Original Article

Optimization and validation of voxel size-related radiomics variability by combatting batch effect harmonization in pulmonary nodules: a phantom and clinical study

Yaoyao Zhuo[1,2#], Jie Shen[2#], Yi Zhan[2], Ye Tian[3], Mingfeng Yu[4], Shuyi Yang[1], Peiyan Ye[5], Li Fan[6], Zhiyong Zhang[1,2,7*], Fei Shan[2,7*]

[1]Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai, China; [2]Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China; [3]Department of Radiology, Beilun Second People's Hospital, Ningbo, China; [4]Department of Thoracic Surgery, Beilun Second People's Hospital, Ningbo, China; [5]Department of Traditional Chinese Medicine, Shanghai Public Health Clinical Center, Fudan University, Shanghai, China; [6]Department of Radiology, Changzheng Hospital, Naval Medical University, Shanghai, China; [7]Research Institute of Big Data, Fudan University, Shanghai, China

Contributions: (I) Conception and design: Y Zhuo, J Shen, Y Zhan, L Fan, F Shan; (II) Administrative support: Y Zhuo, J Shen; (III) Provision of study materials or patients: P Ye, S Yang, Y Tian, M Yu; (IV) Collection and assembly of data: Y Zhuo, J Shen; (V) Data analysis and interpretation: Y Zhuo, J Shen, Y Zhan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

[*]These authors contributed equally to this work and should be considered as co-senior authors.

Correspondence to: Fei Shan, PhD. Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, 2901 Caolang Road, Jinshan, Shanghai 201508, China; Research Institute of Big Data, Fudan University, Shanghai, China. Email: shanfei_2901@163.com; Zhiyong Zhang, PhD. Department of Radiology, Zhongshan Hospital, Fudan University, Shanghai 200032, China; Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, 2901 Caolang Road, Jinshan, Shanghai 201508, China; Research Institute of Big Data, Fudan University, Shanghai 200032, China. Email: zhyzhang@fudan.edu.cn.

Background: Broad generalization of radiomics-assisted models may be impeded by concerns about variability. This study aimed to evaluate the merit of combatting batch effect (ComBat) harmonization in reducing the variability of voxel size-related radiomics in both phantom and clinical study in comparison with image resampling correction method.

Methods: A pulmonary phantom with 22 different types of nodules was scanned by computed tomography (CT) with different voxel sizes. The variability of voxel size-related radiomics features was evaluated using concordance correlation coefficient (CCC), dynamic range (DR), and intraclass correlation coefficient (ICC). ComBat and image resampling compensation methods were used to reduce variability of voxel size-related radiomics. The percentage of robust radiomics features was compared before and after optimization. Pathologically differential diagnosis of invasive adenocarcinoma (IAC) from adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA) (AIS-MIA group) was used for clinical validation in 134 patients.

Results: Before optimization, the number of excellent features in the phantom and clinical data was 26.12% and 32.31%, respectively. The excellent features were increased after image resampling and ComBat correction. For clinical optimization, the effect of the ComBat compensation method was significantly better than that of image resampling, with excellent features reaching 90.96% and poor features only amounting to 4.96%. In addition, the hierarchical clustering analysis showed that the first-order and shape features had better robustness than did texture features. In clinical validation, the area under the curve (AUC) of the testing set was 0.865 after ComBat correction.

Conclusions: The ComBat harmonization can optimize voxel size-related CT radiomics variability in pulmonary nodules more efficiently than image resampling harmonization.

Keywords: Combatting batch effect (ComBat); computed tomography radiomics (CT radiomics); image resampling; pulmonary nodule; radiomics variability

Submitted Sep 19, 2022. Accepted for publication Jun 15, 2023. Published online Jul 13, 2023.
doi: 10.21037/qims-22-992
View this article at: https://dx.doi.org/10.21037/qims-22-992

Introduction

The recent rapid development of machine learning technology and computing power has enabled researchers to use big data of radiomics for precise clinical diagnosis (1). Radiomics refers to the high-throughput extraction of a large number of radiological image features of lesions, that is, the conversion of image information into digital information (2,3). Various studies have shown that the radiomics models have high diagnostic efficacy for pulmonary nodules in terms of differential diagnosis, nodule typing, preoperative prediction, or prognostic analysis (4-7). Lung cancer is one of the most serious diseases in the world, with a 5-year survival rate of around only 19% (8). The most common pathological subtype is lung adenocarcinoma (LUAD). However, it has been reported that the 5-year survival rate after complete surgery for preinvasive LUAD is close to 100% (9). Accurate preoperative diagnosis of invasive LUAD is conducive to clinical decision-making, such as in surgical determination. However, the selected features and radiomics models in previous studies have been inconsistent and with poor reproducibility, which has limited their application in real-world medical practice (10,11).

Advances in computed tomography (CT) technology has revolutionized diagnostic imaging, enabling consistent increases in image quality alongside the decrease of voxel size. Several studies have shown that ultra-high-resolution CT (U-HRCT, including 1,024×1,024, and 2,048×2,048 matrix size) scans can improve the image quality and the assessment of lung diseases in comparison with the 512×512 matrix size (12,13). Several recent studies on the effect of voxel size in the variability of radiomics have achieved inconsistent results. Paul *et al.* and Yang *et al.* found that voxel size strongly affected the reproducibility of the radiomics features (14,15), but Crandall *et al.* reported that the voxel size had a minimal effect in feature value (16).

Shafiq-Ul-Hassan *et al.* optimized the stability of radiomics features using image resampling (17). Image upsampling (increasing the voxel size) was generally used to explore whether it can reduce the radiomics variability, but it was shown that it could result in loss of valuable image information and affect the clinical diagnostic performance. Batch effects, which conceal biological signals and lead to deviations in subsequent data analysis (18), can be generated by either different experimental data sources or data collection processes of images with different voxel sizes. Although the combatting batch effect (ComBat) harmonization can reduce these radiomics variabilities (19,20), few relevant studies have been conducted to date.

In this study, we aimed to evaluate the ability of the ComBat harmonization to reduce variability of voxel size-related radiomics in the differential diagnosis of pulmonary nodules in both pulmonary phantom and actual patients.

Methods

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional ethics review boards of Shanghai Public Health Clinical Center, Shanghai Zhongshan Hospital, Beilun Second People's Hospital, and Shanghai Changzheng Hospital (No. SK2020-174) and the requirement for individual consent for this retrospective analysis was waived.

Phantom image acquisition

The anthropomorphic thorax phantom simulating an intermediate-sized adult (PH-1; Kyoto Kagaku, Kyoto, Japan) was used in this study (https://www.kyotokagaku.com/lineup/). There were 22 artificial nodules embedded in the artificial vascular bundle in the chest phantom: (I) 1 irregular nodule 15 mm in diameter; (II) 7 concentric mixed ground glass nodules (mGGNs) of different diameters

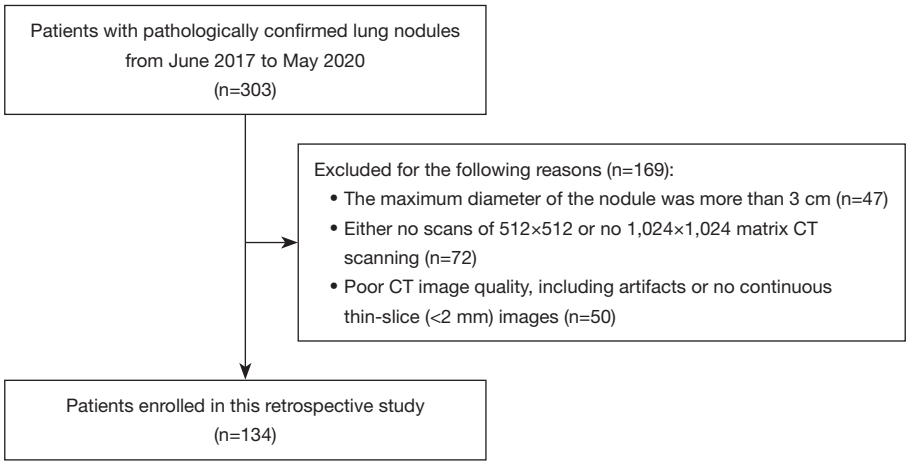

Figure 1 Recruitment pathway for patients in this study. CT, computed tomography.

(15 and 20 mm) with different spherical solid components (3, 5, 7, 9 mm); (III) 2 double eccentric mGGN 20 mm in diameter with different spherical solid components (3, 5, 7 mm); (IV) 12 pure nodules, including 4 pure solid nodules and 8 pure ground glass nodules (pGGNs) (details in Figure S1).

All data acquisitions were performed on a Philips Brilliance 16 slice CT scanner (Philips, Amsterdam, Netherlands). Scanning parameters were divided into two parts based on different voxel sizes: the conventional scanning with 512×512 matrix and the target scanning with 1,024×1,024 matrix. The voxel sizes of phantom data were 0.68×0.68×2.00 and 0.18×0.18×1.00 mm^3, respectively (details in Table S1).

Clinical image acquisition

In this study, a total of 303 patients with pulmonary nodules were retrospectively recruited from June 1, 2017 to May 30, 2020 at Beilun Second People's Hospital. The patients were identified according to the recruitment pathway (*Figure 1*). Finally, a total of 134 patients (46 males and 88 females, age 59.46±10.48 years) who met the inclusion criteria were enrolled in the study (*Table 1*).

As for the lung phantom, all images of patients were performed on the Philips Brilliance 16 slice CT scanner. The scanning parameters were also the same as those of the phantom data: conventional and target scanning. However, the patients were not retested because of the radiation dose. The voxel sizes of clinical data were 0.62×0.62×2.00–0.77×0.77×2.00 and 0.15×0.15×1.00–0.23×0.23×1.00 mm^3, respectively (details in Table S1). A flow diagram of this study is shown in *Figure 2*.

Region of interest segmentation and radiomics features extraction

Deep learning-based pulmonary nodule segmentation (both in phantom and clinical applications) was performed using the uAI Research Portal software (United Imaging Intelligence Inc., Shanghai, China) that was embedded into the widely used package-PyRadiomics (https://pyradiomics.readthedocs.io/en/latest/index.html). All segmented three-dimensional (3D) nodules were reviewed by two chest radiologists (Y Zhuo and Y Zhan) and manually adjusted if necessary. The uAI Research Portal software was used to extract radiomics features, including first-order statistics, shape, and texture features. A total of 14 filters were used during the process of features extraction (details in Appendix 1). Finally, a total of 2,600 radiomic features were extracted for each 3D nodule.

ComBat compensation method

To reduce radiomics variability from batch effects, the ComBat compensation method was performed on the BatchServer (https://lifeinfo.shinyapps.io/batchserver/) (21). Principal variance component analysis (PVCA) fits a mixed linear model to estimate the variation ratio of each factor, and was thus used to evaluate the effectiveness of batch effect correction. The emerging nonlinear dimensionality reduction method, uniform manifold approximation

Table 1 The clinical and CT images characteristics of participants with lung nodules

Characteristics	512×512 matrix (n=134)	1,024×1,024 matrix (n=134)	P value
Age (years)	59.46±10.48		–
Gender (F/M)	88/46		–
Nodule type			
pGGN	54 (40.30)		–
mGGN	75 (55.97)		–
Solid nodule	5 (3.73)		–
Maximum diameter of nodule (mm)	12.79±5.55	13.06±5.54	0.642
Minimum diameter of nodule (mm)	9.89±3.99	10.06±4.07	0.766
Mean diameter of nodule (mm)	11.34±4.53	11.56±4.59	0.676

Data are presented as mean ± standard deviation, n, or n (%). CT, computed tomography; F, female; M, male; pGGN, pure ground glass nodule; mGGN, mixed ground glass nodule.

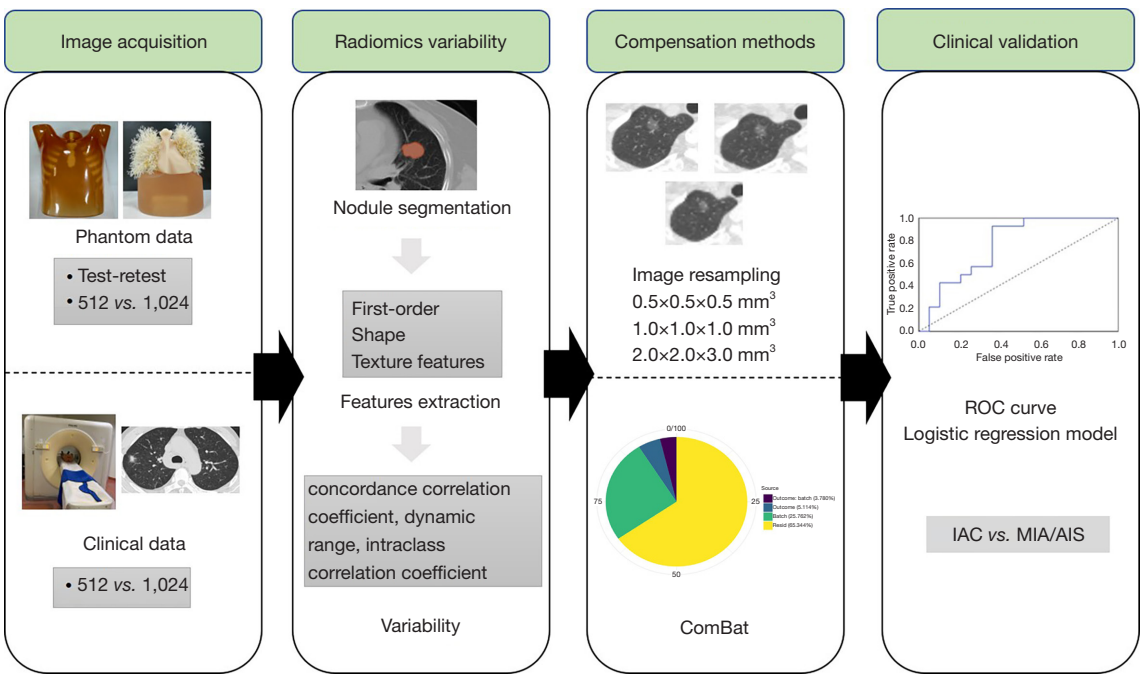

Figure 2 Methodology flowchart of this study. ComBat, combatting batch effect; ROC, receiver operating characteristic; IAC, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; AIS, adenocarcinoma in situ.

and projection (UMAP), was also used to evaluate the effectiveness of batch effect correction in this study.

Image resampling

The original phantom and clinical images had different voxel sizes, 0.15×0.15×1.00–0.77×0.77×2.00 mm^3. To reduce voxel size-related radiomics variability, all images were resampled for 3 different times: 0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×3.0 mm^3, which included not only the conventional upsampling but also the downsampling so as to preserve as much image information as possible.

Table 2 The type of radiomics features based on CCC, DR, and ICC in phantom and clinical data

Type	Phantom data	Clinical data
Excellent	CCC ≥0.90 and DR ≥0.90 and ICC ≥0.75	ICC ≥0.75
Moderate	CCC ≥0.90 and DR ≥0.90 and 0.40≤ ICC <0.75	0.40≤ ICC <0.75
Poor	CCC <0.90 or DR <0.90 or ICC <0.40	ICC <0.40

CCC, concordance correlation coefficient; DR, dynamic range; ICC, intraclass correlation coefficient.

Radiomics variability and correction analysis

In the phantom application, test-retest robustness was calculated on concordance correlation coefficient (CCC) and dynamic range (DR) (22-24). Radiomics features with good reproducibility were defined when CCC and DR ≥0.90 (22).

Both in phantom and clinical applications, radiomics variability was identified by intraclass correlation coefficient (ICC) between different voxel sizes. ICC was divided into 3 levels: ICC ≥0.75, 0.40≤ ICC <0.75, and ICC <0.40 (25,26). The radiomics features were defined as excellent, moderate, and poor based on CCC, DR, and ICC (*Table 2*).

Clinical evaluation

The clinical evaluation dataset was enrolled from July 1, 2018 to June 30, 2020 at Beilun Second People's Hospital and Shanghai Changzheng Hospital. Some studies suggested that sublobectomy could be an alternative approach for LUAD of no more than 2 cm in diameter (27-29). A clinical evaluation dataset with 512×512 matrix containing 186 lung nodules (pathologically confirmed LUAD) with a mean diameter of less than 2 cm was used in this study, including 88 invasive adenocarcinoma (IAC) and 98 adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA) (AIS-MIA group). In the MIA-AIS group, there were 73 cases of MIA and 25 of AIS (Table S2).

The radiomics features of all lung nodules were extracted and a logistic regression model was built after radiomics features selection. The models were also built with the excellent features after ComBat and image resampling correction. The receiver operating characteristic (ROC) curves were used to evaluate the performance of radiomics signature models.

Statistical analysis

The Wilcoxon rank-sum test was used for continuous variables between the two groups, and the categorical variables were compared with χ^2 test. The SPSS software (version 20; IBM Corp., Armonk, NY, USA) was used to perform all statistical analysis. Statistical significance was indicated by a two-tailed P value of less than 0.05. Measurement data were expressed as mean ± standard deviation (SD).

We used two feature selection methods, max-relevance and min-redundancy (mRMR) and least absolute shrinkage and selection operator (LASSO), to select radiomics features. The LASSO method constructed a penalty function by adding constraint conditions, and a prediction model was constructed by performing 10-fold cross-validation. DeLong's test was used between different ROC curves. Heml software (http://hemi.biocuckoo.org) was used to make a heatmap to visually show the radiomics variability. Hierarchical clustering analysis was used to evaluate the redundancy of CT radiomics features.

Results

Information of phantom and clinical data

The 22 artificial nodules of phantom included 8 pGGNs (36.36%), 9 mGGNs (40.91%), and 5 solid nodules (22.73%), of which 21 were spherical and 1 was irregular in shape.

The clinical and CT characteristics of 134 patients are shown in *Table 1* and Figure S2, including 54 pGGNs (40.30%), 75 mGGNs (55.97%), and 5 solid nodules (3.73%).

Voxel size-related radiomics variability

In phantom application, among 2,600 features, 1,788 (68.77%) radiomics features had both CCC and DR ≥0.90. According to ICC value, there were 679 (26.12%) excellent, 325 (12.50%) moderate, and 1,596 (61.38%) poor features. Details are displayed in *Table 3* and *Figure 3*.

In clinical application, there were 840 (32.31%) excellent features, 384 (14.77%) moderate features, and 1,321

Table 3 The number of radiomics features before and after compensation methods both in phantom and clinical data

Data	Type	Before optimization	Image resampling (mm^3)			ComBat
			0.5×0.5×0.5	1.0×1.0×1.0	2.0×2.0×3.0	
Phantom data	Excellent	679 (26.12)	1,655 (63.65)	1,558 (59.92)	1,655 (63.65)	1,085 (41.73)
	Moderate	325 (12.50)	502 (19.31)	661 (25.43)	654 (25.16)	693 (26.65)
	Poor	1,596 (61.38)	443 (17.04)	381 (14.65)	291 (11.19)	822 (31.62)
Clinical data	Excellent	840 (32.31)	1,394 (53.62)	1,519 (58.42)	1,251 (48.12)	2,365 (90.96)
	Moderate	384 (14.77)	637 (24.50)	745 (28.66)	1,116 (42.92)	106 (4.08)
	Poor	1,321 (50.81)	569 (21.88)	336 (12.92)	233 (8.96)	129 (4.96)

Data are presented as n (%). ComBat, combatting batch effect.

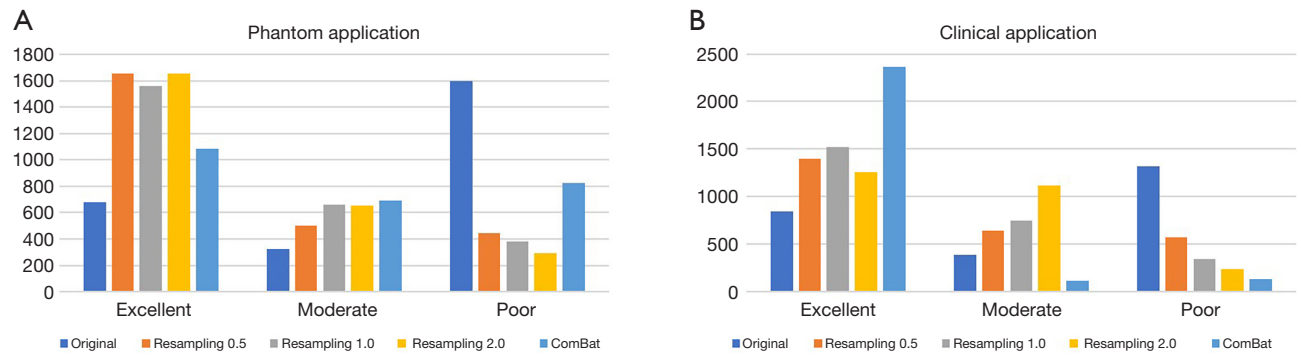

Figure 3 The number of different-type radiomics features before and after optimization. The excellent features were increased after image resampling and ComBat correction, whereas poor features were reduced both in phantom and clinical application. (A) For phantom application, the optimization effect of ComBat compensation method was worse than image resampling; (B) for clinical application, the optimization effect of ComBat compensation method was better than image resampling. Original, before optimization; resampling 0.5, image resampling of 0.5×0.5×0.5 mm^3; resampling 1.0, image resampling of 1.0×1.0×1.0 mm^3; resampling 2.0, image resampling of 2.0×2.0×2.0 mm^3. ComBat, combatting batch effect.

(50.81%) poor features according to ICC value (details in *Table 3* and *Figure 3*).

ComBat compensation

The PVCA and UMAP analyses showed that ComBat optimization reduced batch effects from 46.39% to 14.90% in the phantom and from 25.76% to 14.17% in clinical application (*Figure 4* and Figure S3).

CT scanning on the phantom displayed 1,085 (41.73%) excellent, 693 (26.65%) moderate, and 822 (31.62%) poor features in contrast to 2,365 (90.96%) excellent, 106 (4.08%) moderate, and 129 (4.96%) poor features in patients based on ICC value after ComBat correction (*Table 3* and Figure S4).

Image resampling

The CT resampling images of the phantom and clinical images are shown in *Figure 5*. The 3 different image resampling sizes (0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×3.0 mm^3) in the phantom showed that the numbers of features with CCC and DR ≥0.90 were 2,358 (90.69%), 2,298 (88.38%), and 2,351 (90.42%), respectively, including 1,655 (63.65%) excellent, 502 (19.31%) moderate, and 443 (17.04%) poor features based on ICC value after image resampling of 0.5×0.5×0.5 mm^3. Among these features, the image resampling of 1.0×1.0×1.0 mm^3 displayed 1,558 (59.92%) excellent, 661 (25.43%) moderate, and 381 (14.65%) poor features in contrast to 1,655 (63.65%) excellent, 654 (25.16%) moderate, and 291 (11.19%) poor

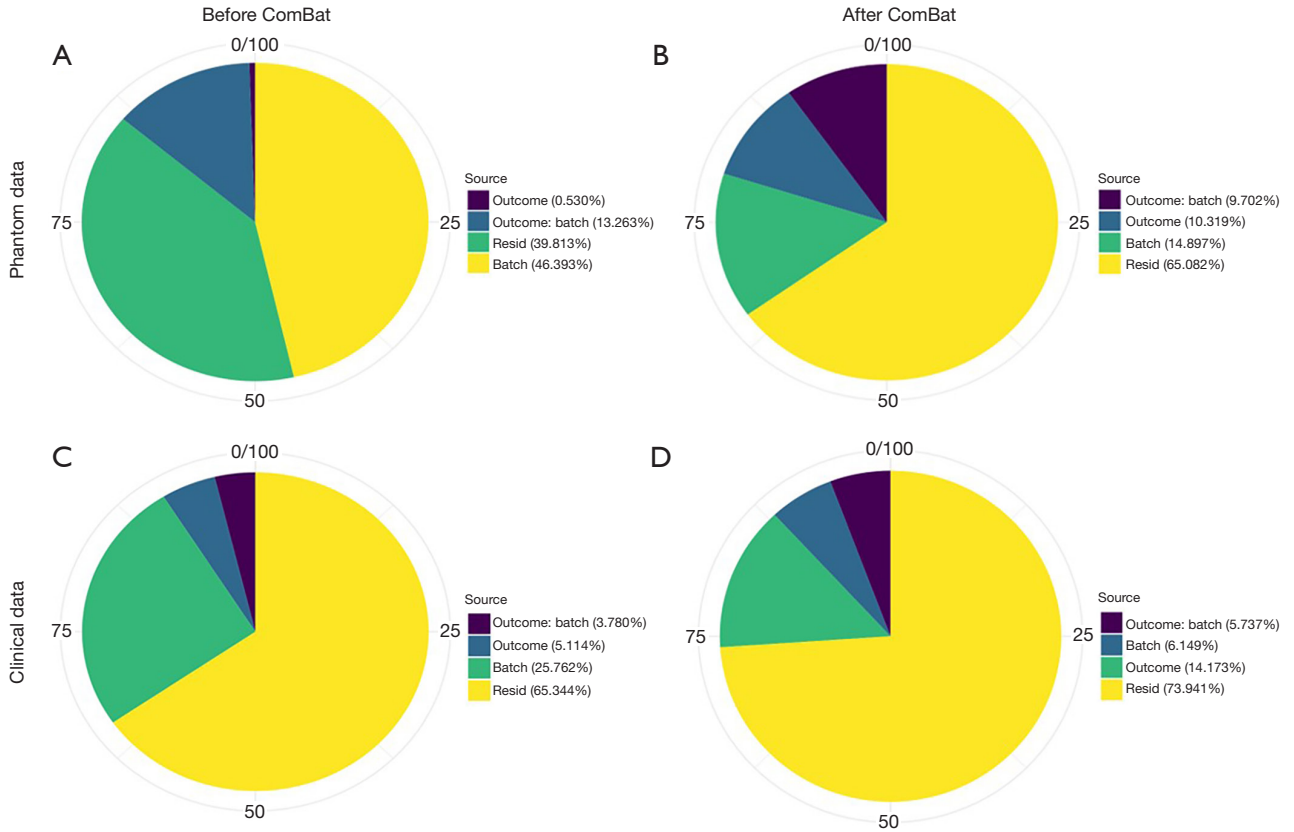

Figure 4 The PVCA results of ComBat compensation method both in phantom and clinical application. (A,B) Batch effect decreased from 46.39% to 14.90% in phantom application; (C,D) batch effect decreased from 25.76% to 14.17% in clinical application. ComBat, combatting batch effect; PVCA, principal variance component analysis.

features image resampling of 2.0×2.0×3.0 mm^3 based on ICC value (*Table 3* and Figure S5).

In clinical application, there were 1,394 (53.62%) excellent, 637 (24.50%) moderate, and 569 (21.88%) poor features after image resampling of 0.5×0.5×0.5 mm^3, in contrast to 1,519 (58.42%) excellent, 745 (28.66%) moderate, and 336 (12.92%) poor features in image resampling of 1.0×1.0×1.0 mm^3, and 1,251 (48.12%) excellent, 1,116 (42.92%) moderate, and 233 (8.96%) poor features in image resampling of 2.0×2.0×2.0 mm^3 based on ICC value (*Table 3* and Figure S6).

Redundancy of CT radiomics features

For excellent radiomics features, the de-redundancy hierarchical clustering analysis divided the original features into 10 groups in contrast to 15 groups by image resampling and ComBat compensation methods, respectively. The heatmap and hierarchical clustering analysis also showed that the first-order, shape, and texture features were 208 (30.63%), 275 (40.50%), and 196 (28.87%), respectively, among 679 excellent features in phantom data in contrast to 391 (46.55%), 250 (29.76%), and 199 (23.69%), respectively, among 840 excellent features in clinical data before optimization (*Figure 6*).

Clinical evaluation

A logistic regression model of pre-optimization was built using 11 selected radiomics features with non-zero coefficients; the area under the curve (AUC) of the training set was 0.998 (accuracy, 0.983; sensitivity, 0.984; specificity, 0.983), whereas that of testing set was 0.763 (accuracy, 0.545; sensitivity, 1.000; specificity, 0.211).

A logistic regression model of ComBat correction was built using 5 selected radiomics features with non-zero coefficients; the AUC of the training set was 0.997 (accuracy, 0.977; sensitivity, 0.976; specificity, 0.977), whereas that

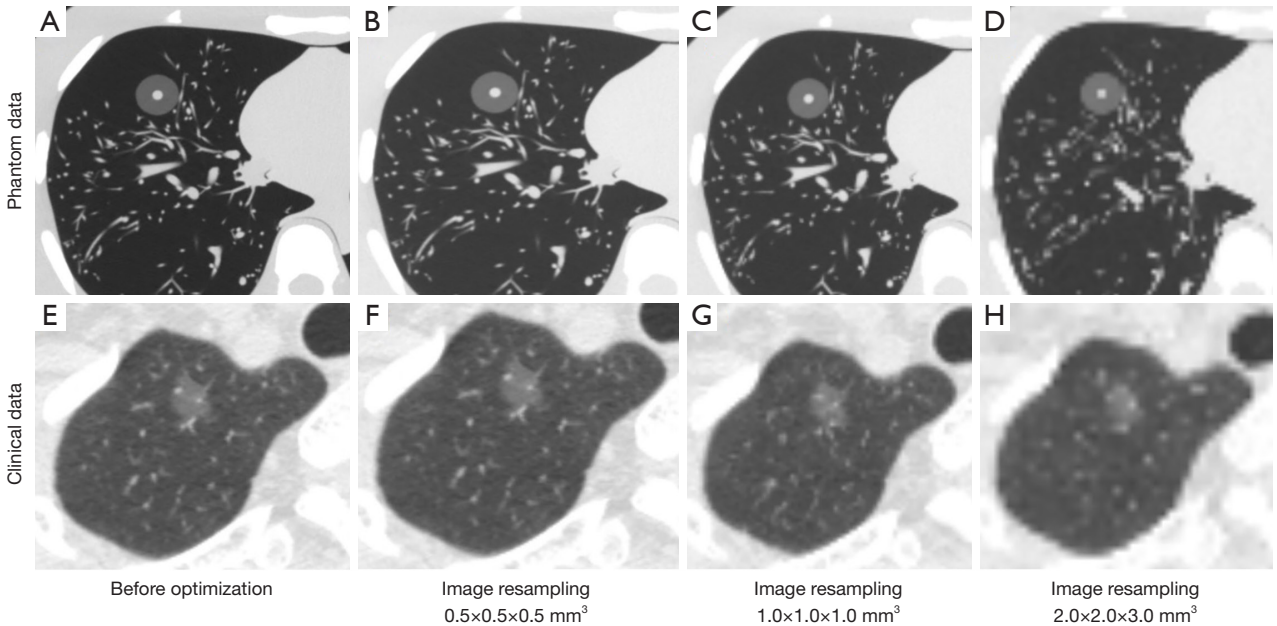

Figure 5 The resampling images of the phantom and clinical images. (A-D) The original voxel sizes of phantom data were 0.68×0.68×2.00 and 0.18×0.18×1.00 mm^3, all images were resampled for 3 different times: 0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×3.0 mm^3; (E-H) the original voxel sizes of clinical data were 0.62×0.62×2.00–0.77×0.77×2.00 and 0.15×0.15×1.00–0.23×0.23×1.00 mm^3, all images were resampled for 3 different times: 0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×3.0 mm^3.

of testing set was 0.865 (accuracy, 0.736; sensitivity, 0.929; specificity, 0.630). The optimization effect of the ComBat compensation method was significantly better than image resampling correction of 0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×2.0 mm^3 (P=0.012, 0.017, and 0.034, respectively; details in *Table 4* and Figure S7).

Discussion

Our study demonstrated that the voxel size affects the variability of radiomics features of phantom and clinical applications, which can be significantly improved by ComBat and image resampling correction. The image resampling was superior to ComBat compensation in phantom application, whereas ComBat compensation was significantly better than image resampling in clinical application. In addition, the hierarchical clustering analysis showed that the first-order and shape features of the original images had better robustness than those of texture features with different filters.

In this study, scanning parameters were divided into two parts based on different voxel sizes: the conventional scanning with 512×512 matrix, and the target scanning with 1,024×1,024 matrix that is a local scanning for nodules with the advantage of resolution improvement to provide more detailed features and simplify qualitative diagnosis, as indicated by previous reports about U-HRCT imaging (12,13).

Radiomics can provide a large amount of information to help clinical decision-making, but variability is a hinderance to its clinical application requiring urgent rectification. The variability in CT scans is mainly due to large differences in CT scan parameters, particularly the differences in voxel size. In our study, more than half of the features were poor after changing the voxel size both in phantom and clinical applications. Yang *et al.* found that voxel size strongly affected the robustness of positron emission tomography/magnetic resonance (PET/MR) radiomics features, and that 0.5×0.5×1.0 mm^3 of voxel size was optimal for features in the primary tumor of nasopharyngeal carcinoma (15). The variability of the voxel size can be obtained by analyzing the pixel size and slice spacing. Paul *et al.* found that most deep features changed with the pixel size, which was the same for many traditional radiomics features (14). Our findings were consistent with those of the above reports. However, Crandall *et al.* found that the reproducibility of most PET/CT features was barely affected by changes in voxel size in cervical cancer (16).

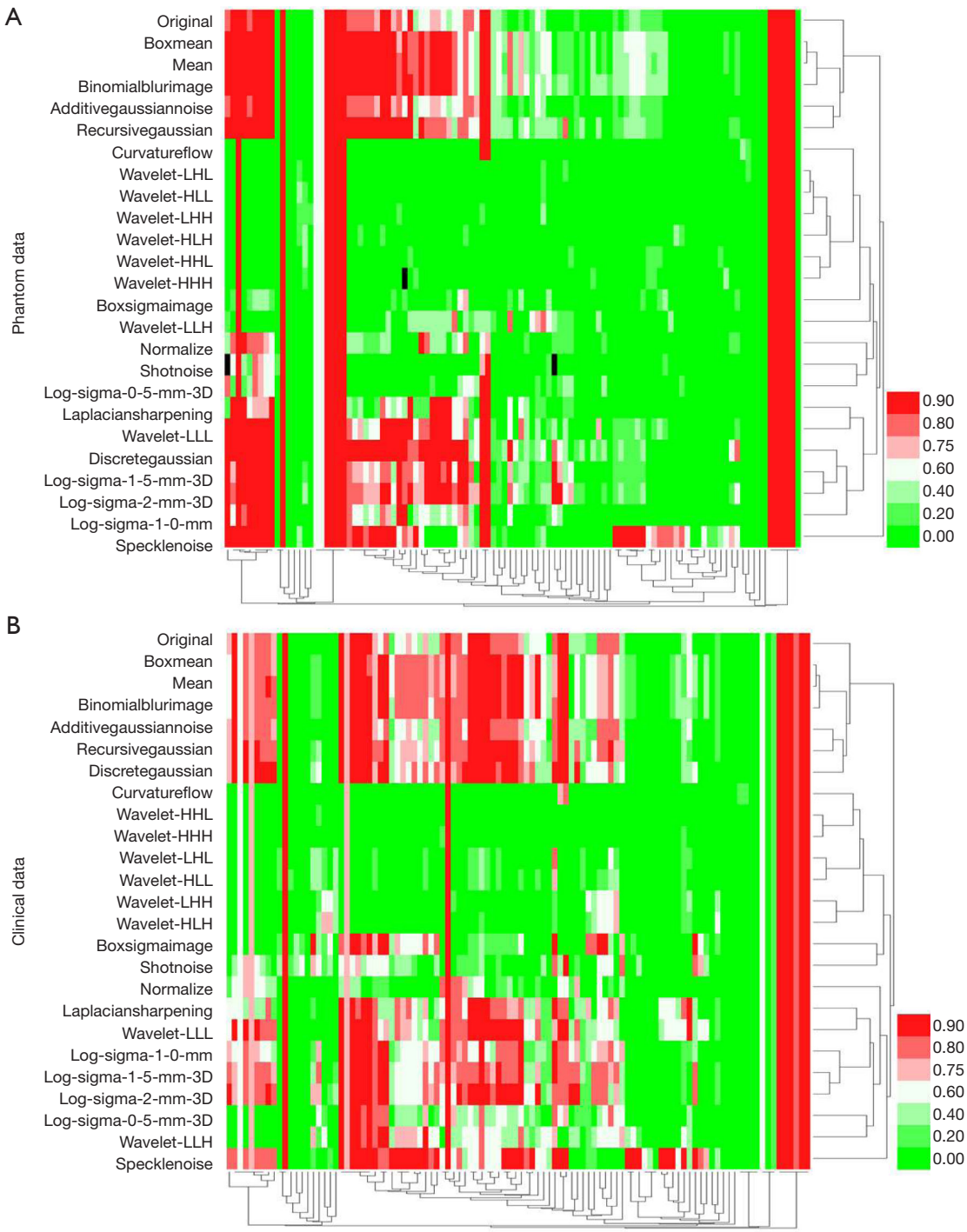

Figure 6 The results of heatmap and hierarchical clustering analysis of phantom and clinical data before optimization. (A) In phantom data, there were 679 (26.12%) excellent features, 325 (12.50%) moderate features, and 1,596 (61.38%) poor features. The hierarchical clustering was stopped arbitrarily at 10 groups of features; (B) in clinical application, there were 840 (32.31%) excellent features, 384 (14.77%) moderate features, and 1,321 (50.81%) poor features. The hierarchical clustering was stopped arbitrarily at 10 groups of features. The color map of the cystogram ranges from 0 to 1, values close to 1 have a red shade and values close to 0 have a green shade. The groups with red shade mean relatively high feature values. Black shade: empty value. H, high; L, low; 3D, three-dimensional.

Table 4 Performance evaluation of models based on radiomics features before and after optimization in clinical validation

ROC	Before optimization	ComBat	Resampling 0.5	Resampling 1.0	Resampling 2.0
Training set					
AUC	0.998	0.997	0.978	0.989	0.991
Accuracy	0.983	0.977	0.930	0.947	0.953
Sensitivity	0.984	0.976	0.929	0.960	0.960
Specificity	0.983	0.977	0.897	0.937	0.949
Testing set					
AUC	0.763	0.865	0.711	0.778	0.812
Accuracy	0.545	0.736	0.485	0.576	0.576
Sensitivity	1.000	0.929	0.897	0.929	1.000
Specificity	0.211	0.630	0.158	0.316	0.263

Resampling 0.5, image resampling of 0.5×0.5×0.5 mm^3; resampling 1.0, image resampling of 1.0×1.0×1.0 mm^3; resampling 2.0, image resampling of 2.0×2.0×3.0 mm^3. ROC, receiver operating characteristic; ComBat, combatting batch effect; AUC, area under the curve.

Both image resampling and the ComBat harmonization could optimize voxel size-related CT radiomics variability in our study, but the diagnostic performance of the model with ComBat compensation method was significantly better than that of models with image resampling correction and model without optimization in the testing set ($P<0.05$) in clinical evaluation. The resampling method was traditionally considered to resolve the source of variation caused by different voxel sizes. Shafiq-Ul-Hassan *et al.* found that image resampling was an appropriate preprocessing step to obtain more reproducible CT radiomics features (17,30). Another study had shown that image resampling accompanied by Butterworth low pass filtering could effectively reduce the variability due to different voxel sizes (31).

The ComBat method was derived from genomics research, and the intended function was to reduce batch effects of data collection processes (32). Orlhac *et al.* compared the radiomics feature distributions before and after application of the ComBat compensation method and showed that the cluster distribution of each texture feature corresponding to different imaging parameters was corrected after fitting (20). Orlhac *et al.* also used the ComBat method to effectively adjust the PET radiomics feature distributions under 3 different imaging parameter settings, making multi-center joint research possible (19). Ligero *et al.* explored resampling, ComBat, and singular value decomposition (SVD) compensation methods for reducing CT radiomics variability, and found that ComBat showed the highest improvement of feature classification (33). In our study, not only the AUC but also the sensitivity and specificity of the ComBat model were higher than those of other models after ComBat optimization, because the radiomics features with small variability were selected after ComBat optimization, and the established model was more stable and more universal. The ComBat optimization method can not only remove the batch effect caused by different voxels, but also remove the batch effect caused by other factors, such as CT vendors (33). This might be one of the reasons for ComBat compensation performing significantly better than image resampling in our clinical application.

Our hierarchical clustering analysis showed that the first-order and shape features of the original images had better robustness than those of texture features with different filters, which was consistent with previous reports (34,35). A systematic review showed that first-order CT features had higher repeatability and reproducibility than shape and texture features according to 41 full-text articles (34). Tunali *et al.* found that statistical, histogram, and a subset of texture features tended to be stable in peritumoral regions of lung cancer lesions (35). The first-order feature describes the distribution of individual voxel values, regardless of the spatial position relationship, whereas the texture features describe the relationship between voxels and regional contrast, which can be used to assess tumor heterogeneity. The high-level features emphasize the use of filters on the image processing and features extraction. Therefore, in

comparison with high-order features and texture, the first-order features obtained from the histogram are more stable.

There were several limitations to this study. First, this study investigated the variability of voxel size-related CT radiomics features, but there would be other variables in the research process, such as the error of nodule segmentation, which could also affect the results. Second, we took voxel size as a whole, and did not conduct a more detailed study of pixel size and slice spacing. Third, different CT vendors and scanning parameters will result in more batch effects in radiomics analysis, which can be reduced by ComBat (33); multi-center validations are needed to evaluate the ability of ComBat harmonization. In addition, there were statistical differences between the gender, age, and nodule type of patients in the IAC and MIA-AIS groups in clinical evaluation. Although these factors may have certain impacts on the results, the data collected in a continuous period of time could better reflect the actual clinical problem.

In conclusion, the ComBat harmonization can optimize voxel size-related variability of CT radiomics both in phantom and clinical application, and the diagnostic efficiency and generalization of the optimized features are improved based on clinical evaluation. According to the results of this prospective multicenter large-scale study in which the variability of the scanning parameter was a major limitation, the ComBat method is easy to apply and endows the potential of radiomics.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (General Program, No. 82172030), the Intelligent Medical Special Research Foundation of the Shanghai Health and Family Planning Commission (No. 2018ZHYL0104), the Clinical Research Plan of SHDC (No. SHDC2020CR3080B), the Ningbo Medical Science and Technology Project (No. 2018A14), the National Key R&D Program of China (Nos. 2016YFE0103000 and 2017YFC1308703), and the National Natural Science Foundation of China (No. 81871321).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-992/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional ethics review boards of Shanghai Public Health Clinical Center, Shanghai Zhongshan Hospital, Beilun Second People's Hospital and Shanghai Changzheng Hospital (No. SK2020-174) and the requirement for individual consent for this retrospective analysis was waived.

References

1. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 2016;375:1216-9.
2. Lee G, Lee HY, Park H, Schiebler ML, van Beek EJR, Ohno Y, Seo JB, Leung A. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. Eur J Radiol 2017;86:297-307.
3. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, Bellomi M. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp 2018;2:36.
4. Fan L, Fang M, Li Z, Tu W, Wang S, Chen W, Tian J, Dong D, Liu S. Radiomics signature: a biomarker for the preoperative discrimination of lung invasive adenocarcinoma manifesting as a ground-glass nodule. Eur Radiol 2019;29:889-97.
5. She Y, Zhang L, Zhu H, Dai C, Xie D, Xie H, Zhang W, Zhao L, Zou L, Fei K, Sun X, Chen C. The predictive value of CT-based radiomics in differentiating indolent from invasive lung adenocarcinoma in patients with pulmonary nodules. Eur Radiol 2018;28:5121-8.
6. Wu G, Woodruff HC, Sanduleanu S, Refaee T, Jochems A, Leijenaar R, Gietema H, Shen J, Wang R, Xiong J,

Bian J, Wu J, Lambin P. Preoperative CT-based radiomics combined with intraoperative frozen section is predictive of invasive adenocarcinoma in pulmonary nodules: a multicenter study. Eur Radiol 2020;30:2680-91.
7. Zhao W, Xu Y, Yang Z, Sun Y, Li C, Jin L, Gao P, He W, Wang P, Shi H, Hua Y, Li M. Development and validation of a radiomics nomogram for identifying invasiveness of pulmonary adenocarcinomas appearing as subcentimeter ground-glass opacity nodules. Eur J Radiol 2019;112:161-8.
8. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70:7-30.
9. Zhang Y, Ma X, Shen X, Wang S, Li Y, Hu H, Chen H. Surgery for pre- and minimally invasive lung adenocarcinoma. J Thorac Cardiovasc Surg 2022;163:456-64.
10. Lee SH, Cho HH, Lee HY, Park H. Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer. Cancer Imaging 2019;19:54.
11. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones AK, Court L. Measuring Computed Tomography Scanner Variability of Radiomics Features. Invest Radiol 2015;50:757-65.
12. Hata A, Yanagawa M, Honda O, Kikuchi N, Miyata T, Tsukagoshi S, Uranishi A, Tomiyama N. Effect of Matrix Size on the Image Quality of Ultra-high-resolution CT of the Lung: Comparison of 512×512, 1024×1024, and 2048×2048. Acad Radiol 2018;25:869-76.
13. Yanagawa M, Tsubamoto M, Satoh Y, Hata A, Miyata T, Yoshida Y, Kikuchi N, Kurakami H, Tomiyama N. Lung Adenocarcinoma at CT with 0.25-mm Section Thickness and a 2048 Matrix: High-Spatial-Resolution Imaging for Predicting Invasiveness. Radiology 2020;297:462-71.
14. Paul R, Shafiq-Ul Hassan M, Moros EG, Gillies RJ, Hall LO, Goldgof DB. Deep Feature Stability Analysis Using CT Images of a Physical Phantom Across Scanner Manufacturers, Cartridges, Pixel Sizes, and Slice Thickness. Tomography 2020;6:250-60.
15. Yang P, Xu L, Cao Z, Wan Y, Xue Y, Jiang Y, Yen E, Luo C, Wang J, Rong Y, Niu T. Extracting and Selecting Robust Radiomic Features from PET/MR Images in Nasopharyngeal Carcinoma. Mol Imaging Biol 2020;22:1581-91.
16. Crandall JP, Fraum TJ, Lee M, Jiang L, Grigsby P, Wahl RL. Repeatability of (18)F-FDG PET Radiomic Features in Cervical Cancer. J Nucl Med 2021;62:707-15.
17. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, Abdalah MA, Schabath MB, Goldgof DG, Mackin D, Court LE, Gillies RJ, Moros EG. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44:1050-62.
18. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís DY, Duque R, Bersini H, Nowé A. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform 2013;14:469-90.
19. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. J Nucl Med 2018;59:1321-8.
20. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. Radiology 2019;291:53-9.
21. Zhu T, Sun R, Zhang F, Chen GB, Yi X, Ruan G, Yuan C, Zhou S, Guo T. BatchServer: A Web Server for Batch Effect Evaluation, Visualization, and Correction. J Proteome Res 2021;20:1079-86.
22. Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, Goldgof DB, Hall LO, Korn R, Zhao B, Schwartz LH, Basu S, Eschrich S, Gatenby RA, Gillies RJ. Test-retest reproducibility analysis of lung CT image features. J Digit Imaging 2014;27:805-23.
23. Schuster DP. The opportunities and challenges of developing imaging biomarkers to study lung function and disease. Am J Respir Crit Care Med 2007;176:224-30.
24. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, Schwartz LH. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep 2016;6:23428.
25. Khan JN, Singh A, Nazir SA, Kanagala P, Gershlick AH, McCann GP. Comparison of cardiovascular magnetic resonance feature tracking and tagging for the assessment of left ventricular systolic strain in acute myocardial infarction. Eur J Radiol 2015;84:840-8.
26. Schmidt B, Dick A, Treutlein M, Schiller P, Bunck AC, Maintz D, Baeßler B. Intra- and inter-observer reproducibility of global and regional magnetic resonance feature tracking derived strain parameters of the left and right ventricle. Eur J Radiol 2017;89:97-105.
27. Nakamura K, Saji H, Nakajima R, Okada M, Asamura H, Shibata T, Nakamura S, Tada H, Tsuboi M. A phase III randomized trial of lobectomy versus limited resection for small-sized peripheral non-small cell lung

cancer (JCOG0802/WJOG4607L). Jpn J Clin Oncol 2010;40:271-4.
28. Stamatis G, Leschber G, Schwarz B, Brintrup DL, Ose C, Weinreich G, Passlick B, Hecker E, Kugler C, Dienemann H, Krbek T, Eggeling S, Hatz R, Müller MR, Weder W, Aigner C, Jöckel KH. Perioperative course and quality of life in a prospective randomized multicenter phase III trial, comparing standard lobectomy versus anatomical segmentectomy in patients with non-small cell lung cancer up to 2 cm, stage IA (7th edition of TNM staging system). Lung Cancer 2019;138:19-26.
29. Wen Z, Zhao Y, Fu F, Hu H, Sun Y, Zhang Y, Chen H. Comparison of outcomes following segmentectomy or lobectomy for patients with clinical N0 invasive lung adenocarcinoma of 2 cm or less in diameter. J Cancer Res Clin Oncol 2020;146:1603-13.
30. Shafiq-Ul-Hassan M, Latifi K, Zhang G, Ullah G, Gillies R, Moros E. Voxel size and gray level normalization of CT radiomic features in lung cancer. Sci Rep 2018;8:10545.
31. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, Court L. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One 2017;12:e0178524.
32. Da-Ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, Rousseau C, Mervoyer A, Reinhold C, Castelli J, De Crevoisier R, Rameé JF, Pradier O, Schick U, Visvikis D, Hatt M. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. Sci Rep 2020;10:10248.
33. Ligero M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, Mast R, Suarez C, Sala-Llonch R, Calvo N, Escobar M, Navarro-Martin A, Villacampa G, Dienstmann R, Perez-Lopez R. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. Eur Radiol 2021;31:1460-70.
34. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys 2018;102:1143-58.
35. Tunali I, Hall LO, Napel S, Cherezov D, Guvenis A, Gillies RJ, Schabath MB. Stability and reproducibility of computed tomography radiomic features extracted from peritumoral regions of lung cancer lesions. Med Phys 2019;46:5075-85.

Cite this article as: Zhuo Y, Shen J, Zhan Y, Tian Y, Yu M, Yang S, Ye P, Fan L, Zhang Z, Shan F. Optimization and validation of voxel size-related radiomics variability by combatting batch effect harmonization in pulmonary nodules: a phantom and clinical study. Quant Imaging Med Surg 2023;13(9):6139-6151. doi: 10.21037/qims-22-992

Appendix 1

Methods

Image acquisition

Scanning parameters were divided into 2 parts based on different voxel sizes both in phantom and clinical data. The first test scanning parameters were as follows: matrix, 512×512; field of view (FOV), 350 mm; slice thickness, 2 mm; tube voltage, 120 kV; tube current, auto mA; pitch, 0.813; collimator width, 16×0.575 mm;; rotation time, 0.5 s; lung window settings (width/level), 1,200/–600 Hounsfield units (HU); and mediastinal window settings (width/level), 350/40 HU. After repositioning, the retest data of the phantom was obtained by repeated scanning. The second test scanning parameters were as follows: matrix, 1,024×1,024; FOV, 180 mm; slice thickness, 1 mm; The remaining scan parameters were set the same as those of the first scanning.

Radiomic features extraction

The uAI Research Portal software was used to extract radiomic features, including first-order statistics, shape, and texture features. Texture features included gray-level co-occurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level run length matrix (GLRLM), gray-level dependence matrix (GLDM), and neighborhood gray tone difference matrix (NGTDM). The features extraction was filtered by 14 parameters including boxmean, additivegaussiannoise, binomialblurimage, curvatureflow, boxsigmaimage, normalize, laplaciansharpening, discretegaussian, mean, specklenoise, recursivegaussian, shotnoise, laplacian of gaussian, and wavelet.

The inclusion/exclusion criteria of validation data

A clinical validation dataset with 512×512 matrix containing 186 lung nodules (pathologically confirmed LUAD) with a mean diameter of less than 2 cm was used in this study from July 1, 2018 to June 30, 2020 at Beilun Second People's Hospital (Zhejiang, China) and Shanghai Changzheng Hospital (Shanghai China). Patients that met any 1 of the following criteria was excluded from the study: absence of CT examination within 1 month before surgery; absence of consecutive CT images with 1.5 mm thickness or less; with maximum diameter more than 2 cm; complications with other tumors or pulmonary disease (such as obstructive pneumonia); with severe respiratory motion artifacts; with preoperative treatment (such as neoadjuvant therapy).

https://dx.doi.org/10.21037/qims-22-992

Table S1 Parameters used for CT images both in phantom and clinical data

Scanning parameters	Phantom data		Clinical data	
	512×512	1,024×1,024	512×512	1,024×1,024
FOV (mm)	350	180	318–395	149–232
Thickness (mm)	2	1	2	1
Pitch	0.891	0.891	0.891	0.891
Rotation time (s)	0.5	0.5	0.5	0.5
Tube voltage (kV)	120	120	120	120
Tube current (mA)	Auto	Auto	Auto	Auto
Voxel size (mm^3)	0.68×0.68×2.00	0.18×0.18×1.00	0.62×0.62×2.00–0.77×0.77×2.00	0.15×0.15×1.00–0.23×0.23×1.00

CT, computed tomography; FOV, field of view.

Table S2 The clinical and CT images characteristics of participants with LUAD in clinical validation

Characteristics	Total (n=186)	IAC group (n=88)	MIA-AIS group (n=98)	P value
Age (years)	57.96±11.12	60.77±10.47	55.43±11.08	0.001
Gender (F/M)	125/61	52/36	73/25	0.019
Nodule type				0.000
pGGN	100 (55.76)	26 (29.55)	74 (75.51)	
mGGN	72 (38.71)	49 (55.68)	23 (23.47)	
Solid nodule	14 (7.53)	13 (14.77)	1 (1.02)	

Data are presented as mean ± standard deviation, n, or n (%). CT, computed tomography; LUAD, lung adenocarcinoma; IAC, invasive adenocarcinoma; MIA, minimally invasive adenocarcinoma; AIS, adenocarcinoma in situ; F, female; M, male; pGGN, pure ground glass nodule; mGGN, mixed ground glass nodule.

https://dx.doi.org/10.21037/qims-22-992

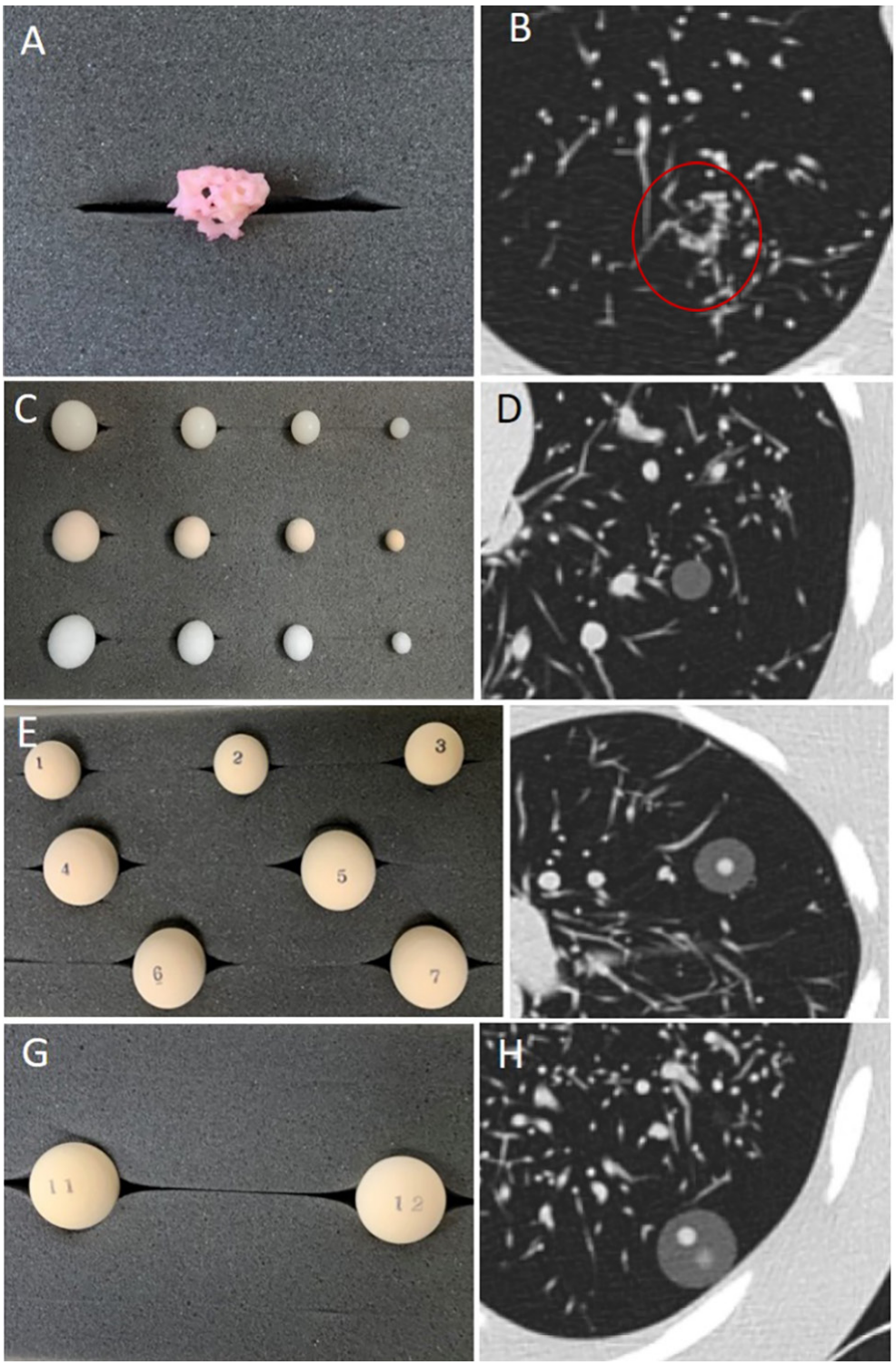

Figure S1 CT scanning of the artificial nodules embedded in the artificial vascular bundle in the chest phantom and. (A,B) Irregular nodule with 15 mm in diameter (red circle); (C,D) pure nodules, including 4 pure solid nodules and 8 pGGNs; (E,F) concentric mGGNs of different diameters (15 and 20 mm) with different spherical solid components (3, 5, 7, 9 mm); (G,H) double eccentric mGGN of 20 mm in diameter with different spherical solid components (3, 5, 7 mm). CT, computed tomography; pGGN, pure ground glass nodule; mGGN, mixed ground glass nodule.

https://dx.doi.org/10.21037/qims-22-992

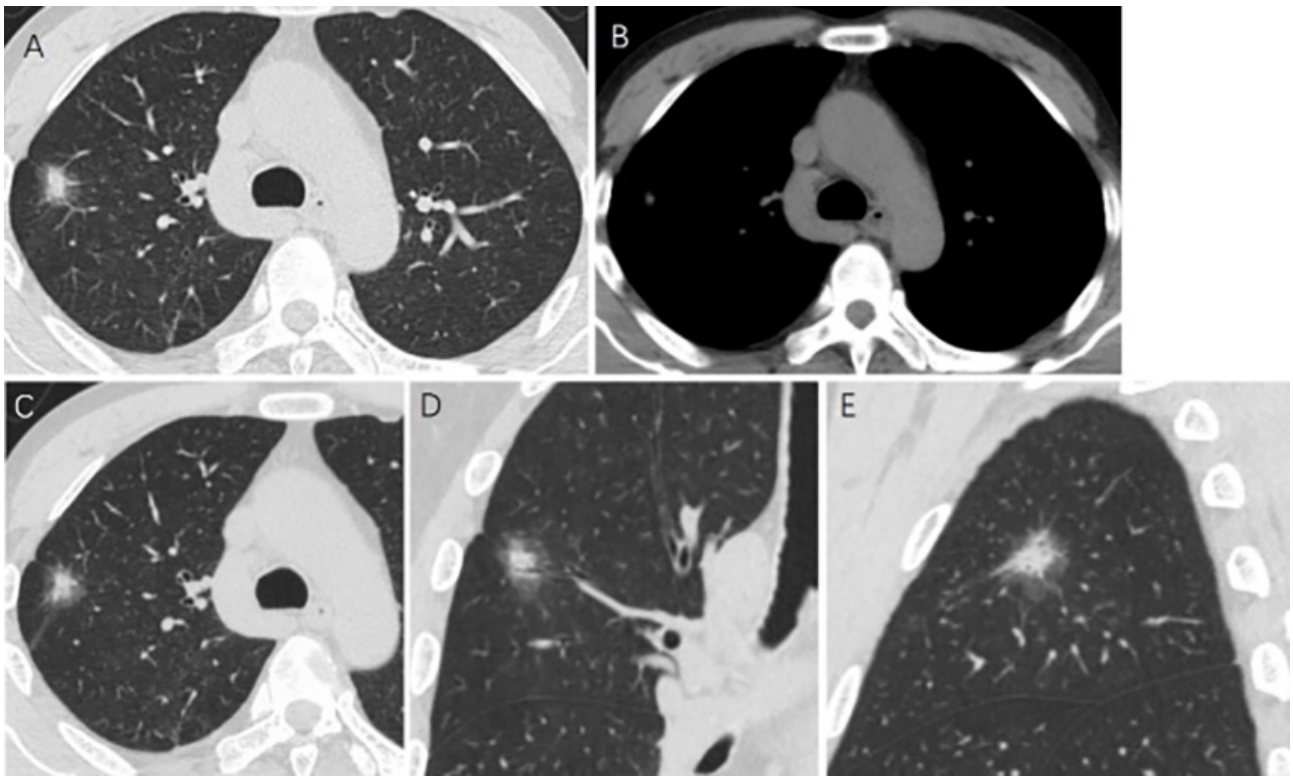

Figure S2 CT scanning of a 52-year-old man with pulmonary adenocarcinoma. (A,B) Images of lung and mediastinal window on conventional CT scanning with 512×512 matrix, the voxel size was 0.68×0.68×2.00 mm^3; (C-E) the transverse, coronal, and sagittal images on CT target scanning with 1,024×1,024 matrix, respectively, and the voxel size was 0.18×0.18×1.00 mm^3. CT, computed tomography.

https://dx.doi.org/10.21037/qims-22-992

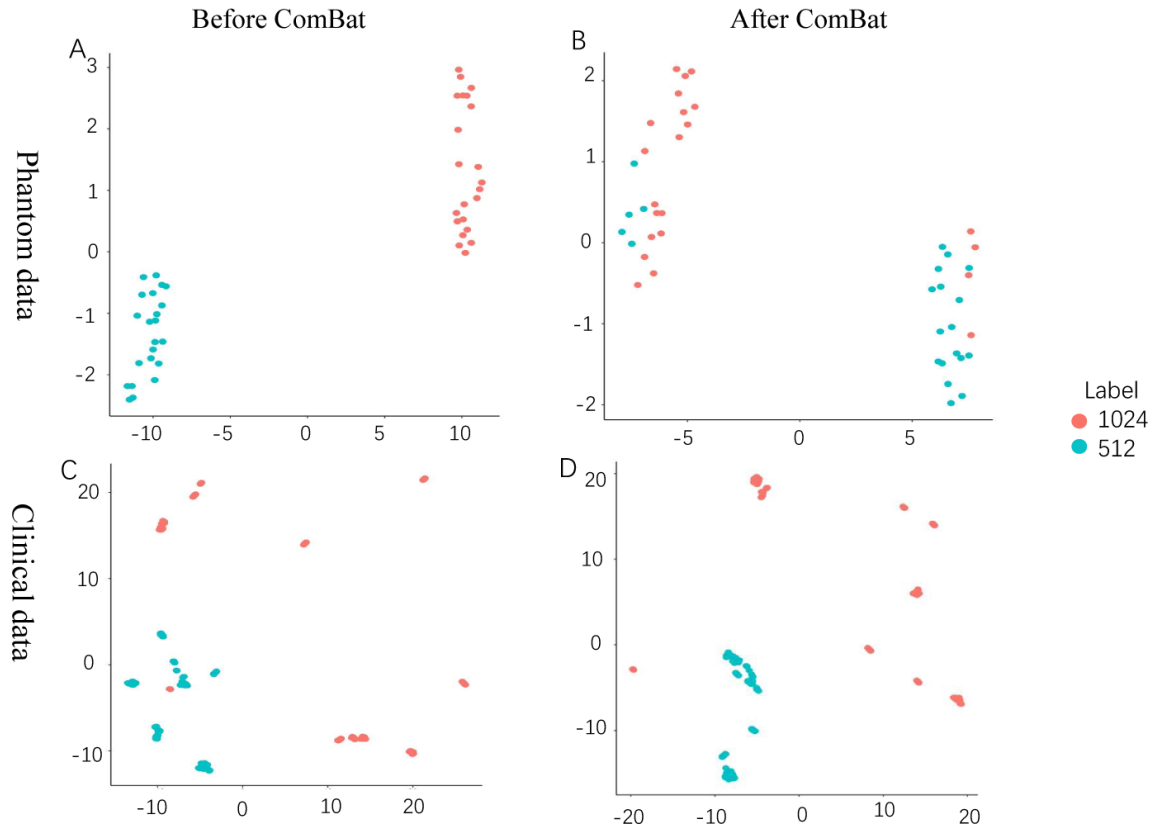

Figure S3 The UMAP by ComBat compensation method both in phantom and clinical application. (A,B) Batch effect decreased from 46.39% to 14.90% in phantom application; (C,D) batch effect decreased from 25.76% to 6.15% in clinical application. ComBat, combatting batch effect; UMAP, manifold approximation and projection.

https://dx.doi.org/10.21037/qims-22-992

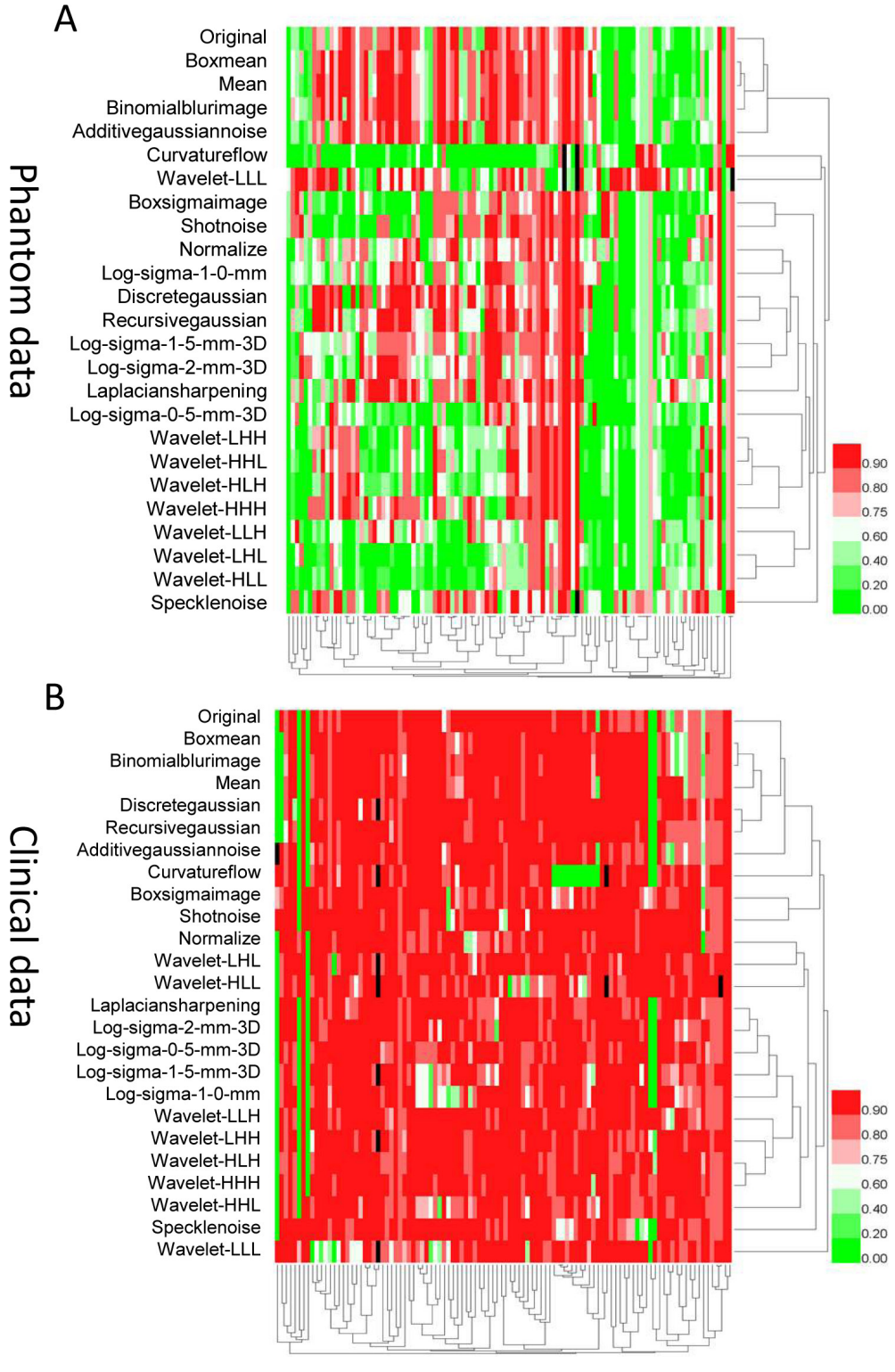

Figure S4 Heatmap of hierarchical clustering analysis of phantom data with ComBat correction. The excellent features were increased after ComBat correction and the poor features were reduced both in phantom and clinical data. (A) Phantom data of 1,085 (41.73%) excellent features, 693 (26.65%) moderate features, and 822 (31.62%) poor features. The hierarchical clustering was stopped arbitrarily at 15 groups of features; (B) clinical data of 2,365 (90.96%) excellent features, 106 (4.08%) moderate features, and 129 (4.92%) poor features. The hierarchical clustering was stopped arbitrarily at 15 groups of features. The color map of the clustogram ranges from 0 to 1, values close to 1 have a red shade and values close to 0 have a green shade. The groups with red shade mean relatively high feature values. Black shade: empty value. ComBat, combatting batch effect.

https://dx.doi.org/10.21037/qims-22-992

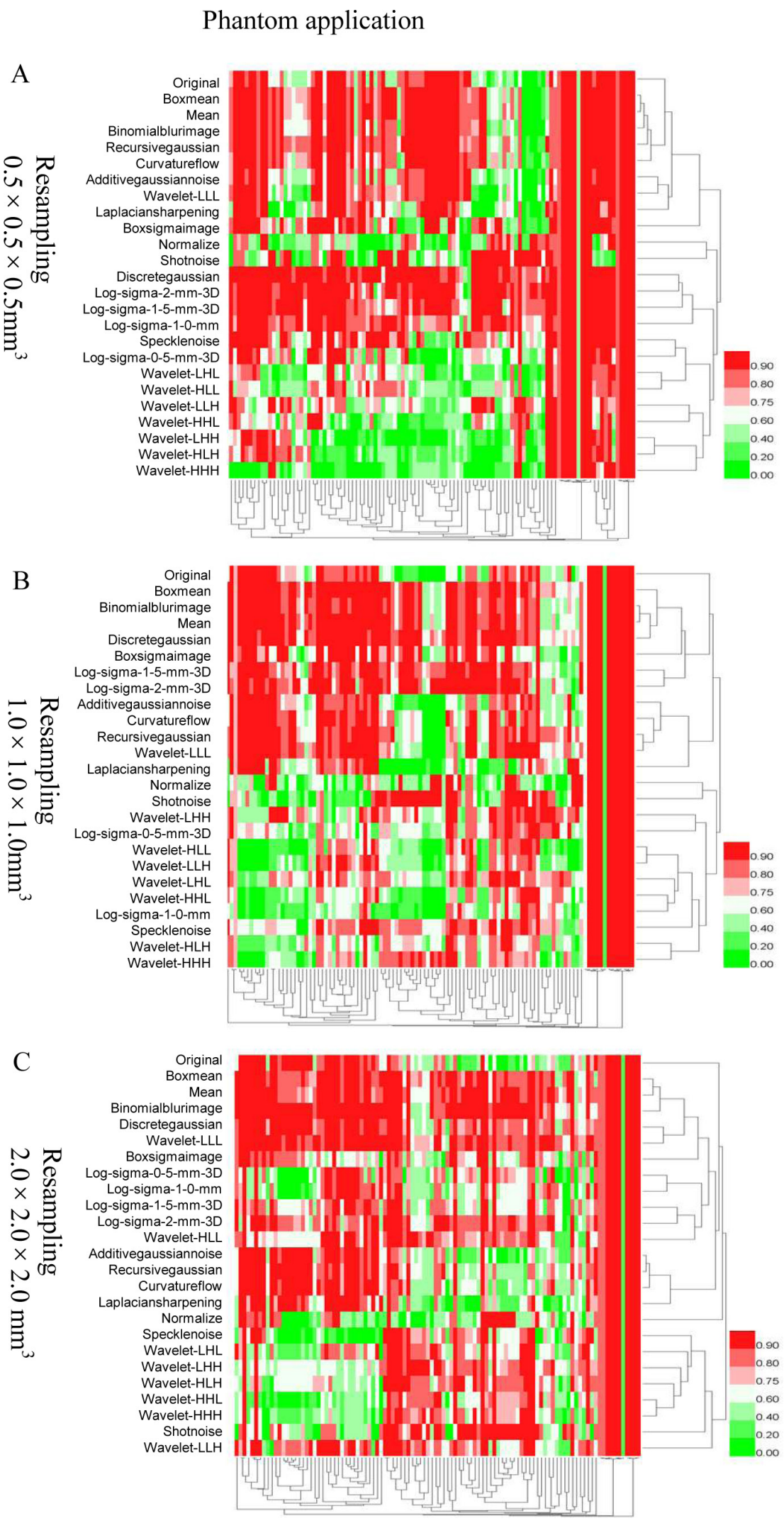

Figure S5 Heatmap of hierarchical clustering analysis of phantom data with image resampling. The excellent features were increased after image resampling and the poor features were reduced. (A) There were 1,655 (63.65%) excellent features, 502 (19.31%) moderate features, and 443 (17.04%) poor features after image resampling 0.5×0.5×0.5 mm^3; (B) there were 1,558 (59.92%) excellent features, 661 (25.43%) moderate features, and 381 (14.65%) poor features after image resampling 1.0×1.0×1.0 mm^3; (C) there were 1,655 (63.65%) excellent features, 654 (25.16%) moderate features, and 291 (11.19%) poor features after image resampling 2.0×2.0×2.0 mm^3. The gradual color changes from green to red in heatmap represent the steady increase in feature values from 0 to 1. H, high; L, low; 3D, three-dimensional.

https://dx.doi.org/10.21037/qims-22-992

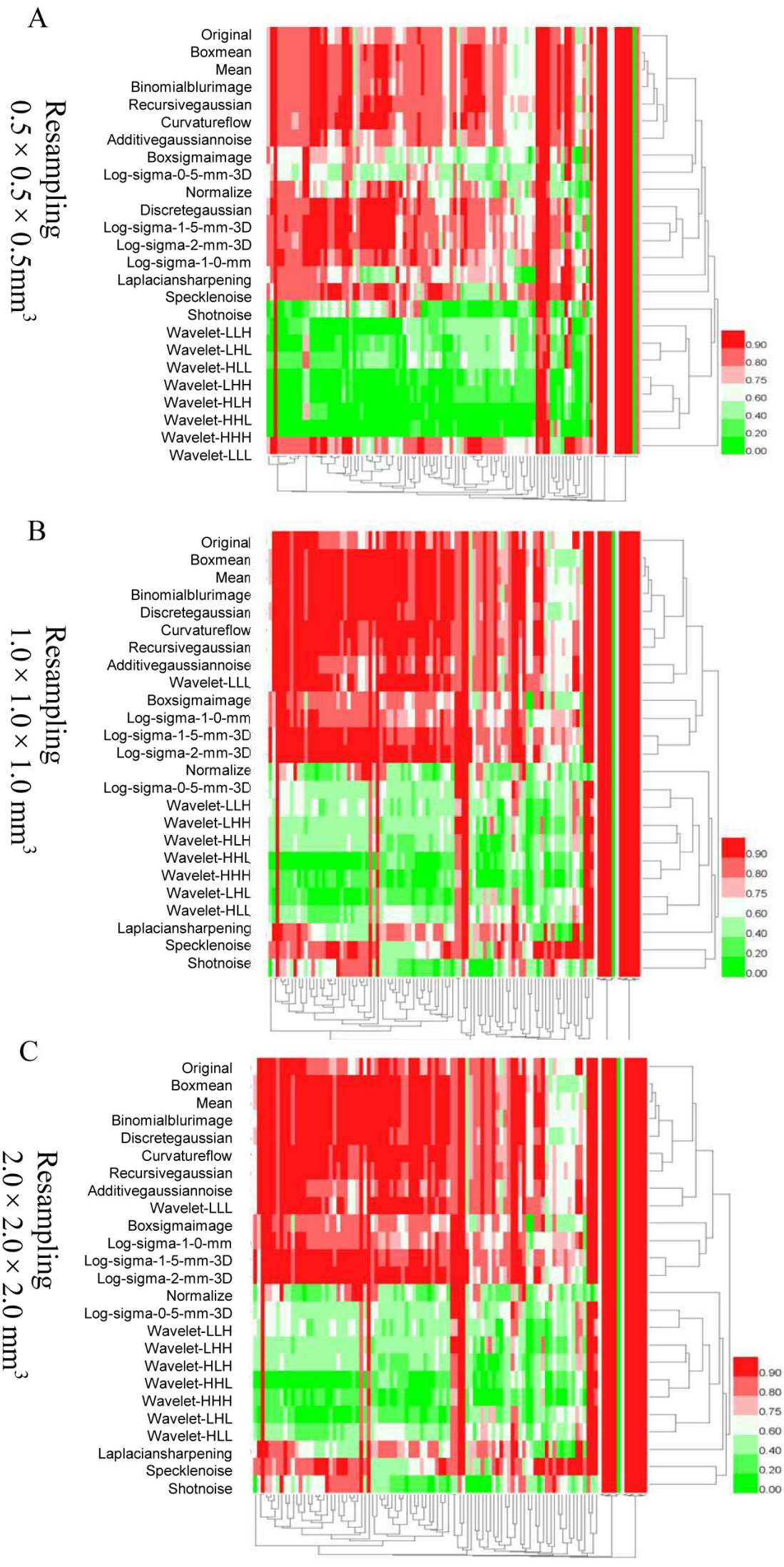

Figure S6 Heatmap of hierarchical clustering analysis of clinical data with image resampling. The excellent features were increased after image resampling, whereas poor features were reduced. (A) There were 1,394 (53.62%) excellent features, 637 (24.50%) moderate features, and 569 (21.88%) poor features after image resampling 0.5×0.5×0.5 mm^3; (B) there were 1,519 (58.42%) excellent features, 745 (28.66%) moderate features, and 336 (12.92%) poor features after image resampling 1.0×1.0×1.0 mm^3; (C) there were 1,251 (48.12%) excellent features, 1,116 (42.92%) moderate features, and 233 (8.96%) poor features after image resampling 2.0×2.0×2.0 mm^3. The gradual color changes from green to red in heatmap represent the steady increase in feature values from 0 to 1. 3D, three-dimensional; H, high; L, low.

https://dx.doi.org/10.21037/qims-22-992

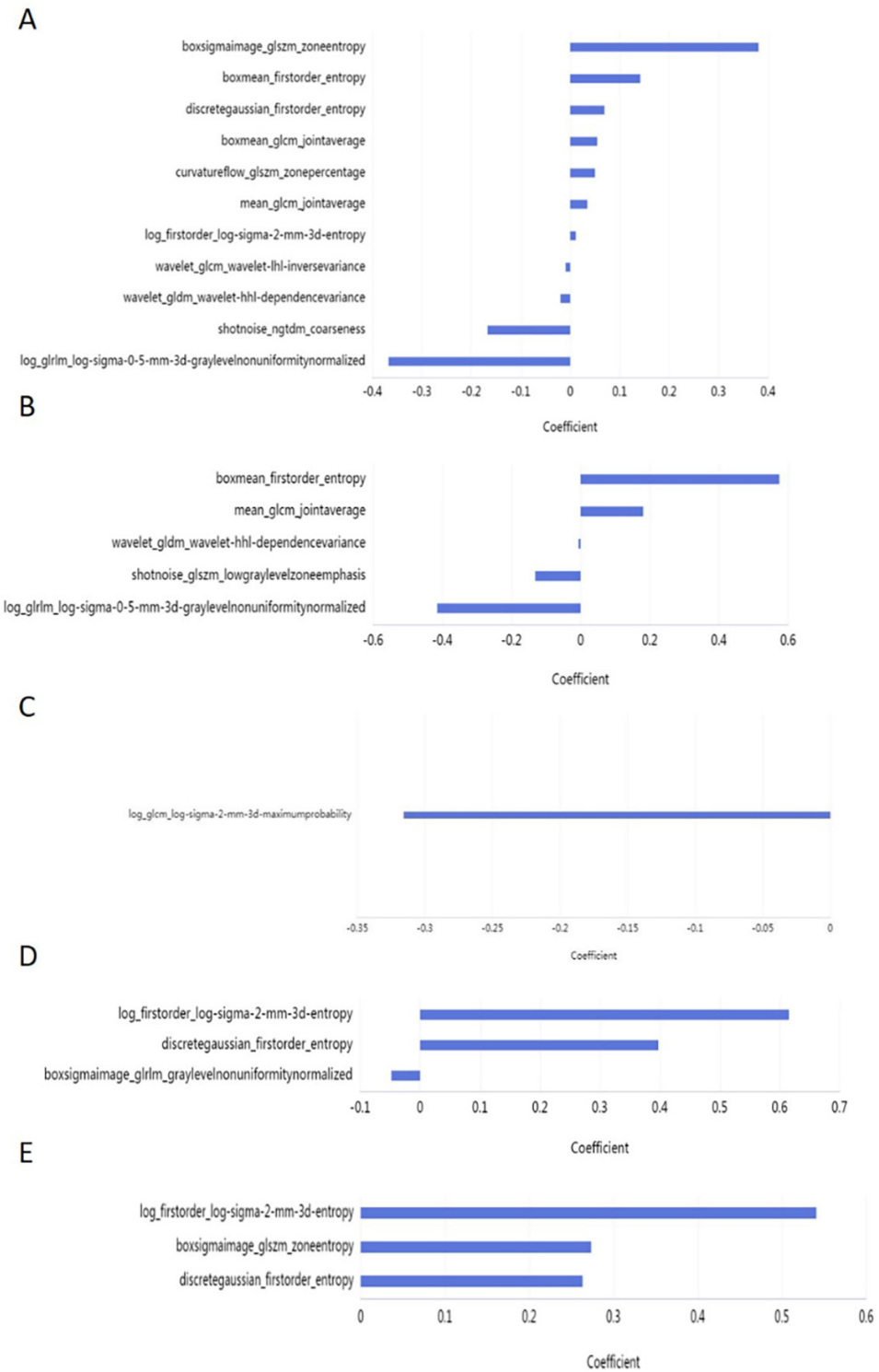

Figure S7 The selected radiomics features with non-zero coefficients. (A) The 11 features of model before optimization; (B) the 5 features of model with ComBat correction; (C-E) the features of models with image resampling correction of 0.5×0.5×0.5, 1.0×1.0×1.0, and 2.0×2.0×2.0 mm^3, respectively. 3D, three-dimensional; ComBat, combatting batch effect.

https://dx.doi.org/10.21037/qims-22-992