PItcHPERFeCT: Primary Intracranial Hemorrhage Probability Estimation using Random Forests on CT

John Muschelli[a,*], Elizabeth M. Sweeney[a], Natalie L. Ullman[b], Paul Vespa[c], Daniel F. Hanley[b], Ciprian M. Crainiceanu[a]

[a]*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA*
[b]*Department of Neurology, Division of Brain Injury Outcomes, Johns Hopkins Medical Institutions, Baltimore, MD, USA*
[c]*Department of Neurosurgery, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*

Abstract

Introduction

Intracerebral hemorrhage (ICH), where a blood vessel ruptures into areas of the brain, accounts for approximately 10-15% of all strokes. X-ray computed tomography (CT) scanning is largely used to assess the location and volume of these hemorrhages.

Manual segmentation of the CT scan using planimetry by an expert reader is the gold standard for volume estimation, but is time-consuming and has within- and across-reader variability. We propose a fully automated segmentation approach using a random forest algorithm with features extracted from X-ray computed tomography (CT) scans.

Methods

The Minimally Invasive Surgery plus rt-PA in ICH Evacuation (MISTIE) trial was a multi-site Phase II clinical trial that tested the safety of hemorrhage removal using recombinant-tissue plasminogen activator (rt-PA). For this analysis, we use 112 baseline CT scans from patients enrolled in the MISTE trial, one CT scan per patient. ICH was manually segmented on these CT scans by expert readers.

We derived a set of imaging predictors from each scan. Using 10 scans, we used a first-pass voxel selection procedure based on quantiles of a set of predictors and then built 4 models estimating the voxel-level probability of ICH. The models used were: 1) logistic regression, 2) logistic regression with a penalty on the model parameters using LASSO, 3) a generalized additive model (GAM) and 4) a random forest classifier. The remaining 102 scans were used for model validation.

For each validation scan, the model predicted the probability of ICH at each voxel. These voxel-level probabilities were then thresholded to produce binary segmentations of the hemorrhage. These masks were compared to the manual segmentations using the Dice Similarity Index (DSI) and the correlation of hemorrhage volume of between the two segmentations. We tested equality of median DSI using the Kruskal-Wallis test across the 4 models. We tested equality of the median DSI from sets of 2 models using a Wilcoxon signed-rank test.

Results

All results presented are for the 102 scans in the validation set. The median DSI for each model was: 0.89 (logistic), 0.885 (LASSO), 0.88 (GAM), and 0.899 (random forest). Using the random forest results in a slightly higher median DSI compared to the other models. After Bonferroni correction, the hypothesis of equality of median DSI was rejected only when comparing the random forest DSI to the DSI from the logistic

*Principal Corresponding Author

Email addresses: jmusche1@jhu.edu (John Muschelli), emsweene1@jhu.edu (Elizabeth M. Sweeney), nullman1@jhmi.edu (Natalie L. Ullman), PVespa@mednet.ucla.edu (Paul Vespa), dhanley@jhmi.edu (Daniel F. Hanley), ccrainic@jhsph.edu (Ciprian M. Crainiceanu)

($p < 0.001$), LASSO ($p < 0.001$), or GAM ($p < 0.001$) models. In practical terms the difference between the random forest and the logistic regression is quite small. The correlation (95% CI) between the volume from manual segmentation and the predicted volume was 0.93 $(0.9, 0.95)$ for the random forest model. These results indicate that random forest approach can achieve accurate segmentation of ICH in a population of patients from a variety of imaging centers. We provide an R package (`https://github.com/muschellij2/ichseg`) and a Shiny R application online (`http://johnmuschelli.com/ich_segment_all.html`) for implementing and testing the proposed approach.

Keywords: CT, ICH Segmentation, Intracerebral Hemorrhage

1. Introduction

Intracerebral hemorrhage (ICH) is a neurological condition that results from a blood vessel rupturing into the tissue and possibly extending into the ventricles of the brain. The use of X-ray computed tomography (CT) scans allows clinicians and researchers to qualitatively and quantitatively describe the characteristics of a hemorrhage to guide interventions and treatments. CT scanning is widely available and is the most commonly used diagnostic tool in patients with ICH [1]. The volume of ICH has been consistently demonstrated to be an important diagnostic predictor of stroke severity, long-term functional outcome, and mortality [2, 3, 4]. ICH volume change is also a common primary outcome [5, 6, 7, 8] and secondary outcome [5, 9, 10] in clinical trials. Moreover, the location of the ICH has been shown to affect functional outcome in patients with stroke [11, 12]. Thus, quantitative measures of ICH (e.g. volume, location, and shape) are increasingly important for treatment and other clinical decision.

ICH volume can be estimated quickly, for example, using the ABC/2 method [2]. In this method, a reader chooses the slice with the largest area of hemorrhage. The length of the intersection between this first axis and the hemorrhage is denoted by A. The next step is to draw an orthogonal line at the middle of the segment of length A in the same plane that contains the largest hemorrhage area. The length of the intersection between this second orthogonal axis and the hemorrhage is denoted by B. The reader then counts the number of slices where hemorrhage is present (C). The volume estimate is $\frac{A \times B \times C}{2}$, which is an approximation of the volume under the assumption that the hemorrhage shape is well approximated by an ellipsoid [13]. As this method is relatively easy to implement in practice, it can be used to quickly produce rough estimates of hemorrhage volume [14].

Although ABC/2 is widely used, Divani et al. [15] found that the measurement error associated with the ABC/2 method were significantly greater than those using planimetry, which requires slice-by-slice hemorrhage segmentation by trained readers. Planimetry is much more labor intensive and time consuming, but it more accurately estimates the true ICH volume compared to the ABC/2 approach, especially for irregularly shaped ICH and for smaller thickness (i.e. higher resolution) scans. Another problem that has not been discussed in the literature is that ICH may change over time. The shape of the ICH may initially be well approximated by an ellipsoid but the approximation may become increasingly inaccurate over time as the lesion changes shape, migrates through the surrounding tissues, or breaks down. Surgical interventions that target the removal of ICH may also change the shape of the ICH or cause additional bleeding. Moreover, the ABC/2 method has been shown to consistently over-estimate infarct volume [16] and may have significant inter-rater variability [17]. Therefore, a rapid, automated, and validated method for estimating hemorrhage location and its volume from CT scans is highly relevant in clinical trials and clinical care. Accuracy is accompanied by increase of both diagnostic and prognostic value.

Methods have been proposed for segmentation of ICH using magnetic resonance images (MRI) [18, 19]. However, in most clinical settings CT, not MRI, is the image of choice. Furthermore, MRI sequences and protocols may vary across sites and there is no general, standardized, agreed-upon MRI protocol for ICH standard-of-care. Thus, there is a need for ICH segmentation that relies only on CT scan information, is reliable, reproducible, available, and well validated against planimetry.

We propose an algorithm that can estimate the probability of ICH at the voxel level, produce a binary image of ICH location, and estimate ICH volume. We will compare our predicted ICH maps to the gold standard – manual segmentation. Several methods have been presented for automated methods for estimating ICH from CT scans [20, 21, 22, 23, 24]. These methods include fuzzy clustering [21, 22], simulated annealing [23], 3-dimensional (3D) mathematical morphology operations [24], and template-based comparisons [20].

Unfortunately, no software for ICH segmentation is publicly available. We provide a completely automated pipeline of analysis from raw images to binary hemorrhage masks and volume estimates, and provide a public webpage to test the software.

2. Methods

2.1. Data

2.2. Participants and Imaging Data

We used CT images from patients enrolled in the MISTIE II (Minimally Invasive Surgery plus recombinant-tissue plasminogen activator for Intracerebral Hemorrhage Evacuation) stroke trial [9]. We analyzed 112 scans taken prior to randomization and treatment, corresponding to the first scan acquired post-stroke for 112 unique patients. Inclusion criteria into the study included: 18 to 80 years of age and spontaneous supratentorial intracerebral hemorrhage above 20 milliliters (mL) in size (for full criteria, see Mould et al. [25]). The population analyzed here had a mean (standard deviation (SD)) age of 60.7 (11.2) years, was 68.8% male, and was 53.6% Caucasian, 31.2% African American, 10.7% Hispanic, and 4.5% Asian or Pacific islander. CT data were collected as part of the Johns Hopkins Medicine IRB-approved MISTIE research studies with written consent from participants.

The study protocol was executed with minor, but important, differences across the 26 sites. Scans were acquired using 4 scanner manufacturers: GE ($N = 46$), Siemens ($N = 38$), Philips ($N = 20$), and Toshiba ($N = 8$). In head CT scanning, the gantry may be tilted for multiple purposes, for example, so that sensitive organs, such as the eyes, are not exposed to X-ray radiation. This causes scan slices to be acquired at an oblique angle with respect to the patient. Gantry tilt was observed in 88 scans. Slice thickness of the image varied within the scan for 14 scans. For example, a scan may have 10 millimeter (mm) slices at the top and bottom of the brain and 5mm slices in the middle of the brain. Therefore, the original scans analyzed had different voxel (volume element) dimensions. These conditions are characteristic of how scan are presented in many diagnostic cases.

2.3. Hemorrhage Segmentation and Location Identification

ICH was manually segmented on CT scans using the OsiriX imaging software by expert readers (OsiriX v. 4.1, Pixmeo; Geneva, Switzerland). After image quality review, continuous, non-overlapping slices of the entire hemorrhage were segmented. Readers employed a semiautomated threshold-based approach using a Hounsfield unit (HU) range of 40 to 80 to select potential regions of hemorrhage [26, 27]; these regions were then further quality controlled and refined by readers using direct inspection of images. Binary hemorrhage masks were created by setting voxel intensity to 1 if the voxel was classified as hemorrhage, regardless of location, and 0 otherwise.

2.4. Image Processing: Brain Extraction, Registration

CT images and binary hemorrhage masks were exported from OsiriX to DICOM (Digital Imaging and Communications in Medicine) format. The image processing pipeline can be seen in Figure 1. Images with gantry tilt were corrected using a customized MATLAB (The Mathworks, Natick, Massachusetts, USA) user-written script (`http://bit.ly/11tIM8c`). Images were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) data format using `dcm2nii` (provided with MRIcro [28]). Images were constrained to values -1024 and 3071 HU to remove potential image rescaling errors and artifacts. No interpolation was done for images with a variable slice thickness. Thickness was determined from the first converted slice and the NIfTI format assumes homogeneous thickness throughout the image. In a future release of `dcm2nii`, called `dcm2niix`, interpolation will be done for scans with variable slice thickness and gantry-tilt correction will be performed automatically.

All image analysis was done in the R statistical software [29], using the `fslr` [30] package to call functions from the FSL [31] neuroimaging software (version 5.0.4), and the `ANTsR` package to call functions from the ANTs (Advanced Normalization Tools) neuroimaging software [32].

Brains were extracted to remove skull, eyes, facial and nasal features, extracranial skin, and non-human elements of the image captured by the CT scanner, such as the gantry, pillows, or medical devices. Removal of these elements was performed using the brain extraction tool (BET) [33], a function of FSL, using a previously published validated CT-specific brain extraction protocol [34].

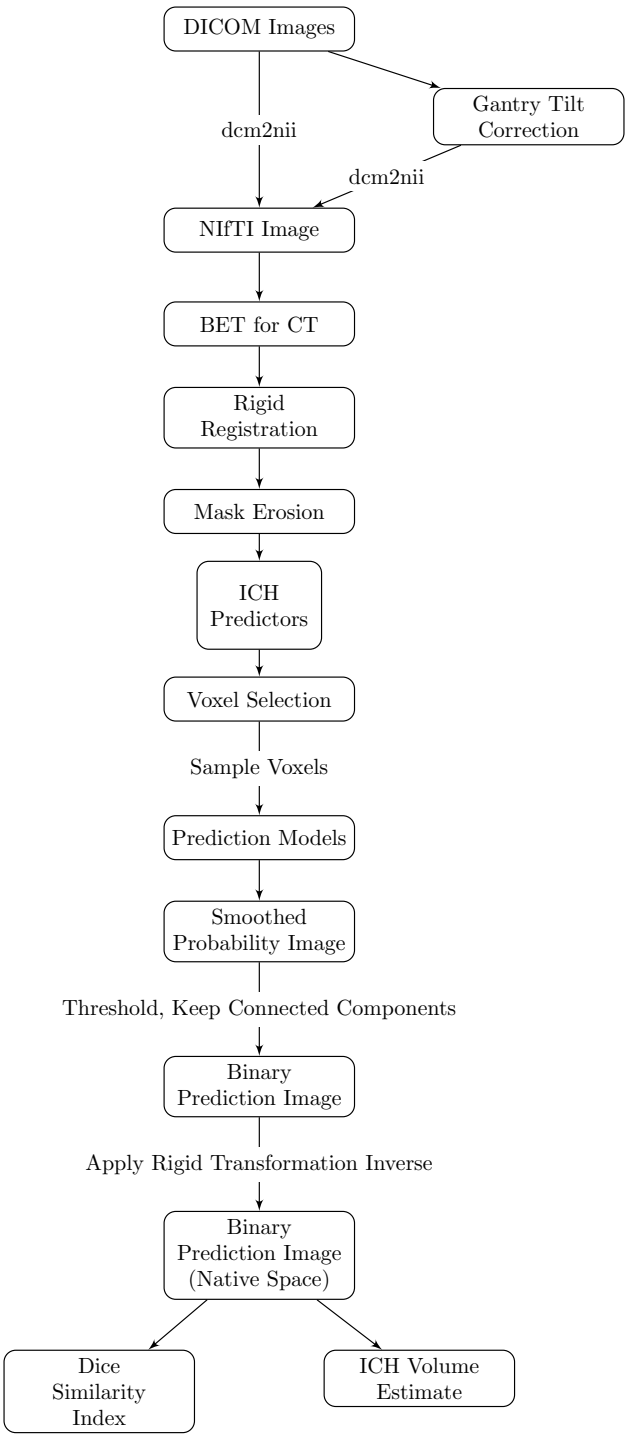

Figure 1: **Processing Pipeline**. Images in DICOM (Digital Imaging and Communications in Medicine) format were gantry tilt corrected if necessary and converted to NIfTI (Neuroimaging Informatics Technology Initiative) format using `dcm2nii`. After NIfTI conversion, the brain extraction tool (BET) was applied to the image using a previously published protocol. The image was registered to a brain CT template using a rigid-body transformation and was interpolated to template resolution. We estimated imaging predictors and used these predictors to estimate the probability of ICH in a prediction model. The probability of ICH was thresholded, connected component below 100 voxels (0.1mL) were discarded, and the image was transformed back into original space of the patient. The ICH volume and the Dice Similarity Index, an overlap measure, were calculated compared to the true estimate from the manual segmentation.

2.5. Image Registration

Rorden et al. [35] introduced a CT template based on 35 individuals who presented with specific neurological deficits that were suspected to be caused by a stroke, but were later found to be due to a metabolic abnormality. This CT template is represented in MNI (Montreal Neurological Institute) space and brain-extraction was performed on the template. Prior to image processing, brain-extracted images were registered to this brain-extracted template using a rigid-body (6 degrees of freedom) and linearly interpolated to a $1 \times 1 \times 1$mm voxel resolution. After interpolation, transformed hemorrhage masks and brain masks are not binary. These transformed masks were re-thresholded using a value of 0.5 to preserve mask volume [36]. Using a nearest-neighbor interpolation for these binary images after registration could also be used as a simpler approach and should be relatively equivalent to the re-thresholding.

This rigid registration does not ensure brains are the same size or that voxels match across subjects. Having voxels registered across subjects is not necessary for our model, as we do not incorporate voxel-level spatial information into the model. Although brains with different shapes and sizes may map to different areas in the template space, the goals of this registration are to reorient the image, ensure isotropic voxel sizes for smoothing and other operations described below, and preserve the relative volume of the ICH. All image preprocessing and analysis are done in MNI space, described as template space, unless otherwise specified.

2.6. Brain Mask Erosion

After registration, each brain mask was eroded by a box kernel ($3 \times 3 \times 1$mm). Though this erosion may exclude voxels from superficial bleeds towards the cortical surface, it excludes voxels with similar ranges as ICH voxels, caused by 1) incomplete skull stripping or 2) partial voluming effects with the skull. If any voxels from the hemorrhage mask were removed due to brain extraction or brain mask erosion, these voxels were included in estimating model performance but their predicted probability of ICH was set to 0. Therefore, these deleted ICH voxels will always be false negatives in our approach.

2.7. Imaging Predictors

We derived a set of imaging predictors from each CT scan. We will describe each predictor together with the rationale for their use. These features make up the potential set of predictors (features) for image segmentation. Below we provide the definition of these predictors, while Figure 2 displays them for one axial slice of one subject.

2.7.1. CT voxel intensity information

The first predictor is the raw voxel intensity value in HU denoted by $x(v)$. This is the main predictor used in visual inspection, with high HU values being indicative of hemorrhage. Based on the voxel intensity we have also created an indicator for the HU intensity value to be between 40 and 80 (inclusive), to mimic the criterion used for screening in manual segmentation. Also, these thresholds have been used in previous ICH semgentation work [21]. More precisely, we have introduced the predictor

$$I_{\text{thresh}}(v) = \begin{cases} 1 & \text{if } 40 \leq x(v) \leq 80 \\ 0 & \text{otherwise} \end{cases}$$

.

2.7.2. Local Moment Information

For each voxel, we extracted a neighborhood of voxels: all adjacent voxels along the 3 dimensions together with the voxel itself, indexed by k: $k = 1, \ldots, N(v) = 27$, where $N(v)$ is the number of voxels in the neighborhood. If $x_k(v)$ denotes the voxel intensity in HU for voxel neighbor k, then the local mean intensity is defined as:

$$\bar{x}(v) = \frac{1}{N(v)} \sum_{k \in N(v)} x_k(v). \tag{1}$$

We also calculate statistics based on higher order moments and define the local SD, skew, and kurtosis as:

$$\begin{aligned}
\text{SD}(v) &= \sqrt{\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^2} \\
\text{Skew}(v) &= \frac{\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^3}{\left[\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^2\right]^{3/2}} = \frac{\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^3}{SD(v)^3} \\
\text{Kurtosis}(v) &= \frac{\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^4}{\left[\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^2\right]^{2}} = \frac{\frac{1}{N(v)} \sum_{k \in N(v)} \{x_k(v) - \bar{x}(v)\}^4}{SD(v)^4}
\end{aligned}$$

We did not divide by $\{N(v) - 1\}$ in standard deviation and skew formula and did not subtract by 3 for kurtosis. As $N(v)$ is the same at every voxel, these simplified choices will have no effect on modeling or prediction.

Voxels with a larger local mean have higher HU neighboring voxels, which increases their likelihood to be in or adjacent to the ICH. The higher order moments can provide information about how homogeneous the intensities in the neighborhood are and where edges may be located. We also introduce the variable of the percentage of voxels in each neighborhood that have HU values between 40 and 80:

$$p_{\text{thresh}}(v) = \frac{1}{N(v)} \sum_{k \in N(v)} I\{40 \leq x_k(v) \leq 80\} \tag{2}$$

which should be higher for ICH voxels as they are surrounded by neighbors with higher HU values.

Voxels that are on the surface or are surrounded by non-brain tissue are less likely to be in the ICH. Thus, voxels not in eroded mask are set to 0. We also introduce the variable percentage of voxels that have neighbors of value of 0:

$$p_0(v) = \frac{1}{N(v)} \sum_{k \in N(v)} I\{x_k(v) = 0\}, \tag{3}$$

and an indicator of whether any voxels in the neighborhood had a value of 0:

$$\bar{I}_0(v) = I\{p_0(v) > 0\}. \tag{4}$$

The reason for introducing these predictors is that we expect that voxels that have neighbors with intensity zero are less likely to be ICH. Our approach will not assume that the probability of voxels with neighbors with HU intensity equal to zero are not in the ICH. Instead, we will model the probability of belonging to the ICH as a function of the predictors described in this section.

2.7.3. Within-plane Standard Scores

Some brain structures have high HU values but are not ICH, such as the falx cerebri, which lies largely on the mid-sagittal plane. Moreover, raw CT images may contain substantial inhomogeneity. For example, tissues closer to the top of the brain may have higher observed intensities (measured in HU) than those in the middle or bottom of the brain. Thus, if values are standardized within each plane (axial, sagittal, coronal), for each scan separately, the resulting plane-specific z-scores may discriminate better high relative values within the plane, which may attenuate the effect of HU intensity inhomogeneities.

Thus, for each voxel and slice (axial, sagittal, and coronal) planes, we defined

$$z_o(v) = \frac{x(v) - \bar{x}(v, o)}{\hat{\sigma}(v, o)} \tag{5}$$

where $o \in \{\text{axial, sagittal, and coronal}\}$, $\bar{x}(v, o)$ and $\hat{\sigma}(v, o)$ denote the mean and standard deviation of the intensities of voxels in the plane o that contains the voxel v, excluding voxels outside the brain mask. In addition to the standardized images within each plane we have also calculated standardized scores based on the Winsorized mean and standard deviation. More precisely, we used the same formula as in equation (5), we set any HU values below the 20^{th} percentile to the 20^{th} percentile value and above the 80^{th} percentile to the 80^{th} percentile value within that slice and calculated the slice-specific mean and standard deviation,. This approach is expected to be more robust to small and moderate artifacts in the image.

2.7.4. Initial Segmentation

A major advantage of our approach is that it can use the results of other segmentation algorithms as covariates in our model. Consider, for example, Atropos [37], a previously published, open source, general segmentation tool based on Markov random fields for image segmentation. Atropos combines an initial segmentation based on k-means with an expectation-maximization algorithm for a finite mixture model along with a Markov random field prior. We used Atropos to conduct a 4-tissue class segmentation which provides the probability for each class. We combined the top 2 probability classes into one class as Atropos orders the classes by the mean intensity and hemorrhages have higher HU values. This combined probability was then used as a predictor, denoted by Atropos(v). Although Atropos has been shown to perform well in other studies for tissue-class segmentation [37, 38], the Atropos segmentation did not perform adequately in our ICH CT data. However, using the Atropos segmentation probabilities as predictors can be done seamlessly in our approach. Similarly, the results of any other segmentation approach can be incorporated in our approach and the relative performance of methods can be compared.

2.7.5. Contralateral Difference Images

As most hemorrhages are constrained to one side of the brain, the contralateral side tends to have lower HU values. In contrast, for non-hemorrhage voxels, the contralateral voxels tend to have similar HU values due to the quasi-symmetry of the brain and its adjacent tissues such as bone. To take advantage of this property, we right-left flipped the registered image, and computed a difference image

$$f(v) = x(v) - x(v^*), \tag{6}$$

where v^* is the contralateral voxel of v.

2.7.6. Global Head Information

Another potential predictor was the distance to the center of the brain, $d(v)$, account for voxels that are far from the brain center but may contain artifacts. We also created 3 images by smoothing the original image using large Gaussian kernels ($\sigma = 5mm^3, 10mm^3, 20mm^3$) to account for potential heterogeneity in intensity. These smooth images we denoted by $s_5(v)$, $s_{10}(v)$ and $s_{20}(v)$, respectively.

2.7.7. Standardized-to-template Intensity

We have also incorporated predictors that contrast the scan HU intensities with those of an average brain obtained from healthy individuals. Using 30 CT images from non-stroke patients from Dr. Rorden (personal communication), we registered the brain-extracted scans to a CT template, and created a voxel-wise mean image M and voxel-wise standard deviation S image across registered images in template space. Each scan in our ICH study, we registered (using affine transformations followed by SyN [39]) it to the same CT template. We then created a standardized voxel intensity with respect to this population, z_{template}, using the following equation:

$$z_{\text{template}}(v) = \frac{x(v) - M(v)}{S(v)}$$

The image was then returned to the original space to align this predictor with the other predictors. This predictor is similar to that used in Gillebert, Humphreys, and Mantini [20] and the authors have shown that this predictor can detect voxels outside of a standard range such as the hemorrhage.

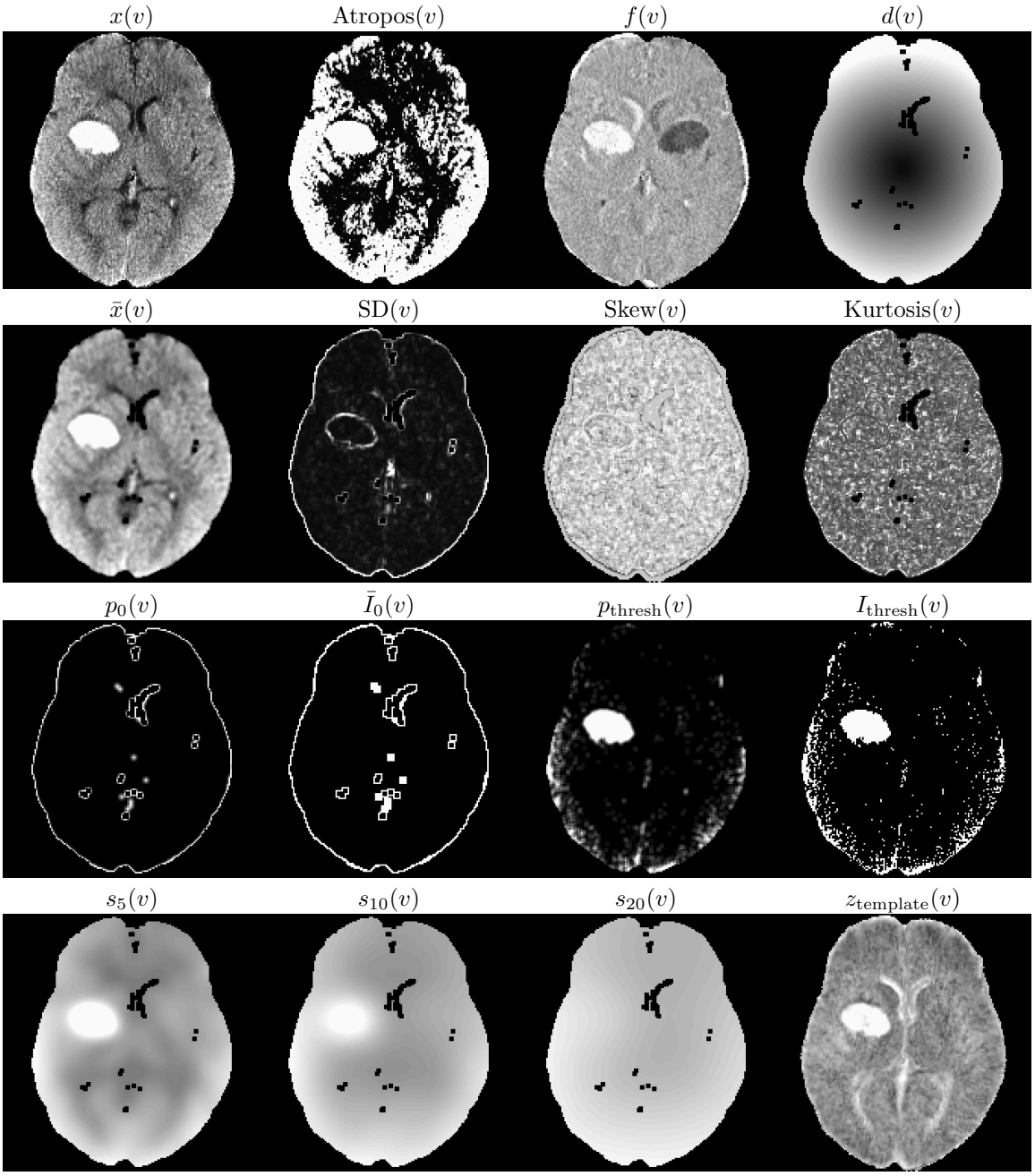

Figure 2: **Predictor Images**. Here we display one axial slice of predictor images from one patient. The within-plane standardized and Winsorized predictor images were not shown as they are within-subject scaled versions of the image $x(v)$ and appear very similar. Although they appear similar at a subject level, the distribution of these predictors is different across patients. Images that visually separate the areas of ICH compared to the rest of the images are likely to be better predictors.

2.8. Voxel Selection Procedure

We chose 10 scans from 10 patients to perform exploratory data analysis, model fitting, and estimation of model cutoffs; these data will be referred to as the training data. These scans were randomly selected. These selected patients had a mean (SD) hemorrhage volume of 37.8 (6) mL, had an average HU intensity of 58.8 HU and were scanned on the following scanners: GE (N = 6), Philips (N = 2), Siemens (N = 1), and Toshiba (N = 1). We used the 102 remaining scans as test data to evaluate the performance of the proposed approaches.

Using the training data, we estimated the 0.5% and 99.5% quantiles for all predictors across ICH voxels. The voxel selection procedure consisted of choosing all voxels that had all of predictors z_{axial}, z_{coronal}, and p_{thresh} within the corresponding 0.5 and 99.5 quantiles as well as values of HU intensity between 30 and 100. Voxels that did not meet these criteria were assigned a 0 probability of ICH. These cutoffs were found empirically to work well in the test scans. This approach excluded a mean of 63.6 (min: 37.1, max: 89.8) percentage of non-ICH voxels and included a mean of 97.9 (min: 91.6, max: 99.9) percentage of ICH voxels. We have found that this voxel selection procedure improves computational speed as well as the performance of the algorithms.

2.9. Models

Using the 10 training scans we obtained all voxels passing the voxel selection procedure described in Section 2.8. We then randomly sub-sampled 100,000 voxels, which were used for model fitting, model selection, and exploratory analysis, to reduce computational burden and increase speed of model fitting and exploratory analysis. The rest of the remaining voxels from the training data were used for model calibration. All models were fit with all the predictors described in Sections 2.7.

We fit several different models on the 100,000 sub-sampled voxels: 1) logistic regression with all covariates used as main effects (without interactions or nonlinear effects), 2) logistic regression model with a penalty on the model parameters to reduce the potential effect of high correlations between predictors, 3) generalized additive model (GAM) [40, 41], which is similar to the logistic regression, but allows for non-linear effects on the linear predictor scale, and 4) random forest classifier [42]. All models were fit using R.

For the standard and penalized logistic regression model, we used all predictors. The penalized model was fit using the LASSO (Least Absolute Shrinkage and Selection Operator) penalty [43] using the `glmnet` package [44]. The tuning parameter, λ, was chosen using 10-fold cross-validation on the training voxels; the cost function used was the misclassification rate. The parameter was chosen using the largest value of λ where that the misclassification rate is within 1 standard error of its minimum; this approach led to superior out-of-sample stability.

The generalized additive model (GAM) [40, 41] was also fit using indicator variables for binary variables and thin-plate splines for all continuous measures. The model was fit using fast-estimation of the restricted maximum likelihood (fREML) implemented in the `mgcv` package [45, 46]. Detailed model specifications are provided in Section 5.2.

The random forest [42] classification algorithm was implemented using the `randomForest` package in R [47] with the default pruning parameters and number of trees (`ntree=500`, `mtry=4`).

2.10. Estimating a Cutoff for Model Probability

Each model described in Section 2.9 provides an estimate of the probability for each voxel to be in ICH.

To choose probability thresholds to create a binary segmentation, we chose the Dice Similarity Index (DSI) [48] as a measure of the quality of segmentation. The DSI is a measure of overlap that is insensitive to areas where neither the true nor the predicted segmentation were labeled ICH, and will be used as a performance measure when comparing models on the test data. DSI for scan i is calculated by

$$DSI_i = \frac{2 \times TP}{2 \times TP + FN + FP}$$

where TP denotes the number of "true positive" voxels, where the manual and predicted segmentation agree that the voxel is in ICH, FP denotes the number of "false positive" voxels, where the predicted segmentation indicates that there is no lesion when the manual segmentation indicates lesion, and FN denotes the number of "false negative" voxels, where the predicted segmentation indicates that there is lesion when the manual

segmentation indicates that there is not. DSI ranges from 0 to 1, where 0 indicates no overlap and 1 denotes perfect overlap.

For each model, the probability image was smoothed by taking the average over the neighborhood voxels (1 voxel in every direction). For each threshold, we used the voxels in the training data that were not used for estimating the model, and found the threshold that maximized the DSI compared to the manual segmentation at those voxels. This threshold was applied to the smoothed probability maps to produce binary ICH images.

After thresholding the smoothed image using these DSI-optimized thresholds, we discarded regions with fewer than 100 (0.1mL) connected voxels. This removal was done to eliminate speckling, which helped improve the false positive rate of images. For each model, the predicted ICH binary mask was transformed back to the original (i.e. native) space using the inverse of the rigid-body transformation. As the linear interpolation associated with this step results in a non-binary mask, we thresholded the image at 0.5 to preserve volume [36].

2.11. Testing ICH Prediction and Measuring Model Performance

For each of the remaining 102 scans in the test data, the voxel selection described in Section 2.8 was applied. Using the prediction process described in Section 2.9 an ICH probability map was estimated using each of the four models. Using the probability thresholds calculated on the training data, as described in Section 2.10, we obtained a binary ICH map for each of the four methods. For evaluation, all results are provided in the native space of the patient, not in the template space, as the manual segmentation was done in the native space.

Model performance was evaluated on the validation data using the DSI and the ICH volume. The ICH volume was estimated counting the number of ICH voxels multiplied by product of the voxel sizes, divided by 1000 (1000 mm^3 per 1 mL), provided in mL. The large number of true negatives (non-ICH voxels) artificially inflates specificity and overall accuracy measures, which are not reported. The global equality median DSI across models was tested using the Kruskal-Wallis test. If a difference was present, we tested the null hypothesis of no difference in the medians for each combination of models (6 combinations) using a Wilcoxon signed-rank test, and corrected the p-value using a Bonferroni correction.

After choosing a single model, we explore the DSI over certain factors. First we will investigate the DSI by different scanner manufacturers and test median equality using the Kruskal-Wallis test, performing a similar pairwise Wilcoxon signed-rank test procedure with multiplicity correction. We similarly explore differences of DSI based on 3 categories of hemorrhage size (based on manual segmentation): using $0-30$mL, >30 to 60mL, and >60mL cutoffs, similar to Hemphill et al. [3].

The performance of total ICH volume prediction was compared to the manual segmentation using the Pearson correlation and root mean squared errors (RMSE) between volumes measures. Similarly, we performed the Kruskal-Wallis test for the null hypothesis of equality of medians of the absolute value of the difference between the estimated volume from each model and the true volume. If the null hypothesis is rejected, pairwise tests were conducted as in the case of DSI. For DSI and correlation, higher values indicate better agreement with the manual segmentation. For RMSE, lower values indicate better agreement.

3. Results

3.1. Dice Similarity Index

In Figure 3, we show the DSI distributions based on the test data for each model. DSI is high on average for all models, with a few scans having a very small DSI (i.e. failures). The median DSI for each model was: 0.89 (logistic), 0.885 (LASSO), 0.88 (GAM), and 0.899 (random forest). Using the random forest results in a slightly higher median DSI compared to the other models, and there was a statistically significant difference across medians ($\chi^2(3) = 13.49$, $p < 0.05$). Indeed, after Bonferroni correction, the hypothesis of equality of median DSI was rejected only when comparing the random forest DSI to the DSI from the logistic ($p < 0.001$), LASSO ($p < 0.001$), or GAM ($p < 0.001$) models. Based on visual inspection, the difference between the random forest and the logistic regression is quite small.

To better understand the DSI measurements in our data, Figure 4 displays the CT scan of the patient in the test data that has the median DSI in the test scans. The image depicts the brain-extracted CT scan and

the CT scan indicating different types of classification properties using overlaid colors. Green indicates a correct classification of ICH from the model (true positive), blue indicates a false negative, and red indicates a false positive. The image has a finer resolution along the axial plane (0.5mm by 0.5mm) than in the sagittal and coronal planes (5mm), as is commonly used for radiological evaluation of hemorrhages. Patients with the lowest, 25^{th}, 75^{th}, and highest DSI are shown in Supplemental Figures S1, S2, S3, and S4, respectively.

3.2. ICH Volume Estimation

In Figure 5, we show the estimated ICH volume versus that from the manual segmentation. The pink line represents the $X = Y$ line, where the estimated and true volume are identical. The blue line represents the linear fit; the estimated linear regression equation and correlation are printed on the plot. The farther away the slope of the equation is from 1 the larger the bias with values larger than 1 representing over-estimated volumes. The correlation (95% confidence interval (CI)) between the true volume and the predicted volume were 0.92 (95% CI: $0.88, 0.95$) for the logistic model, 0.92 $(0.88, 0.94)$ for the LASSO, 0.91 (95% CI: $0.87, 0.94$) for the GAM, and 0.93 (95% CI: $0.9, 0.95$) for the random forest. The RMSE for logistic (RMSE: 10.7 mL), LASSO (10.8 mL), and random forest (10.3 mL) models were relatively close, but was slightly higher for the GAM model (11.4 mL). The Kruskal-Wallis test indicated no significant difference in the median absolute value of the difference in estimated versus true volume over models ($\chi^2(3) = 2.3$, $p = 0.51$).

3.3. Model Choice

In Supplemental Table 1, we report the estimated coefficients, standard errors, and z-statistics from the logistic regression model. In Figure S5, we present the variable importance plot, representing the mean decrease in the Gini coefficient for each variable. The standardized-to-template and neighborhood mean seem to be the strongest predictors in the random forest. To reduce complexity and decrease computation time, we will remove some predictors in future implementations of this method, such as the thresholded image (I_{thresh}) as the this information can be encoded in thresholding the voxel intensity in a decision tree of the forest.

Overall, all models perform well for ICH segmentation. A small percentage (1.0%) of failures were observed with DSI < 0.5 ($N = 1$ out of 102 scans), but the random forest algorithm had a slightly higher median DSI, slightly lower RMSE, and a higher slightly correlation than the other models. Therefore, when implementing the algorithm, we will use the random forest model.

3.4. DSI by Scanner Manufacturer, Average HU, and Hemorrhage Size

In the following section, we will use the DSI from the random forest model. In the 102 test scans, the median DSI was 0.89 for patients scanned in a GE scanner, 0.9 for Philips, 0.9 for Toshiba, and 0.92 for Siemens (Supplemental Figure S6). The patient with a failed segmentation was scanned using a GE scanner. The Kruskal-Wallis test indicated a difference in medians ($\chi^2(3) = 11.6$, $p = 0.009$). After Bonferroni correction, the difference of DSI of patients scanned with Siemens versus GE scanners was the only statistically significant comparison ($W = 428$, $p = 0.008$, corrected).

In Figure 6, we see the effect of average HU over the hemorrhage versus DSI. We note the failed scan had a much lower average HU (44.4 HU) in the hemorrhage compared to the rest of the images (panel A). Overall, however, above 50 HU, there does not seem to be a strong effect on the average HU and DSI (panel B).

After categorization of the hemorrhage volume, 34 (33.3%) patients had a volume of $0 - 30$mL (small), 46 (45.1%) had volumes > 30 to 60mL (medium), and 22 (21.6%) had > 60mL (large). The Kruskal-Wallis test indicated a difference in medians ($\chi^2(2) = 6.5$, $p = 0.04$, see Supplemental Figure S7). After Bonferroni correction, there was no statistically significant difference in median DSI, but the strongest comparison was for the small versus medium hemorrhage sizes ($W = 537$, $p = 0.0502$, corrected).

4. Discussion

We have presented a novel, fully automated method for segmentation of ICH from CT scans. Our method uses only CT scans from patients with acute ICH, from the MISTIE II trial. MRI was not used because MRI procedures for ICH have not been standardized and are not performed per the standard of care for the

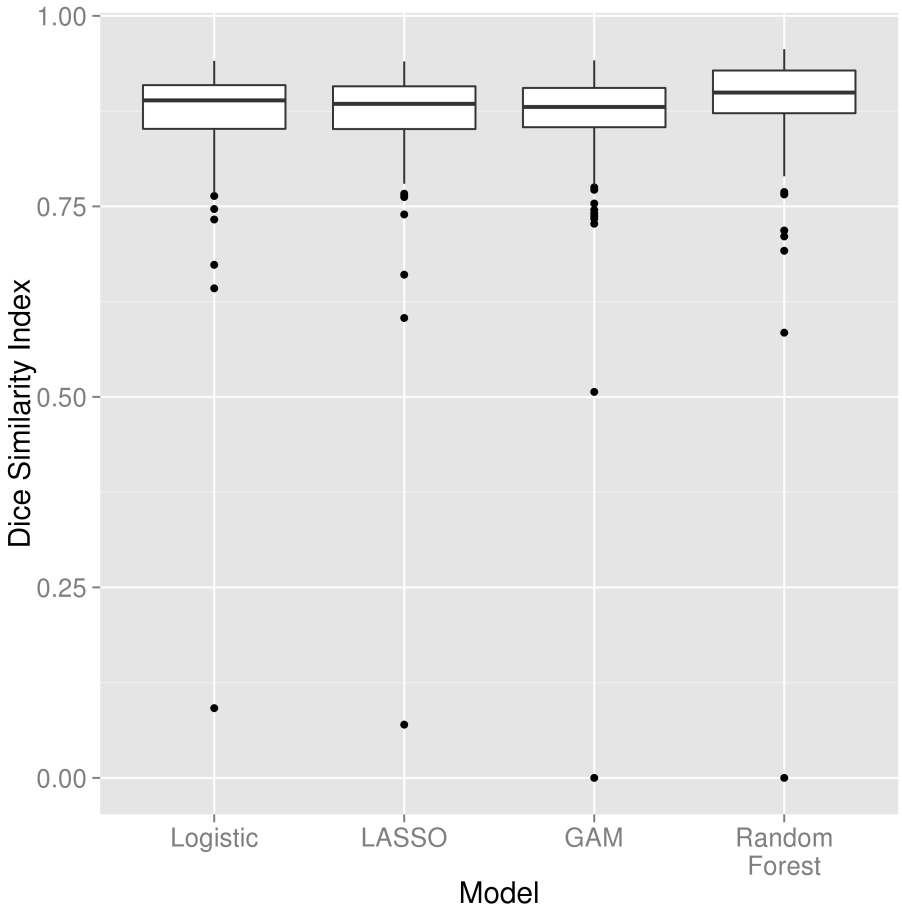

Figure 3: **Distribution of Dice Similarity Index in Test Scans**. Here we display the boxplot of the Dice Similarity Index (DSI), a measure of spatial overlap between the estimated hemorrhage mask and the manually delineated hemorrhage mask, in the 102 test scans. We present the DSI distribution for each model fit: a logistic regression, a logistic model penalized with the Least Absolute Shrinkage and Selection Operator penalty (LASSO), a generalized additive model (GAM), and a random forest algorithm. Overall, we see high agreement between the manual and estimated hemorrhage masks with the median of 0.89 for the logistic model, 0.885 for the LASSO, 0.88 for the GAM, and 0.899 for the random forest. The median DSI for the random forest was significantly higher than those of the other 3 models, after adjusting for multiplicity using a Bonferroni correction (all $p < 0.05$).

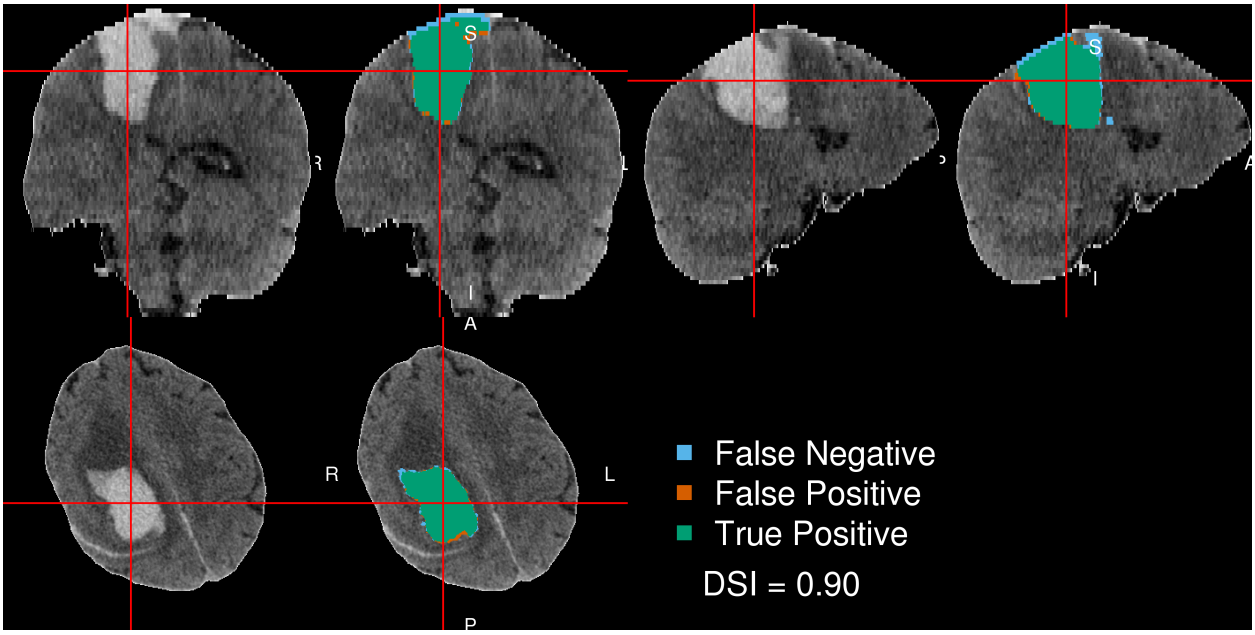

Figure 4: **Patient with Median Dice Similarity Index**. We present the patient with the median Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The median DSI was 0.899, which indicates high spatial overlap. The green indicates a correct classification of ICH from the model, blue indicates a false negative, where the manual segmentation denoted the area to be ICH but the predicted one did not, and red indicates a false positive, where the predicted segmentation denoted the area to be ICH but the manual one did not.

disease. We validated this method against manual segmentation. We used the Dice Similarity Index and correlation between the volume of ICH from manual and automatic segmentation as measures of algorithm performance.

We started by creating a rich set of predictors that are likely to capture the most discriminating features between ICH and non-ICH voxels and described the rationale for each predictor. Models of these predictors using logistic regression, logistic regression penalized with the LASSO, GAM, and random forests result in high values for the DSI and high correlations between the total ICH volume obtained from manual and automatic segmentations. Random forest was chosen as the algorithm, as it slightly outperformed the other approaches on the test data.

The approach failed on a very small subset of the validation set ($N = 1$ out of 102 with DSI below 0.5). Most discrepancies observed occur at the edges of the hemorrhage. This "edge effect" may be due to the isotropic, non edge-preserving smoothing. Anisotropic smoothing, as proposed by Perona, Shiota, and Malik [49], may improve segmentation.

Although we have shown good performance in patients with ICH, intraventricular hemorrhage (IVH) and subarachnoid hemorrhages (SAH) are two other forms of hemorrhagic stroke that may occur clinically. As obstructive IVH requiring external ventricular drainage was in the exclusion criteria in MISTIE II [25], no cases with primary IVH were included. The 3 cases with the largest overall volume had some considerable IVH extensions, the largest being around 20mL. This leads to an overall underestimation of the hemorrhage volume in our models. We believe the model may perform reasonably well with patients with IVH, yet may require a separate model, but not necessarily SAH.

Intraventricular hemorrhages have similar HU ranges for large portions of the hemorrhage, but may have lower HU values of the hemorrhage which are less dense and surrounded by cerebrospinal fluid. Moreover, cases with larger IVH volumes may cause larger deformations than seen in cases with ICH only, which may negatively affect any registration process. In contrast, for SAH, the process of mask erosion may considerably remove voxels from the hemorrhage, which occurs largely on the cortical surface. This step is important to remove false positives, which may occur from partial volume effects with areas of the skull and voxels

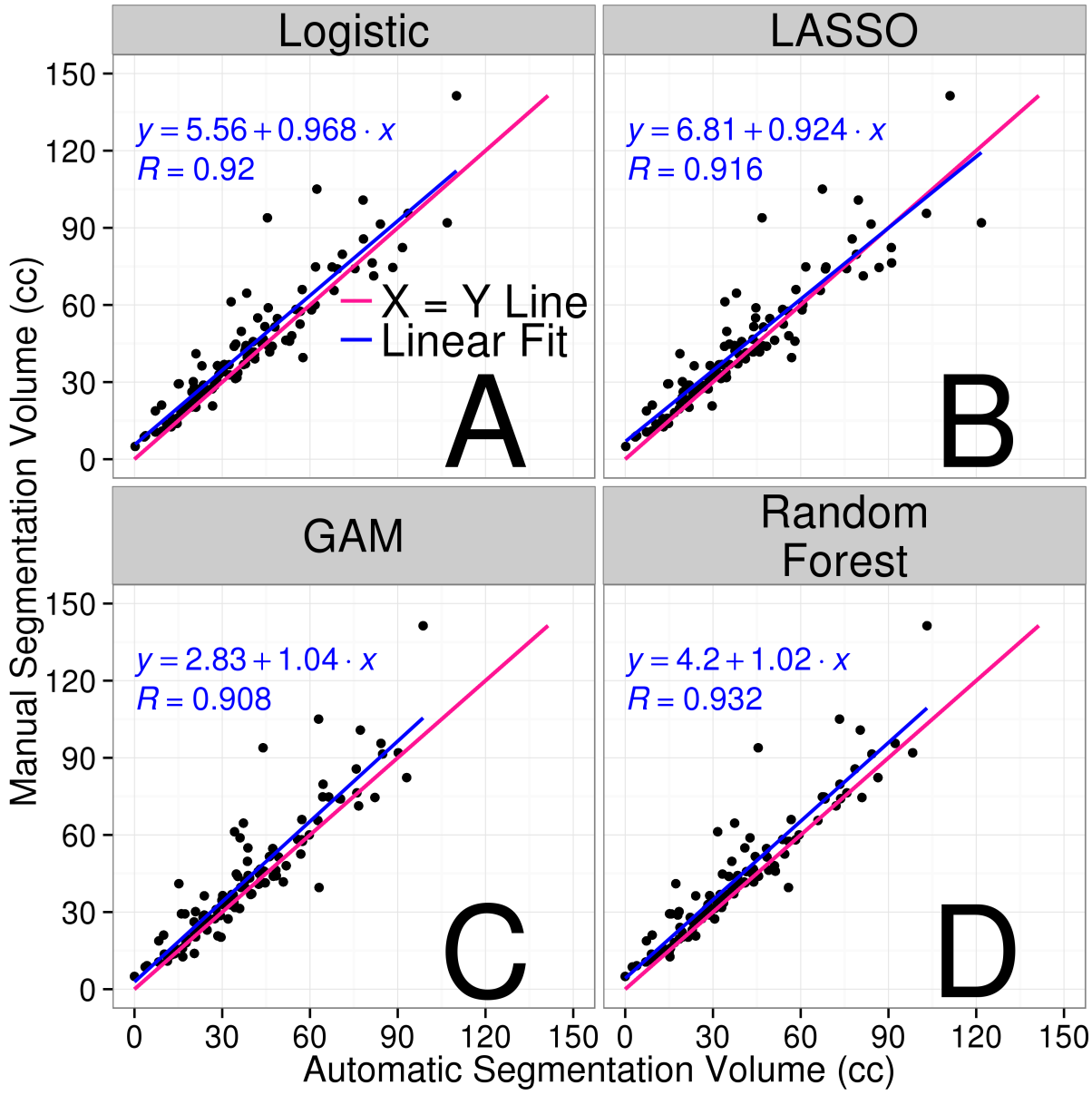

Figure 5: **Comparison of Estimated and Manual Intracerebral Hemorrhage Volume for Each Model**. In each panel, we show the volume of intracerebral hemorrhage (ICH) estimated from each model (x-axis) versus that from the gold-standard manual segmentation (y-axis) in the 102 test scans. The pink line represents the $X = Y$ line, which represents perfect agreement. The blue line represents a linear fit of the data, and the estimated slope equation is displayed along with the Pearson correlation. Panel A represents the volume from the logistic regression model, B represents that from logistic model penalized with the LASSO, C represents that from a generalized additive model (GAM), and D represents that from a random forest algorithm. Overall, we see high agreement between the estimated volumes from automated segmentation from each model as all correlations are above 0.9. The farther away the slope of the equation is from 1 represents a multiplicative bias, where values greater than 1 represents larger estimated volumes. The farther the intercept is from 0 represents and additive bias in the estimated volume, where values greater than 0 again represent larger estimated volumes.

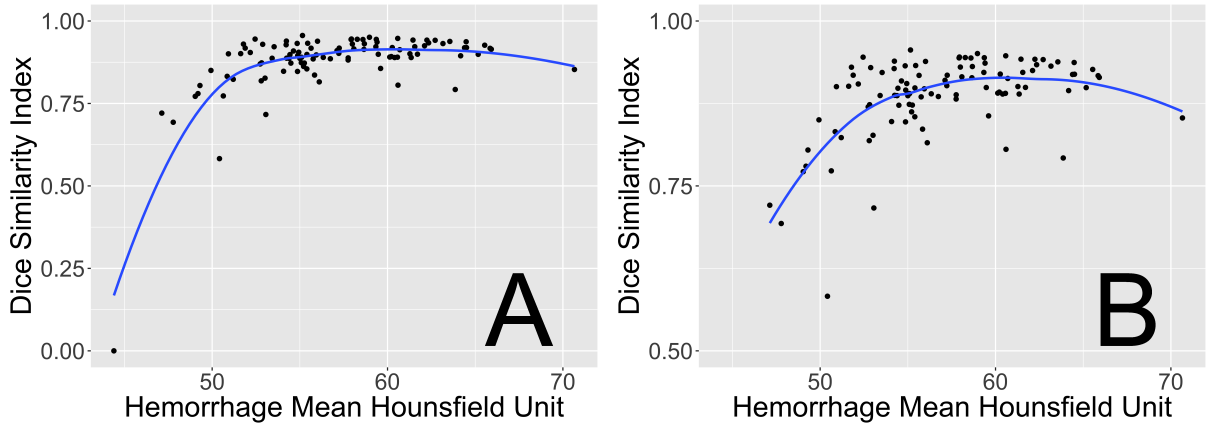

Figure 6: **Dice Similarity Index (DSI) by Average Voxel HU in the Hemorrhage**. These values represent the mean Hounsfield Unit (HU) over all voxels in the manual hemorrhage mask for each patient in the test data set versus the DSI for that patient compared to automatic segmentation. The blue line represents a locally-weighted regression smoother (loess). We see that the case with the lowest DSI had a much lower average HU in the hemorrhage. Overall, however, above 50 HU, there does not seem to be a strong effect on the average HU and DSI.

towards the cortical surface. If the methods for registration and erosion perform similarly to the cases with ICH, we can perform the same procedure and re-fit the model using additional cases of IVH and/or SAH.

Several other methods have been proposed for segmentation of ICH from CT scans [20, 21, 22, 23, 24]. Loncaric, Cosic, and Dhawan [22] performed the analysis on only one 2-dimensional scan and could not be compared with our approach. The method proposed by Pérez et al. [24] is semiautomated and was only validated by visual inspection. They reported segmentation failure in 6 out of 36 scans (16.7%) compared to 1 out of 102 scans (1.0%) for our method. Our reported median DSI (0.899) is much larger than the one reported by Gillebert, Humphreys, and Mantini [20] (approximately 0.62 and 0.78 for hemorrhagic strokes as read from their graphs). Our results were comparable to those reported by Prakash et al. [21] (0.897, 0.858, and 0.9173 for different groups with hemorrhage). Loncaric et al. [23] did not compare hemorrhage and manual segmentation masks; instead they compared ICH volumes from 5 subjects measured at 3 time points. Their reported Pearson correlation was 0.917 for the 15 scans, similar to our results using the random forest ($R = 0.93$).

Only Gillebert, Humphreys, and Mantini [20] responded to our requests for segmentation software to perform the segmentation and we could not find any software online. Although the method of Gillebert, Humphreys, and Mantini [20] is comparable to ours, their approach has not been packaged for general use. We have released an open source package that can perform ICH segmentation (`https://github.com/muschellij2/ichseg`). Our software includes the models for prediction, the CT template from Rorden et al. [35], template-level standardized mean and standard deviation images, as well as functions to register the images, create the predictors, predict from the models, and return a binary hemorrhage mask. Although an R package is ideal for prediction on a large number of images and for researchers who prefer scripting, releasing easy to use graphic user interfaces may increase the appeal of the methods proposed. Therefore, we have also released a Shiny [50] R application online (`http://johnmuschelli.com/ich_segment_all.html`) that takes an input CT scan and outputs ICH segmentation mask and provides a representation of each processing step.

A potential issue for CT images that contain ICH is image registration. Indeed, methods developed for registration of healthy brains can fail in brains exhibiting pathology. The only predictor that used non-linear registration was the standardized-to-template intensity. The potential problems associated with this transformation are mitigated by the transformation back to the native space. Thus, we use non-linear registration, but do not rely on a highly accurate image registration to template to compare voxels across patients; instead we use registration to obtain potentially noisy predictors in the native space.

Another potential concern could be that training data consisted of only 10 patient scans and using only

100,000 randomly sub-sampled voxels that passed the voxel selection procedure from these scans. Remarkably, the models have shown to have high out-of-sample accuracy. The training and test sets were kept unchanged to avoid overfitting the models to the test set. Validation of the method on additional data would be useful, while rater studies may provide more insight into the clinical differences between various segmentation approaches. However, we would like to note that our data was highly heterogeneous and contains test scans from multiple sites and scanners. Moreover, the location and size of hemorrhages is also highly heterogeneous. Thus, we expect our method to have a good out-of-sample accuracy in a heterogeneous population of CT images with ICH.

The proposed approach provides estimated binary hemorrhage masks, which can be used to automatically estimate quantitative measures of hemorrhage location [51]. Our results would also allow automated shape analysis, which require a binary mask. The subject-specific hemorrhage masks can be used for other voxel-based analyses that could yield novel insights into the relationship between hemorrhage characteristics and patient outcomes.

4.1. Conclusions

We have implemented and validated a fully automated segmentation algorithm of ICH in CT scans and published the associated software both as an R package and as a GUI. The method relies on a series of processing steps and on creating a set of relevant predictors. This method has been shown to have very good agreement with the gold standard of manual delineation of hemorrhages. As an automated process, it is much faster, does not require extensive radiologic image experience, is scalable to thousands of images, and does not have inter-reader variability. As our methods and software produce binary hemorrhage masks that can be localized both in the native and template space [51], quantitative voxel- and region-level analyses could be conducted to assess the association between ICH characteristics and health outcomes. Methods also provide an estimator of the ICH volume, which can be used in standard statistical analyses, as it has been shown to be associated with long-term functional outcomes [2, 4, 52].

Acknowledgments

We would like to thank the patients and families who volunteered for this study, Genentech Inc. for the donation of the study drug (Alteplase), and the readers who manually segmented the ICH (W. Andrew Mould, Tim Morgan, Natalie Ullman, Saman Nekoovaght-Tak). Dr. Chris Rorden was also extremely helpful in adapting his `dcm2nii` software to some issues specific to CT scans and the data for the population moment images.

Sources of Funding

The project described was supported by the NIH grant RO1EB012547 from the National Institute of Biomedical Imaging and Bioengineering, T32AG000247 from the National Institute on Aging, R01NS046309, RO1NS060910, RO1NS085211, R01NS046309, U01NS080824 and U01NS062851 from the National Institute of Neurological Disorders and Stroke, and RO1MH095836 from the National Institute of Mental Health. Minimally Invasive Surgery and rt-PA in ICH Evacuation Phase II (MISTIE II) was supported by grants R01NS046309 and U01NS062851 awarded to Dr. Daniel Hanley from the National Institutes of Health (NIH)/National Institute of Neurological Disorders and Stroke (NINDS). Minimally Invasive Surgery and rt-PA in ICH Evacuation Phase III (MISTIE III) is supported by the grant U01 NS080824 awarded to Dr. Daniel Hanley from the National Institutes of Health (NIH)/National Institute of Neurological Disorders and Stroke (NINDS). Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Hemorrhage Phase III (CLEAR III) is supported by the grant U01 NS062851 awarded to Dr. Daniel Hanley from the National Institutes of Health (NIH)/National Institute of Neurological Disorders and Stroke (NINDS).

References

[1] Ramandeep Sahni and Jesse Weinberger. "Management of intracerebral hemorrhage". In: *Vascular Health and Risk Management* 3.5 (Oct. 2007), pp. 701–709.

[2] J. P. Broderick et al. "Volume of intracerebral hemorrhage. A powerful and easy-to-use predictor of 30-day mortality." In: *Stroke* 24.7 (July 1, 1993), pp. 987–993.

[3] J. Claude Hemphill et al. "The ICH Score A Simple, Reliable Grading Scale for Intracerebral Hemorrhage". In: *Stroke* 32.4 (Apr. 1, 2001), pp. 891–897.

[4] Stanley Tuhrim et al. "Volume of ventricular blood is an important determinant of outcome in supratentorial intracerebral hemorrhage". In: *Critical care medicine* 27.3 (1999), pp. 617–621.

[5] Craig S. Anderson et al. "Intensive blood pressure reduction in acute cerebral haemorrhage trial (INTERACT): a randomised pilot trial". In: *The Lancet Neurology* 7.5 (2008), pp. 391–399.

[6] Craig S. Anderson et al. "Effects of Early Intensive Blood Pressure-Lowering Treatment on the Growth of Hematoma and Perihematomal Edema in Acute Intracerebral Hemorrhage The Intensive Blood Pressure Reduction in Acute Cerebral Haemorrhage Trial (INTERACT)". In: *Stroke* 41.2 (Feb. 1, 2010), pp. 307–312.

[7] Adnan I. Qureshi et al. "Association of Serum Glucose Concentrations During Acute Hospitalization with Hematoma Expansion, Perihematomal Edema, and Three Month Outcome Among Patients with Intracerebral Hemorrhage". In: *Neurocritical Care* 15.3 (Dec. 1, 2011), pp. 428–435.

[8] Stephan A. Mayer et al. "Recombinant Activated Factor VII for Acute Intracerebral Hemorrhage". In: *New England Journal of Medicine* 352.8 (Feb. 24, 2005), pp. 777–785.

[9] T. Morgan et al. "Preliminary findings of the minimally-invasive surgery plus rtPA for intracerebral hemorrhage evacuation (MISTIE) clinical trial". In: *Cerebral Hemorrhage*. Springer, 2008, pp. 147–151.

[10] T Morgan et al. "Preliminary report of the clot lysis evaluating accelerated resolution of intraventricular hemorrhage (CLEAR-IVH) clinical trial". In: *Cerebral Hemorrhage*. Springer, 2008, pp. 217–220.

[11] Natalia S. Rost et al. "Prediction of Functional Outcome in Patients With Primary Intracerebral Hemorrhage The FUNC Score". In: *Stroke* 39.8 (Aug. 1, 2008), pp. 2304–2309.

[12] M. Castellanos et al. "Predictors of good outcome in medium to large spontaneous supratentorial intracerebral haemorrhages". In: *Journal of Neurology, Neurosurgery & Psychiatry* 76.5 (May 1, 2005), pp. 691–695.

[13] Rashmi U. Kothari et al. "The ABCs of Measuring Intracerebral Hemorrhage Volumes". In: *Stroke* 27.8 (Aug. 1, 1996), pp. 1304–1305.

[14] Alastair J. S. Webb et al. "Accuracy of the ABC/2 Score for Intracerebral Hemorrhage Systematic Review and Analysis of MISTIE, CLEAR-IVH, and CLEAR III". In: *Stroke* 46.9 (Sept. 1, 2015), pp. 2470–2476.

[15] Afshin A. Divani et al. "The ABCs of accurate volumetric measurement of cerebral hematoma". In: *Stroke* 42.6 (2011), pp. 1569–1574.

[16] Salvador Pedraza et al. "Reliability of the ABC/2 Method in Determining Acute Infarct Volume". In: *Journal of Neuroimaging* 22.2 (2012), pp. 155–159.

[17] Haitham M. Hussein et al. "Reliability of Hematoma Volume Measurement at Local Sites in a Multicenter Acute Intracerebral Hemorrhage Clinical Trial". In: *Stroke* 44.1 (Jan. 1, 2013), pp. 237–239.

[18] Shuo Wang et al. "Hematoma Volume Measurement in Gradient Echo MRI Using Quantitative Susceptibility Mapping". In: *Stroke* 44.8 (Aug. 1, 2013), pp. 2315–2317.

[19] Ricardo J Carhuapoma et al. "Brain Edema After Human Cerebral Hemorrhage A Magnetic Resonance Imaging Volumetric Analysis". In: *Journal of neurosurgical anesthesiology* 15.3 (2003), pp. 230–233.

[20] Céline R. Gillebert, Glyn W. Humphreys, and Dante Mantini. "Automated delineation of stroke lesions using brain CT images". In: *NeuroImage: Clinical* 4 (2014), pp. 540–548.

[21] K. N. Bhanu Prakash et al. “Segmentation and quantification of intra-ventricular/cerebral hemorrhage in CT scans by modified distance regularized level set evolution technique”. In: *International Journal of Computer Assisted Radiology and Surgery* 7.5 (Sept. 1, 2012), pp. 785–798.

[22] Sven Loncaric, Dubravko Cosic, and Atam P. Dhawan. “Hierarchical segmentation of CT head images”. In: *Proc IEEE EMBS. doi* 10 (1996), p. 1109.

[23] Sven Loncaric et al. “Quantitative intracerebral brain hemorrhage analysis”. In: *Medical Imaging'99*. International Society for Optics and Photonics, 1999, pp. 886–894.

[24] Noel Pérez et al. “Set of methods for spontaneous ICH segmentation and tracking from CT head images”. In: *Progress in Pattern Recognition, Image Analysis and Applications*. Springer, 2007, pp. 212–220.

[25] W. Andrew Mould et al. “Minimally Invasive Surgery Plus Recombinant Tissue-type Plasminogen Activator for Intracerebral Hemorrhage Evacuation Decreases Perihematomal Edema”. In: *Stroke* 44.3 (Mar. 1, 2013), pp. 627–634.

[26] M Bergström et al. “Variation with time of the attenuation values of intracranial hematomas”. In: *Journal of computer assisted tomography* 1.1 (Jan. 1977), pp. 57–63.

[27] Eric E. Smith, Jonathan Rosand, and Steven M. Greenberg. “Imaging of Hemorrhagic Stroke”. In: *Magnetic Resonance Imaging Clinics of North America* 14.2 (May 2006), pp. 127–140.

[28] Chris Rorden and Matthew Brett. “Stereotaxic Display of Brain Lesions”. In: *Behavioural Neurology* 12.4 (2000), pp. 191–200.

[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015.

[30] John Muschelli et al. “fslr: Connecting the FSL Software with R”. In: *R Journal* 7.1 (2015), pp. 163–175.

[31] Mark Jenkinson et al. “FSL”. In: *NeuroImage* 62.2 (Aug. 15, 2012), pp. 782–790.

[32] Brian B. Avants et al. “A reproducible evaluation of ANTs similarity metric performance in brain image registration”. In: *NeuroImage* 54.3 (Feb. 1, 2011), pp. 2033–2044.

[33] Stephen M. Smith. “Fast robust automated brain extraction”. In: *Human Brain Mapping* 17.3 (2002), pp. 143–155.

[34] John Muschelli et al. “Validated automatic brain extraction of head CT images”. In: *NeuroImage* (2015).

[35] Christopher Rorden et al. “Age-specific CT and MRI templates for spatial normalization”. In: *NeuroImage* 61.4 (July 16, 2012), pp. 957–965.

[36] *Frequently Asked Questions for FLIRT*. `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT/FAQ`.

[37] Brian B Avants et al. “An open source multivariate framework for n-tissue segmentation with evaluation on public data”. In: *Neuroinformatics* 9.4 (2011), pp. 381–400.

[38] Bjoern H Menze et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *Medical Imaging, IEEE Transactions on* 34.10 (2015), pp. 1993–2024.

[39] B. B. Avants et al. “Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain”. In: *Medical Image Analysis*. Special Issue on The Third International Workshop on Biomedical Image Registration - WBIR 2006 12.1 (Feb. 2008), pp. 26–41.

[40] Trevor Hastie and Robert Tibshirani. “Generalized additive models”. In: *Statistical science* (1986), pp. 297–310.

[41] Trevor J. Hastie and Robert J. Tibshirani. *Generalized additive models*. Vol. 43. CRC Press, 1990.

[42] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[43] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

[44] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of statistical software* 33.1 (2010), pp. 1–22.

[45] Simon N. Wood. "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1 (2011), pp. 3–36.

[46] Simon N. Wood, Yannig Goude, and Simon Shaw. "Generalized additive models for large data sets". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1 (Jan. 1, 2015), pp. 139–155.

[47] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22.

[48] Lee R. Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

[49] Pietro Perona, Takahiro Shiota, and Jitendra Malik. "Anisotropic diffusion". In: *Geometry-driven diffusion in computer vision*. Springer, 1994, pp. 73–92.

[50] Winston Chang et al. *shiny: Web Application Framework for R*. 2015.

[51] John Muschelli et al. "Quantitative Intracerebral Hemorrhage Localization". In: *Stroke* 46.11 (2015), pp. 3270–3273.

[52] Lori C Jordan, Jonathan T Kleinman, and Argye E Hillis. "Intracerebral hemorrhage volume predicts poor neurologic outcome in children". In: *Stroke* 40.5 (2009), pp. 1666–1671.

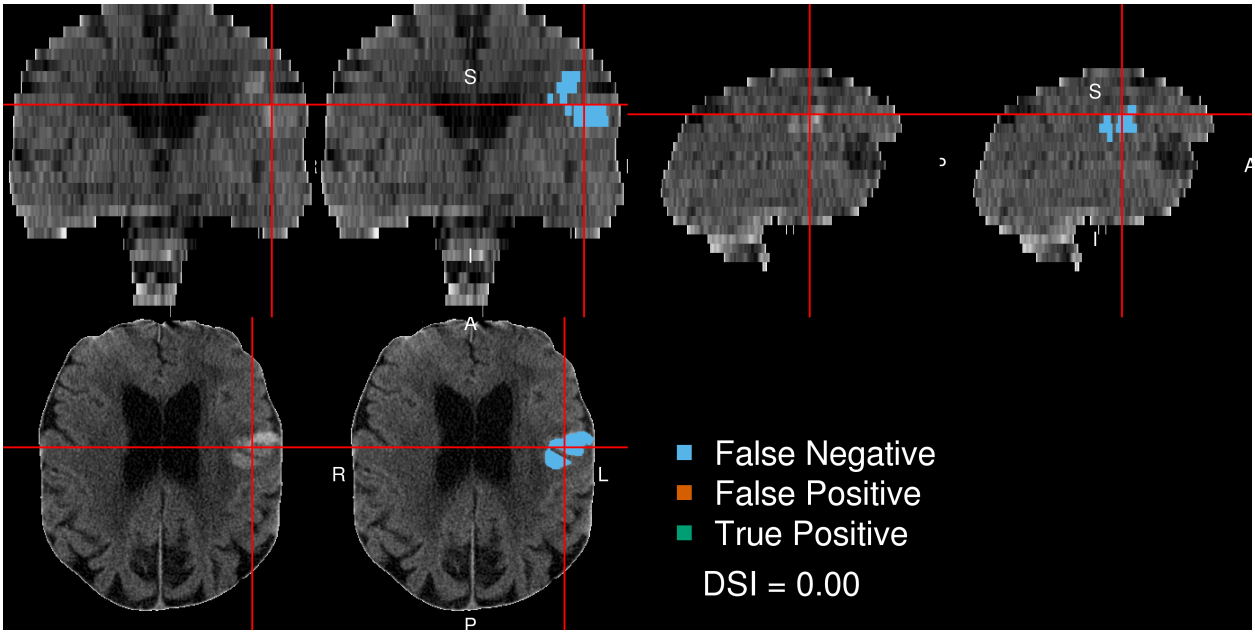

Figure S1: **Patient with Lowest Dice Similarity Index**. We present the patient with the lowest Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The lowest DSI was 0. The green indicates a correct classification of ICH from the model, blue indicates a false negative, where the manual segmentation denoted the area to be ICH but the predicted one did not, and red indicates a false positive, where the predicted segmentation denoted the area to be ICH but the manual one did not.

5. Supplemental Material

5.1. Examples of Dice Similarity Index in Test Scans

5.2. Model Specification

Let $Y_i(v)$ represent the binary hemorrhage mask indicator for voxel v, from patient i, and $x_{i,v}(k)$ represent the predictor image for image j, $j = 1, \ldots 21$.

$$\text{logit}\left(P(Y_i(v) = 1)\right) = \beta_0 + \sum_{j=1}^{21} x_{i,j}(v)\beta_j$$

The coefficients for the logistic model are (in log odds or log odds ratios):

The specification for the functional form of the model fit with the LASSO penalty, is the same, but optimizes the following criteria (`https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#log`):

$$\min_{\beta} - \left(\frac{1}{\sum_i V_i} \sum_i Y_i(v) \times X_i(v)\beta - \log\left(1 + e^{X_i(v)\beta}\right) \right) + \lambda \sum_k |\beta_k|$$

5.3. Variable Importance Plot

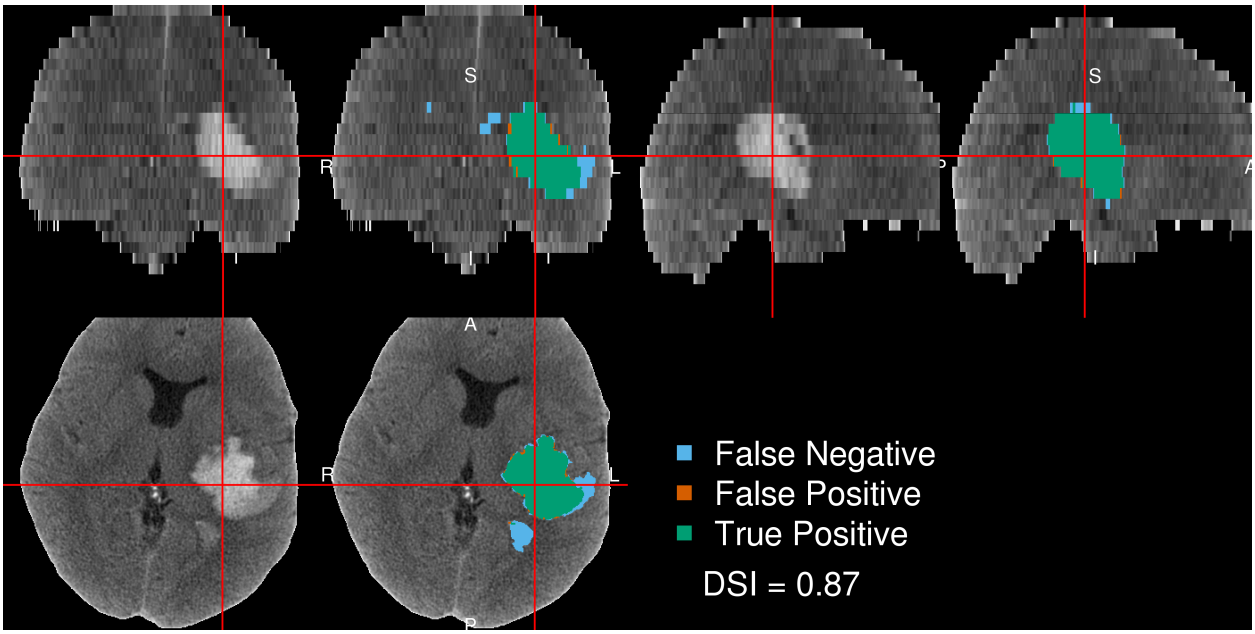

Figure S2: **Patient with 25$^{\text{th}}$ Quantile Dice Similarity Index**. We present the patient with the 25$^{\text{th}}$ quantile Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The 25$^{\text{th}}$ quantile DSI was 0.87. The green indicates a correct classification of ICH from the model, blue indicates a false negative, where the manual segmentation denoted the area to be ICH but the predicted one did not, and red indicates a false positive, where the predicted segmentation denoted the area to be ICH but the manual one did not.

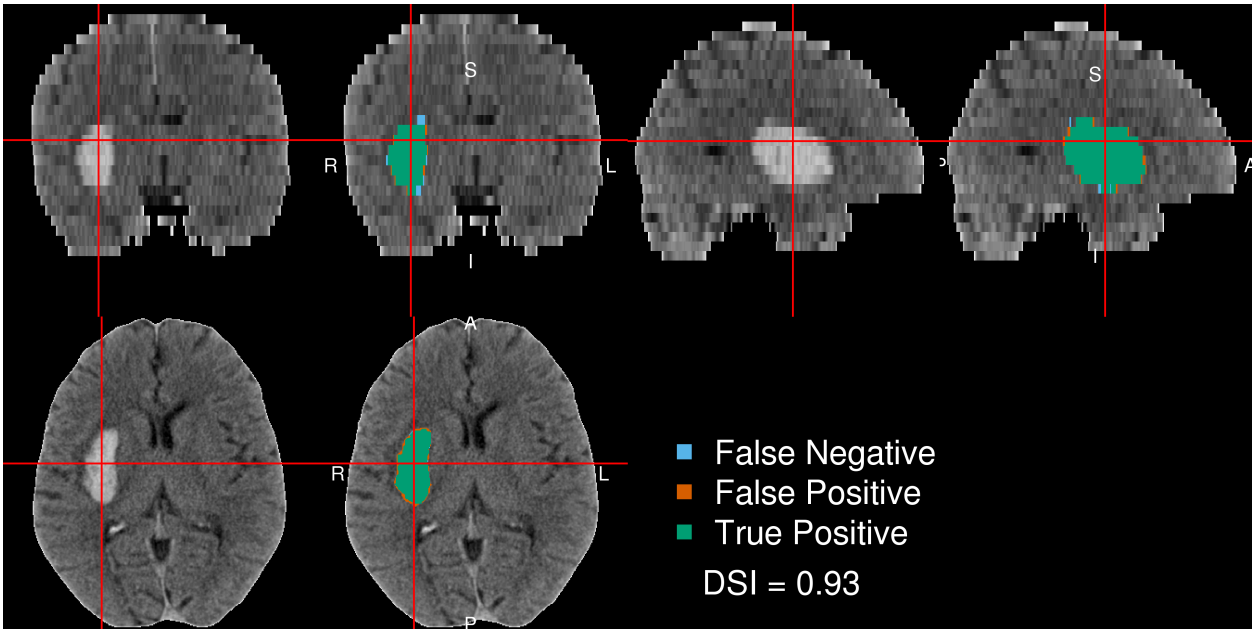

Figure S3: **Patient with 75$^{\text{th}}$ Quantile Dice Similarity Index**. We present the patient with the 75$^{\text{th}}$ quantile Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The 75$^{\text{th}}$ quantile DSI was 0.93. The green indicates a correct classification of ICH from the model, blue indicates a false negative, where the manual segmentation denoted the area to be ICH but the predicted one did not, and red indicates a false positive, where the predicted segmentation denoted the area to be ICH but the manual one did not.

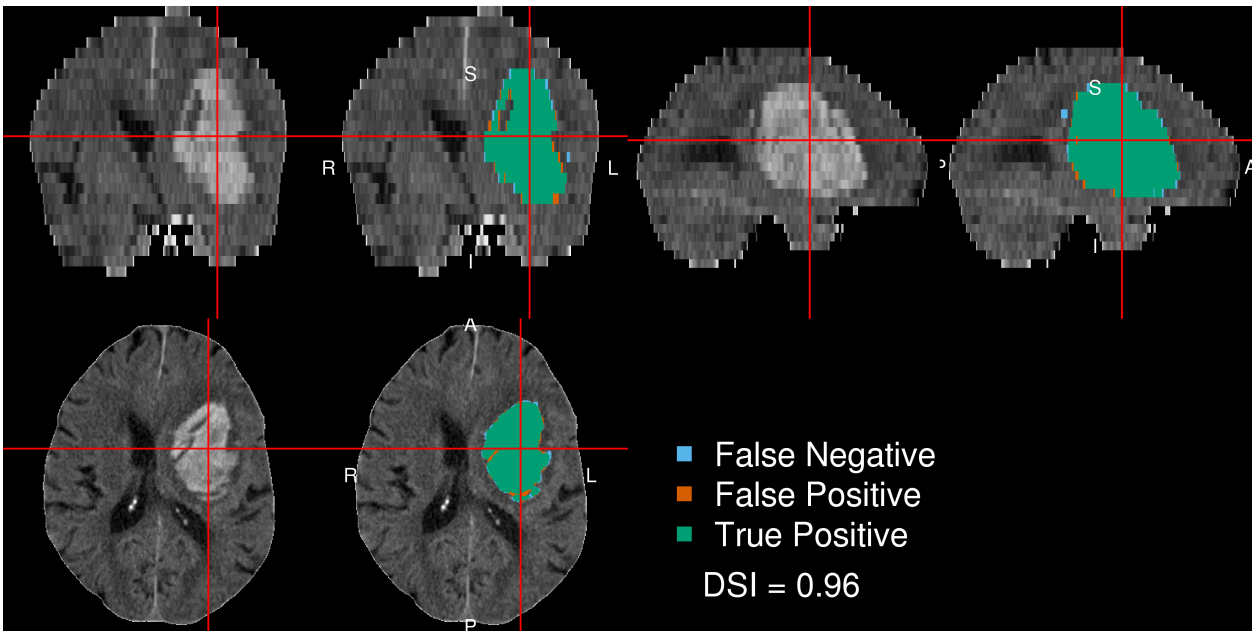

Figure S4: **Patient with Highest Dice Similarity Index**. We present the patient with the highest Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The highest DSI was 0.96. The green indicates a correct classification of ICH from the model, blue indicates a false negative, where the manual segmentation denoted the area to be ICH but the predicted one did not, and red indicates a false positive, where the predicted segmentation denoted the area to be ICH but the manual one did not.

Predictor	Beta	SE	Z
Intercept	1.008	0.331	3.046
Neighborhood mean	0.051	0.010	4.964
Neighborhood sd	0.000	0.000	0.304
Neighborhood skew	0.065	0.046	1.415
Neighborhood kurtosis	-0.352	0.026	-13.357
Image intensity (HU)	-0.172	0.012	-14.741
Threshold ($\geq$ 40 and $\leq$ 80)	-0.151	0.072	-2.090
Within-plane coronal	-0.632	0.050	-12.537
Within-plane sagittal	-0.249	0.057	-4.381
Within-plane axial	1.037	0.056	18.354
Winsorized standardized (20% trim)	0.547	0.041	13.518
Percentage thresholded neighbors	2.061	0.172	11.955
Atropos probability image	0.150	0.092	1.635
Percent of zero neighbors	-9.180	1.437	-6.387
Indicator of any zero neighbors	0.071	0.345	0.205
Distance to image centroid	-0.087	0.002	-45.265
Gaussian smooth ($\sigma = 5\text{mm}^3$)	-0.051	0.014	-3.591
Gaussian smooth ($\sigma = 10\text{mm}^3$)	0.550	0.022	25.416
Gaussian smooth ($\sigma = 20\text{mm}^3$)	-0.390	0.020	-19.757
Standardized-to-template intensity	1.460	0.034	43.100
Contralateral difference	0.033	0.002	17.530

Table 1: Beta coefficients (log odds ratio) for the logistic regression model for all coefficients. Combining these for each voxel value and using the inverse logit transformation yields the probability that voxel is ICH. After smoothing by 1 voxel in all 3 directions, the probability cutoff for thresholding was 0.5481. We note the standardized-to-template intensity and the neighborhood mean appear to be the strongest predictors.

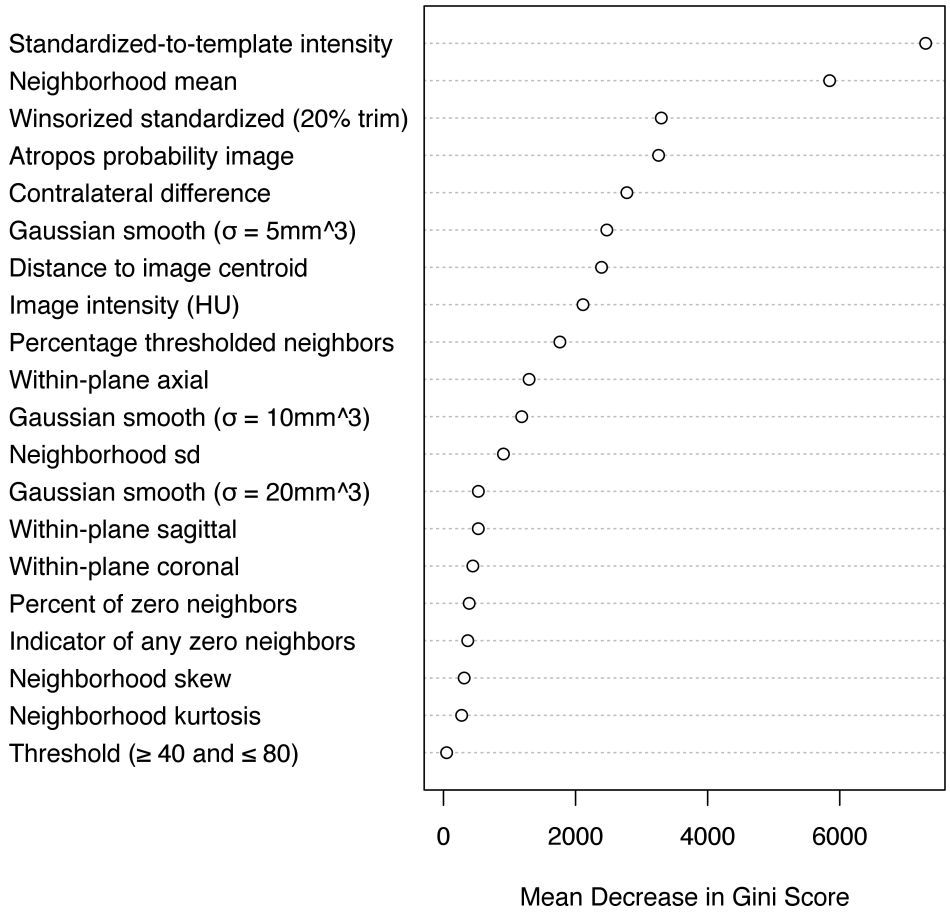

Figure S5: **Variable Importance Plot of Random Forest Classifier**. These numbers represent the mean decrease in the Gini coefficient in the random forest classifier for all coefficients. After smoothing by 1 voxel in all 3 directions, the probability cutoff for thresholding with this classifier was 0.509.

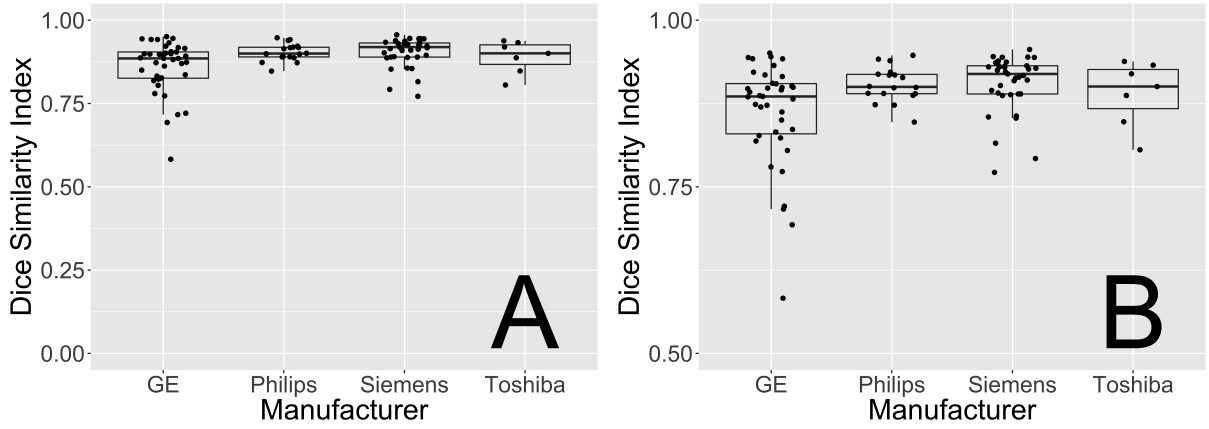

Figure S6: **Dice Similarity Index (DSI) by CT Scanner Manufacturer**. Here we present the DSI for all patients in the test set for the different scanner manufacturers (panel A). We note that the failed segmentation was a patient scanned with a GE scanner. We present the same data in panel B without that patient to illustrate the distributions of the DSI by scanner (note the y-axis begins at 0.5 DSI). All median DSI is relatively high. We see the lowest median DSI for patients scanned in GE scanners, comparable median DSI for Toshiba and Philips, slightly higher DSI for those scanned in a Siemens machine.

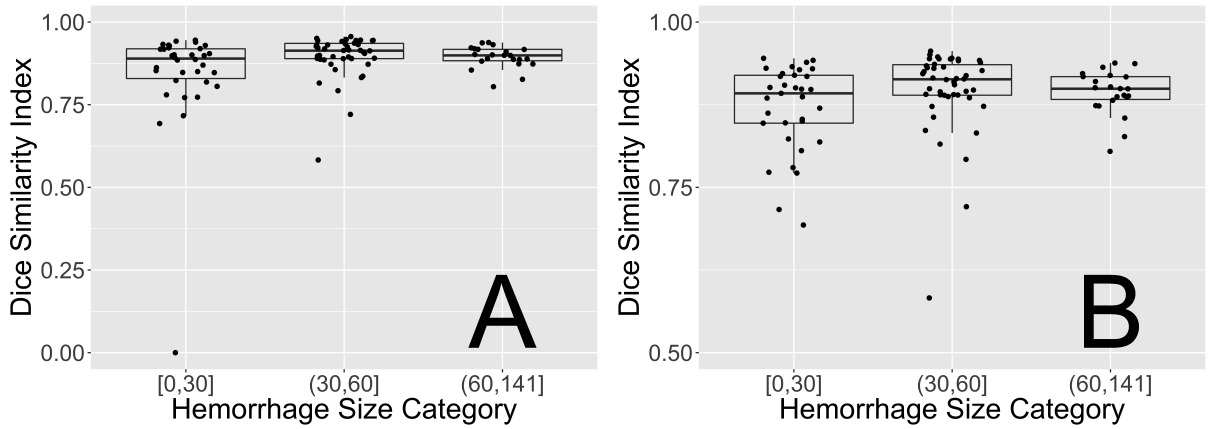

Figure S7: **Dice Similarity Index (DSI) by Hemorrhage Volume Category**. Here we present the DSI for all patients in the test set for the different hemorrhage voxel categories, described above (panel A). We will denote the categories as small, medium and large. We note that the failed segmentation was in the small category. We present the same data in panel B without that patient to illustrate the distributions of the DSI by category (note the y-axis begins at 0.5 DSI). All median DSI is relatively high. We see the lowest median DSI for patients with small hemorrhages, followed by large hemorrhages, and the highest median DSI is in the medium hemorrhages.