Artificial Intelligence-assisted reader evaluation in acute CT head interpretation (AI-REACT): a multireader multicase study.

Statistical Plan

Sample size and power calculation

A sample of 30 readers and minimum 135 scans (82 with presence of critical findings and 53 with no critical findings) was estimated to have a minimum 80% power at a type I error rate of 5% to detect a minimum difference in readers' AUC of 5%, assuming a large inter-reader and intra-reader variability of 0.3 and 0.05, respectively, a 0.35 conservative correlation between readers, and anticipated average readers' AUC of 0.75, guided by previous literature.[16,17]

Statistical analyses

The stand-alone performance of qER algorithm was compared with the ground truth generated by the neuroradiologists, using the continuous probability score from the algorithm for the AUC analyses, and binary classification results for the evaluation of sensitivity, specificity, positive predictive value and negative predictive value.

The difference in AUC of readers with and without AI was tested based on the Obuchowski-Rockette model for MRMC analysis which models the data using a two-way mixed effects analysis of variance model treating readers and cases (images) as random effects and effect of AI as a fixed effect with recommended adjustment to df by Hillis et al. Sensitivity and specificity was analysed as part of this model. The main analysis was performed as a single pool including all groups and sites. Prespecified subgroup analyses were performed for the following variables: professional group (radiologist vs ED clinician vs radiographer), post-graduation experience level (junior <5 years, middle grade 5 to 10 years, and senior >10 years), pathological finding and difficulty of image.[16-18]

The median review time per scan with vs without AI was compared using a non-parametric Wilcoxon sign-rank test.