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

John Muschelli[a,*], Elizabeth M. Sweeney[a], 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

Keywords: CT, ICH Segmentation

1. Introduction

Intracerebral hemorrhage (ICH) is a neurological condition that results from a blood vessel rupturing into the tissue and possibly 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 [44]. The volume of ICH has been consistently demonstrated to be an important diagnostic predictor of stroke severity, long-term functional outcome, and mortality [8, 19, 48]. ICH volume change is also common primary outcome [1, 2, 27, 39] and secondary outcome [1, 29, 30] in clinical trials. Moreover, the location of the ICH has been shown to affect functional outcome in patients with stroke [10, 43].

ICH volume can be estimated quickly, for example, using the ABC/2 method [8]. 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 [23]. As this method is relatively easy to implement in practice, it can be used to quickly produce estimates of hemorrhage volume.

Although ABC/2 is widely used, Divani et al. [13] 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 tends to significantly outperform the ABC/2 approach, especially for irregularly shaped ICH and for smaller thickness (i.e. higher resolution) scans. Recently, Webb et al. [50] found that ABC/2 measurements at a clinical site, 81% of the 4,369 scans were within 5 milliliters (mL) of ICH volume compared to planimetry methods, but only 41% were within 20%. 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. This is particularly worrying in clinical trials where the ICH exposure may be measured with substantial bias and reduce the power of the tests. An additional problem is that the ABC/2 method does not take into account intraventricular hemorrhage (IVH),

*Principal Corresponding Author
Email addresses: jmusche1@jhu.edu (John Muschelli), emsweene1@jhu.edu (Elizabeth M. Sweeney), PVespa@mednet.ucla.edu (Paul Vespa), dhanley@jhmi.edu (Daniel F. Hanley), ccrainic@jhsph.edu (Ciprian M. Crainiceanu)

which has been shown to be prognostic of 30-day mortality [19, 48]. Moreover, the ABC/2 method has been shown to consistently over-estimate infarct volume [35] and may have significant inter-rater variability [20]. Therefore, a rapid, automated, and validated method for estimating hemorrhage location and its volume from CT scans is highly relevant in clinical trials and has both diagnostic and prognostic value.

Methods have been proposed for segmentation of ICH using magnetic resonance images (MRI) [9, 49]. However, in most clinical settings, the protocol is not standardized for MRI scans. MRI sequences and protocols may vary across sites and there is no general, standardized, agreed-upon MRI modality 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, a binary ICH image, and the 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 [16, 25, 26, 36, 38]. These methods include fuzzy clustering [25, 38], simulated annealing [26], 3-dimensional (3D) mathematical morphology operations [36], and template-based comparisons [16]. Unfortunately, no software for ICH segmentation is publicly available. The only software that we were able to obtain after requesting it was the one presented in Gillebert, Humphreys, and Mantini [16]. This software requires a moderate level of technical knowledge. 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 (Minimally Invasive Surgery plus recombinant-tissue plasminogen activator for Intracerebral Evacuation) stroke trial [29]. 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, spontaneous supratentorial intracerebral hemorrhage above 20 milliliters (mL) in size (for full criteria, see Mould et al. [31]). The population analyzed here had a mean (SD) age was 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 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). Readers employed a semiautomated threshold-based approach using a Hounsfield unit (HU) range of 40 to 80 to select potential regions of ICH [6, 45]; 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/1ltIM8c`). Images were converted to the Neuroimaging Informatics Technology

Initiative (NIfTI) data format using `dcm2nii` (provided with MRIcro [41]). 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.

All image analysis was done in the R statistical software [40], using the `fslr` [32] package to call functions from the FSL [21] neuroimaging software (version 5.0.4), and the `ANTsR` package to call functions from the ANTs (Advanced Normalization Tools) neuroimaging software [4].

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) [46], a function of FSL, using a previously published validated CT-specific brain extraction protocol [34].

2.5. Image Registration

Rorden et al. [42] 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. Transformed hemorrhage masks and brain masks were thresholded using a value of 0.5 to preserve mask volume [14]. This operation reoriented the image, ensured isotropic voxel sizes for smoothing and other operations described below, and preserved 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

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 was 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 incorrectly predicted by our approach as not ICH.

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 thess 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. 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, denoted $N(v)$, of all adjacent voxels along the 3 dimensions together with the voxel itself. If $x_k(v)$ denotes the voxel intensity in HU for voxel neighbor k, where $k = 1, \ldots, N(v) = 27$, then the local mean intensity is defined as:

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

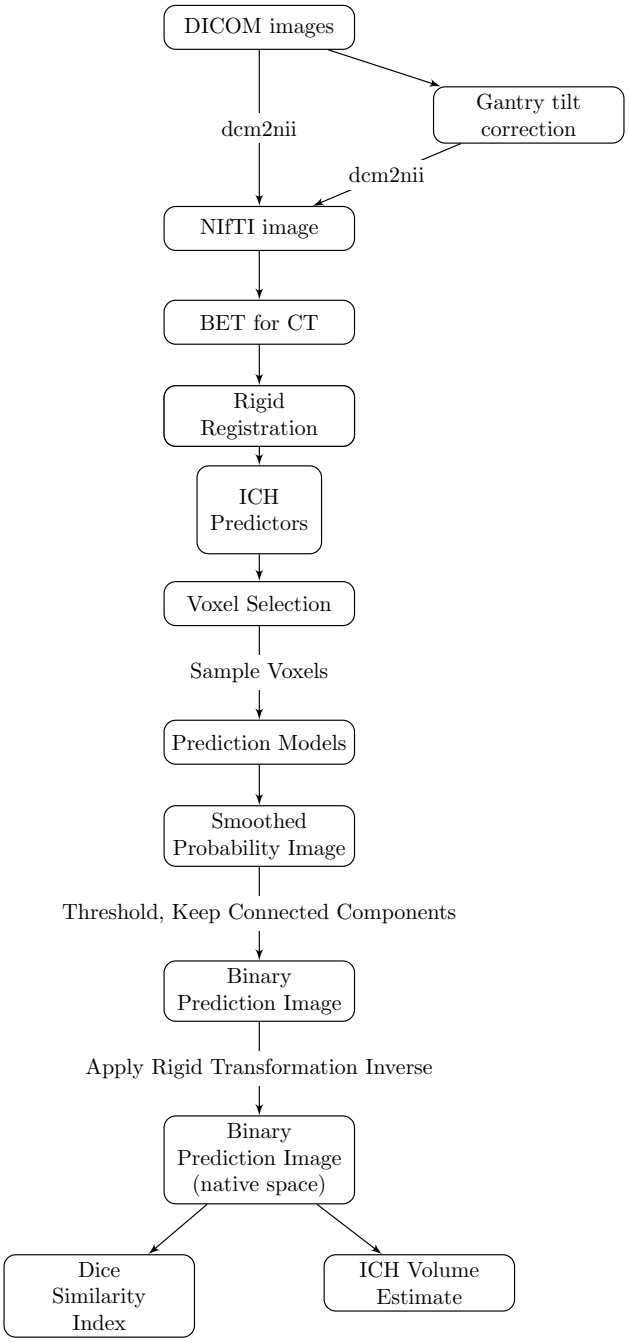

Figure 1: **Processing Pipeline**. Images in DICOM (Digital Imaging and Communications in Medicine) format were gantry tilt corrected if necessary and converted to NIfTI (Neuroimaging Informatics Technology Initiative) format using `dcm2nii`. After NIfTI conversion, the brain extraction tool (BET) was applied to the image using a previously published protocol. The image was rigidly registered to a brain CT template. 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.

Based on similar ideas we have also calculated statistics based on higher order moments and define the local standard deviation (SD), skew, and kurtosis as:

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

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 a 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), the resulting plane-specific z-scores may discriminate better high relative values within the plane, which may attenuate the effect of low-frequency HU intensity inhomogeneities.

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

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

where $o \in \{\text{axial, sagittal, and coronal}\}$, $\bar{x}(v, o)$ and $\sigma(v, o)$ 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 calculated the same formula as in equation (5), but used only voxels with HU values between the 10th and 90th percentile of intensity within the slice to calculate 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. First-pass 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 [5], a previously published, open source, general segmentation tool based on Markov random fields for image segmentation. We used used Atropos to conduct a 4-tissue class segmentation and combined the top 2 probability classes into one class. The probability of the resulting class was then used a predictor, denoted by Atropos(v). Although Atropos has been shown to perform well in other studies for tissue-class segmentation [5, 28], 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. To take advantage of this property, we right-left flipped the 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 [3]) 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 warped back into the original space to align this predictor with the other predictors. This predictor is similar to that used in Gillebert, Humphreys, and Mantini [16].

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. 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

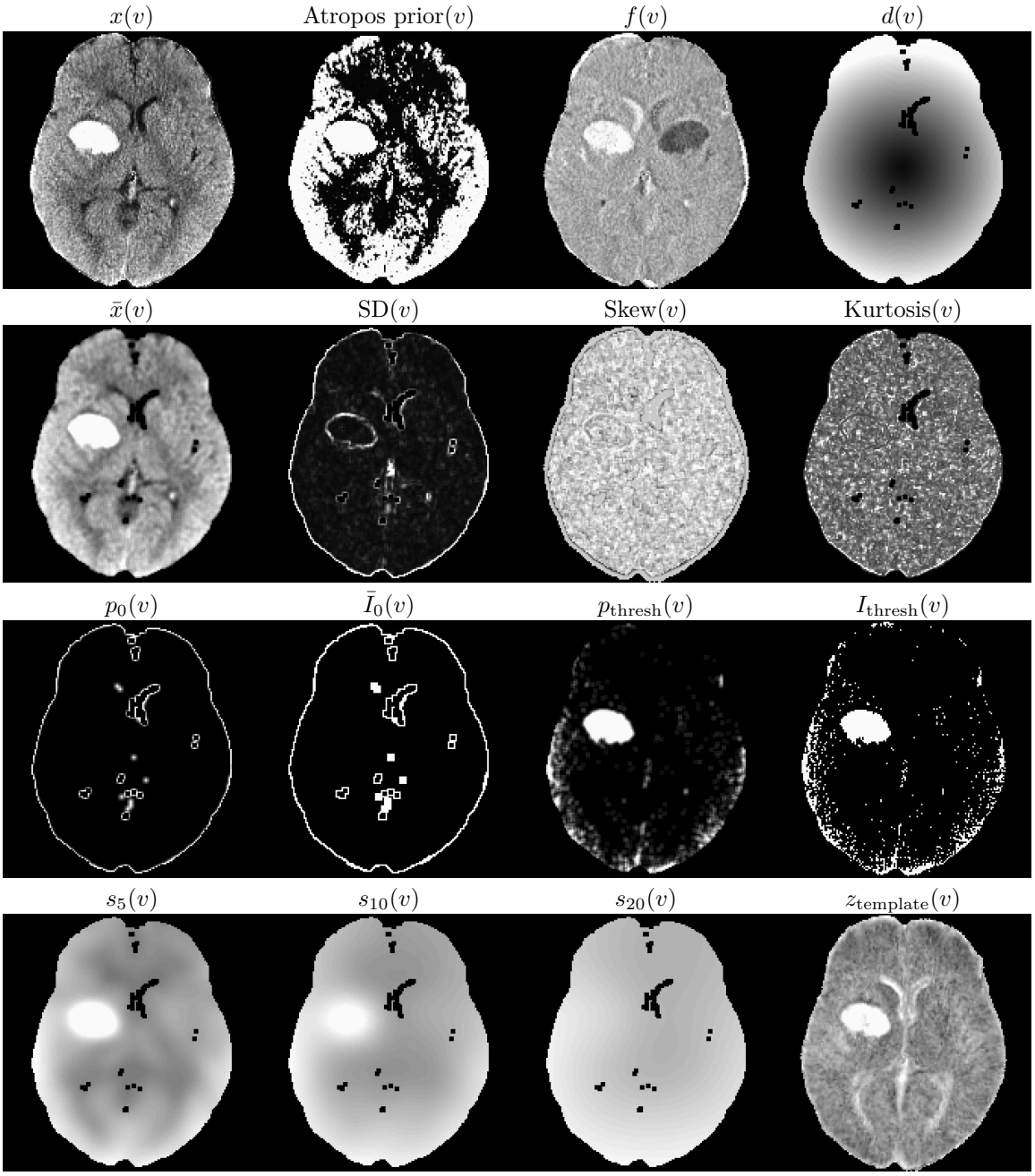

Figure 2: **Predictor Images**. Here we display one slice from one patient for predictor images. The within-plane standardized and Winsorized predictor images were not shown as they are within-subject scaled versions of the image $x(v)$ and appear very similar. Although they appear similar at a subject 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.

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

From these 10 scans, we collapsed all the voxels passing the voxel selection procedure. From these voxels, we randomly sampled 100,000 to fit the models and do data exploration, and used the remaining voxels to estimate the model cutoffs of the probability of ICH. All models were fit with all predictors.

Using the sampled training voxels, we created predictions for the probability of ICH using a 1) logistic regression model, 2) logistic regression model with a penalty, 3) generalized additive model (GAM), and 4) classifier fit with the random forest algorithm.

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 [47] using the `glmnet` package [15]. The tuning parameter (λ) for the penalization was chosen using 10-fold cross-validation (of only the training voxels in the model), with the cost function of misclassification rate. The parameter was chosen using the largest value of λ such that the error is within 1 standard error of the minimum, chosen for out-of-sample performance stability.

The generalized additive model (GAM) [17, 18] was also created using indicator variables for binary variables and thin-plate splines for all continuous measures, fit with fast-estimation of restricted maximum likelihood (fREML) using the `mgcv` package [51, 52]. Model specifications can be seen in Section 5.2.

We fit a random forest [7] using the `randomForest` package in R [24], using the default pruning parameters and number of trees (`ntree=500`, `mtry=4`).

2.10. Estimating a Cutoff for Model Probability

Each model gives the estimate of the probability a voxel is ICH. In the end, however, we want a predicted binary hemorrhage mask to compare to the manual segmentation. Using the voxels from the training data, we estimated the probability of ICH from each model. For each model probability, we smoothed this probability image using the neighborhood voxels (1 voxel in every direction). To choose a probability cutoff to make a binary image, we used the voxels in the training data that were not sampled for estimating the model. Using these voxels in the smoothed images, we estimated the probability cutoff that minimized the Dice Similarity Index (DSI) [12] between the prediction and true value. The DSI is measure of overlap insensitive to values where neither the true segmentation nor predicted segmentation were considered ICH, and will be used as a performance measure when comparing models. 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 positives", where the manual and predicted segmentation are both 1, FP denotes the number of "false positives", where the manual segmentation is 0 and predicted segmentation is 1, and FN denotes the number of "false negatives", where the manual segmentation is 1 and predicted segmentation is 0. DSI ranges from 0 to 1; 0 indicates no overlap and 1 denotes perfect overlap.

After thresholding the smoothed image using this probability cutoff, we discarded regions with fewer than 100 (0.1mL) connected voxels. We then transformed this binary mask back to original (i.e. native) space using the inverse from the previously estimated rigid-body transformation. As the linear interpolation results in a non-binary mask, we thresholded this image at 0.5 to preserve volume [14].

For each scan in the test data, this voxel selection and prediction process was performed and each scan has a corresponding binary prediction image. For evaluation of all measures, the comparison was done in the native space of the patient, not in the template space, as this was where the manual segmentation was done.

2.11. Measuring and Testing ICH Prediction Performance

For all measures, we used the binary prediction masks from the test scans. We measured performance for each model using the DSI and also estimated the ICH volume of the prediction. The number of true negatives inflates specificity and overall accuracy, since most of the voxels within the brain are non-ICH, and are not reported. We tested the difference between median DSI across models using a Kruskal-Wallis test. If a difference was present, we tested each combination of models (6 combinations) using a Wilcoxon signed rank test, and corrected the p-value using a Bonferroni correction.

We also calculated the relationship of estimated volume of ICH compared to the manual segmentation using the Pearson correlation and root mean squared (RMSE) between volumes measures. Similarly, we did performed a Kruskal-Wallis test on the medians of the absolute value of the difference in estimated volume from each model and the true volume and performed pairwise tests if necessary. For DSI and correlation, higher values indicate better agreement with the manual segmentation. For RMSE, lower indicates better agreement.

3. Results

3.1. Dice Similarity Index

In Figure 3, we show the DSI distributions from the test data for each model. We see that 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). We also note that using the random forest results in a slightly higher median DSI compared to the other models, and there was a difference across medians ($\chi^2(3) = 13.49$, $p < 0.05$). Indeed, after Bonferroni correction, the only comparisons that were significant were comparing the random forest DSI to the DSI from the logistic ($p < 0.001$), LASSO ($p < 0.001$), or GAM ($p < 0.001$) models.

In Figure 4, we display 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 then the CT scan with overlaid colors. The green indicates a correct classification of ICH from the model (true positive), blue indicates a false negative, and red indicates a false positive. Note, the image has finer resolution in the axial slice (0.5mm by 0.5mm) than in the inferior/superior 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 6, 7, 8, and 9, 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 line equation and correlation are printed on the plot. 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. The correlation (95% confidence interval (CI)) between the true volume and the volume predicted volume were 0.92 (95% CI: $0.884, 0.945$) for the logistic model, 0.916 ($0.878, 0.942$) for the LASSO, 0.908 (95% CI: $0.866, 0.937$) for the GAM, and 0.932 (95% CI: $0.901, 0.954$) for the random forest. The RMSE for all the logistic (RMSE: 10.67 mL), LASSO (10.83 mL), and random forest (10.27 mL) models, but was slightly higher for the GAM model (11.36 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

Overall, all models perform well for ICH segmentation. Some failures exist, but the algorithm using the random forest to fit the probability of ICH had higher median DSI, the lowest RMSE, and the highest correlation compared to the other models. Therefore, when implementing the algorithm, the model used will be the random forest.

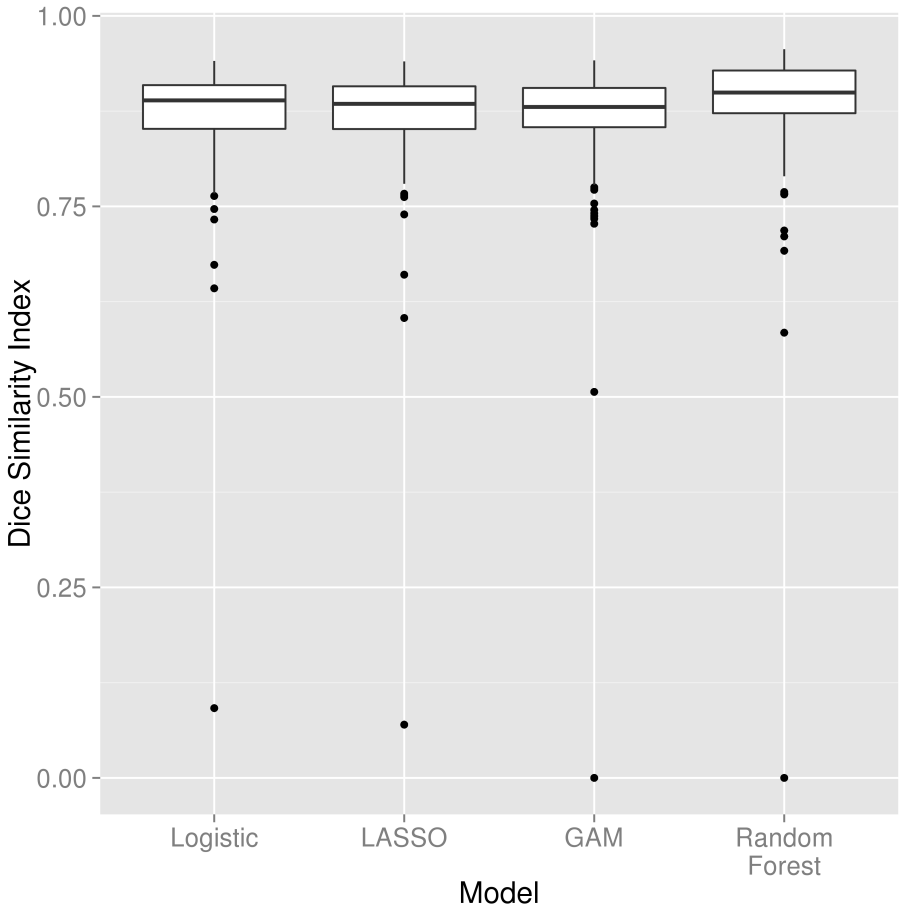

Figure 3: **Distribution of Dice Similarity in Test Scans**. Here we display the boxplot of the Dice Similarity Index (DSI), a measure of spatial overlap between the estimated hemorrhage mask and the manually delineated hemorrhage mask, in the 102 test scans. We present the DSI distribution for each 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 $p < 0.05$).

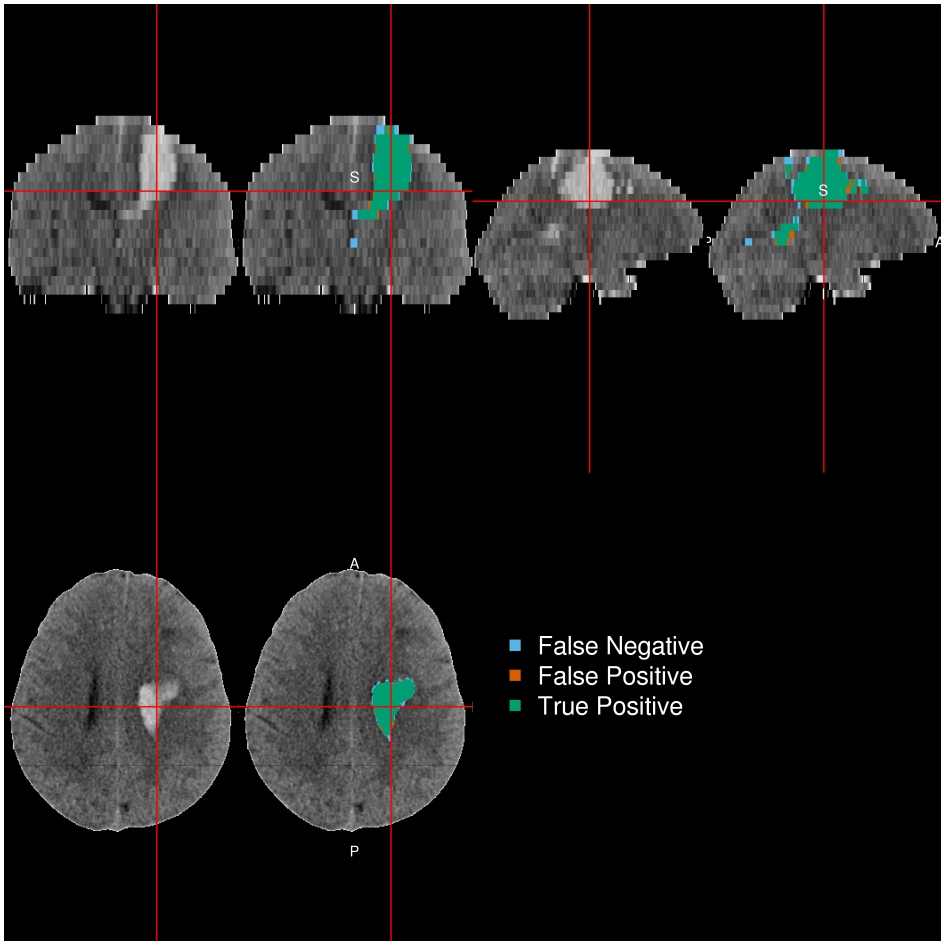

Figure 4: **Patient with Median Dice Similarity Index**. We present the patient with the median Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The median DSI was 0.899, which indicates high spatial overlap. The green indicates a correct 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.

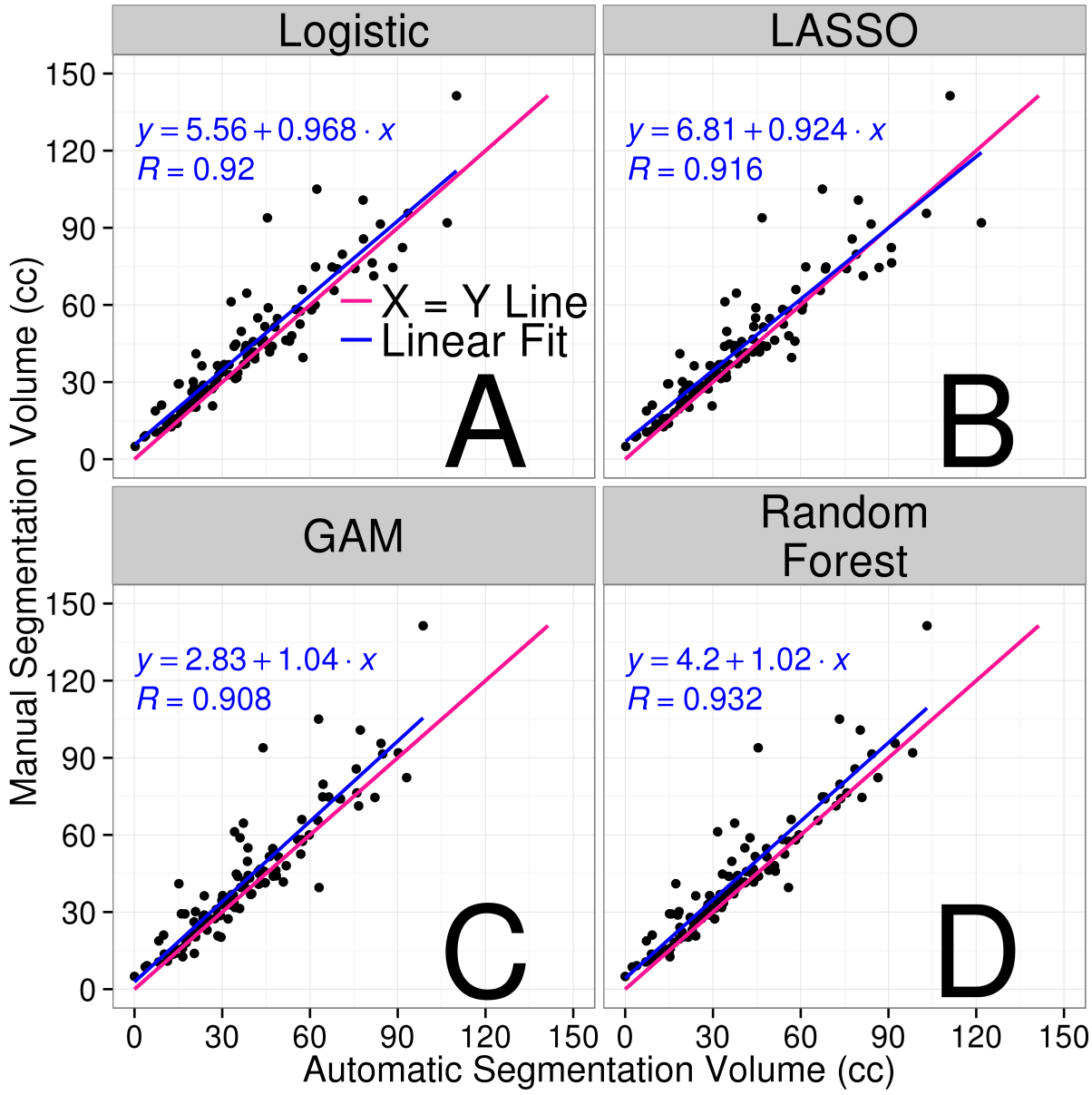

Figure 5: **Comparison of Estimated and Manual Intracerebral Hemorrhage Volume for Each Model**. In each panel, we show the volume of intracerebral hemorrhage (ICH) estimated from each model (x-axis) versus that from the gold-standard manual segmentation (y-axis) in the 102 test scans. The pink line 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.

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 stroke as additional information from MRI was not standardized nor is the standard of care with the disease. We validated this method against the manual segmentation of ICH as a gold standard. We used the Dice Similarity Index and correlation between the volume of ICH from manual and automatic segmentation as measures to determine how well an algorithm performed.

We create a set of predictors that attempt to capture the features relevant to distinguishing ICH versus non-ICH areas and described the rationale for each predictor. We have used these predictors in a series of models, including those fit with logistic regression, logistic regression penalized with the LASSO, generalized additive models, and random forests. All models result in high values for the DSI and high correlations for the volume of ICH from manual and automatic segmentation. Estimating the probability of ICH using random forests appeared was chosen as the final algorithm based on DSI; that algorithm also had high correlations of automated versus manual ICH volume.

Comparing the manual to the automatic segmentation in each quantile of DSI showed that some cases may fail ($N = 1$ with DSI below 0.5), but much of the discrepancy occurs at the edges of the hemorrhage. This may be due to the smoothing operations performed in the model, which are isotropic and do not preserve edges. Anisotropic smoothing, such as those presented in Perona, Shiota, and Malik [37], may help preserve edges and result in better segmentation.

Other methods have been presented for segmentation of ICH from CT scans [16, 25, 26, 36, 38]. Loncaric, Cosic, and Dhawan [25] performed an analysis on only one 2-dimensional scan and could not be compared. The method in Pérez et al. [36] was semiautomated and was validated by visual inspection, which did not succeed in 6 out of 36 scans (16.7%). If we use the criteria of a DSI below 0.5 as a failure, we had 1 fail out of 102 test scans (1%). We have higher median DSI (0.899) than those reported in Gillebert, Humphreys, and Mantini [16] (approximately 0.62 and 0.78 for hemorrhagic strokes from graphs) and were comparable to [38] (0.897, 0.858, and 0.9173 for different groups with hemorrhage). Loncaric et al. [26] did not compare hemorrhage masks to the manual segmentation, but compared ICH volumes on the manual and automatic segmentation on 5 subjects measured 3 times, which had a Pearson correlation of 0.917 on all 15 scans, similar to that reported from the random forest model selected above ($R = 0.932$).

Only Gillebert, Humphreys, and Mantini [16] responded with any requests for software to perform the segmentation. Although we believe that the method from Gillebert, Humphreys, and Mantini [16] is comparable, it has not been packaged for general use. As the analysis above was done in R, we have released a package that can perform ICH segmentation (`https://github.com/muschellij2/ichseg`), including the models for prediction, CT template from Rorden et al. [42], standardized mean and standard deviation images, and functions to register the images, create the predictors, predict from the models, and return a binary hemorrhage mask. Although a package is ideal for prediction on a large number of images, we believe that this limits those who can test or try the software. Therefore, we have released a Shiny [11] R application online (`http://bit.ly/ICH_SEG`) for researchers to input CT scans and the application will output ICH segmentation masks, while giving an image depicting each processing step.

One of the potential concerns with image registration with patients with ICH is image registration. Although this method uses non-linear registration for one predictor, the standardized-to-template intensity, this predictor is transformed back to the rigid-to-template space after standardization. The method generally relies only rigid-body registration to a template. This registration is largely done for head reorientation and resampling to isotropic voxel sizes across patients. Thus, when morphological operations are performed, such as smoothing using millimeter specifications, they do not depend on the original voxel sizes. Although the models were fit in the rigid-to-template space, voxels are not compared across patients in the models. Also, the method returns hemorrhage masks in the native space of the patient, so they are easily comparable to any segmentation method performed on the original scan, such as clustering. Thus, we use non-linear registration in one step, but do not rely on highly accurate image registration for comparability of voxels across patients, which can be problematic with patients with large hemorrhages.

Another concern from the process above is that the algorithm was only fit on 10 patient scans. Moreover, it was only fit with the sampled 100,000 voxels that passed the voxel selection procedure from these scans. Although this is a small sample, the models have shown to have high out-of-sample accuracy. As this was the original procedure done, the training and test sets were not changed after this analysis to help avoid

overfitting the models to the test set. If a cross-validation approach was done across the entire set of scans, the models must also be combined in some way to give a final prediction. Validation of this method on additional data is required, but the test scans used are from multiple sites, different scanners, and have patients with heterogeneous hemorrhages both heterogeneous in size and location. Thus, we believe this method to have good out-of-sample accuracy in a similar population.

As the process results in binary hemorrhage masks, one can use methods described in Muschelli et al. [33] to estimate quantitative measures of hemorrhage location. Other automated analysis of hemorrhages can be done, such as shape analysis, which could not be done without a binary mask. Overall, this process allows for researchers to use these hemorrhage masks for other voxel-based analyses that can 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. The method relies on series of processing steps and creating a set of relevant predictors based on morphological operations such as smoothing, registration, and intensity normalization. This method has been shown to agree with the gold standard of manual delineation of hemorrhages. As an automated process, it is faster, does not require extensive radiologic image experience, is scalable to thousands of images, and does not have inter-reader variability. As the process results in binary hemorrhage masks, one can then use these to define or test characteristics of ICH compared to patient-level information, such as those described in Muschelli et al. [33]) to estimate quantitative measures of hemorrhage location. Most importantly, the volume of ICH is automatically calculated and can be used as a covariate in analysis, as this has been shown to be associated with long-term functional outcome [8, 22, 48].

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. Dr. Chris Rorden was also extremely helpful in adapting his `dcm2nii` software to some issues specific to CT scans.

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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

[15] 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.

[16] 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.

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

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

[19] 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.

[20] 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.

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

[22] 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.

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

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

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

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

[43] 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.

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

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

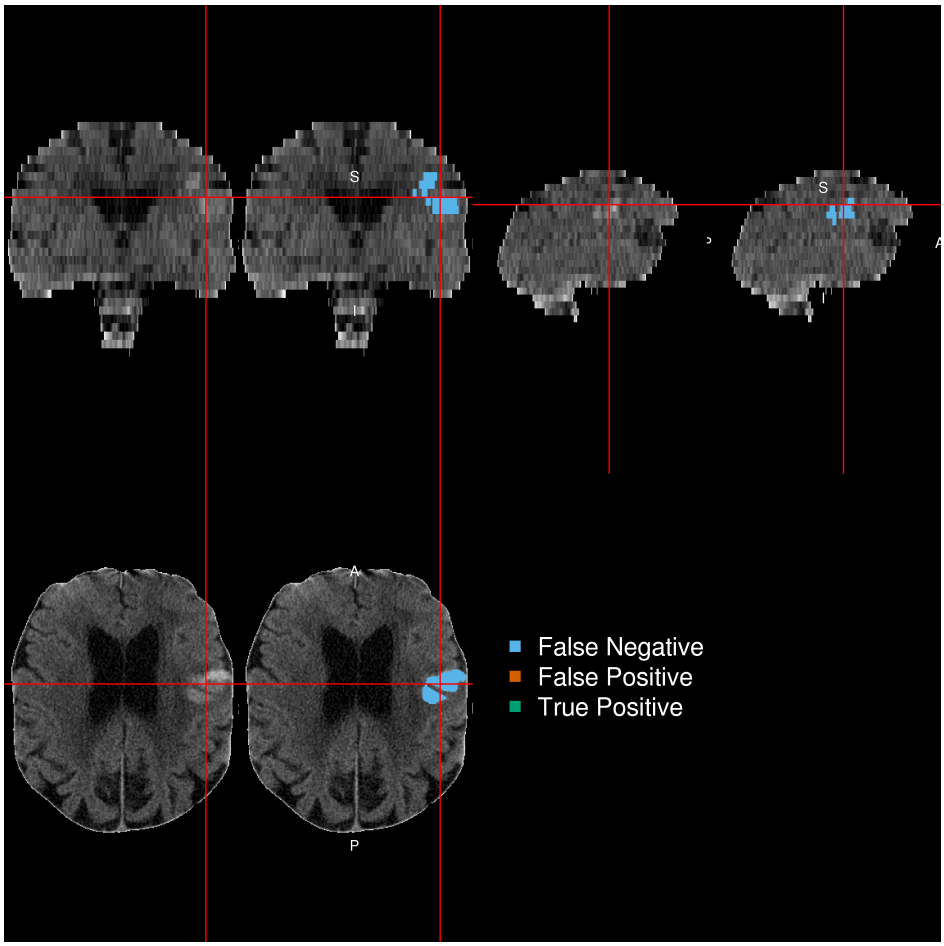

Figure 6: **Patient with Lowest Dice Similarity Index**. We present the patient with the lowest Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The lowest DSI was 0. The green indicates a correct classification of ICH from the model, 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 k, $k = 1, \ldots 21$.

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

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:

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

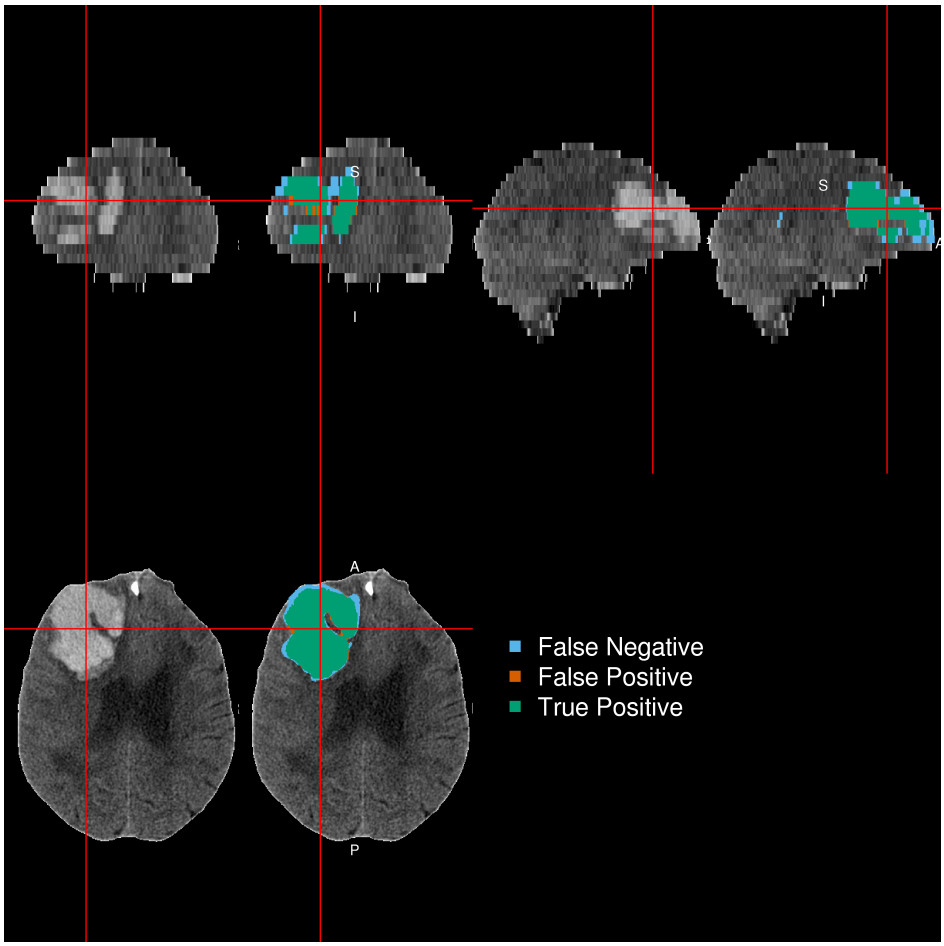

Figure 7: **Patient with 25$^{\text{th}}$ Quantile Dice Similarity Index**. We present the patient with the 25$^{\text{th}}$ quantile Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The 25$^{\text{th}}$ quantile DSI was 0.872. The green 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.

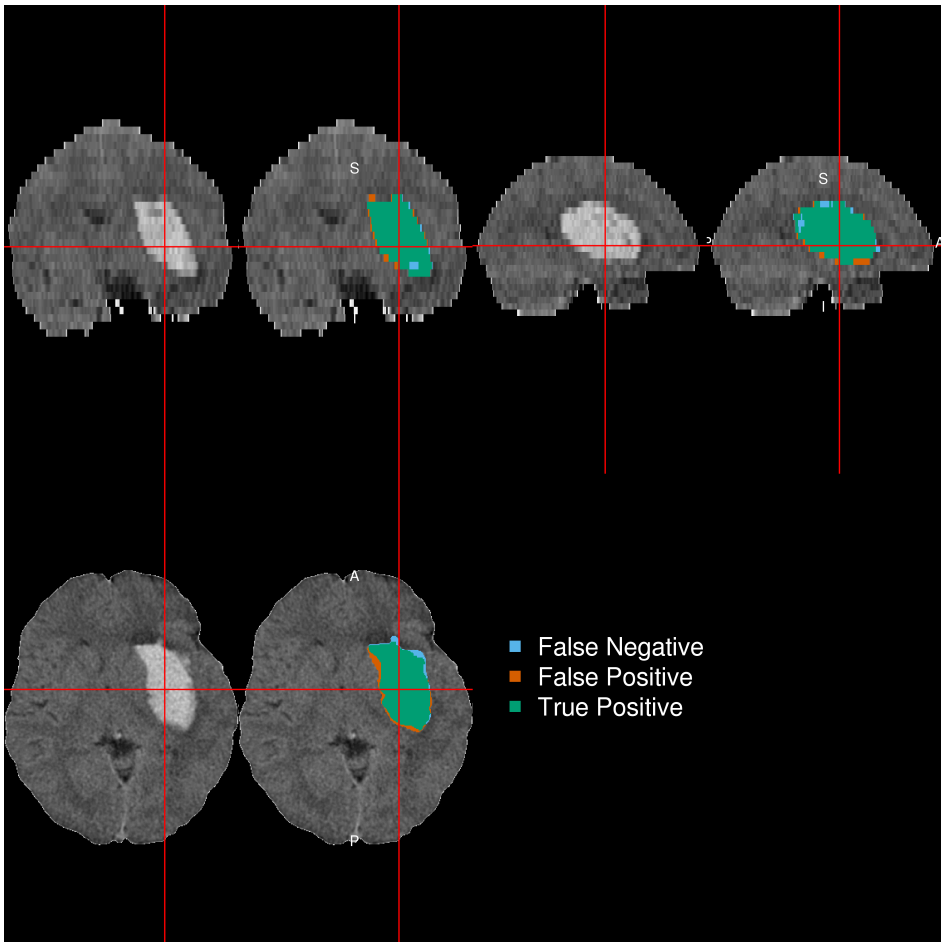

Figure 8: **Patient with 75$^{\text{th}}$ Quantile Dice Similarity Index**. We present the patient with the 75$^{\text{th}}$ quantile Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The 75$^{\text{th}}$ quantile DSI was 0.928. The green 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.

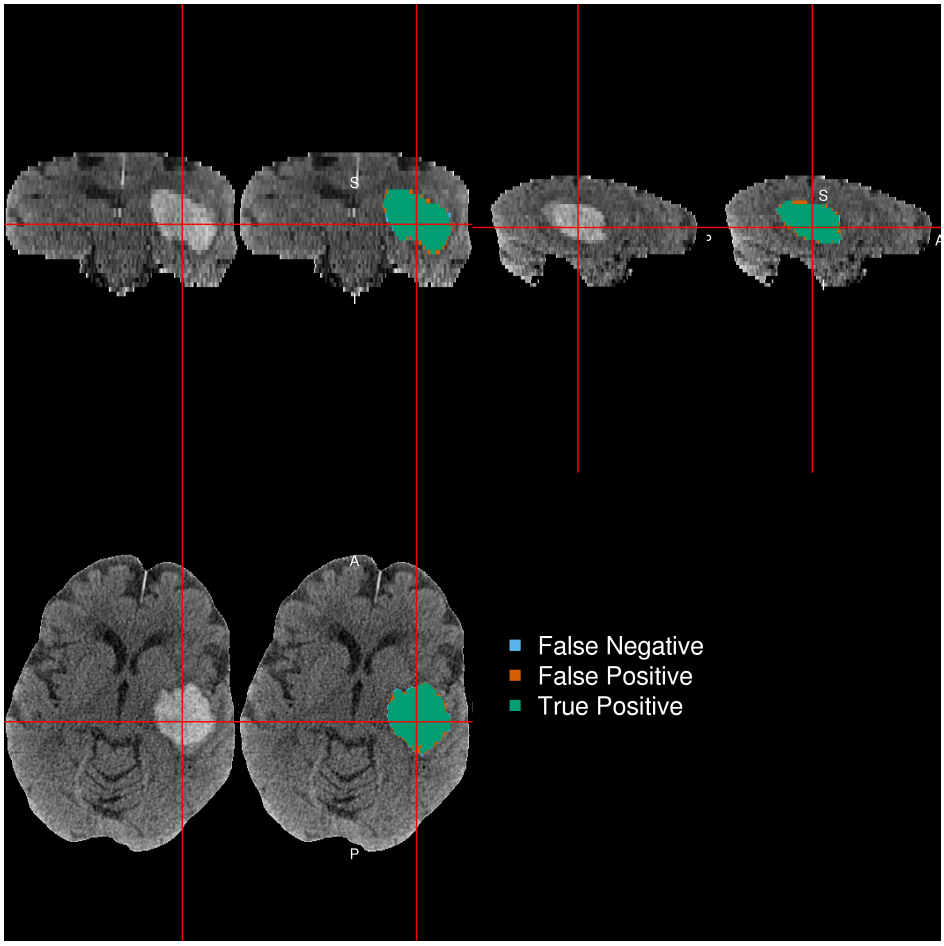

Figure 9: **Patient with Highest Dice Similarity Index**. We present the patient with the highest Dice Similarity Index (DSI), a measure of spatial overlap, from the chosen predictor model fit with a random forest. The highest DSI was 0.956. The green indicates a correct classification of ICH from the 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
Intercept	1.008
Neighborhood mean	0.051
Neighborhood sd	0.000
Neighborhood skew	0.065
Neighborhood kurtosis	-0.352
Image intensity (HU)	-0.172
Threshold (>= 40 and <= 80)	-0.151
Within-plane Coronal	-0.632
Within-plane Sagittal	-0.249
Within-plane Axial	1.037
Winsorized standardized (20\% trim)	0.547
Percentage thresholded neighbors	2.061
Atropos probability image	0.150
Percent of zero neighbors	-9.180
Indicator of any zero neighbors	0.071
Distance to Image centroid	-0.087
Gaussian smooth (5mm^3)	-0.051
Gaussian smooth (10mm^3)	0.550
Gaussian smooth (20mm^3)	-0.390
Standardized-to-template Intensity	1.460
Contralateral difference	0.033