**Figure legends**

**Figure 1. A recently developed semi-quantitative score system for acute lung injury is suited to generate ground truth labels for training deep neural networks. (A)** Schematics of the manual scoring pipeline. The lettering signifying the original (A), validation (B) and Matute-Bello (C) cohorts is used throughout the other figure panels **(B)** Heatmap of median total score per slide for the selected cohorts. Scores were normalized to lie in a range of 0-100 for this and the following figure panels. **(C)** Violin plot comparing the distributions of the normalized median total scores from the three cohorts. **(D)** Scatter plot comparing the normalized median total scores of the two new cohorts with the original Silva cohort. N refers to the number of scorers. Slope refers to the slope of the regression line. **(E)** Histogram demonstrating the distribution of the normalized median total scores of the selected cohorts.

**Figure 2. Low agreement between human scorers reinforces the need for automated histological scoring of acute lung injury by deep neural networks.**

**(A)** Heatmap of normalized total scores from the reviewers in different cohorts that performed scoring based on the Silva score system. The lettering on top signifies the original (A) and validation cohort (B). **(B)** Violin plot of the normalized total scores from the reviewers in different cohorts that performed scoring based on the Silva score system. **(C)** Correlation matrix indicating the correlation between the total scores from the reviewers in different cohorts that performed scoring based on the Silva score system. **(D)** Heatmap of normalized total scores from the reviewers that performed scoring based on the Matute-Bello score system. **(E)** Violin plot of the normalized total scores from the reviewers that performed scoring based on the Matute-Bello score system. **(F)** Correlation matrix indicating the correlation between the total scores from the reviewers that performed scoring based on the Matute-Bello score system. **(G)** Range of the total normalized scores from the Silva score system given by scorers of the original cohort with the same level of expertise. For this and panels H and I, the range was calculated for each slide by subtracting the lowest score given by a reviewer in the respective expertise group from the highest score given by a reviewer in the same expertise group. **(H)** Range of the total normalized scores from the Silva score system given by scorers of the validation cohort with the same level of expertise. **(I)** Range of the total normalized scores from the Matute-Bello score system given by scorers with the same level of expertise. **(J)** Correlation Matrix indicating the correlation between the total normalized median scores from each cohort as well as the group of two certified pathologists the . **(K)** Scatter plot comparing the self-agreement for the total normalized scores from selected reviewer with moderate expertise level (M3), **(L)** pre-clinical expert (pE3) and **(M)** second pre-clinical expert (pE4) for scoring performed on identical slides but over an at least one month time period. Slope refers to the slope of the regression lines.

**Figure 3. Development of deep neural networks for automated histology scoring. (A)**  Overview over the development of the deep learning scoring pipeline. CNNs were trained on 224 x 224 tiles (with or without augmentation) cut from up to 10 images from each slide with ground truth labels for low, medium and high damage derived from the median slide scores of five reviewers from the original cohort. After training, models were used to make predictions on the tile level which were then aggregated to slide level by “majority vote”. **(B)** Histogram displaying the distribution of tile scores in the dataset used to train the CNN models. Each tile received its score, based on which the ground-truth class label (“low”, “medium” or “high” damage) was later assigned, by calculating the median of the scores given by the five reviewers to the corresponding slide. **(C)** Distribution of scores across classes and treatment groups (MV stands for mechanical ventilation). Median tile scores were converted to 3-class labels, corresponding to low (≤10), medium (10-25) and high damage (>25). **(D)** Slide scores assigned by the reviewers from the original cohort in comparison with predictions by CNNs. Individual reviewer scores, their average and median (used as ground truth label for training the CNNs) and predicted scores by the 2 individual CNN models trained on raw dataset (EN\_r and V\_r),  3 best individual CNN models trained with augmentation (EN\_m\_a,V\_r\_a, and EN\_r\_a) , two ensemble predictions obtained by combining predictions from the three models on tile level (T\_3\_t) and slide level (T\_3\_s), and the binary model in tile level expanded to 3-class model in slide level (V\_bin\_3). Slides are identified by treatment group and slide number.

**Figure 4. Evaluation of the trained CNNs. (A)** Summary of CNN evaluations.Results obtained from combining the evaluation for all three folds are shown for Efficient (indicated with EN\_ prefixes), and VGG16 ("V\_") CNNs trained on the raw dataset (EN\_r and V\_r), augmented dataset (two bests are shown, EN\_m\_a and V\_r\_a) and two ensemble models. For the ensemble models a majority vote on tile level (T\_3\_t) or slide level (T\_3\_s) was used to aggregate the slide label prediction of the three best models (En\_m\_a, En\_r\_a, and V\_r\_a).The V\_bin\_3 model was trained utilizing extreme tiles (either low or high), and was subsequently refined to classify three levels of damage at the slide level through the assignment of thresholds. The confusion matrices show true tile labels vs predicted tile labels for all damage classes. The total number of tiles calculated by adding the corresponding matrices from the three folds is shown in the top of each cell and the cell shaded correspondingly as indicated by the color bar. The number of tiles normalized as percentage of the true labels is shown in the bottom of each cell. Values were calculated with the “confusion\_matrix” function from scikit-learn (v. 1.0.1). The tables underneath the matrices show the averages of the F1 score for each class, of the micro averaged F1 score and of the weighted average F1 score. Note that the values in the tables are not calculated form the confusion matrices but instead by averaging the respective slide-level metrics from the three folds. **(B)** Tile scores predicted by the best individual model (EN\_m\_a) for each slide. Each bar represents one slide with multiple images from which the tiles were cut. The height of each bar represents the number of tiles that received the respective score. Bar heights vary because the number of images and their dimensions differed between slides resulting in a different overall number of tiles.

**Figure 5. Assessment of generalizability of CNNs by conducting inference on unseen dataset. A)** The overall schematic of collecting publicly available dataset from different resources included pig tissue, COVID-19 patients and healthy samples from cancer patients. **B)** The slide damage of two groups of COVID-19 and healthy samples predicted by En\_m\_a and T\_3\_t models were depicted. **C)** Few examples from the public datasets were displayed here, depicting grayscale damage predictions made by the models.