Appendix 1 Feature extraction

Four groups of radiomic features were extracted from IBEX: (1I) intensity histogram, including 9 statistical features; (2II) gray level co-occurrence matrix (GLCM), including 22 texture features; (3III) gray level run length matrix (GLRLM), including 11 texture features; and (4IV) shape, including 16 texture features.

For the intensity histogram, percentile and percentile area were calculated ranging from 5% to 95% (intervals, 5%); quantile was calculated by 0.25, 0.50, 0.75 and 0.95; the remaining 6 features were calculated by one parameter. Thus, in total 48 features were generated. For GLCM, 4 directions (θ=0°, 45°, 90°, 135°) and 3 offsets (d=1, 4, 7) were measured. Thus, in total 264 features were generated. For GLRLM, 2 directions (θ=0°, 90°) and 1 offset (d=1) were measured. Thus, in total 22 features were generated. All shape features were measured by one parameter, and there were 16 features in total. In total, 350 features were included in our study.

Appendix 2 Preprocessing methods for the data

Z-score standardization

Different radiomics features have different ranges of values, and features of different magnitudes are difficult to compare. Before further analysis, we used z-score standardization to eliminate the effects of different dimensions by scaling the values to a mean of 0 and a standard deviation of 1, using the following formula:

$$z = \frac{\chi - \mu}{\sigma}$$

where μ is the population mean and σ is the population standard deviation.

Appendix 3 Features for modeling

After principal component analysis (PCA), characteristics with the largest variation and a cumulative variance of 90% of each individual feature set were selected for model building. Therefore, there were five principal component features (including PCA1, PCA2, PCA3, PCA4, and PCA5) in the arterial phase model and four principal component features (including PCA1, PCA2, PCA3, and PCA4) in the portal phase model. A combined model picked out the most discriminative feature of each feature set, including all of the principal component features, and clinical data, including the size and the differentiation degrees of the tumor were generated for model building. A logistic regression was applied for modeling, which is a regression method that eliminates variables that do not contribute much to the linear model.

https://dx.doi.org/10.21037/qims-22-112