Appendix 1 Threshold-based approach of automated lung involvement assessment using two thresholds

The main manuscript introduces an arbitrary threshold at –522 HU for best assessment of lung involvement in COVID pneumonia. This supplementary material explores a dual-threshold approach, using an upper and lower margin to determine the infiltrated lung tissue.

Analogously to the full manuscript, the optimal location of both thresholds was investigated by maximizing the fit of a linear regression model of deep-learning and threshold-based results in a five-fold cross validation strategy. The goodness of fit of the regression model was reported as r^2. During the training process, the same five training folds as in the main manuscript were used, which ensures comparability of the mono- and dual-threshold methods.

In contrast to the full manuscript, which uses an arbitrary threshold at 250 locations (–1,000 HU to 250 HU in steps of 5 HU), this supplementary material introduces a thresholding window with an upper and lower margin. Dual-threshold-assessed lung involvement was defined as the voxels inside this window divided by the total lung volume. The lower threshold margin was consecutively set at 400 locations (–1,500 HU to 500 HU in steps of 5 HU), while for each window position, 200 window widths were assessed (width of 5 HU to 1,000 HU in steps of 5 HU). This resulted in 80,000 consecutively fitted linear regression models for each training fold, covering –1,500 HU to 1500 HU. Since a positive slope of the regression line was assumed (increasing deep-learning-based lung involvement corresponds to increasing threshold-based lung involvement), only those regression models with a positive slope coefficient were adopted for further analysis. *Figure S1* illustrates the training process of the dual-threshold approach (analogously to *Figure 5* of the main manuscript).

During training, the best accuracy of fit of the dual-threshold model was almost identical to the training results of the mono-threshold model (best r^2 for the five training folds 0.84, 0.86, 0.83, 0.82, 0.84, and 0.84, 0.86, 0.83, 0.81, 0.84, for the dual- and mono-threshold models, respectively). Similar to the mono-threshold approach of the full manuscript, the best fitting dual-threshold windows were then tested on five non-overlapping test sets. The five models achieving the highest accuracy of fit and their consecutive testing is reported in *Table S1* (analogously to *Table 2* of the main manuscript).

After inclusion of a second threshold, we observed only minimal changes of the regression model, compared to the mono-threshold model. The threshold for identification of infiltrated lung was marginally higher (–504 *vs.* –522 HU in the dual- and mono-threshold models, respectively), and the slope of the regression line slightly steeper (1.03 *vs.* 0.96 in the dual- and mono-threshold models, respectively). Yet, most importantly, the performance of a threshold-based model to quantify COVID-19 lung involvement did not benefit from introduction of a second threshold, compared to the mono-threshold approach: The goodness of fit and the confidence of the regression model, as represented by the r^2 and the width of the prediction interval, did not improve after addition of a second threshold (mean r^2=0.84, mean width of 95% prediction interval =0.23).

https://dx.doi.org/10.21037/qims-22-175

Table S1 Five-fold cross validation of a linear regression model to predict deep-learning-based lung involvement by a dual-threshold-based approach

Fold	1	2	3	4	5	Mean
Best lower threshold identified in training split (best arbitrary threshold)	–510 HU (–525 HU)	–505 HU (–520 HU)	–510 HU (–525 HU)	–500 HU (–520 HU)	–495 HU (–520 HU)	–504 HU (–522 HU)
Best upper threshold identified in training split (not applicable for the mono-threshold approach)	5 HU	–5 HU	10 HU	5 HU	5 HU	4 HU
Intercept of regression line	–0.05 (–0.05)	–0.05 (–0.05)	–0.04 (–0.04)	–0.03 (–0.04)	–0.05 (–0.06)	–0.04 (–0.05)
Slope of regression line	0.99 (0.93)	1.09 (0.96)	0.94 (0.89)	1.02 (0.97)	1.12 (1.03)	1.03 (0.96)
Width of 95% pred. interval	0.20 (0.20)	0.30 (0.30)	0.20 (0.20)	0.18 (0.18)	0.26 (0.26)	0.23 (0.23)
r^2	0.82 (0.82)	0.74 (0.74)	0.88 (0.88)	0.91 (0.91)	0.83 (0.83)	0.84 (0.84)

The best upper- and lower threshold margins were identified in five training splits (200 CT scans each, second and third row) and consecutively tested on five non-overlapping test folds (50 CT scans each, bottom rows). Identical training and testing sets were used for evaluation of the mono- and dual-threshold models. Corresponding results of the mono-threshold approach are complemented in parentheses for better comparability.

https://dx.doi.org/10.21037/qims-22-175

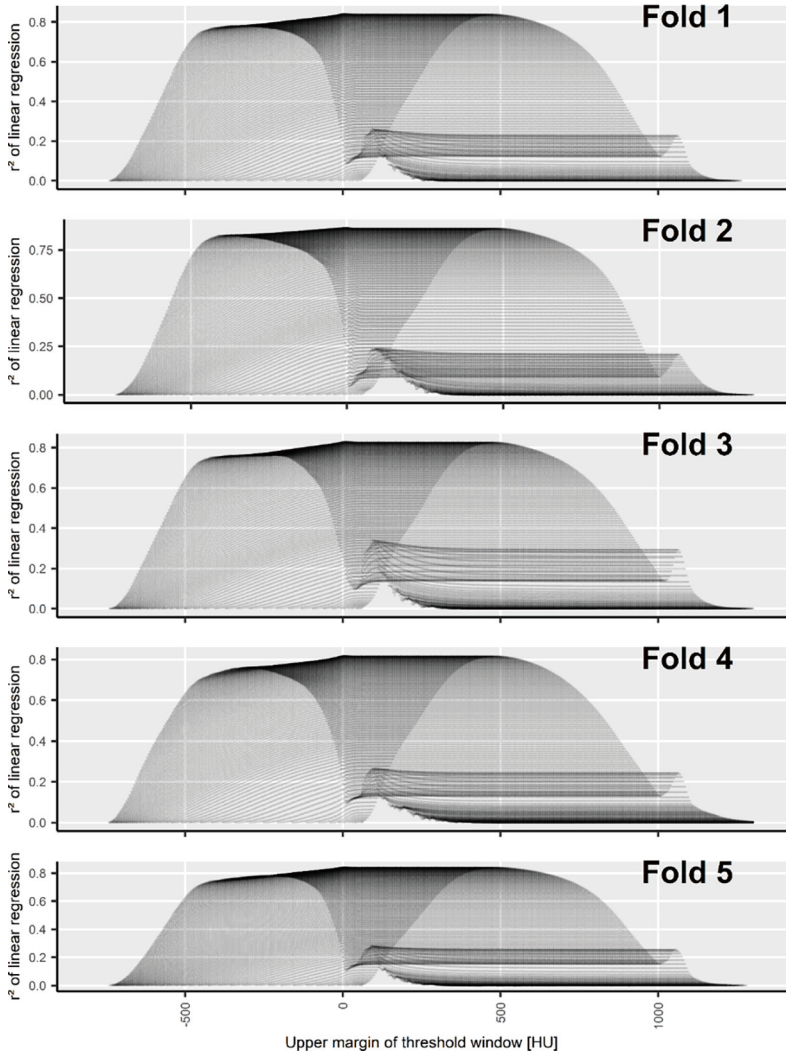

Figure S1 Five-fold training of a dual-threshold approach to assess COVID-19 lung involvement. A window of attenuation with a width between 5 to 1,000 HU was defined by two threshold margins. The portion of voxels inside this window was then divided by the total lung volume. This portion was calculated for 80,000 possible windows covering attenuation values between –1,500 to 1,500 HU. Consecutively, for each of five training folds (panel 1-5), 80,000 linear regression models were fit to the deep-learning assessed portion of lung involvement. All models with a positive slope coefficient are reported by their goodness of fit (r^2, y-axis), ordered by the upper margin of their corresponding threshold window (x-axis). The best fitting models are reported in *Table S1*.

https://dx.doi.org/10.21037/qims-22-175