hyperintensity segmentation in  
531 brain MR images. *Med Image Anal* 2021;74:102215.
- 532 32. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep  
533 convolutional neural networks. *Advances in neural information processing systems* 2016;29.
- 534 33. Bang OY, Ovbiagele B, Liebeskind DS, Restrepo L, Yoon SR, Saver JL. Clinical  
535 determinants of infarct pattern subtypes in large vessel atherosclerotic stroke. *J Neurol*  
536 2009;256:591-599.
- 537

Figure 1

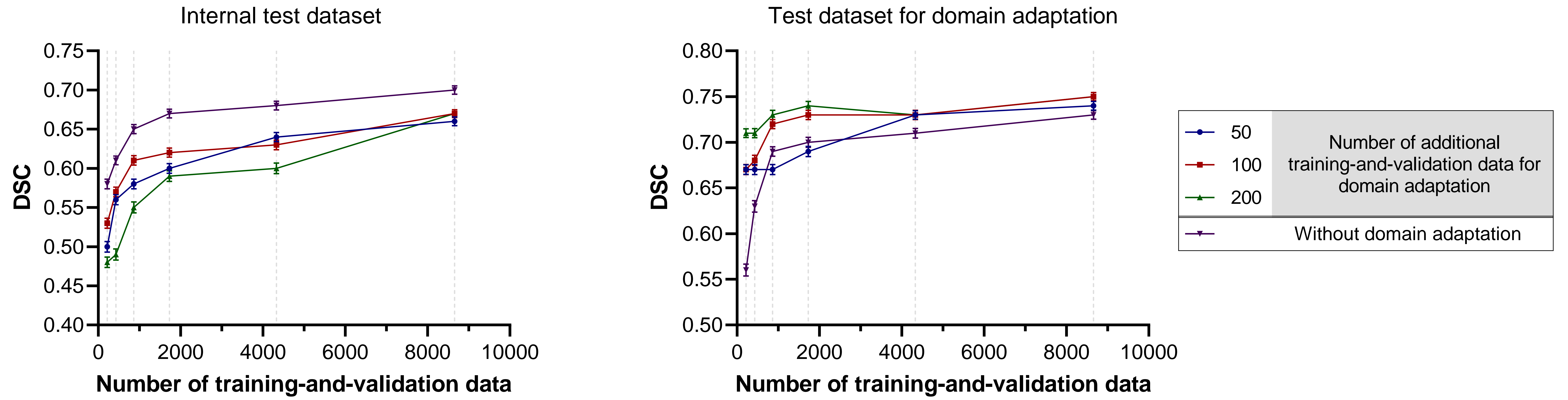


**Figure 2**

**Figure 3**



**Figure 4**





## **Tables and Figure legends for**

Deep learning algorithms for automatic segmentation of acute cerebral infarcts on diffusion-weighted images: Effects of training data sample size, transfer learning, and data features (Noh et al.)

**Table 1. Baseline demographic and imaging characteristics of subjects, whose diffusion weighted magnetic resonances images were used for the Training-and-validation dataset, Internal test dataset, or External dataset**

Variable	Training-and-validation dataset (n = 8,661)	Internal test dataset (n = 2,159)	External dataset (n = 2,777)	P-value
Age (year) <sup>a</sup>	67.9 ± 12.9	68.2 ± 12.7	68.2 ± 12.4	.55
Male <sup>a</sup>	4,431 (58.9%)	1,144 (60.4%)	1,571 (58.0%)	.26
BMI <sup>b</sup>	23.5 ± 3.4	23.4 ± 3.2	23.5 ± 3.4	.23
Admission NIHSS <sup>a</sup> , median (IQR)	4 (2 – 9)	4 (2 – 9)	4 (2 – 10)	.79
Subtype <sup>c</sup>				<.001
LAA	2,775 (37.2%)	688 (36.5%)	1,080 (40.0%)	
SVO	1,421 (19.0%)	377 (20.0%)	227 (8.4%)	
CE	1,606 (21.5%)	394 (20.9%)	663 (24.6%)	
Undetermined	1,507 (20.2%)	381 (20.2%)	685 (25.4%)	
Other determined	160 (2.1%)	46 (2.4%)	42 (1.6%)	
Previous stroke <sup>a</sup>	1,720 (22.9%)	440 (23.2%)	409 (15.1%)	<.001
Hypertension <sup>a</sup>	5,302 (70.5%)	1,353 (71.4%)	1,668 (61.5%)	<.001
Diabetes mellitus <sup>a</sup>	2,622 (34.9%)	624 (32.9%)	765 (28.2%)	<.001
Hyperlipidemia <sup>a</sup>	2,853 (37.9%)	720 (38.0%)	386 (14.2%)	<.001
Smoking <sup>a</sup>	3,061 (40.7%)	758 (40.0%)	1,037 (38.3%)	.09
Atrial fibrillation <sup>a</sup>	1,561 (20.7%)	378 (20.0%)	657 (24.2%)	<.001
Time from LKW to imaging <sup>d</sup> , median (IQR, hour)	20.48 (5.3 – 49.6)	19.41 (5.0 – 48.0)	11.41 (4.0 – 35.9)	<.001
Infarct volume, median (IQR, mL)	1.95 (0.47 – 11.05)	1.89 (0.51 – 10.9)	4.19 (0.76 – 19.35)	<.001 <sup>e</sup>
MRI vendor				<.001
Phillips	3,435 (40.7%)	868 (40.2%)	3 (0.1%)	
GE	1,709 (20.2%)	438 (20.3%)	2,706 (97.4%)	
Siemens	3,292 (39.0%)	851 (39.4%)	60 (2.2%)	
Other	7 (0.1%)	2 (0.1%)	8 (0.3%)	
Magnetic field strength <sup>f</sup>				<.001
1.5T	5,129 (69.3%)	1,231 (66.9%)	2,724 (98.5%)	
3.0T	2,273(30.7%)	609 (33.1%)	41 (1.5%)	
Pixel spacing (mm) <sup>g</sup>				<.001

< 0.8	1,311 (15.1%)	335 (15.6%)	11 (0.4%)
0.8 ~ 0.849	1,373 (15.9%)	359 (16.6%)	11 (0.4%)
0.85 ~ 0.899	2,181 (25.2%)	544 (25.2%)	10 (0.4%)
0.9 ~ 0.949	1,073 (12.4%)	257 (11.9%)	12 (0.4%)
0.95 ~ 0.999	515 (5.9%)	137 (6.3%)	55 (2.0%)
≥ 1.0	2,208 (25.5%)	527 (24.4%)	2,676 (96.4%)
Slice thickness (mm) <sup>h</sup>			<.001
3.0 ~ 3.9	2,335 (31.5%)	573 (31.1%)	1 (0.0%)
4.0 ~ 4.9	625 (8.5%)	156 (8.5%)	2,699 (97.3%)
5.0 ~ 5.9	4,417 (59.6%)	1,109 (60.2%)	66 (2.4%)
≥ 6.0	32 (0.4%)	4 (0.2%)	8 (0.3%)

BMI, body mass index; NIHSS, National Institutes of Health Stroke Scale; IQR, interquartile range; LAA, large artery atherosclerosis; SVO, small vessel occlusion; CE, cardioembolism; LKW, Last-known-well. Data are presented as mean ± standard deviation, number (percentage), or median (interquartile range). See Figure 1 for a better understanding of datasets.

<sup>a</sup>Data of age, sex, BMI, admission NIHSS, previous stroke, hypertension, diabetes, hyperlipidemia, smoking, and atrial fibrillation were missing for 1,138, 266, and 67 patients in Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

<sup>b</sup>Data of BMI were missing for 1,218, 285, and 485 patients of Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

<sup>c</sup>Data of stroke subtype were missing for 1,192, 274, and 90 patients of Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

<sup>d</sup>Data of LKW to imaging time were missing for 4,373, 1,078, and 1,849 patients in Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

<sup>e</sup>Kruskal-Wallis test was used.

<sup>f</sup>Data of magnetic field strength were missing for 1,259, 319, and 12 patients in Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

<sup>g</sup>Data of pixel spacing were missing for 2 patients in External dataset.

<sup>h</sup>Data of slice thickness were missing for 1,252, 317, and 3 patients in Training-and-validation dataset, Internal test dataset, and External dataset, respectively.

**Table 2. Lesion segmentation performance after domain adaptation using the Training-and-validation dataset for domain adaptation**

Metric	Before domain adaptation	After domain adaptation				
		50 cases	100 cases	200 cases	500 cases	1000 cases
<b>Dice similarity coefficient</b>						
Internal test dataset ( $n = 2,159$ )	0.70 (0.25)	0.66 (0.26)	0.67 (0.24)	0.67 (0.24)	0.68 (0.24)	0.67 (0.24)
$P$ -value <sup>a</sup>	Reference	< .001	< .001	< .001	.007	< .001
Test dataset for domain adaptation ( $n = 1,777$ )	0.73 (0.21)	0.74 (0.21)	0.75 (0.19)	0.75 (0.19)	0.75 (0.19)	0.76 (0.19)
$P$ -value <sup>a</sup>	Reference	.15	.002	.002	.002	.002
<b>Sensitivity<sup>b</sup></b>						
Internal test dataset ( $n = 2,159$ )	0.69 (0.27)	0.69 (0.30)	0.73 (0.26)	0.71 (0.26)	0.72 (0.25)	0.72 (0.26)
$P$ -value <sup>a</sup>	Reference	>.99	< .001	.001	< .001	< .001
Test dataset for domain adaptation ( $n = 1,777$ )	0.69 (0.23)	0.73 (0.24)	0.74 (0.21)	0.73 (0.21)	0.75 (0.21)	0.75 (0.21)
$P$ -value <sup>a</sup>	Reference	< .001	< .001	< .001	< .001	< .001
<b>Precision<sup>b</sup></b>						
Internal test dataset ( $n = 2,159$ )	0.78 (0.21)	0.72 (0.22)	0.68 (0.24)	0.69 (0.25)	0.70 (0.24)	0.69 (0.25)
$P$ -value <sup>a</sup>	Reference	< .001	< .001	< .001	< .001	< .001
Test dataset for domain adaptation ( $n = 1,777$ )	0.86 (0.16)	0.82 (0.16)	0.80 (0.20)	0.82 (0.19)	0.82 (0.19)	0.82 (0.19)
$P$ -value <sup>a</sup>	Reference	< .001	< .001	< .001	< .001	< .001

Data are presented as mean (standard deviation). See Figure 1 for a better understanding of datasets.

<sup>a</sup> $P$ -value for difference compared to the value of before domain adaptation.

<sup>b</sup>Sensitivity and precision were computed voxel-wise.

## Figure 1. Study flow chart

**Figure 2. Lesion segmentation performance of deep learning algorithm as training data increase with stratification by infarct volumes and onset-to-imaging time.**

(A) Dice similarity coefficient (DSC) in all patients. (B) Pixel-level sensitivity in all patients. (C) Pixel-level precision in all patients. (D-F) DSC stratified by infarct volume ( $< 1.7$ ,  $1.7 - 14$ , and  $\geq 14$  mL). (G-H) DSC stratified by time from last-known-well to image time. Dot and bar indicate mean and standard error, respectively. Data of time from onset to imaging were missing for 565 and 1,849 patients in Internal test dataset and External dataset, respectively. Gray dot lines indicate data points of 217, 433, 866, 1,732, 4,330, and 8,661. Sensitivity and precision were calculated voxel-wise. Compared with DSC in the model trained with 217 patients, all DSCs in the model trained with 433, 866, 1,732, 4,330, and 8,661 were significantly higher. See Figure 1 for a better understanding of datasets. LKW = last-known-well.

**Figure 3. Lesion segmentation performance in the Internal test dataset and the External dataset with stratification by lesion location**

(A) Cortex. (B) Corona radiata. (C) Basal ganglia & internal capsule. (D) Thalamus. (E) Midbrain. (F) Pons. (G) Medulla. (H) Cerebellum. (I) Multiple. Dot and bar indicate mean and standard error, respectively. Gray dot lines indicate data points of 217, 433, 866, 1,732, 4,330, and 8,661. Sensitivity and precision were calculated voxel-wise. Note that Y-axis ranges varied in each figure. Compared with supratentorial lesions (A-C), infratentorial lesion except for cerebellum had lower dice similarity coefficient (DSC). See Figure 1 for a better understanding of datasets.

**Figure 4. Lesion segmentation performance before and after domain adaptation using the Training-and-validation dataset for domain adaptation.**

(A) Dice similarity coefficient (DSC) in Internal test dataset. (B) DSC in Test dataset for domain adaptation. Data are presented as mean and stranded error. Gray dot lines indicate data points of 217, 433, 866, 1,732, 4,330, and 8,661. See Figure 1 for a better understanding of datasets.