Supplementary Material

Edge-of-distribution Training

This experiment was performed with nnU-Net only. **Table 6** shows the test set performance on protocol 1 and on protocols 2, 3 and 4 under different training data sets. Set 2 refers to the scheme in which we include 12 examples from protocols 2, 3 and 4 in the training data (as per **Table 2**) and is the scheme used in all other sections of this paper. To facilitate investigation of the effectiveness of this scheme, we repeated the training without these 12 (out-of-distribution) examples and we refer to this as set 1. When set 2 is used there is a statistically significant increase in precision and decrease in recall for both WMH and stroke (Wilcoxon signed-rank test, $p<0.05$, **Table 9**) in the protocols 2, 3 and 4 data. Upon visual inspection of the data, protocols 2, 3 and 4 exhibit lower contrast, making the differentiation of WMH or stroke lesions from surrounding white matter more difficult (**Fig 6b**). We suggest that the improvement in precision is due to exposure to these lower contrast areas during training, enabling the model to make less false positives on the areas adjacent to the pathology at test time. The decrease in recall is likely a direct consequence of this. The result of the change in precision and recall is a statistically significant improvement in WMH DSC and minimal change in the stroke DSC. There is also a large and statistically significant reduction in the 95% HD for stroke in these protocols. With the exception of an increase in precision for stroke, the changes in scores when evaluating on protocol 1 test data are not significantly impacted by the addition of the 12 training examples from protocols 2, 3 and 4.

While a key motivation for this experiment was to investigate whether we could achieve a level of robustness to lower resolution scans, our observation that there is noticeably lower contrast in this data highlights the multitude of different ways in which dataset characteristics can vary.

In conclusion, we found that adopting the approach of Dorent et al. [5] and including even a small number of would-be out-of-distribution examples in the training data resulted in a statistically significant improvement in the WMH DSC and stroke 95% HD on this now edge-of-distribution data. We hence suggest that it is feasible to use high quality already-labelled research datasets to train models for use on poorer quality data, provided that a small sample of this data, or similar data, is collected and labelled.

Table 6: Comparison of two training schemes for nnU-Net. Set 2 is as shown in **Table 2** and all other experiments while set 1 is the same, only without the 12 protocols 2, 3 and 4 examples. 95% HD in mm.

Data	Training scheme	DSC		Precision		Recall		95% HD	
		WMH	Stroke	WMH	Stroke	WMH	Stroke	WMH	Stroke
Protocol 1	Set 1	0.7800	0.4804	0.8116	0.6039	0.7681	0.4420	9.534	26.98
	Set 2	0.7778	0.4868	0.8096	0.6216	0.7657	0.4484	9.554	30.01
Protocols 2, 3, 4	Set 1	0.6873	0.4736	0.6658	0.5670	0.7629	0.4448	13.23	46.63
	Set 2	0.7093	0.4715	0.7320	0.6854	0.7323	0.4051	10.89	30.44

Appendix

Table 7: The performance of each model tested on cases with low, medium and high WMH and stroke lesion volume. Low, medium and high correspond to the first, second and third tertiles within the entire test set respectively.

WMH vol	Stroke vol	Model	DSC		Precision		Recall		95% HD	
			WMH	Stroke	WMH	Stroke	WMH	Stroke	WMH	Stroke
Low	Low	nnU-Net	0.6997	0.3816	**0.7325**	**0.4714**	0.7064	0.3510	13.97	**30.24**
		Ensemble 2	**0.6850**	**0.3899**	0.6849	0.3987	**0.7179**	**0.4126**	**11.53**	35.323
	Medium	nnU-Net	**0.6408**	0.5401	**0.6119**	**0.7615**	0.7315	0.4578	**13.12**	28.26
		Ensemble 2	0.6325	**0.5495**	0.5834	**0.8088**	**0.7545**	**0.4745**	14.01	**27.65**
	High	nnU-Net	**0.5990**	**0.7775**	**0.6314**	**0.8703**	0.6162	0.7405	17.83	**8.621**
		Ensemble 2	0.5893	0.7710	0.5754	0.8594	**0.6634**	**0.7408**	**16.99**	14.41
Medium	Low	nnU-Net	**0.8148**	**0.2603**	**0.8376**	**0.2982**	0.8135	**0.2544**	8.741	**49.10**
		Ensemble 2	0.8032	0.2019	0.7993	0.2450	**0.8259**	0.1968	**6.038**	53.65
	Medium	nnU-Net	**0.7875**	**0.6667**	**0.8134**	**0.8207**	0.7744	0.5896	7.184	**23.28**
		Ensemble 2	0.7702	0.6494	**0.7604**	0.8203	**0.7980**	**0.5977**	**6.715**	30.71
	High	nnU-Net	**0.6606**	**0.5461**	**0.6762**	0.8032	**0.7008**	**0.4617**	**11.90**	28.25
		Ensemble 2	0.6585	0.5433	0.6705	**0.8256**	0.6951	**0.4585**	12.49	**27.75**
High	Low	nnU-Net	**0.8794**	**0.2903**	**0.9252**	0.3094	0.8464	0.2890	4.514	**11.36**
		Ensemble 2	0.8713	0.2772	0.8944	**0.3914**	**0.8562**	**0.3623**	**4.011**	40.25
	Medium	nnU-Net	**0.8333**	0.3425	**0.8908**	**0.5981**	0.7910	0.2901	**6.740**	**41.18**
		Ensemble 2	0.8300	**0.4046**	0.8691	0.5404	**0.8018**	**0.3507**	8.259	47.503
	High	nnU-Net	**0.8195**	0.4913	**0.8549**	0.8197	0.7939	0.4208	6.807	44.52
		Ensemble 2	0.8180	**0.5901**	0.8490	**0.8756**	**0.7973**	**0.5053**	**6.600**	**30.10**

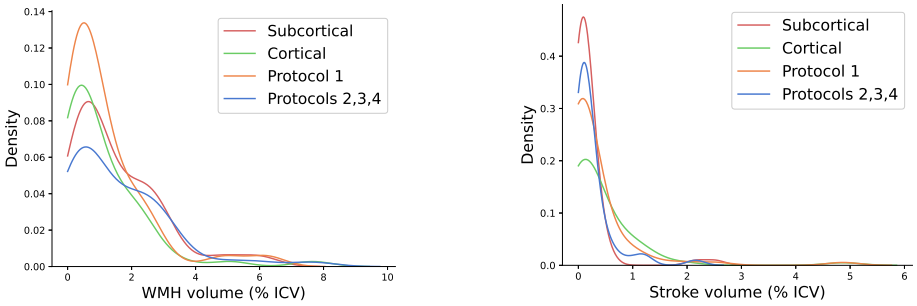

Fig. 7: Kernel density estimates portraying distribution of WMH and stroke volumes across different sections of the test set.

Table 8: Mann-Whitney U test showing differences in distributions of scores under two different stroke subtypes (subcortical and cortical). U statistic for cortical stroke with p-value in parentheses. P-values <0.05 in bold.

	DSC		Precision		Recall		95% HD	
	WMH	Stroke	WMH	Stroke	WMH	Stroke	WMH	Stroke
nnU-Net	1468 (0.113)	1544 (0.238)	**1350 (0.026)**	1649 (0.533)	1786 (0.922)	1530 (0.209)	2087 (0.089)	1526 (0.076)
Ensemble 2	1468 (0.113)	1520 (0.191)	**1286 (0.010)**	1681 (0.650)	1856 (0.639)	1511 (0.175)	1994 (0.229)	**1789 (0.039)**

Table 9: Wilcoxon signed-rank test showing differences in distributions of scores under the two different training data splits (with and without the 12 protocols 2, 3 and 4 examples) for nnU-Net. W statistic with p-value in parentheses. P-values <0.05 in bold. Note: Mann-Whitney U test (with U statistic for split 1) is done for stroke 95% HD because there are some NaN values due to empty predictions (<20%).

	DSC		Precision		Recall		95% HD	
	WMH	Stroke	WMH	Stroke	WMH	Stroke	WMH	Stroke
Protocol 1	1084 (0.103)	488 (0.101)	1308 (0.672)	**369 (0.006)**	1064 (0.082)	616 (0.836)	654 (0.240)	1458 (0.644)
Protocols 2, 3, 4	**482 (0.001)**	590 (0.647)	**16 (0.000)**	**211 (0.000)**	**395 (0.000)**	**365 (0.022)**	687 (0.093)	**1961 (0.006)**