ON-LINE APPENDIX: METHODS

Subjects and Data

Subjects were initially identified by searching the radiology archives of our tertiary care university hospital for the diagnoses included in the study (see below). This set of studies was acquired between January 2008 and January 2018 at our own institution or uploaded to our PACS for secondary interpretation from another institution during the same timeframe. Specific diagnoses were then confirmed by comprehensive review of individual electronic medical records, using pathologic data when available or using follow-up clinical and radiologic assessments across time as the ground truth. Diagnoses were also independently confirmed by 2 neuroradiologists. Exclusion criteria for the validation sample included multiple diagnoses causing abnormalities on FLAIR within an individual, a history of cranial surgery causing abnormalities on FLAIR, or excessive imaging artifacts precluding accurate radiologic interpretation.

Diseases

The 19 diseases included in the validation sample were the following: low-grade glioma, high-grade glioma (glioblastoma), primary CNS lymphoma, metastatic disease, vascular disease (ischemia), SVID, Susac syndrome, active multiple sclerosis, inactive multiple sclerosis, tumefactive multiple sclerosis, neuromyelitis optica, acute disseminated encephalomyelitis, adrenoleukodystrophy, CADASIL, HIV encephalopathy, progressive multifocal leukoencephalopathy, toxic leukoencephalopathy, posterior reversible encephalopathy syndrome, and migraine.

Ground Truth Segmentation

All FLAIR lesions on all native subject space images were hand-segmented by a radiologist (neuroradiology fellow with extensive segmentation experience) using ITK-SNAP,[1] to provide segmentation masks for the training data and a basis for calculating criterion standard lesion volumes and performance measures for the validation data. A second radiologist (fourth-year radiology resident, neuroradiology mini-fellow, also with extensive segmentation experience) independently hand-segmented all validation cases in native subject space to provide a measure of interrater reliability and as a basis for comparison with human performance. Diagnoses were not available to the radiologists at the time of hand segmentation, and only the FLAIR sequence was used.

Detailed Description of Comparison Automated Algorithms

The LST is an automated segmentation algorithm designed for segmenting MS lesions from input of either 3D gradient-echo T1-weighted and FLAIR images or from FLAIR images alone.[2] We used the version of LST that uses FLAIR images alone, the "lesion prediction algorithm," to make it most analogous to our U-Net, which also only requires FLAIR images. Furthermore, because only some of the 19 diseases in our sample are characterized by abnormal T1 signal, while others have normal T1 signal, this version of LST qualitatively provided the best overall performance across all diseases at the cost of some disease-specific performances. The LST method does not require or allow training data, so we applied the algorithm to all our study subjects.

BIANCA is a method applying a k-nearest neighbors approach for white matter hyperintensity detection on FLAIR MR imaging.[3] This method can use any number of sequences but requires at least FLAIR images. Again, to make the algorithm most analogous to our U-Net and to not bias the algorithm for particular diseases, we provided the algorithm with only skull-stripped FLAIR images. We used the same training/validation data split for the BIANCA algorithm as was used for our CNN/U-Net. LST and BIANCA were both implemented on an iMac Pro (2017; Apple, Cupertino, California), with a 3.2-GHz Intel Xeon CPU and 64-GB RAM, running Matlab R2017b (MathWorks, Natick, Massachusetts), and Python 3.7.

Statistical Analysis

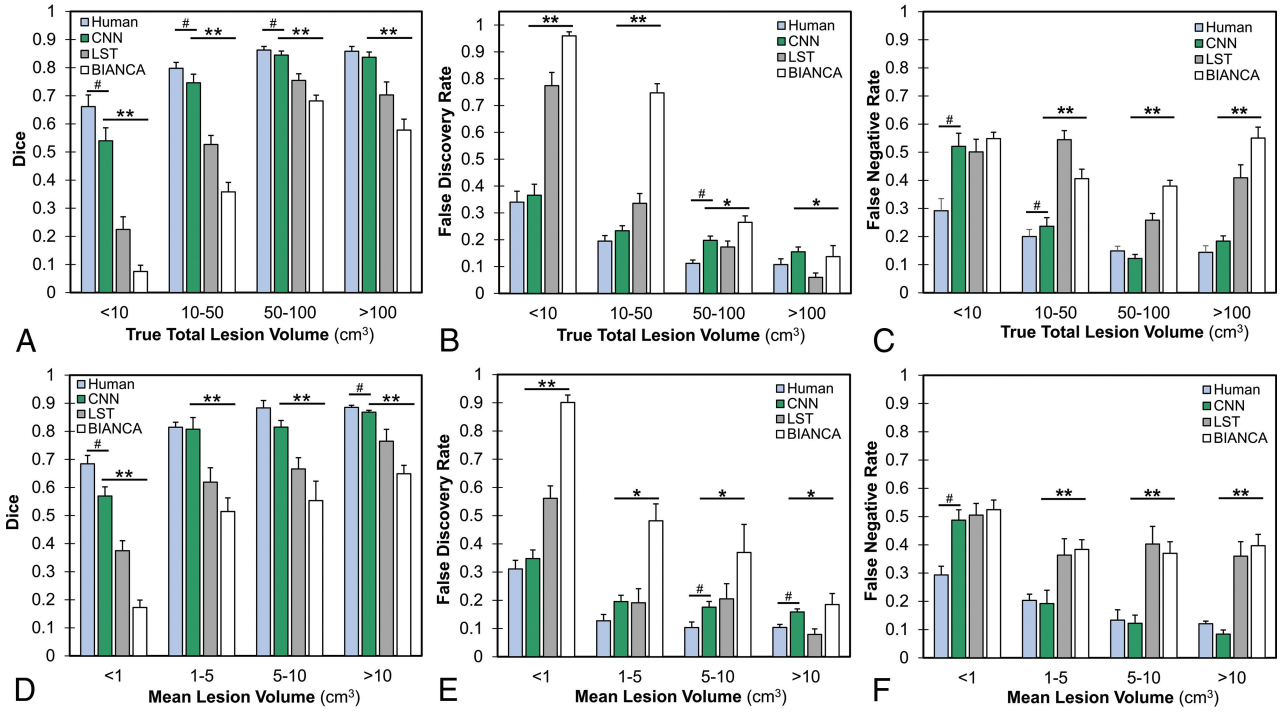
Comparisons of performance across methods used paired 2-tailed *t* tests of Dice scores. We were interested in the performance on a variety of diseases and lesion appearances, so Dice scores and other performance measures were stratified across diseases and lesion volumes. Relationships between performance and lesion volumes were expressed using linear regressions and Spearman correlations. Finally, given the importance of the method for extracting volumetric lesion data, we evaluated the algorithms' estimations of total lesion volume using subject-by-subject correlations with and forecasting deviations (root mean/median square percent error) from ground truth. In evaluating the effect of technical factors on algorithm performance, we split the validation data according to various variables of interest using a 1-way ANOVA or 2-tailed *t* tests for comparisons among groups, depending on the number of groups.

Results

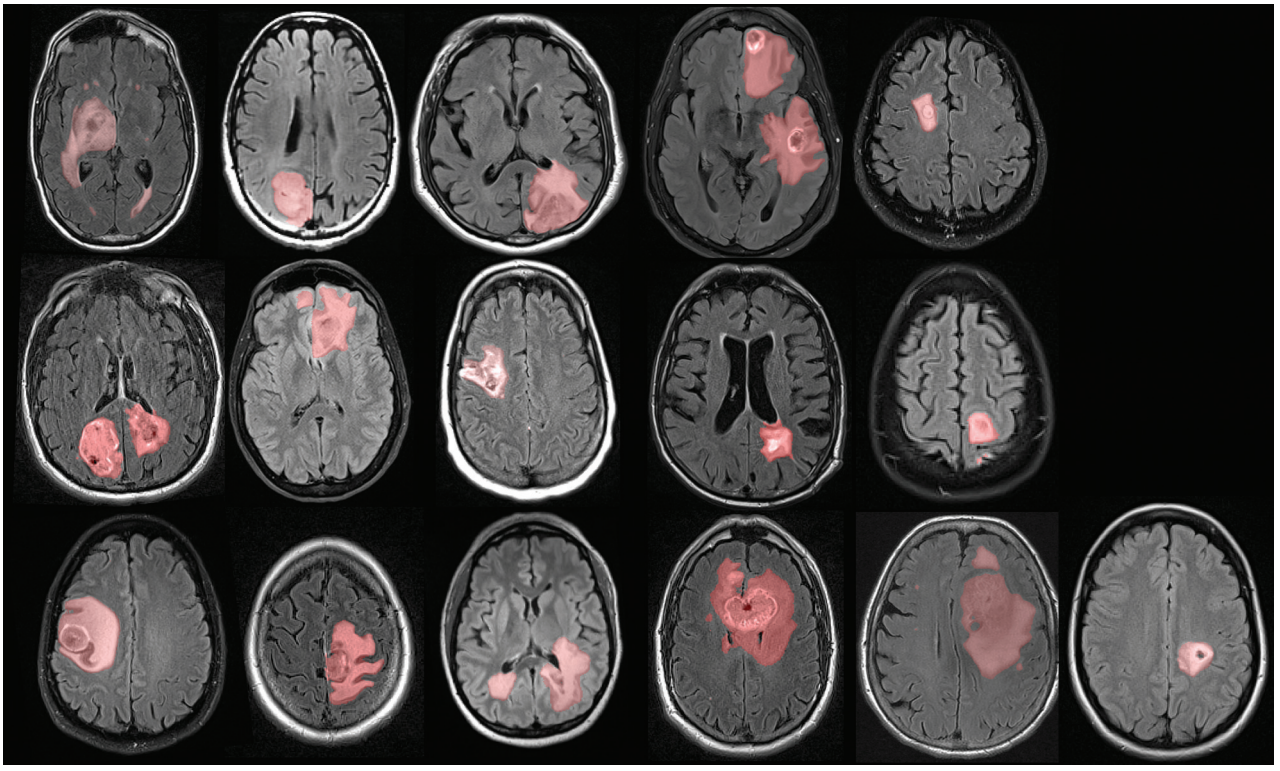
We analyzed performance as a function of total lesion volume and mean lesion volume across all validation cases for all 4 methodologies (second human radiologist, CNN, LST, and BIANCA). To fully evaluate the contribution of false-positives and false-negatives on overall performance (as measured by Dice), we plotted the false discovery rate and false-negative rate as a function of lesion volume (On-line Fig 1). In general, both false-positives and false-negatives decrease with large average and total lesion volume, though this effect of volume is smaller with the CNN than with LST or BIANCA. In other words, the CNN is more specific at low lesion volumes (On-line Fig 1*B*, -*E*) and more sensitive at high lesion volumes (On-line Fig 1*C*, -*F*) than the other automated methods. A similar effect is seen in human performance, though it is less pronounced for false-negatives—that is, relative to the automated algorithms, humans have fewer false-negatives at low lesion volumes (On-line Fig 1*C*, -*F*).

REFERENCES

1. Yushkevich PA, Piven J, Hazlett HC, et al. **User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability.** *Neuroimage* 2006;31:1116–28 CrossRef Medline
2. Schmidt P, Gaser C, Arsic M, et al. **An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis.** *Neuroimage* 2012;59:3774–83 CrossRef Medline
3. Griffanti L, Zamboni G, Khan A, et al. **BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities.** *Neuroimage* 2006;141:191–205 Medline



ON-LINE FIG 1. Complete performance profile of segmentation methods according to lesion characteristics, demonstrating the effect of false-positives and false-negatives on Dice scores, depending on lesion volumes. *A*, Median Dice scores of cases stratified by total lesion volume, as in Fig 4*B*. *B*, False discovery rate stratified by total lesion volume, as in Fig 4*C*. *C*, The false-negative rate stratified by total lesion volume. *D–F*, Same measures as in *A*, *B*, and *C*, but cases are grouped according to mean individual lesion volumes. *D* is same as Fig 4*E*. *Error bars* in all panels represent ±1 standard error of the mean across cases. The *asterisk* denotes $P < .01$ for the CNN compared with 1 method, and *double asterisks* denote $P < .01$ for the CNN compared with both methods using 1-way group ANOVA and paired 2-tailed *t* tests. The *hashtag* separately denotes $P < .05$ for human performance compared with CNN.



ON-LINE FIG 2. CNN segmentations (overlaid in transparent red) of abnormal findings on FLAIR in cases with heterogeneous FLAIR signal. In these 16 cases, abnormal FLAIR signal contained areas of heterogeneity. In nearly all cases, the entire region of abnormality on FLAIR was correctly segmented despite the internal heterogeneity, at the same time avoiding incorrect segmentation of the lateral ventricles despite similar signal intensity to necrotic areas. In 2 cases (last 2 cases, *lower right*), a few voxels of abnormality were missed because of the apparently normal signal intensity. Cases with heterogeneous FLAIR signal included 4 cases of primary CNS lymphoma, 4 cases of high-grade glioma, 4 cases of metastatic disease, 3 cases of tumefactive multiple sclerosis, and 1 case of an ischemic vascular lesion.