# Appendix S4 : Further detail on the statistical methods

Statistical analysis was performed in three phases by an investigator (AV) who neither knew nor observed study participants.

In the first phase, BRadSTAT scores were assessed visually, overall and question by question. Descriptive models were produced with total and question scores linearly regressed on the group. The BRadSTAT score’s association with the groups was tested using the Jonckheere-Terpstra procedure. Questions were assessed for internal consistency using Cronbach’s 𝛼.

A second phase analysis was performed to determine if BRadSTAT could be refined. Different weightings to each question were assessed. Thirteen different methods were assessed, including the original BRadSTAT score weighting which allotted equal weight to each question.

Eight of the weighting methods produced data-dependent weights:

• 3 sets of log-odds ratios from the proportional odds model: only the positive estimated coefficients in an unconstrained model, coefficients constrained to be non-negative, and coefficients constrained to be non-negative and to have equal coefficients for questions laterally paired (3 and 4; 5 and 6; 9 and 10)

• inverse of estimated question variances in a mixed linear regression model of the question scores with participants as random effects, with either distinct coefficients for each question and equal coefficients for laterally paired questions

• loadings or factor scores from an exploratory factor analysis

• estimated discriminant indices from a graded response model.

Five weighting methods were based on fixed weights, some of the values of which were informed by preliminary results:

• BRadSTAT score, obtained from equal weights of 1 on all questions

• BRadSTAT-9 score, obtained from excluding Question 1 and equal weights of 10/9 on the remaining questions

• CT-only score, obtained from equal weights of 10/7 on questions regarding CT only

• BRadSTAT-CT score, consisting of weights of 1 on Questions 2,4,5,6,8 and 2.5 on each of Questions 9 and 10.

• BRadSTAT-CT-alternative score, consisting of weights of 1 on Questions 2,6,8, 1.5 on Questions 3 and 4, and 2 on Questions 9 and 10.

The assessment criterion was misclassification error, i.e. the proportion of misclassified participants per total number of participants. Weights were applied as regression coefficients in an adjacent-category logit proportional odds model. This model called for the estimation of thresholds on the log-odds scale between every pair of adjacent categories. Thresholds could be translated as maximum likelihood cut points for the corresponding scores, conditionally on the weighting scheme.

Because some weights were data-dependent and because the misclassification error depended upon the estimated cut point selection, misclassification was minimized using leave-one-out cross-validation in all cases. Each of the weightings was subjected to scaling using multipliers on a grid between 0.5 and 1.5. Such scaling changes the classification properties of the score even with fixed weights because of the non-linearity of the logit transformation. For each scaling, the classification was calibrated by finding the best probability value with which to compare the fitted probabilities to minimize misclassification. The differences between the logit of this best value and the optimal thresholds thus form calibrated maximum likelihood estimates of the logit of the cut points on the score scale. Classification according to the corresponding cut points on the score’s original scale is equivalent to this shrunk, thresholded and calibrated process.

In the third phase, we selected the weighting method yielding the smallest misclassification error as well as being most clinically relevant. The first phase was then repeated with the corresponding score. The conditional maximum likelihood cut points of this score and the BRadSTAT score were assessed by producing cross-validated sensitivity, specificity and accuracy estimates by comparing Beginner to non- Beginner, Beginner or Intermediate to Experienced or Expert, and non-Expert to Expert respectively. Sensitivity and specificity here correspond to correct classification in the lower-ability and the higher-ability groups respectively.

Conditional maximum likelihood cut points do not have proven classification optimality properties of themselves, which is why we used cross-validation to find the best weighting and cut points simultaneously. We compared the cross-validated misclassification proportions of the selected weightings to those obtained from an algorithm designed to minimize misclassification when using cut points (which will do so within a given data set, but not necessarily do so in new data). These results are not shown in the paper, but uniformly favoured the conditional maximum likelihood cut point approach.