medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Ensemble SHAP Aggregation and Attribution Variability in Clinical Machine Learning: A COVID-19 Mortality Study

Na Dai [*1], Ruben K. Briceno [*2,3,4,5], Alex N. Castaneda[3,5], Miguel A. Tresierra[3,5], Maribel L. Esteban[6], Moises E. Rosas[5], and Rene C. Hinojosa[2]

[1]Independent Researcher, MI, USA

[2]Global Health Institute, Michigan State University, East Lansing, MI, USA

[3]Hospital Victor Lazarte Echegaray, Trujillo, Peru

[4]College of Osteopathic Medicine, Michigan State University, East Lansing, MI, USA

[5]Universidad Cesar Vallejo, Trujillo, Peru

[6]Hospital de Alta Complejidad Virgen de la Puerta

Abstract

Objective: To combine performance-weighted ensemble SHAP aggregation with resampling-based assessment of feature-importance variability and patient-level attribution alignment for interpreting COVID-19 mortality predictions.

Methods: We analyzed a prospective cohort of 1,857 patients hospitalized with COVID-19 at two hospitals in Peru. Ten predictive algorithms were evaluated using five-fold cross-validation, and their mean AUROC values determined their attribution-aggregation weights. Within each model and training trial, signed SHAP values were normalized by the mean total absolute attribution across evaluation patients before performance-weighted aggregation. Feature importance was summarized across 30 resampled training trials using means and normal-approximation 95% confidence intervals. Patient-level attribution alignment was assessed using cosine similarity between signed feature-attribution vectors for corresponding patients, with whole-trial patient-correspondence randomization and Benjamini–Hochberg correction. Subgroup observations were reweighted toward the evaluated population's feature distributions for secondary comparisons of existing absolute SHAP values.

Results: Among the 1,857 included patients, 982 (52.9%) died during hospitalization. Random forest achieved the highest mean AUROC (0.925 $\pm$ 0.010), followed by AdaBoost (0.919 $\pm$ 0.017) and logistic regression (0.918 $\pm$ 0.007), with variability reported as the SEM. At-

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Co-corresponding authors. Emails: na.dai.research@gmail.com (N. Dai), briceode@msu.edu (R. K. Briceno).

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

tribution profiles were more similar across repeated training trials of the same algorithm (mean Pearson $r = 0.783$; SEM, 0.004) than across algorithms within the same trial (mean Pearson $r = 0.369$; SEM, 0.006). The largest normalized attribution magnitudes were observed for dexamethasone use at home without oxygen support (6.186%; 95% CI, 5.795–6.577), PaO_2/FiO_2 ratio (3.014%; 95% CI, 2.787–3.242), shortness of breath (2.774%; 95% CI, 2.592–2.956), and FiO_2 (2.560%; 95% CI, 2.284–2.836). Features differed in patient-level attribution alignment, indicating that importance magnitude and robustness provided complementary information. Among 1,763 eligible subgroup–feature comparisons, 64 had nominal one-sided $p < 0.05$, although none remained below 0.05 after Benjamini–Hochberg adjustment.

Conclusions: Algorithms with similar predictive performance produced substantially different feature-attribution profiles. Normalized performance-weighted SHAP aggregation provided a representative relative-importance summary across algorithms, while resampling-based alignment assessment qualified the consistency of individual feature contributions. The resulting attribution patterns describe fitted-model behavior and should not be interpreted as causal clinical effects.

Keywords: Clinical machine learning, Explainable artificial intelligence (XAI), SHapley Additive exPlanations (SHAP), Ensemble learning, Model interpretability, Subgroup analysis, Attribution variability

1 Introduction

Clinical prediction models can estimate patient risk, but interpreting that risk requires understanding which clinical features contribute to it. SHapley Additive exPlanations (SHAP) support this interpretation through feature attribution [14]. However, well-performing models may assign different importance to the same variables [4], potentially leading to different clinical interpretations. This motivates assessing the representativeness and robustness of feature-importance profiles across modeling algorithms and training samples.

To reduce dependence on individual models, we draw on the rationale of ensemble learning: combining information from multiple fitted models to construct a representative summary. Previous work has demonstrated that aggregating feature-based explanations can reduce their sensitivity [1]. Building on this rationale, we use ensemble SHAP aggregation to integrate attributions from diverse predictive algorithms using performance-based weights, providing a

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

representative feature-importance profile for clinical interpretation.

However, an aggregated profile can conceal differences in feature-level robustness. Features with similar average importance may contribute consistently across fits or receive large contributions in only a subset of models and trials. We therefore complement aggregation with resampling-based robustness analysis to characterize the variability underlying each importance estimate. Considering magnitude and robustness together distinguishes prominent but variable contributions from those supported more consistently.

This assessment can help place model findings within a coherent clinical account. Alignment with established knowledge supports clinical plausibility, while less familiar patterns may suggest hypotheses about mortality-associated physiology. Robustness helps qualify the support for these interpretations, which describe model behavior and do not independently establish biological causality [11].

We propose an integrated approach that combines performance-weighted ensemble SHAP aggregation with resampling-based robustness assessment to support clinical interpretation of feature importance. We investigate this approach in a prospective cohort of 1,857 patients hospitalized with COVID-19 at two hospitals in Peru. Building on previous applications of machine learning to mortality prediction and clinical feature identification [23], we aggregate SHAP attributions from ten predictive algorithms, assess variability across repeated training trials, and examine subgroup patterns through distribution-adjusted reweighting.

This study makes three contributions:

- An integrated workflow that pairs a representative ensemble feature-importance profile with assessment of the robustness of its underlying contributions.
- An empirical evaluation of how feature importance varies across predictive algorithms and repeated training trials.
- A clinical analysis of COVID-19 mortality that interprets feature importance alongside robustness, including distribution-adjusted subgroup comparisons.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

2 Related Work

2.1 Attribution Variability Across Models and Data

Variation across models. Different predictive models can assign different importance to the same clinical features, even when evaluated on the same patients. Fisher et al. introduced model class reliance to characterize the range of variable importance across well-performing models [4]. Extending this perspective to Shapley-based importance, Ning et al. developed ShapleyVIC to pool importance estimates across models and quantify their uncertainty, including in a clinical mortality application [16]. Laberge et al. identified local and global importance relationships supported across a Rashomon set of similarly performing models, allowing features to remain unordered when models disagree [12]. Collectively, these approaches characterize importance ranges, pooled estimates, and attribution consensus across plausible models. Our study builds on this foundation by comparing SHAP attributions across heterogeneous clinical prediction algorithms and integrating them into a performance-weighted profile.

Variation among training sets. Feature importance can also change when the same modeling approach is trained on different samples, even when explanations are evaluated on the same patients. Yasodhara et al. demonstrated sensitivity of global SHAP importance to input noise, random seeds, and hyperparameter changes [24]. Addressing sampling variability more directly, Xiang et al. [22] used bootstrap attribution distributions to construct RoSHAP, a ranking criterion incorporating attribution frequency, strength, and stability. These studies characterize perturbation sensitivity and distributional variation, but global importance and ranking summaries can obscure changes in signed contributions across individual observations. We examine these changes by comparing feature-specific attribution patterns across the same patients after repeated resampling and retraining, complementing assessments based on average importance and ranking consistency.

2.2 Explanation Aggregation and Uncertainty Assessment

Aggregation offers a means of summarizing explanations beyond individual fitted models. Bhatt et al. developed explanation-aggregation procedures to reduce explanation sensitivity and improve other explanation properties [1]. More directly related to model combination, Haghish's WMSHAP software provides weighted SHAP summaries and confidence intervals across fitted

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

models [7]. These approaches establish a basis for combining attributions, but an aggregated estimate alone does not reveal whether its underlying patient-level contributions remain consistent across training samples. Our study therefore pairs performance-weighted SHAP aggregation with resampling-based assessment of feature-level robustness.

Previous work has also proposed Shapley-based methods to assess distinct sources of uncertainty. Li et al. developed DistDeepSHAP to quantify reference-dependent attribution variation by sampling background observations for a fixed model and input [13]. Johnsen et al. introduced Sub-SAGE to measure global feature importance through reductions in expected prediction loss, using paired bootstrap resampling of independent test observations to estimate sampling uncertainty conditional on a fitted model [9]. In these uncertainty assessments, the predictive model is held fixed: the former examines variation across reference observations, whereas the latter estimates sampling uncertainty from evaluation data. Our study complements these assessments by examining SHAP attribution variability across predictive algorithms and resampled training datasets.

2.3 Clinical Applications and Subgroup Interpretation

Feature-importance analysis has been used to connect COVID-19 mortality predictions with clinically interpretable patient characteristics. Among the early studies in this area, Yadaw et al. [23] identified age, minimum oxygen saturation, and encounter type as informative predictors of mortality. Moving from predictor identification to attribution analysis, Booth et al. [2] used SHAP to examine how laboratory measurements contributed to mortality predictions from a support-vector-machine model. Smith and Alvarez [18] further examined global and patient-level Shapley explanations across machine-learning models, relating predicted mortality to individual patient characteristics. These studies demonstrate how feature interpretation can characterize mortality-associated patterns, motivating further assessment of how those interpretations depend on model choice and training data.

Beyond COVID-19, feature interpretation has also been applied in other clinical settings. Lundberg et al. [15] developed a system that paired intraoperative hypoxaemia predictions with explanations of contributing factors, examined whether those explanations agreed with clinical knowledge, and evaluated anaesthesiologists' predictions with system assistance. Incorporating variability assessment into clinical modeling, Huang and Huang [8] examined distributions

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

of predictive performance and feature-gain statistics across repeated training and test splits, alongside SHAP-based interpretation of heart-disease predictors. Their resampling summaries concerned performance and gain statistics rather than patient-level SHAP consistency. Together, these studies demonstrate the clinical value of interpreting feature contributions while considering the variability underlying model-derived findings.

Clinical research has also begun to examine whether model explanations remain stable across modeling conditions. Riley and Collins [17] used bootstrap model redevelopment to evaluate instability in clinical predictions and cautioned that post hoc explanations may be misleading when the underlying models are unstable. Connecting resampling with subgroup assessment, Ellis and Polberg-Riener [3] evaluated how female-targeted resampling affected predictive utility, calibration, and explanation stability across clinical datasets and model families, measuring stability through feature-importance drift and partial-dependence divergence. Focusing directly on SHAP explanations, Guillén and Frias-Martinez [5] assessed their coherence and stability across model architectures, disease-stage boundaries, and diagnostic and prognostic tasks in Alzheimer's disease.

A related line of research has compared attribution profiles across patient groups. Wang et al. [21] examined SHAP importance across sex, race and ethnicity, and insurance subgroups in an emergency-department prediction model, identifying shared and subgroup-specific predictors. Such comparisons use the observed feature distributions within each subgroup, so apparent differences in importance may partly reflect differences in subgroup composition.

Taken together, existing studies have examined prediction stability, explanation consistency across modeling settings, and SHAP differences across observed patient groups. However, they do not jointly connect an aggregated attribution profile with the consistency of each feature's signed contributions for the same patients across resampled training sets or assess subgroup differences after accounting for differences in feature distributions. Our study brings these elements together through ensemble SHAP aggregation, resampling-based robustness assessment, and distribution-adjusted subgroup comparisons.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

3 Methods

3.1 Study Design and Population

We conducted a prospective cohort study of patients hospitalized with coronavirus disease 2019 (COVID-19) in the inpatient wards of Hospital de Alta Complejidad Virgen de la Puerta and Hospital Víctor Lazarte Echegaray (EsSalud) in Peru between March 1 and December 31, 2020. Patients were eligible if they were admitted for inpatient hospitalization with COVID-19 at either participating hospital during the study period. Of 2,000 admitted patients, 143 with documented psychiatric conditions were excluded, leaving 1,857 patients in the analysis.

Clinical data were collected prospectively from enrollment and throughout hospitalization until recovery and discharge or in-hospital death. COVID-19 was diagnosed using molecular testing, antibody or antigen assays, or the treating clinician's assessment based on the clinical presentation. The primary outcome was in-hospital mortality.

3.2 Feature Construction and Preprocessing

A total of 279 candidate clinical variables were extracted. These covered demographics, history of presenting illness, ABO blood type, past medical history, evidence of bacterial infection, symptoms, laboratory and physiologic measurements, imaging findings, treatments, and medications. Continuous measurements were retained numerically and, where clinically relevant, were also represented using predefined categories, such as $PaO_2/FiO_2 > 300$ and PaO_2/FiO_2 between 200 and 300. Feature encoding produced 283 model inputs.

For the reported analyses, missing values were imputed by k-nearest-neighbor imputation using two neighbors with uniform weights. Predictor values were subsequently transformed to the range $[0, 1]$. Imputation and scaling were fitted on each trial-specific training subset and then applied without refitting to its evaluation observations. Vaccination indicators were retained during model fitting but omitted from the reported feature-importance results because only one patient had a recorded COVID-19 vaccination, providing insufficient exposure variation for meaningful interpretation.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

3.3 Predictive Models and Cross-Validation

Ten predictive algorithms were evaluated: logistic regression (LR), linear support vector machine (Linear SVM), support vector machine with a radial basis function kernel (RBF SVM), random forest, decision tree, multilayer perceptron (MLP), AdaBoost, Gaussian naïve Bayes (NB), quadratic discriminant analysis (QDA), and a convolutional neural network (CNN). Logistic regression used the SAGA solver with an elastic-net penalty, an ℓ_1 ratio of 0.5, and a tolerance of 0.02. Both SVM models used $C = 0.5$. The random forest contained 300 trees, and the MLP used $\alpha = 3$ with a maximum of 100 iterations. Other scikit-learn models used default hyperparameters unless otherwise specified. The CNN contained two one-dimensional convolutional layers with 256 and 64 filters, respectively, each followed by average pooling; the resulting representation was flattened and passed through a 32-unit rectified-linear dense layer and a one-unit sigmoid output layer.

Predictive performance was evaluated by five-fold cross-validation on the full cohort. The area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were summarized as the mean and standard error of the mean (SEM) across folds. Let A_k denote the mean cross-validated AUROC for algorithm k. Its attribution weight was

$$w_k = \frac{A_k}{\sum_{\ell=1}^{K} A_\ell}, \tag{1}$$

where $K = 10$ and $\sum_{k=1}^{K} w_k = 1$. These weights were fixed for all subsequent attribution trials. Because the cross-validation used the full cohort, observations later included in the attribution evaluation subsets also contributed to weight estimation.

3.4 Resampled Training Trials

The attribution analysis comprised 30 training trials per algorithm, organized into three evaluation batches of ten trials. Each batch used a fixed evaluation subset containing 10% of the cohort (186 patients). For each trial, 30% of the available observations were randomly selected as the training subset, without replacement within that trial and using a trial-specific random seed. A patient could therefore appear in multiple trials but no more than once in a given training subset. The training and evaluation subsets were selected independently; consequently,

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

evaluation observations were not guaranteed to be excluded from a trial's training subset.

Within an evaluation batch, all ten algorithms and all ten training trials were explained on the same patients in the same order. This correspondence enabled patient-level comparison of attributions across fitted models. The three batches contained different evaluation subsets, so patient-level comparisons were restricted to trials within the same batch.

3.5 SHAP Attribution

SHapley Additive exPlanations (SHAP) were used to describe the contribution of each feature to an individual model output. For a fitted model f and patient x, the SHAP representation is

$$f(x) = \phi_0 + \sum_{i=1}^{P} \phi_i(x), \tag{2}$$

where ϕ_0 is the explainer baseline, $\phi_i(x)$ is the signed attribution for feature i, and $P = 283$ is the number of encoded inputs. Positive and negative values indicate contributions that increase or decrease the explained output relative to its baseline.

TreeExplainer was used for the decision-tree and random-forest models and was initialized without an explicit background dataset. KernelExplainer was used for the other algorithms. For KernelExplainer, the corresponding evaluation matrix served both as the background reference and as the observations to be explained. The non-tree scikit-learn models were explained using their `predict_proba` output, while the CNN was explained using its `predict` output. The retained class index was 1 for the non-CNN models and 0 for the single-output CNN; both represented the mortality output used in the analysis. KernelExplainer used the identity link. No additional transformation was applied to the raw SHAP values before the normalization described below.

3.6 Normalized Ensemble SHAP Aggregation

Raw SHAP magnitudes can differ across algorithms because their fitted functions, output representations, and explainer baselines differ. We therefore standardized the overall attribution magnitude of each model–trial before combining models. For algorithm k, batch b, and trial t,

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the attribution budget was defined as

$$B_{kbt} = \frac{1}{N_b} \sum_{j=1}^{N_b} \sum_{i=1}^{P} \left| \phi_{ij}^{(kbt)} \right|, \tag{3}$$

where $N_b = 186$ is the number of evaluation patients in batch b. The normalized signed attribution was

$$\psi_{ij}^{(kbt)} = \frac{\phi_{ij}^{(kbt)}}{B_{kbt}}. \tag{4}$$

This operation set the mean patient-level total absolute attribution to one within each model–trial while retaining the relative magnitudes and signs of its patient-level contributions.

The normalized attributions were then aggregated across algorithms using the fixed AUROC weights:

$$\bar{\psi}_{ij}^{(bt)} = \sum_{k=1}^{K} w_k \psi_{ij}^{(kbt)}. \tag{5}$$

This quantity is an empirical performance-weighted summary of attributions from heterogeneous fitted models. It is not the exact SHAP decomposition of a single ensemble prediction function.

For feature i in trial t of batch b, ensemble importance was defined as

$$I_i^{(bt)} = \frac{100}{N_b} \sum_{j=1}^{N_b} \left| \bar{\psi}_{ij}^{(bt)} \right|. \tag{6}$$

Multiplication by 100 expresses the result relative to the standardized attribution budget. Because signed contributions were combined across models before their absolute values were taken, feature importances need not sum to 100. For each feature, the mean, standard deviation, and SEM were calculated from the 30 trial-specific estimates. Normal-approximation 95% confidence intervals were calculated as the mean $\pm 1.96\times$ SEM, with negative lower limits truncated at zero because importance is nonnegative.

3.7 Attribution Variation Across Algorithms and Training Trials

To compare global attribution profiles, model-specific normalized importance was calculated as

$$G_i^{(kbt)} = \frac{1}{N_b} \sum_{j=1}^{N_b} \left| \psi_{ij}^{(kbt)} \right|. \tag{7}$$

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Pearson correlations across all encoded features were then calculated for two types of comparison: the same algorithm across unique pairs of training trials within an evaluation batch, and different algorithms within the same trial and batch. Each comparison comprised 1,350 profile pairs: $10 \times 3 \times \binom{10}{2}$ for the former and $3 \times 10 \times \binom{10}{2}$ for the latter. Correlations were summarized by their mean and SEM.

3.8 Patient-Level Attribution Alignment

Feature importance describes average attribution magnitude but does not indicate whether a feature contributes similarly for the same patients across training trials. For each feature i, batch b, and trial t, we therefore formed the signed ensemble-attribution vector

$$\mathbf{s}_{ib}^{(t)} = \left(\bar{\psi}_{i1}^{(bt)}, \ldots, \bar{\psi}_{iN_b}^{(bt)}\right). \tag{8}$$

For each unique pair of trials $a < b$ within the same batch, alignment was quantified by cosine similarity:

$$C_{ib}^{(a,b)} = \frac{\mathbf{s}_{ib}^{(a)} \cdot \mathbf{s}_{ib}^{(b)}}{\left\|\mathbf{s}_{ib}^{(a)}\right\|_2 \left\|\mathbf{s}_{ib}^{(b)}\right\|_2}. \tag{9}$$

Cosine similarity ranges from -1 to 1. Positive values indicate that the signed attribution patterns tend to align across corresponding patients, values near zero indicate little patient-specific alignment, and negative values indicate opposing patterns. The reported alignment for each feature was the mean across the 45 unique trial pairs in each batch and then across the three batches, giving 135 pairwise comparisons per feature.

A one-sided patient-correspondence randomization test evaluated whether the observed mean cosine alignment exceeded that expected without consistent patient matching across trials. For each of 2,000 randomizations, the patient labels of each whole trial vector were independently permuted within its evaluation batch. This preserved the attribution values and within-trial vector structure while breaking correspondence for the same patients across trials. The complete alignment statistic was recalculated after each randomization. If T_i^{obs} is the observed mean alignment and T_{ir}^{null} is its value in randomization r, the p-value was

$$p_i = \frac{1 + \sum_{r=1}^{R} \mathbf{1}\left(T_{ir}^{\text{null}} \geq T_i^{\text{obs}}\right)}{R+1}, \qquad R = 2000. \tag{10}$$

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The resulting p-values were adjusted across all 283 encoded features using the Benjamini–Hochberg false-discovery-rate procedure. Table 3 reports the adjusted q-values. These tests concern patient-specific alignment of signed model attributions; they do not test association with mortality, importance magnitude, or a clinical effect.

3.9 Visualization of Signed Ensemble Attributions

For the supplementary beeswarm plots, normalized ensemble attributions were averaged across the ten trials within each fixed evaluation batch and then pooled across the three batches, yielding 558 plotted evaluation observations per feature. The horizontal axis represents signed normalized ensemble attribution, expressed on the same percentage scale as Equation 6. Feature color represents the within-batch percentile rank of the observed feature value, allowing visual comparison despite preprocessing differences. Separate plots show up to the 20 most important features within each of 16 clinical categories; categories containing fewer than 20 reported features show all available features.

3.10 Distribution-Adjusted Subgroup Analysis

Distribution-adjusted subgroup analysis used the normalized ensemble attributions generated for Table 4. The archived attribution data comprised three evaluation batches of 186 observations each, yielding 558 pooled evaluation observations. Original cohort identifiers were not retained across batches; the pooled records are therefore described as evaluation observations rather than unique patients. For every subgroup G, the pooled evaluation reference R comprised all 558 available evaluation observations, including observations belonging to G.

Evaluation observations were grouped by age 50 years or older, sex, ABO blood group, asthma or chronic obstructive pulmonary disease, metabolic or cardiovascular disease, kidney disease, chronic viral or liver disease, and the presence of at least two documented comorbidities. Metabolic or cardiovascular disease comprised type 2 diabetes, hypertension, ischemic heart disease, or obesity. Kidney disease comprised acute kidney injury or chronic kidney disease stages 1–5. Chronic viral or liver disease comprised HIV, hepatitis C, compensated or decompensated cirrhosis, or hepatocellular carcinoma.

For feature i and evaluation observation j in batch b, patient-level attribution magnitude was calculated by averaging the absolute normalized ensemble attribution across the ten training

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

trials in that batch:

$$a_{ij}^{(b)} = \frac{100}{10} \sum_{t=1}^{10} \left| \bar{\psi} ij^{(bt)} \right|. \tag{11}$$

The batch superscript is omitted below when observations from the three batches are pooled. Pooled-reference importance was

$$Ii, R = \frac{1}{N_R} \sum_{j \in R} a_{ij}, \qquad N_R = 558. \tag{12}$$

$\sum_{j \in G} v_{ij} a_{ij} \overline{\sum_{j \in G} v_{ij},}$ where v_{ij} is the feature-specific weight. Absolute and relative differences were calculated as

$$\Delta I_{i,G} = I_{i,R} - I_{i,G}^{(\text{adj})}, \ \Delta \qquad = 100 \times \frac{Ii, R - I_{i,G}^{(\text{adj})}}{I_{i,R}}. \tag{13}$$

Positive differences therefore indicated greater attribution magnitude in the pooled evaluation reference, whereas negative differences indicated greater magnitude in the distribution-adjusted subgroup.

$\left(\sum_{j \in G} v_{ij}\right)^2 \overline{\sum_{j \in G} v_{ij}^2 (13)}$ was below 10. Features with zero patient-level attribution variance, for which the Student's t statistic was undefined, were also excluded from inference.

For eligible features, nominal p-values were calculated from the patient-level absolute SHAP magnitudes a_{ij} using a weighted, pooled-variance Student's t test. Let $s_{i,R}^2$ denote the variance of a_{ij} in the pooled evaluation reference and let $s_{i,G,w}^2$ denote the feature-weighted variance in subgroup G. The pooled variance and test statistic were

$$s_{i,p}^2 = \frac{(N_R - 1) s_{i,R}^2 + \left(n_{i,G}^{(\text{eff})} - 1\right) s_{i,G,w}^2}{N_R + n_{i,G}^{(\text{eff})} - 2}, \ t_{i,G} \qquad = \frac{I_{i,R} - I_{i,G}^{(\text{adj})}}{\sqrt{s_{i,p}^2 \left(N_R^{-1} + \left[n_{i,G}^{(\text{eff})}\right]^{-1}\right)}}. \tag{14}$$

The degrees of freedom were $N_R + n_{i,G}^{(\text{eff})} - 2$, and the reported one-sided tail probability in the direction of the observed difference was

$$p_{i,G} = 1 - F_{t,,df} \left(|t_{i,G}| \right). \tag{15}$$

The test was conducted on attribution magnitudes rather than signed SHAP values because Table 4 compares feature importance. Because subgroup observations also occurred in the pooled

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

evaluation reference and uncertainty in the estimated reweighting weights was not propagated, these p-values were treated as nominal and exploratory.

All 283 encoded attributes entered the subgroup calculations. Attributes defining a particular subgroup and attributes failing the balance or effective-sample-size criteria were excluded from inference only for the affected subgroup. Subgroups containing fewer than 20 evaluation observations were summarized descriptively without inferential values. Within each subgroup, Table 4 displays up to 20 attributes with the largest absolute differences; selection was based on $|\Delta I|$, rather than on the nominal p-value. Benjamini–Hochberg adjustments were calculated both within each subgroup and across all eligible subgroup–attribute comparisons and were retained in the complete supplementary data as sensitivity summaries. These comparisons describe differences in model attribution after marginal distribution adjustment and do not establish subgroup-specific biological or treatment effects.

4 Results

4.1 Performance on COVID-19 Mortality Prediction

Among the 1,857 included patients, 982 (52.9%) died during hospitalization and 875 (47.1%) survived to discharge. Table 1 summarizes predictive performance under fivefold cross-validation. Seven algorithms achieved mean AUROC values between 0.907 and 0.925. Random forest had the highest mean AUROC (0.925 $\pm$ 0.010 SEM), followed by AdaBoost (0.919 $\pm$ 0.017), logistic regression (0.918 $\pm$ 0.007), the multilayer perceptron (0.915 $\pm$ 0.012), linear SVM (0.913 $\pm$ 0.013), the CNN (0.911 $\pm$ 0.006), and RBF SVM (0.907 $\pm$ 0.008). Gaussian naïve Bayes, decision tree, and quadratic discriminant analysis achieved lower mean AUROCs of 0.854, 0.775, and 0.768, respectively.

Operating characteristics also differed across algorithms. Gaussian naïve Bayes had high specificity (0.931) but low sensitivity (0.378), whereas AdaBoost and random forest showed more balanced sensitivity and specificity. No algorithm had the highest value for every metric. The combination of broadly similar AUROC values among several algorithms and differences in their operating characteristics supported using the normalized mean-AUROC values as weights in the ensemble attribution analysis.

Table 1: COVID-19 mortality prediction performance (mean ± SEM[1]) across ten predictive algorithms under fivefold cross-validation.

Algorithm	AUROC[5]	Accuracy	Sensitivity	Specificity	PPV[6]	NPV[7]
LR[3]	0.918±0.007	0.841±0.009	0.853±0.025	0.827±0.032	0.848±0.020	0.835±0.020
Linear SVM ($C = 0.5$)	0.913±0.013	0.836±0.013	0.848±0.021	0.822±0.032	0.843±0.023	0.829±0.017
RBF SVM ($C = 0.5$)	0.907±0.008	0.815±0.008	0.804±0.023	0.829±0.027	0.841±0.019	0.790±0.016
Random forest (300 trees)[4]	0.925±0.010	0.844±0.010	0.828±0.021	0.863±0.023	0.872±0.017	0.818±0.016
Decision tree[4]	0.775±0.017	0.776±0.017	0.799±0.021	0.750±0.042	0.783±0.028	0.769±0.016
MLP ($\alpha = 3$, maximum 100 iterations)[4]	0.915±0.012	0.827±0.012	0.806±0.052	0.850±0.043	0.860±0.029	0.799±0.035
AdaBoost[4]	0.919±0.017	0.850±0.017	0.861±0.010	0.838±0.030	0.857±0.023	0.843±0.013
Gaussian NB[4]	0.854±0.064	0.639±0.064	0.378±0.139	0.931±0.023	0.858±0.015	0.577±0.053
QDA[4]	0.768±0.059	0.699±0.059	0.614±0.066	0.795±0.075	0.773±0.073	0.648±0.050
CNN[2]	0.911±0.006	0.827±0.006	0.840±0.044	0.811±0.051	0.836±0.033	0.822±0.033

[1] Values are mean ± SEM across five folds. Missing values were imputed using k-nearest-neighbor imputation, and predictors were scaled to $[0, 1]$.
[2] CNN, convolutional neural network. Two one-dimensional convolutional layers with 256 and 64 filters, each followed by average pooling, were connected to a 32-unit rectified-linear dense layer and a one-unit sigmoid output.
[3] LR, logistic regression; SAGA solver, elastic-net penalty, ℓ_1 ratio 0.5, and tolerance 0.02.
[4] Non-CNN algorithms were implemented using scikit-learn. Default hyperparameters were used unless otherwise shown.
[5] AUROC, area under the receiver operating characteristic curve.
[6] PPV, positive predictive value.
[7] NPV, negative predictive value.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

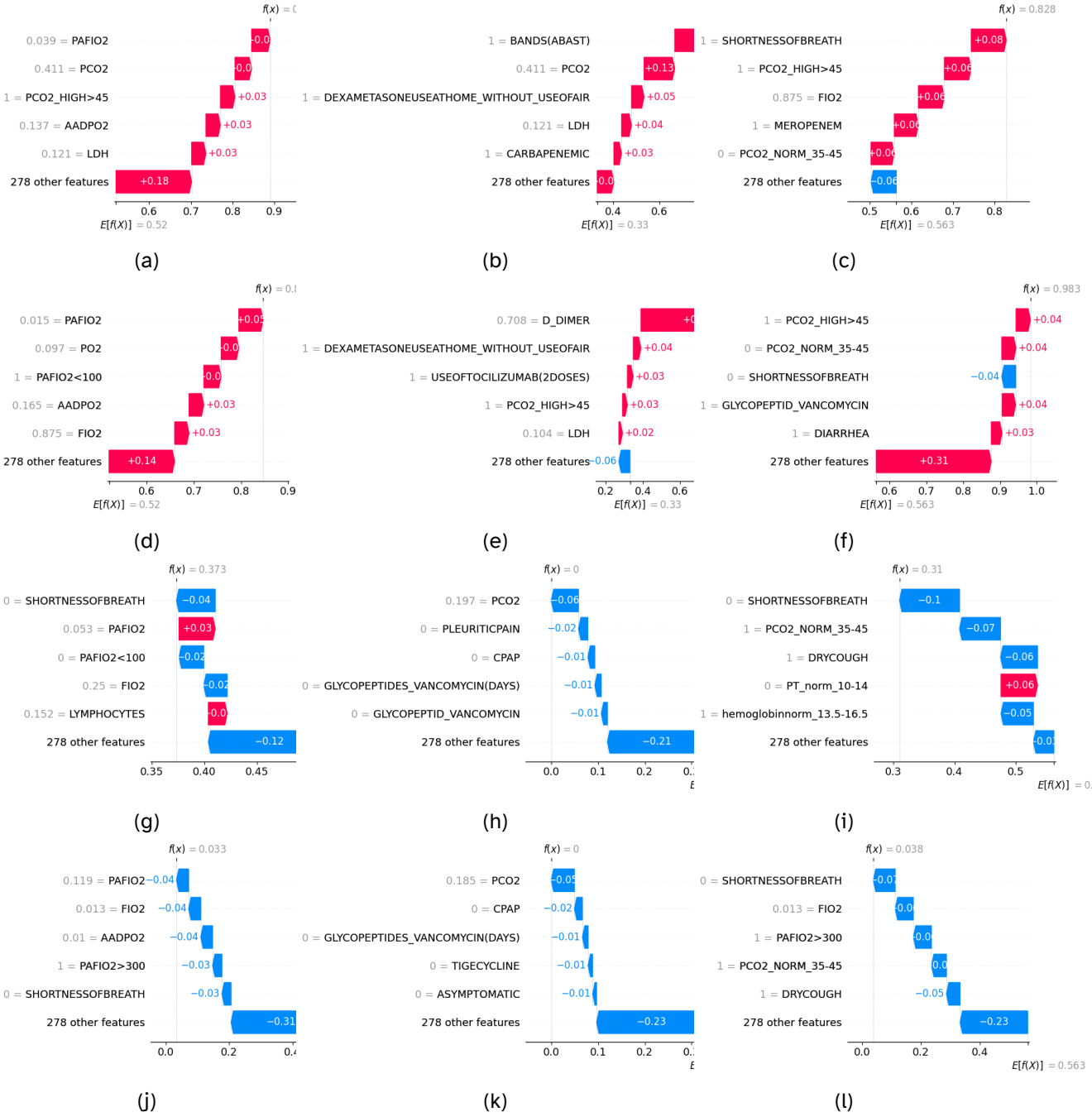


Figure 1: Patient-level SHAP explanations for four illustrative patients across three predictive algorithms. Each row corresponds to one patient: the first two died during hospitalization and the latter two survived to discharge. Columns show random forest, Gaussian naïve Bayes, and logistic regression, respectively. Positive SHAP values increase the explained model output relative to its baseline, whereas negative values decrease it.

4.2 Variation in Feature Attribution Across Trials and Algorithms

Figure 1 shows patient-level SHAP explanations for four illustrative patients across random forest, Gaussian naïve Bayes, and logistic regression. For the same patient, the algorithms differed in the features receiving the largest attributions and in their attribution magnitudes. Because

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

these patients and algorithms were selected for illustration, the figure does not provide a quantitative estimate of agreement.

Table 2: Similarity of normalized SHAP feature-importance profiles across training trials and predictive algorithms.

Comparison	Number of pairs	Pearson r, mean $\pm$ SEM
Same algorithm across training trials	1,350	0.783 ± 0.004
Different algorithms within a trial	1,350	0.369 ± 0.006

Quantitative comparison of the normalized global importance profiles supported the pattern illustrated in Figure 1. Across 1,350 unique comparisons, repeated trials of the same algorithm had a mean Pearson correlation of 0.783 ± 0.004 SEM. Across 1,350 comparisons of different algorithms fitted within the same trial, the mean correlation was 0.369 ± 0.006 SEM (Table 2). Thus, profiles were more similar across changes in training samples within an algorithm than across algorithm choices, although both sources contributed to attribution variation.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4.3 Normalized Ensemble Feature Importance and Attribution Alignment

Table 3: Normalized ensemble SHAP feature importance and patient-level attribution alignment across 30 resampled training trials, organized into 16 clinical categories.

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Group 1: Demographics			
Age 60+	1.148	[1.073, 1.223]	0.990 (< 0.001)
Age	1.018	[0.923, 1.113]	0.804 (< 0.001)
Age 18+	0.412	[0.324, 0.501]	0.958 (< 0.001)
Female	0.247	[0.192, 0.301]	0.634 (< 0.001)
Male	0.222	[0.172, 0.273]	0.251 (< 0.001)
Age 30+	0.245	[0.186, 0.305]	0.898 (< 0.001)
Age 40+	0.203	[0.163, 0.243]	0.489 (< 0.001)
Age 50+	0.132	[0.109, 0.155]	0.155 (< 0.001)
Age 0+	0.040	[0.014, 0.065]	0.364 (< 0.001)
Group 2: HPI			
Days (onset, ADM.)	0.102	[0.085, 0.118]	0.083 (< 0.001)
Group 3: Blood Group			
type_not_mentioned	0.399	[0.322, 0.476]	0.723 (< 0.001)
O_positive	0.181	[0.146, 0.217]	0.199 (< 0.001)
A_positive	0.078	[0.062, 0.095]	0.058 (< 0.001)
B_positive	0.064	[0.043, 0.085]	0.162 (< 0.001)
A_negative	0.061	[0.041, 0.081]	0.124 (< 0.001)
O_negative	0.029	[0.010, 0.047]	-0.011 (1.000)
AB_positive	1.25e-04	[7.32e-05, 1.77e-04]	0.004 (0.048)
Group 4: Past Medical History			
Hypertension	0.208	[0.159, 0.257]	0.192 (< 0.001)
No 2+ Comorbid	0.189	[0.142, 0.237]	-0.048 (1.000)
Diabetes type 2	0.157	[0.126, 0.187]	0.051 (< 0.001)
Other co-morbids	0.206	[0.164, 0.248]	0.495 (< 0.001)
Obesity	0.181	[0.133, 0.228]	0.272 (< 0.001)
No comorbidities	0.156	[0.129, 0.184]	0.454 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
CKD stage 5[4]	0.160	[0.102, 0.218]	0.631 (< 0.001)
Hyperglycemia	0.169	[0.146, 0.191]	0.838 (< 0.001)
Ischemic HD	0.129	[0.085, 0.173]	0.163 (< 0.001)
CKD Hemodialysis[5]	0.134	[0.087, 0.181]	0.649 (< 0.001)
CKD stage 3[4]	0.125	[0.083, 0.167]	0.742 (< 0.001)
CKD stage 2[4]	0.107	[0.077, 0.137]	0.205 (< 0.001)
AKI[6]	0.106	[0.085, 0.127]	0.853 (< 0.001)
Tuberculosis	0.093	[0.074, 0.112]	-0.057 (1.000)
Asthma	0.084	[0.068, 0.099]	0.625 (< 0.001)
COPD[7]	0.070	[0.059, 0.081]	0.362 (< 0.001)
No CLD/cirrhosis	0.072	[0.051, 0.094]	0.802 (< 0.001)
CKD stage 1[4]	0.048	[0.033, 0.062]	-0.026 (1.000)
HIV (undetected)	0.038	[0.027, 0.048]	0.146 (< 0.001)
HIV (PLWHV)	0.040	[0.029, 0.050]	0.381 (< 0.001)
CKD stage 4[4]	0.027	[0.012, 0.042]	0.055 (< 0.001)
HCV (Hepatitis)	0.023	[0.010, 0.036]	-0.019 (1.000)
Decomp. cirrhosis	0.023	[0.011, 0.036]	0.142 (< 0.001)
Pregnancy	0.020	[0.009, 0.031]	0.114 (< 0.001)
Pregnancy VAG Del.[8]	1.61e-04	[6.70e-05, 2.56e-04]	0.012 (< 0.001)
Pregnancy in 3rd Trim	1.53e-04	[6.05e-05, 2.45e-04]	0.068 (< 0.001)
Liver cancer (HCC)[9]	9.78e-05	[2.06e-05, 1.75e-04]	0.027 (< 0.001)
Group 5: Bacterial Infection			
CRAB[10]	0.119	[0.051, 0.187]	-9.03e-04 (0.613)
S. *aureus*	9.11e-04	[0.000, 0.002]	0.010 (< 0.001)
K. *pneumoniae*	8.16e-04	[0.000, 0.002]	0.069 (< 0.001)
Bacter. infection	1.36e-04	[7.42e-05, 1.99e-04]	0.079 (< 0.001)
Pseudomona	3.50e-05	[7.85e-06, 6.21e-05]	-0.006 (0.025)
Group 6: Clinical Symptoms			
Shortness of breath	2.774	[2.592, 2.956]	0.976 (< 0.001)
Productive cough	1.026	[0.914, 1.138]	0.980 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Dry cough	0.868	[0.777, 0.959]	0.960 (< 0.001)
Fever	0.472	[0.414, 0.530]	0.952 (< 0.001)
Fatigue or lethargy	0.439	[0.380, 0.499]	0.935 (< 0.001)
Septic shock	0.416	[0.380, 0.452]	0.952 (< 0.001)
Headache (cephalgia)	0.376	[0.325, 0.428]	0.906 (< 0.001)
Sore throat	0.337	[0.274, 0.399]	0.839 (< 0.001)
Anosmia	0.263	[0.211, 0.315]	0.330 (< 0.001)
Diarrhea	0.286	[0.226, 0.346]	0.721 (< 0.001)
Disgeusia	0.189	[0.154, 0.224]	-0.012 (1.000)
Rhinorrhea	0.186	[0.144, 0.227]	0.023 (< 0.001)
ARDS/resp. failure	0.315	[0.281, 0.349]	0.903 (< 0.001)
Total symptoms	0.229	[0.200, 0.259]	0.407 (< 0.001)
Muscle ache	0.194	[0.153, 0.234]	0.549 (< 0.001)
Malaise	0.187	[0.138, 0.237]	0.634 (< 0.001)
Other systems	0.149	[0.065, 0.233]	0.117 (< 0.001)
Asymptomatic	0.151	[0.121, 0.181]	0.778 (< 0.001)
Pleuritic pain	0.099	[0.076, 0.122]	0.156 (< 0.001)
Hemoptysis	0.099	[0.079, 0.118]	0.524 (< 0.001)
Pneumothorax	0.081	[0.064, 0.099]	0.281 (< 0.001)
Total symptoms 2+	0.081	[0.052, 0.110]	0.528 (< 0.001)
Shock liver	0.043	[0.016, 0.071]	0.060 (< 0.001)
Group 7: Imaging			
CT ground-glass B/L	0.541	[0.429, 0.653]	0.760 (< 0.001)
CT consolidation B/L	0.427	[0.360, 0.495]	0.838 (< 0.001)
COVID RADS 5[12]	0.312	[0.252, 0.372]	0.808 (< 0.001)
X-ray ground-glass B/L	0.134	[0.109, 0.158]	0.462 (< 0.001)
CT consolidation unilat.	0.117	[0.088, 0.146]	0.515 (< 0.001)
X-ray consolidation B/L	0.095	[0.076, 0.114]	0.276 (< 0.001)
COVID RADS 4[12]	0.071	[0.063, 0.080]	0.005 (0.246)
X-ray consolidation unilat.	0.082	[0.060, 0.103]	0.314 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
COVID RADS 3[12]	0.073	[0.055, 0.091]	-0.040 (1.000)
X-ray ground-glass unilat.	0.068	[0.037, 0.098]	0.069 (0.103)
CT ground-glass unilat.	0.052	[0.048, 0.056]	0.030 (< 0.001)
COVID RADS 1[12]	0.033	[0.014, 0.052]	0.344 (< 0.001)
Group 8: Liver Function (Blood)			
AST (Norm: 0-38)	0.401	[0.339, 0.462]	0.854 (< 0.001)
LDH	0.361	[0.315, 0.408]	0.703 (< 0.001)
ALT	0.256	[0.211, 0.300]	0.345 (< 0.001)
ALT (High: >42)	0.304	[0.258, 0.350]	0.750 (< 0.001)
AST (High: >38)	0.253	[0.213, 0.292]	0.552 (< 0.001)
ALP	0.217	[0.173, 0.261]	0.266 (< 0.001)
ALT (Norm: 0-42)	0.200	[0.162, 0.239]	0.284 (< 0.001)
TBIL	0.187	[0.157, 0.217]	0.081 (< 0.001)
AST	0.186	[0.161, 0.211]	0.418 (< 0.001)
TBIL (Norm: 0.1-1.2)	0.163	[0.135, 0.192]	0.554 (< 0.001)
TBIL (High: >1.2)	0.144	[0.118, 0.170]	0.541 (< 0.001)
LDH (High: >333)	0.109	[0.098, 0.120]	0.220 (< 0.001)
LDH (Norm: 0-333)	0.100	[0.089, 0.111]	0.106 (< 0.001)
ALP (High: >147)	0.093	[0.081, 0.104]	0.140 (< 0.001)
ALP (Norm: 44-147)	0.104	[0.093, 0.116]	0.451 (< 0.001)
ALP (Low: <44)	0.053	[0.033, 0.074]	0.028 (0.015)
Group 9: Complete blood count (CBC)			
Lymphocytes	0.566	[0.495, 0.638]	0.671 (< 0.001)
WBC leukocytes ($\times$10)	0.584	[0.516, 0.653]	0.765 (< 0.001)
Lymphocytes (L: <0.8)	0.480	[0.408, 0.552]	0.904 (< 0.001)
Hematocrit	0.459	[0.381, 0.537]	0.674 (< 0.001)
Lymphocytes (N: 0.8-4.6)	0.499	[0.444, 0.553]	0.920 (< 0.001)
Neutrophils	0.422	[0.368, 0.477]	0.421 (< 0.001)
Neutrophils (H: >6000)	0.520	[0.452, 0.588]	0.948 (< 0.001)
Hemoglobin	0.420	[0.331, 0.508]	0.624 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Neutrophils (N: 1-6$\times$10^3)	0.205	[0.157, 0.253]	0.265 (< 0.001)
Hemoglobin (N: 13.5-16.5)	0.300	[0.249, 0.350]	0.786 (< 0.001)
Hemoglobin (L: <13.5)	0.210	[0.169, 0.250]	0.330 (< 0.001)
Hematocrit (N: 37-47)	0.175	[0.141, 0.208]	0.085 (< 0.001)
Hematocrit (L: <37)	0.159	[0.130, 0.189]	0.051 (< 0.001)
Bands (ABAST)	0.161	[0.106, 0.215]	0.668 (< 0.001)
PLT (H: >450)[13]	0.129	[0.110, 0.148]	0.613 (< 0.001)
PLT (N: 150-450)[13]	0.106	[0.093, 0.118]	0.236 (< 0.001)
PLT (L: <150)[13]	0.127	[0.105, 0.149]	0.499 (< 0.001)
Hematocrit (H: >47)	0.083	[0.071, 0.095]	0.028 (< 0.001)
Neutrophils (L: <1000)	0.107	[0.097, 0.116]	0.760 (< 0.001)
Hemoglobin (H: >16.5)	0.079	[0.068, 0.089]	0.421 (< 0.001)
Lymphocytes (H: >4.6)	0.055	[0.039, 0.072]	0.086 (< 0.001)
Group 10: Septic Markers			
CRP[14]	0.243	[0.178, 0.309]	0.059 (< 0.001)
PCT[15]	0.230	[0.162, 0.299]	0.132 (< 0.001)
CRP (N: <5)[14]	0.167	[0.138, 0.196]	0.779 (< 0.001)
CRP (H: >5)[14]	0.154	[0.132, 0.177]	0.718 (< 0.001)
PCT (H: >0.5)[15]	0.068	[0.062, 0.074]	0.003 (0.341)
Group 11: Comprehensive Metabolic Panel (CMP)			
Urea (H: >50)	0.741	[0.646, 0.836]	0.935 (< 0.001)
Urea (N: 10-50)	0.544	[0.473, 0.614]	0.884 (< 0.001)
Urea	0.529	[0.453, 0.604]	0.759 (< 0.001)
Creatinine (N: 0.7-1.3)	0.527	[0.466, 0.589]	0.925 (< 0.001)
Creatinine (H: >1.3)	0.339	[0.264, 0.413]	0.839 (< 0.001)
Albumin (L: <3.5)	0.190	[0.141, 0.239]	0.161 (< 0.001)
Creatinine	0.280	[0.206, 0.354]	0.309 (< 0.001)
Albumin (N: 3.5-5.0)	0.165	[0.136, 0.193]	0.096 (< 0.001)
Creatinine (L: <0.7)	0.212	[0.173, 0.250]	0.334 (< 0.001)
Albumin	0.166	[0.130, 0.203]	0.199 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Urea (L: <10)	0.093	[0.072, 0.114]	0.480 (< 0.001)
Albumin (H: >5.0)	0.075	[0.067, 0.084]	0.034 (< 0.001)
Group 12: Airway Management			
Total #. of treatments	0.710	[0.630, 0.789]	0.780 (< 0.001)
Reservoir mask	0.300	[0.232, 0.368]	0.433 (< 0.001)
Binasal cannula	0.349	[0.292, 0.405]	0.769 (< 0.001)
High flow cannula	0.383	[0.335, 0.430]	0.934 (< 0.001)
CPAP[16]	0.158	[0.116, 0.200]	0.758 (< 0.001)
VMNI[17]	0.128	[0.102, 0.153]	0.641 (< 0.001)
Venturi mask	0.119	[0.098, 0.140]	0.621 (< 0.001)
Invasive ventilation	0.085	[0.065, 0.105]	0.223 (< 0.001)
Nebulizer	0.086	[0.070, 0.102]	0.095 (< 0.001)
Tracheostomy	9.75e-04	[0.000, 0.002]	0.321 (< 0.001)
Group 13: Tests for diagnostics			
Molecular[18]	0.567	[0.504, 0.629]	0.956 (< 0.001)
Antibodies (IgG/IgM)[19]	0.189	[0.154, 0.224]	-0.028 (1.000)
Antigen[20]	0.126	[0.101, 0.151]	0.537 (< 0.001)
Clinical[21]	1.04e-04	[3.60e-05, 1.72e-04]	0.045 (< 0.001)
Group 14: Coagulation (Blood)			
APTT[22]	0.404	[0.343, 0.466]	0.697 (< 0.001)
D dimer (L:<0.5)	0.352	[0.317, 0.386]	0.683 (< 0.001)
Fibrinogen (N:200-400)	0.294	[0.266, 0.322]	0.368 (< 0.001)
Fibrinogen (H:>400)	0.261	[0.231, 0.291]	0.305 (< 0.001)
Ferritin	0.345	[0.250, 0.439]	0.658 (< 0.001)
D dimer (H:>0.5)	0.254	[0.226, 0.282]	0.562 (< 0.001)
PT (N:10-14)[23]	0.226	[0.196, 0.257]	0.501 (< 0.001)
APTT (N:26-47)[22]	0.201	[0.163, 0.238]	0.433 (< 0.001)
INR[24]	0.212	[0.177, 0.248]	0.113 (< 0.001)
INR (N:0.8-1.1)[24]	0.174	[0.144, 0.204]	0.145 (< 0.001)
INR (H:>1.1)[24]	0.166	[0.136, 0.197]	0.108 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
PT (H:>14)[23]	0.204	[0.168, 0.241]	0.352 (< 0.001)
APTT (L:<26)[22]	0.145	[0.120, 0.170]	0.038 (< 0.001)
PT[23]	0.144	[0.117, 0.172]	0.083 (< 0.001)
INR (L:<0.8)[24]	0.118	[0.102, 0.134]	0.071 (< 0.001)
Ferritin (H:>300)	0.109	[0.095, 0.124]	0.046 (< 0.001)
Ferritin (N:30-300)	0.111	[0.100, 0.122]	0.095 (< 0.001)
Fibrinogen (L:<200)	0.116	[0.072, 0.161]	0.058 (< 0.001)
D dimer	0.131	[0.071, 0.192]	0.153 (< 0.001)
Fibrinogen	0.117	[0.090, 0.144]	0.029 (< 0.001)
APTT (H:>47)[22]	0.122	[0.091, 0.153]	0.499 (< 0.001)
PT (L:<10)[23]	0.090	[0.080, 0.100]	0.202 (< 0.001)
Ferritin (L:<30)	0.091	[0.065, 0.117]	0.040 (< 0.001)
Group 15: Arterial Blood Gas			
$PAFIO_2$[25]	3.014	[2.787, 3.242]	0.851 (< 0.001)
FIO_2[26]	2.560	[2.284, 2.836]	0.944 (< 0.001)
$PAFIO_2$ (L:<100)[25]	1.980	[1.850, 2.110]	0.989 (< 0.001)
$FAFIO_2$ (H:>300)[25]	2.124	[2.036, 2.212]	0.994 (< 0.001)
pH	1.578	[1.364, 1.792]	0.796 (< 0.001)
$AADPO_2$[27]	1.529	[1.371, 1.686]	0.750 (< 0.001)
PCO_2	1.328	[0.971, 1.684]	0.833 (< 0.001)
PCO_2 (H:>45)	1.109	[1.046, 1.171]	0.975 (< 0.001)
pH (L:<7.35)	0.808	[0.763, 0.852]	0.972 (< 0.001)
PO_2	0.865	[0.768, 0.962]	0.772 (< 0.001)
PCO_2 (N:35-45)	1.002	[0.943, 1.060]	0.978 (< 0.001)
pH (N:7.35-7.45)	0.914	[0.838, 0.989]	0.963 (< 0.001)
$PAFIO_2$ (N:200-300)[25]	0.606	[0.561, 0.652]	0.974 (< 0.001)
HCO_3 (N:22-28)	0.462	[0.390, 0.533]	0.891 (< 0.001)
SO_2	0.382	[0.335, 0.430]	0.586 (< 0.001)
Lactate (N:0.1-2)	0.382	[0.333, 0.431]	0.744 (< 0.001)
PO_2 (N:75-100)	0.403	[0.354, 0.451]	0.778 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Lactate (H:>2)	0.412	[0.366, 0.459]	0.899 (< 0.001)
HCO_3	0.335	[0.281, 0.389]	0.486 (< 0.001)
SO_2 (N:94-100)	0.324	[0.281, 0.368]	0.825 (< 0.001)
PO_2 (L:<75)	0.172	[0.140, 0.204]	0.361 (< 0.001)
SO_2 (L:<94)	0.285	[0.241, 0.329]	0.734 (< 0.001)
pH (H:>7.45)	0.371	[0.340, 0.402]	0.924 (< 0.001)
$PAFIO_2$ (N:100-200)[25]	0.207	[0.168, 0.245]	0.675 (< 0.001)
HCO_3 (H:>28)	0.286	[0.238, 0.335]	0.841 (< 0.001)
PCO_2 (L:<35)	0.273	[0.214, 0.331]	0.602 (< 0.001)
HCO_3 (L:<22)	0.172	[0.135, 0.209]	0.292 (< 0.001)
Lactate	0.224	[0.187, 0.260]	0.468 (< 0.001)
PO_2 (H:>100)	0.092	[0.076, 0.107]	0.294 (< 0.001)
$AADPO_2$ (Gradient>15)[27]	0.110	[0.083, 0.136]	0.402 (< 0.001)
Group 16: Medications			
DEX (home) w/o air[28]	6.186	[5.795, 6.577]	0.996 (< 0.001)
DEX (hospital) w. air[28]	1.254	[1.138, 1.370]	0.969 (< 0.001)
3rd CEP (days)[29]	0.491	[0.446, 0.537]	0.838 (< 0.001)
IVM (hospital)[30]	0.261	[0.206, 0.316]	0.395 (< 0.001)
Carbapenem	0.332	[0.290, 0.374]	0.870 (< 0.001)
cefepime (4th CEP)[29]	0.276	[0.216, 0.336]	0.741 (< 0.001)
3rd CEP[29]	0.212	[0.165, 0.259]	0.487 (< 0.001)
Medrol (pulse for >5 days)[31]	0.152	[0.108, 0.196]	0.016 (0.005)
IVM (home)[30]	0.156	[0.130, 0.182]	0.401 (< 0.001)
HCQ[32]	0.158	[0.117, 0.198]	0.175 (< 0.001)
Meropenem	0.239	[0.204, 0.274]	0.822 (< 0.001)
Meropenem (days)	0.194	[0.131, 0.257]	0.612 (< 0.001)
cefepime (4th CEP, days)[29]	0.126	[0.103, 0.148]	0.223 (< 0.001)
Acetylcysteine	0.128	[0.090, 0.165]	0.040 (< 0.001)
Colchicine	0.159	[0.133, 0.184]	0.430 (< 0.001)
Vancomycin	0.172	[0.141, 0.202]	0.802 (< 0.001)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3 – continued

Feature	Importance (%)[1,3]	95% CI[1]	Alignment (FDR q)[2]
Heparin (prophylaxis)	0.139	[0.115, 0.164]	0.589 (< 0.001)
Colistine (alone)	0.117	[0.092, 0.143]	0.684 (< 0.001)
Ureidopenicillin (Zosyn)	0.095	[0.082, 0.107]	0.092 (< 0.001)
Enoxaparin (prophylaxis)	0.079	[0.068, 0.090]	0.062 (< 0.001)
Colistin	0.110	[0.089, 0.131]	0.805 (< 0.001)
Broad spectrum (ABX)	0.103	[0.065, 0.140]	0.413 (< 0.001)
Tocilizumab (2 doses)	0.099	[0.074, 0.124]	0.316 (< 0.001)
Vancomycin (days)	0.077	[0.060, 0.094]	0.187 (< 0.001)
Ceftazidima (3rd CEP)[29]	0.086	[0.070, 0.101]	0.728 (< 0.001)
Imipenem (carbapenem)	0.071	[0.061, 0.080]	0.125 (< 0.001)
Chlorine dioxide	0.060	[0.056, 0.064]	0.161 (< 0.001)
Linezolid (days)	0.025	[0.012, 0.039]	-0.020 (1.000)
anti fungal	0.023	[0.011, 0.035]	0.051 (< 0.001)
Fluconazole (days)	0.020	[0.009, 0.031]	0.058 (< 0.001)
Fluconazole	0.018	[0.008, 0.027]	0.062 (< 0.001)
Colistin comb. (days)[33]	3.85e-05	[1.05e-05, 6.64e-05]	-0.007 (0.027)
Tigecycline	0.100	[0.077, 0.124]	0.335 (< 0.001)
Colistin (alone, #. days)	0.112	[0.053, 0.170]	0.070 (< 0.001)
Ceftazidima (days)	0.095	[0.074, 0.116]	0.757 (< 0.001)
Tigecycline (days)	0.090	[0.066, 0.113]	0.008 (0.114)
Ureidopenicillin (days)	0.081	[0.055, 0.107]	-0.005 (0.849)
Lopinavir/Ritonavir	0.077	[0.061, 0.092]	-0.024 (1.000)
SAM (days)[34]	0.075	[0.057, 0.093]	0.047 (< 0.001)
Imipenem (days)	0.074	[0.061, 0.087]	0.125 (< 0.001)
Vitamin C (IV)	0.070	[0.061, 0.078]	0.280 (< 0.001)
SAM[34]	0.067	[0.045, 0.088]	0.323 (< 0.001)
Codeine	0.055	[0.038, 0.073]	0.392 (< 0.001)
Linezolid	0.047	[0.033, 0.060]	0.140 (< 0.001)
Colistin comb.	1.15e-04	[5.21e-05, 1.77e-04]	0.027 (0.002)

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Notes:

[1] Within each model–trial, signed SHAP values were divided by the mean total absolute attribution across evaluation patients. The normalized values were aggregated across the ten algorithms using normalized mean-AUROC weights. Importance is the mean absolute aggregated attribution across evaluation patients, expressed as a percentage of the standardized attribution budget. Means and normal-approximation 95% confidence intervals (mean $\pm$ 1.96 SEM) were calculated from 30 trial-specific estimates.

[2] Alignment is the mean cosine similarity between signed feature-attribution vectors for corresponding evaluation patients across unique trial pairs. Parentheses contain Benjamini–Hochberg FDR-adjusted q-values from a one-sided whole-trial patient-correspondence randomization test using 2,000 permutations. These values assess patient-specific attribution alignment, not association with mortality, importance magnitude, or a clinical effect.

[3] Absolute SHAP values describe attribution magnitude; the sign, displayed in Supplementary Figures 1-16, describes direction relative to the model baseline.

[4] CKD stages 1–5 denote chronic kidney disease stages from least to most severe.

[5] CKD hemodialysis denotes chronic kidney disease with receipt of hemodialysis.

[6] AKI, acute kidney injury.

[7] COPD, chronic obstructive pulmonary disease.

[8] Pregnancy with vaginal delivery.

[9] HCC, hepatocellular carcinoma.

[10] CRAB, carbapenem-resistant *Acinetobacter baumannii*.

[12] CO-RADS categories 1–5 indicate increasing suspicion for pulmonary involvement by COVID-19 on chest computed tomography, from very low to very high.

[13] PLT, platelet count.

[14] CRP, C-reactive protein.

[15] PCT, procalcitonin.

[16] CPAP, continuous positive airway pressure.

[17] VMNI, non-invasive mechanical ventilation.

[18] Molecular testing denotes nucleic acid amplification testing, including polymerase chain reaction.

[19] Antibodies denote SARS-CoV-2 IgG/IgM rapid testing.

[20] Antigen testing detects SARS-CoV-2 proteins.

[21] Clinical diagnosis denotes the treating physician's clinical assessment.

[22] APTT, activated partial thromboplastin time.

[23] PT, prothrombin time.

[24] INR, international normalized ratio.

[25] PaO_2/FiO_2, also called the P/F ratio, is arterial oxygen partial pressure divided by the fraction of inspired oxygen.

[26] FiO_2, fraction of inspired oxygen.

[27] $AADPO_2$ denotes the alveolar–arterial oxygen gradient.

[28] DEX, dexamethasone.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[29] 3rd CEP and 4th CEP, third- and fourth-generation cephalosporins.

[30] IVM, ivermectin.

[31] Medrol, methylprednisolone.

[32] HCQ, hydroxychloroquine.

[33] Colistin comb., colistin combination therapy.

[34] SAM, ampicillin/sulbactam.

[*] Vaccination indicators were omitted because only one participant had received a COVID-19 vaccine.

Table 3 reports two complementary summaries for 250 non-vaccination features. Importance is the mean absolute normalized ensemble attribution across 30 trials, with trial-based 95% confidence intervals. Alignment is the mean cosine similarity between signed attribution vectors for corresponding patients across unique trial pairs; the accompanying q-value comes from the patient-correspondence randomization analysis after false-discovery-rate adjustment. Importance therefore measures average magnitude, whereas alignment measures the consistency of the patient-level attribution pattern.

The largest normalized ensemble importance was observed for dexamethasone use at home without oxygen support (6.186%; 95% CI, 5.795–6.577), followed by the continuous PaO_2/FiO_2 ratio (3.014%; 95% CI, 2.787–3.242), shortness of breath (2.774%; 95% CI, 2.592–2.956), and FiO_2 (2.560%; 95% CI, 2.284–2.836). These features also had high mean patient-level alignment, ranging from 0.851 for PaO_2/FiO_2 to 0.996 for home dexamethasone use, with $q < 0.001$ for each.

4.3.1 Demographic and Baseline Health Indicators

The indicator for age 60 years or older had an importance of 1.148% (95% CI, 1.073–1.223), and continuous age had an importance of 1.018% (95% CI, 0.923–1.113). Their mean alignments were 0.990 and 0.804, respectively ($q < 0.001$ for both). The sex indicators had lower importance: 0.247% for female and 0.222% for male. Among past medical history variables, other recorded comorbidities (0.206%), hypertension (0.208%), obesity (0.181%), hyperglycemia (0.169%), stage 5 chronic kidney disease (0.160%), and type 2 diabetes (0.157%) contributed to the ensemble profile at broadly similar magnitudes.

Importance magnitude did not uniformly imply patient-level alignment. For example, the indicator for fewer than two comorbidities had an importance of 0.189% but a mean alignment of -0.048 ($q = 1.000$), whereas hyperglycemia had a similar importance of 0.169% and an align-

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ment of 0.838 ($q < 0.001$). This difference illustrates the distinct information provided by the two columns of Table 3.

4.3.2 Clinical Presentation and Symptoms

Shortness of breath was the leading symptom feature (2.774%; 95% CI, 2.592–2.956), followed by productive cough (1.026%), dry cough (0.868%), fever (0.472%), fatigue or lethargy (0.439%), and septic shock (0.416%). These features showed high mean patient-level alignment, ranging from 0.935 to 0.980, with $q < 0.001$ for each. Dysgeusia had a lower importance of 0.189% and did not show corresponding alignment (mean cosine similarity, -0.012; $q = 1.000$).

4.3.3 Laboratory Measurements

Laboratory contributions spanned hematologic, renal, hepatic, inflammatory, coagulation, and blood-gas measurements. Elevated urea had an importance of 0.741%, while continuous urea and normal-range urea had values of 0.529% and 0.544%, respectively. Normal-range creatinine had an importance of 0.527%. Among complete blood count variables, leukocyte count and lymphocyte count had importance values of 0.584% and 0.566%. Other contributions included APTT (0.404%), normal-range AST (0.401%), LDH (0.361%), low D-dimer (0.352%), ferritin (0.345%), C-reactive protein (0.243%), and procalcitonin (0.230%). These values identify information used by the fitted models and do not establish organ injury or a temporal sequence of physiologic deterioration.

4.3.4 Imaging and Physiologic Measurements

Oxygenation and arterial blood-gas variables contained several of the largest attribution magnitudes. In addition to continuous PaO_2/FiO_2 and FiO_2, $PaO_2/FiO_2 < 100$ had an importance of 1.980%, pH 1.578%, the alveolar–arterial oxygen gradient 1.529%, and PCO_2 1.328%. Bilateral ground-glass opacity and bilateral consolidation on computed tomography had importance values of 0.541% and 0.427%, respectively. These results show that physiologic and imaging measurements both contributed to the mortality predictions, with larger attribution magnitudes for several oxygenation measures.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4.3.5 Acute Care Interventions and Pharmacotherapy

Treatment and respiratory-support variables also contributed to the ensemble profile. In addition to home dexamethasone use without oxygen support, dexamethasone administered in hospital with oxygen support had an importance of 1.254%, total number of treatments 0.710%, third-generation cephalosporin duration 0.491%, high-flow cannula 0.383%, and binasal cannula 0.349%. These attributions describe patterns captured by the models and may reflect disease severity, treatment indication, timing, access to care, and treatment selection. They do not estimate treatment benefit or harm.

4.4 Distribution-Adjusted Subgroup Differences

Table 4 presents, within each subgroup, up to 20 features with the largest absolute distribution-adjusted differences in normalized ensemble SHAP importance. The pooled evaluation reference for every comparison comprised all 558 available evaluation observations, including observations belonging to the subgroup. Differences were calculated as pooled-reference importance minus distribution-adjusted subgroup importance. Positive values therefore indicate greater attribution magnitude in the pooled evaluation reference, whereas negative values indicate greater magnitude in the distribution-adjusted subgroup. All 283 encoded features entered the calculations; rows were selected by $|\Delta|$, rather than by the nominal p-value.

Table 4: The largest absolute distribution-adjusted differences in normalized ensemble SHAP importance between each subgroup and the pooled evaluation reference. Within each subgroup, up to 20 features were selected by absolute Δ, rather than by nominal p-value.

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
Age ≥ 50 *years* ($n_G = 109$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	5.634	0.552	8.9%	0.260
2	pH	1.578	1.392	0.187	11.8%	0.145
3	Dexamethasone in hospital with oxygen	1.254	1.423	-0.168	-13.4%	0.195
4	A_negative	0.061	0.161	-0.100	-163.0%	0.030
5	Fever	0.472	0.570	-0.098	-20.7%	0.057
6	FiO_2	2.560	2.655	-0.095	-3.7%	0.195
7	Total symptoms 2+	0.081	0.173	-0.092	-113.7%	0.072
8	PaO_2/FiO_2	3.014	3.102	-0.088	-2.9%	0.200
9	Tigecycline	0.100	0.180	-0.080	-79.4%	0.044
10	HCO_3	0.335	0.410	-0.074	-22.2%	0.117
11	Tuberculosis	0.093	0.163	-0.070	-75.0%	0.082

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
12	Tocilizumab (2 doses)	0.099	0.163	-0.064	-64.2%	0.044
13	Vancomycin	0.172	0.234	-0.062	-36.4%	0.038
14	CKD stage 3	0.125	0.069	0.057	45.2%	0.157
15	Ischemic HD	0.129	0.076	0.053	40.9%	0.098
16	$AADPO_2$	1.529	1.581	-0.052	-3.4%	0.150
17	PCT	0.230	0.179	0.051	22.3%	0.065
18	X-ray ground-glass B/L	0.134	0.183	-0.050	-37.4%	0.033
19	CKD stage 2	0.107	0.153	-0.046	-42.9%	0.203
20	CKD stage 1	0.048	0.092	-0.044	-93.2%	0.050
Male ($n_G = 374$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	6.125	0.061	1.0%	0.456
2	FiO_2	2.560	2.518	0.042	1.6%	0.271
3	High flow cannula	0.383	0.424	-0.042	-10.9%	0.229
4	CKD stage 5	0.160	0.123	0.037	23.2%	0.074
5	$PaO_2/FiO_2 < 100$	1.980	1.945	0.035	1.8%	0.267
6	Age 18+	0.412	0.439	-0.027	-6.5%	0.335
7	Total #. of treatments	0.710	0.685	0.024	3.4%	0.298
8	pH	1.578	1.601	-0.023	-1.5%	0.423
9	Meropenem	0.239	0.261	-0.021	-8.9%	0.106
10	AKI	0.106	0.127	-0.021	-19.4%	0.272
11	Carbapenem	0.332	0.352	-0.020	-6.0%	0.243
12	Urea (H: >50)	0.741	0.761	-0.020	-2.7%	0.246
13	CKD stage 2	0.107	0.126	-0.020	-18.3%	0.293
14	Total symptoms 2+	0.081	0.061	0.020	24.1%	0.256
15	$AADPO_2$	1.529	1.509	0.019	1.3%	0.275
16	CKD stage 3	0.125	0.144	-0.019	-15.4%	0.323
17	Septic shock	0.416	0.434	-0.018	-4.3%	0.311
18	Urea (N: 10-50)	0.544	0.561	-0.018	-3.2%	0.182
19	Pneumothorax	0.081	0.064	0.017	21.2%	0.127
20	Age 0+	0.040	0.023	0.017	42.2%	0.125
Female ($n_G = 177$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	6.408	-0.222	-3.6%	0.379
2	$PaO_2/FiO_2 < 100$	1.980	2.066	-0.086	-4.4%	0.126
3	FiO_2	2.560	2.644	-0.084	-3.3%	0.165
4	High flow cannula	0.383	0.299	0.084	21.9%	0.102
5	CKD stage 5	0.160	0.239	-0.079	-49.1%	0.036
6	Dexamethasone in hospital with oxygen	1.254	1.176	0.078	6.2%	0.307
7	Carbapenem	0.332	0.274	0.058	17.4%	0.040

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
8	Age 18+	0.412	0.361	0.052	12.5%	0.250
9	PCO_2 (H:>45)	1.109	1.058	0.051	4.6%	0.248
10	pH	1.578	1.530	0.048	3.1%	0.375
11	Meropenem	0.239	0.192	0.047	19.7%	0.010
12	Urea (H: >50)	0.741	0.695	0.046	6.2%	0.103
13	Hyperglycemia	0.169	0.126	0.043	25.6%	0.048
14	AKI	0.106	0.063	0.043	40.5%	0.110
15	Total symptoms 2+	0.081	0.123	-0.042	-52.4%	0.196
16	CKD stage 2	0.107	0.067	0.040	37.5%	0.140
17	Urea (N: 10-50)	0.544	0.504	0.040	7.3%	0.051
18	CKD stage 3	0.125	0.086	0.039	31.0%	0.198
19	INR	0.212	0.175	0.037	17.4%	0.044
20	Pneumothorax	0.081	0.118	-0.037	-45.2%	0.088
Blood type A ($n_G = 10$, $n_R = 558$; descriptive)						
1	Dexamethasone at home without oxygen	6.186	9.048	-2.862	-46.3%	–
2	FiO_2	2.560	3.368	-0.808	-31.6%	–
3	HCO_3	0.335	1.007	-0.672	-200.7%	–
4	Dexamethasone in hospital with oxygen	1.254	0.643	0.611	48.7%	–
5	CKD stage 5	0.160	0.742	-0.582	-363.6%	–
6	pH	1.578	2.154	-0.576	-36.5%	–
7	Fever	0.472	0.781	-0.309	-65.5%	–
8	Platelet count ($\times 10^3$)	0.337	0.644	-0.308	-91.3%	–
9	$PaO_2/FiO_2 > 300$	2.124	1.846	0.278	13.1%	–
10	PCO_2	1.328	1.556	-0.228	-17.2%	–
11	Age 18+	0.412	0.190	0.222	53.8%	–
12	Age	1.018	0.805	0.212	20.9%	–
13	$AADPO_2$	1.529	1.335	0.193	12.6%	–
14	PO_2	0.865	0.676	0.189	21.8%	–
15	Ischemic HD	0.129	0.311	-0.182	-140.8%	–
16	PaO_2/FiO_2 (200–300)	0.606	0.786	-0.180	-29.7%	–
17	Carbapenem	0.332	0.511	-0.179	-54.0%	–
18	Total #. of treatments	0.710	0.534	0.175	24.7%	–
19	Shortness of breath	2.774	2.619	0.155	5.6%	–
20	Creatinine	0.280	0.426	-0.146	-52.0%	–
Blood type B ($n_G = 4$, $n_R = 558$; descriptive)						
1	Dexamethasone at home without oxygen	6.186	3.485	2.701	43.7%	–
2	X-ray consolidation unilat.	0.082	1.567	-1.485	-1815.4%	–
3	FiO_2	2.560	1.320	1.240	48.4%	–

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
4	PaO_2/FiO_2	3.014	1.846	1.168	38.8%	–
5	High flow cannula	0.383	1.478	-1.096	-286.5%	–
6	$PaO_2/FiO_2 < 100$	1.980	0.957	1.023	51.7%	–
7	Dexamethasone in hospital with oxygen	1.254	0.505	0.749	59.7%	–
8	AST	0.186	0.867	-0.681	-366.2%	–
9	Other systems	0.149	0.825	-0.677	-455.5%	–
10	CKD Hemodialysis	0.134	0.714	-0.580	-433.7%	–
11	PCO_2	1.328	1.892	-0.564	-42.5%	–
12	Platelet count ($\times 10^3$)	0.337	0.898	-0.562	-166.8%	–
13	$AADPO_2$	1.529	1.000	0.528	34.6%	–
14	PCO_2 (H:>45)	1.109	1.549	-0.441	-39.7%	–
15	Lopinavir/Ritonavir	0.077	0.513	-0.436	-569.3%	–
16	CKD stage 5	0.160	0.579	-0.419	-261.8%	–
17	Colistin	0.110	0.508	-0.398	-360.8%	–
18	X-ray ground-glass B/L	0.134	0.523	-0.389	-291.7%	–
19	VMNI	0.128	0.503	-0.375	-294.4%	–
20	Colistine (alone)	0.117	0.427	-0.310	-263.6%	–
Blood type O ($n_G = 143$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	5.542	0.644	10.4%	0.198
2	Dexamethasone in hospital with oxygen	1.254	1.502	-0.248	-19.8%	0.081
3	Age 18+	0.412	0.589	-0.177	-42.8%	0.027
4	pH	1.578	1.409	0.169	10.7%	0.136
5	Age 30+	0.245	0.403	-0.158	-64.4%	0.004
6	High flow cannula	0.383	0.233	0.150	39.2%	0.016
7	Total symptoms 2+	0.081	0.189	-0.108	-133.6%	0.034
8	PCO_2	1.328	1.405	-0.077	-5.8%	0.307
9	Ischemic HD	0.129	0.194	-0.064	-49.9%	0.076
10	CPAP	0.158	0.222	-0.064	-40.4%	0.055
11	Ferritin	0.345	0.282	0.062	18.0%	0.148
12	Other systems	0.149	0.206	-0.057	-38.6%	0.111
13	Malaise	0.187	0.231	-0.044	-23.3%	0.040
14	Neutrophils (H: >6000)	0.520	0.562	-0.042	-8.1%	0.059
15	FiO_2	2.560	2.518	0.042	1.6%	0.325
16	AKI	0.106	0.146	-0.040	-37.6%	0.191
17	Hyperglycemia	0.169	0.205	-0.037	-21.7%	0.124
18	CKD Hemodialysis	0.134	0.097	0.037	27.3%	0.140
19	CRP	0.243	0.280	-0.036	-14.9%	0.182
20	CKD stage 5	0.160	0.124	0.036	22.4%	0.181

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
Asthma or COPD ($n_G = 11$, $n_R = 558$; descriptive)						
1	AKI	0.106	0.655	-0.549	-517.6%	–
2	Ceftazidima (3rd CEP)	0.086	0.515	-0.429	-500.9%	–
3	PaO_2/FiO_2	3.014	2.590	0.425	14.1%	–
4	Tuberculosis	0.093	0.488	-0.395	-424.4%	–
5	Age 0+	0.040	0.383	-0.344	-867.4%	–
6	$PaO_2/FiO_2 < 100$	1.980	1.637	0.343	17.3%	–
7	AADPO$_2$	1.529	1.187	0.342	22.3%	–
8	Dexamethasone in hospital with oxygen	1.254	1.549	-0.295	-23.5%	–
9	Ceftazidima (days)	0.095	0.387	-0.292	-307.0%	–
10	pH	1.578	1.308	0.270	17.1%	–
11	Total #. of treatments	0.710	0.472	0.237	33.5%	–
12	Lymphocytes	0.566	0.335	0.232	40.9%	–
13	Age	1.018	1.249	-0.231	-22.7%	–
14	Ischemic HD	0.129	0.354	-0.225	-174.2%	–
15	PO$_2$	0.865	0.648	0.217	25.1%	–
16	Age 18+	0.412	0.201	0.212	51.4%	–
17	PCO$_2$	1.328	1.530	-0.202	-15.2%	–
18	Acetylcysteine	0.128	0.322	-0.194	-152.3%	–
19	Broad spectrum (ABX)	0.103	0.290	-0.187	-182.5%	–
20	Dexamethasone at home without oxygen	6.186	6.010	0.176	2.8%	–
Metabolic/cardiovascular ($n_G = 303$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	5.679	0.507	8.2%	0.189
2	pH	1.578	1.475	0.104	6.6%	0.191
3	Dexamethasone in hospital with oxygen	1.254	1.352	-0.098	-7.8%	0.234
4	FiO_2	2.560	2.605	-0.045	-1.8%	0.261
5	Hyperglycemia	0.169	0.211	-0.043	-25.3%	0.045
6	$PaO_2/FiO_2 < 100$	1.980	2.022	-0.042	-2.1%	0.249
7	Septic shock	0.416	0.457	-0.041	-9.9%	0.146
8	PaO_2/FiO_2	3.014	3.051	-0.037	-1.2%	0.303
9	Other systems	0.149	0.185	-0.037	-24.7%	0.153
10	CKD stage 2	0.107	0.143	-0.036	-33.8%	0.183
11	HCO$_3$ (H:>28)	0.286	0.252	0.035	12.2%	0.043
12	AKI	0.106	0.140	-0.034	-32.5%	0.180
13	CPAP	0.158	0.193	-0.034	-21.7%	0.130
14	HCO$_3$	0.335	0.301	0.034	10.1%	0.153
15	CKD Hemodialysis	0.134	0.167	-0.033	-25.0%	0.139
16	Neutrophils (H: >6000)	0.520	0.553	-0.033	-6.4%	0.054

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
17	PCO_2	1.328	1.296	0.032	2.4%	0.392
18	Total symptoms 2+	0.081	0.112	-0.031	-37.6%	0.229
19	CRP (N: <5)	0.167	0.196	-0.029	-17.7%	0.103
20	Fatigue or lethargy	0.439	0.410	0.029	6.6%	0.161
Kidney disease ($n_G = 30$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	4.736	1.450	23.4%	0.172
2	PCO_2	1.328	0.885	0.443	33.4%	0.109
3	$PaO_2/FiO_2 < 100$	1.980	1.595	0.385	19.4%	0.008
4	FiO_2	2.560	2.276	0.284	11.1%	0.071
5	Age	1.018	0.801	0.217	21.3%	0.135
6	X-ray consolidation unilat.	0.082	0.295	-0.213	-261.0%	0.001
7	Age 18+	0.412	0.213	0.200	48.4%	0.118
8	$AADPO_2$	1.529	1.333	0.196	12.8%	0.024
9	PCO_2 (H:>45)	1.109	0.914	0.195	17.6%	0.118
10	Fever	0.472	0.666	-0.194	-41.2%	0.041
11	Total symptoms 2+	0.081	0.259	-0.178	-219.9%	0.044
12	COPD	0.070	0.248	-0.178	-254.2%	< 0.001
13	High flow cannula	0.383	0.548	-0.166	-43.4%	0.142
14	$PaO_2/FiO_2 > 300$	2.124	1.960	0.164	7.7%	0.085
15	B_positive	0.064	0.216	-0.152	-239.6%	0.008
16	Ferritin	0.345	0.467	-0.122	-35.5%	0.186
17	Age 30+	0.245	0.123	0.122	49.7%	0.120
18	Dexamethasone in hospital with oxygen	1.254	1.142	0.113	9.0%	0.370
19	WBC leukocytes ($\times 10$)	0.584	0.475	0.110	18.8%	0.086
20	Ferritin (L:<30)	0.091	0.199	-0.108	-118.5%	0.013
Chronic viral or liver disease ($n_G = 4$, $n_R = 558$; descriptive)						
1	Dexamethasone at home without oxygen	6.186	3.357	2.829	45.7%	–
2	Dexamethasone in hospital with oxygen	1.254	3.199	-1.945	-155.1%	–
3	Hemoptysis	0.099	1.106	-1.008	-1022.6%	–
4	FiO_2	2.560	1.636	0.924	36.1%	–
5	Ceftazidima (3rd CEP)	0.086	0.869	-0.783	-913.6%	–
6	$PaO_2/FiO_2 > 300$	2.124	1.509	0.615	29.0%	–
7	Age 18+	0.412	1.014	-0.602	-145.9%	–
8	Broad spectrum (ABX)	0.103	0.702	-0.599	-584.2%	–
9	Shortness of breath	2.774	2.260	0.514	18.5%	–
10	Age 30+	0.245	0.721	-0.476	-194.0%	–
11	Fluconazole (days)	0.020	0.492	-0.472	-2362.3%	–
12	Ceftazidima (days)	0.095	0.549	-0.454	-477.9%	–

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4 – continued

Rank	Feature	Pooled reference	Adjusted	Δ	Δ%	Nominal p
13	PCO_2	1.328	1.771	-0.444	-33.4%	–
14	Age	1.018	1.459	-0.441	-43.3%	–
15	$AADPO_2$	1.529	1.138	0.391	25.6%	–
16	Age 60+	1.148	0.825	0.323	28.1%	–
17	PaO_2/FiO_2 (200–300)	0.606	0.287	0.319	52.6%	–
18	CRAB	0.119	0.411	-0.292	-245.1%	–
19	PCO_2 (H:>45)	1.109	1.396	-0.287	-25.9%	–
20	X-ray ground-glass B/L	0.134	0.394	-0.260	-194.8%	–
≥ 2 *comorbidities* ($n_G = 214$, $n_R = 558$)						
1	Dexamethasone at home without oxygen	6.186	5.213	0.973	15.7%	0.063
2	Dexamethasone in hospital with oxygen	1.254	1.388	-0.133	-10.6%	0.189
3	Septic shock	0.416	0.511	-0.095	-22.8%	0.017
4	CKD stage 3	0.125	0.218	-0.093	-74.4%	0.048
5	Hyperglycemia	0.169	0.254	-0.085	-50.6%	0.002
6	Other systems	0.149	0.229	-0.081	-54.4%	0.028
7	CKD Hemodialysis	0.134	0.212	-0.078	-58.5%	0.018
8	Fatigue or lethargy	0.439	0.367	0.072	16.4%	0.013
9	pH	1.578	1.511	0.067	4.2%	0.309
10	AKI	0.106	0.172	-0.066	-62.5%	0.069
11	CKD stage 2	0.107	0.173	-0.066	-61.5%	0.083
12	Total symptoms 2+	0.081	0.140	-0.059	-73.3%	0.111
13	Shortness of breath	2.774	2.719	0.055	2.0%	0.070
14	Tigecycline	0.100	0.153	-0.053	-52.5%	0.072
15	Ceftazidima (days)	0.095	0.147	-0.052	-55.1%	0.107
16	High flow cannula	0.383	0.335	0.048	12.5%	0.224
17	CT consolidation unilat.	0.117	0.164	-0.047	-40.3%	0.047
18	Ischemic HD	0.129	0.176	-0.047	-36.1%	0.108
19	HCO_3 (H:>28)	0.286	0.240	0.047	16.2%	0.021
20	Ceftazidima (3rd CEP)	0.086	0.132	-0.046	-54.1%	0.081

Notes. The pooled analysis comprises all 558 available evaluation observations from three fixed evaluation batches. For every subgroup G, the pooled evaluation reference ($N_R = 558$) includes observations belonging to G. Patient-level importance was calculated as the mean absolute normalized ensemble attribution across ten trials within each batch. For each feature separately, subgroup observations were reweighted toward the feature's marginal distribution in the pooled evaluation reference using up to 40 quantile-based bins and Gaussian smoothing with a width of two bins. Δ is pooled-reference importance minus distribution-adjusted subgroup importance; Δ% is the corresponding

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

difference relative to pooled-reference importance. Positive values indicate greater importance in the pooled evaluation reference, whereas negative values indicate greater importance in the distribution-adjusted subgroup. All 283 encoded features entered the calculation; subgroup-defining features, features with a post-reweighting weighted Kolmogorov–Smirnov distance > 0.10, features with an effective subgroup size < 10, and features with zero attribution variance were excluded from inference for the affected subgroup. Displayed p-values are nominal values from weighted, pooled-variance Student's t tests based on patient-level absolute SHAP magnitudes, using a one-sided tail in the direction of the observed difference. Because each subgroup overlaps the pooled evaluation reference and the feature-specific weights were estimated, these p-values are exploratory. Within each subgroup, up to 20 rows were selected by absolute Δ, rather than by p-value. Groups with fewer than 20 evaluation observations are descriptive and have no inferential values. Benjamini–Hochberg adjustments were calculated both within each subgroup and across all eligible subgroup–feature comparisons. The resulting q-values are provided with the complete supplementary results as sensitivity summaries.

Several of the largest absolute differences occurred in the kidney-disease subgroup. Home dexamethasone use without oxygen support had lower adjusted-subgroup importance than pooled-reference importance (4.736% versus 6.186%; $\Delta = 1.450$; 23.4%), although its nominal one-sided p-value was 0.172. Adjusted importance for $PaO_2/FiO_2 < 100$ was also lower than pooled-reference importance in this subgroup (1.595% versus 1.980%; $\Delta = 0.385$; 19.4%; nominal $p = 0.008$).

Among evaluation observations from patients with at least two documented comorbidities, adjusted importance was lower than pooled-reference importance for home dexamethasone use without oxygen support ($\Delta = 0.973$; 15.7%; nominal $p = 0.063$) and higher for septic shock ($\Delta = -0.095$; -22.8%; nominal $p = 0.017$) and hyperglycemia ($\Delta = -0.085$; -50.6%; nominal $p = 0.002$). In the type O subgroup, age 30 years or older had greater adjusted importance than in the pooled evaluation reference ($\Delta = -0.158$; -64.4%; nominal $p = 0.004$). Results for blood types A and B, asthma or COPD, and chronic viral or liver disease were descriptive because these subgroups contained fewer than 20 evaluation observations.

Across 1,763 eligible subgroup–feature comparisons, 64 had nominal one-sided $p < 0.05$. None remained below 0.05 after Benjamini–Hochberg adjustment within each subgroup or across all comparisons. Because these tests do not fully account for subgroup–reference overlap or uncertainty in the estimated reweighting weights, the findings are exploratory and do not establish clinical mechanisms or treatment effects.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

5 Discussion

5.1 Principal Findings

Algorithms with broadly similar predictive performance produced different feature-attribution profiles. Mean correlations were higher across repeated training trials of the same algorithm ($r = 0.783$) than across different algorithms within the same trial ($r = 0.369$), indicating that model choice contributed more strongly to global attribution variation than changes in the sampled training set. Performance-weighted ensemble SHAP aggregation provided a representative relative-importance profile across the included algorithms, while trial-based confidence intervals and patient-level alignment characterized complementary aspects of attribution variability.

The aggregated profile emphasized oxygenation measures, blood-gas variables, symptoms, renal and metabolic indicators, imaging findings, and treatment-related variables. Distribution-adjusted subgroup comparisons provided an additional, exploratory view of how attribution magnitude varied across patient groups after adjusting each feature's marginal distribution toward the pooled evaluation reference. Of 1,763 eligible subgroup–feature comparisons, 64 had nominal one-sided $p < 0.05$, but none remained below 0.05 after Benjamini–Hochberg adjustment. The subgroup results should therefore be interpreted as candidate patterns for further study rather than evidence of established subgroup-specific effects.

5.2 Clinical Interpretation of the Attribution Profile

Oxygenation and blood-gas measures appeared alongside renal, metabolic, and coagulation indicators, suggesting that the fitted models drew on information from several physiologic domains. This pattern is compatible with the pulmonary and extrapulmonary manifestations described in COVID-19 [6]. The prominence of oxygenation and urea measurements also parallels their inclusion in the validated 4C Mortality Score [10], while the contribution of coagulation variables is consistent with reports linking coagulation abnormalities to adverse outcomes [19]. These comparisons support the clinical plausibility of the attribution profile but do not establish a temporal sequence of organ dysfunction or identify causal pathways.

Shortness of breath and cough received substantial attribution alongside blood-gas and laboratory measurements. Clinical cohort studies have similarly considered presenting symptoms

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

together with physiologic and laboratory findings when characterizing COVID-19 outcomes [25]. Their joint prominence in our analysis suggests that clinical presentation and objective measurements provided complementary predictive information. The analysis did not, however, quantify the independent predictive gain attributable to each feature group.

Treatment-related variables, including dexamethasone use and respiratory support, require contextual interpretation because their attributions may reflect disease severity, treatment selection, timing, and care setting. In the RECOVERY trial, dexamethasone reduced mortality among patients receiving oxygen or invasive mechanical ventilation, whereas no mortality benefit was demonstrated among patients receiving no respiratory support [20]. This evidence emphasizes the importance of treatment context; it does not validate the direction or magnitude of treatment attributions in our cohort. The present findings describe treatment and care patterns captured by the models and do not estimate treatment benefit or harm.

5.3 Subgroup Patterns and Hypotheses

The subgroup comparisons suggest that the relative predictive information contributed by individual features may vary with patient characteristics. Differences involving oxygenation, symptoms, renal measures, and metabolic indicators motivate the hypothesis that baseline health and physiologic reserve influence which measurements are most informative during hospitalization. However, none of the subgroup–feature comparisons remained below 0.05 after false-discovery-rate adjustment, and several clinically defined subgroups were too small for inferential analysis. These patterns therefore require evaluation in larger independent cohorts with longitudinal measurements and prespecified subgroup hypotheses before they can inform clinical monitoring or treatment.

5.4 Relationship to Prior Clinical and Machine Learning Studies

Previous COVID-19 studies used predictive models and feature interpretation to identify informative patient characteristics and examine their contributions to mortality predictions [23, 2, 18]. Our findings similarly emphasize age, oxygenation, symptoms, and laboratory measurements. The present study extends these applications by examining attribution profiles across algorithms and repeated training trials and by comparing subgroup profiles after feature-specific distribution adjustment.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Methodologically, prior work has characterized importance across well-performing models, pooled Shapley-based estimates, and aggregated explanations [4, 16, 1]. Our contribution lies in integrating performance-weighted SHAP aggregation with trial-based importance summaries, patient-level attribution alignment, and distribution-adjusted subgroup comparison in a clinical cohort. This workflow places a representative importance profile alongside information about its variation across fitted models and patient groups, providing interpretive context that is unavailable from a single fitted-model explanation.

5.5 Strengths, Limitations, and Future Directions

Strengths include a prospectively collected cohort from two hospitals in Peru, evaluation of ten predictive algorithms, and comparison of feature attributions across repeated training trials. The analysis distinguishes global importance magnitude from the consistency of signed patient-level attribution patterns. Distribution-adjusted subgroup analysis provides a further descriptive view of heterogeneity in model explanations.

Several limitations concern the clinical data and study setting. The cohort was drawn from two hospitals during the early pandemic, and patients with documented psychiatric conditions were excluded, limiting generalizability to other populations, care settings, and pandemic periods. Clinical variables were collected from enrollment throughout hospitalization until discharge or death. Because the analysis did not impose a common prediction time or account for when each variable became available, the reported performance and attributions characterize information observed during the hospital course rather than a deployable admission-time prediction model. External validation with a prespecified prediction time is needed.

The resampling design also limits interpretation of attribution robustness. Training and evaluation subsets were selected independently, so some evaluation observations may also have contributed to model fitting. This overlap may have increased apparent agreement across trials. In addition, original patient identifiers were unavailable across the three evaluation batches; repeated individuals between batches, if present, could therefore not be identified or accounted for. Cross-validated AUROC weights were estimated from the full study cohort, including observations later used in attribution evaluation. The resulting estimates characterize internal model behavior and should not be interpreted as out-of-sample evidence of explanation transportability.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The uncertainty summaries have additional limitations. The normal-approximation 95% confidence intervals in Table 3 were calculated from 30 trial-specific importance estimates and do not capture every source of model-development uncertainty. Patient-level alignment was assessed using cosine similarity across unique trial pairs and a whole-trial patient-correspondence randomization procedure. This approach preserves trial-level attribution structure but remains an internal measure of alignment and requires confirmation on independently trained models and external patients.

The subgroup analysis adjusts one feature's marginal distribution at a time and does not remove multivariable differences between patient groups. Its nominal weighted Student's t tests compare each reweighted subgroup with a pooled evaluation reference that includes the subgroup itself. They also do not propagate uncertainty from estimating the feature-specific weights, and the one-sided tail was evaluated in the direction of the observed difference. These results are therefore exploratory. The absence of false-discovery-rate-adjusted findings and the small sizes of several subgroups further limit subgroup-specific interpretation.

Finally, model-specific SHAP explainers operated on the implemented prediction functions, and their outputs were not transformed to a common probability or log-odds scale. Normalization made relative attribution magnitudes comparable for aggregation, but the resulting performance-weighted profile is an empirical summary of model-specific attributions rather than the exact SHAP decomposition of a single probabilistic ensemble. SHAP describes fitted-model behavior and does not independently identify causal effects. Future work should use a fixed external evaluation cohort, preserve patient identifiers across trials, define a clinically meaningful prediction time, explain a common mortality-probability output across algorithms, and evaluate whether the aggregated attribution profile generalizes to other clinical populations.

6 Conclusions

This study integrates normalized performance-weighted ensemble SHAP aggregation with repeated training trials and distribution-adjusted subgroup analysis to interpret COVID-19 mortality predictions. Algorithms with similar predictive performance produced different attribution profiles, supporting the use of information from multiple fitted models when constructing a representative summary of feature importance.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

The analysis distinguishes feature-importance magnitude, summarized across 30 trials using means and normal-approximation 95% confidence intervals, from patient-level attribution alignment, assessed using cosine similarity and patient-correspondence randomization. The aggregated profile identified clinically plausible patterns involving oxygenation, symptoms, laboratory measurements, and treatment context, while the subgroup comparisons generated exploratory patterns that did not remain below 0.05 after false-discovery-rate adjustment. These findings describe model-captured relationships rather than causal clinical effects and require validation using independent cohorts, a prespecified prediction time, and evaluation patients excluded from model training.

References

[1] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

[2] Adam L. Booth, Elizabeth Abels, and Peter McCaffrey. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Modern Pathology*, 34(3):522–531, 2021. doi: 10.1038/s41379-020-00700-x.

[3] Paul Ellis and Sylwia Polberg-Riener. Resampling for clinical predictions: Balancing subgroup utility, calibration, and explanation stability. In Hao Ni and Daniele Cafolla, editors, *Artificial Intelligence in Healthcare*, volume 16875 of *Lecture Notes in Computer Science*, pages 3–16. Springer, Cham, 2027. doi: 10.1007/978-3-032-35387-0_1. Published online August 12, 2026.

[4] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL http://jmlr.org/papers/v20/18-760.html.

[5] Pablo Guillén and Enrique Frias-Martinez. Enhancing SHAP explainability for diagnostic and prognostic ML models in alzheimer's disease. *Computers, Materials & Continua*, 87 (2):95, 2026. doi: 10.32604/cmc.2026.076400.

[6] Aakriti Gupta, Mahesh V Madhavan, Kartik Sehgal, Nandini Nair, Shiwani Mahajan, Te-

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

jasav S Sehrawat, Behnood Bikdeli, Neha Ahluwalia, John C Ausiello, Elaine Y Wan, Daniel E Freedberg, Ajay J Kirtane, Sahil A Parikh, Mathew S Maurer, Anna S Nordvig, Domenico Accili, Joan M Bathon, Sumit Mohan, Kenneth A Bauer, Martin B Leon, Harlan M Krumholz, Nir Uriel, Mandeep R Mehra, Mitchell S V Elkind, Gregg W Stone, Allan Schwartz, David D Ho, John P Bilezikian, and Donald W Landry. Extrapulmonary manifestations of COVID-19. *Nat. Med.*, 26(7):1017–1032, July 2020.

[7] E. F. Haghish. shapley: Weighted mean SHAP (WMSHAP) for feature assessment in ML grid and ensemble, 2023. URL https://github.com/haghish/shapley. Computer software.

[8] Alexander A. Huang and Samuel Y. Huang. Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. *PLOS ONE*, 18(2): e0281922, 2023. doi: 10.1371/journal.pone.0281922.

[9] Pål Vegard Johnsen, Inga Strümke, Mette Langaas, Andrew Thomas DeWan, and Signe Riemer-Sørensen. Inferring feature importance with uncertainties with application to large genotype data. *PLOS Computational Biology*, 19(3):e1010963, 2023. doi: 10.1371/journal.pcbi.1010963.

[10] Stephen R. Knight, Antonia Ho, Riinu Pius, et al. Risk stratification of patients admitted to hospital with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ*, 370:m3339, 2020. doi: 10.1136/bmj.m3339.

[11] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5491–5500. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kumar20e.html.

[12] Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023. URL http://jmlr.org/papers/v24/23-0149.html.

[13] Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, Yufeng Gu, Pamela Ventola, and James S. Dun-

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

can. Efficient shapley explanation for features importance estimation under uncertainty. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12261 of *Lecture Notes in Computer Science*, pages 792–801. Springer, 2020. doi: 10.1007/978-3-030-59710-8_77.

[14] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

[15] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018. doi: 10.1038/s41551-018-0304-0.

[16] Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Benjamin Alan Goldstein, Daniel Shu Wei Ting, Roger Vaughan, and Nan Liu. Shapley variable importance cloud for interpretable machine learning. *Patterns*, 3(4):100452, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100452. URL https://www.sciencedirect.com/science/article/pii/S2666389922000253.

[17] Richard D. Riley and Gary S. Collins. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, 65(8):2200302, 2023. doi: 10.1002/bimj.202200302.

[18] Matthew Smith and Francisco Alvarez. Identifying mortality factors from machine learning using Shapley values—a case of COVID-19. *Expert Systems with Applications*, 176:114832, 2021. doi: 10.1016/j.eswa.2021.114832.

[19] Ning Tang, Dengju Li, Xiong Wang, and Ziyong Sun. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. *Journal of Thrombosis and Haemostasis*, 18(4):844–847, 2020. doi: 10.1111/jth.14768.

[20] The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. *New England Journal of Medicine*, 384(8):693–704, 2021. doi: 10.1056/NEJMoa2021436.

[21] Hao Wang, Nethra Sambamoorthi, Nathan Hoot, David Bryant, and Usha Sambamoorthi. Evaluating fairness of machine learning prediction of prolonged wait times in emer-

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

gency department with interpretable extreme gradient boosting. *PLOS Digital Health*, 4 (3):e0000751, 2025. doi: 10.1371/journal.pdig.0000751.

[22] Lanxin Xiang, Liang Shi, Youhui Ye, Boyu Jiang, Dawei Zhou, and Feng Guo. Roshap: A distributional framework and robust metric for stable feature attribution, 2026. URL https://arxiv.org/abs/2605.15154.

[23] Arjun S Yadaw, Yan-Chak Li, Sonali Bose, Ravi Iyengar, Supinda Bunyavanich, and Gaurav Pandey. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit. Health*, 2(10):e516–e525, October 2020.

[24] Angeline Yasodhara, Azin Asgarian, Diego Huang, and Parinaz Sobhani. On the trustworthiness of tree ensemble explainability methods. In *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings*, page 293–308, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-84059-4. doi: 10.1007/978-3-030-84060-0_19. URL https://doi.org/10.1007/978-3-030-84060-0_19.

[25] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in wuhan, china: a retrospective cohort study. *Lancet*, 395 (10229):1054–1062, March 2020.

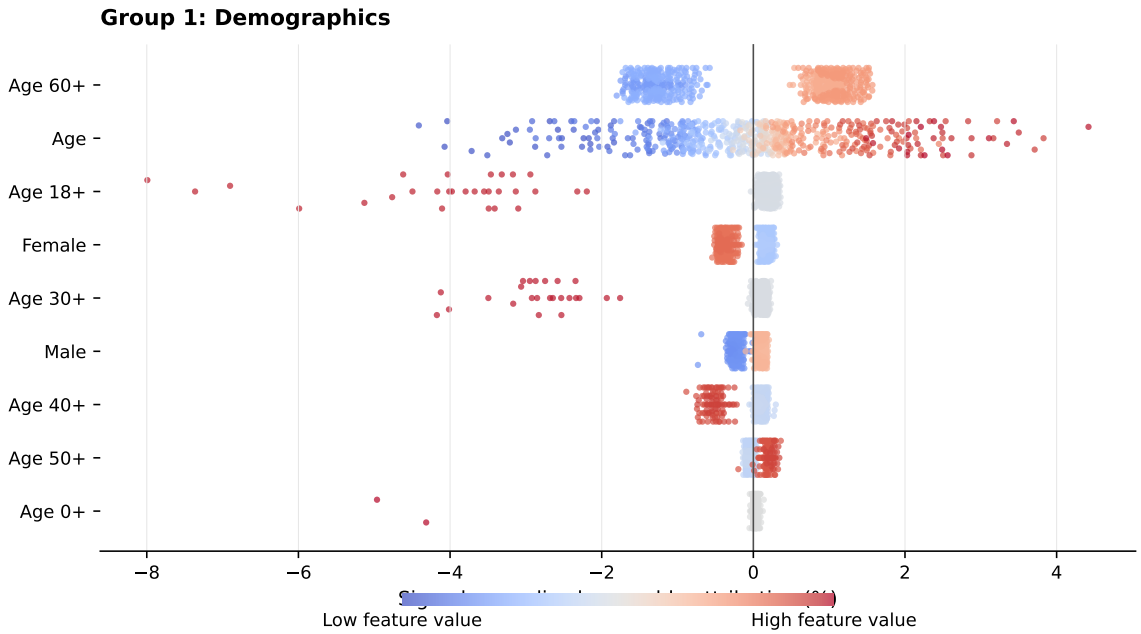
medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Information

Supplementary Figures

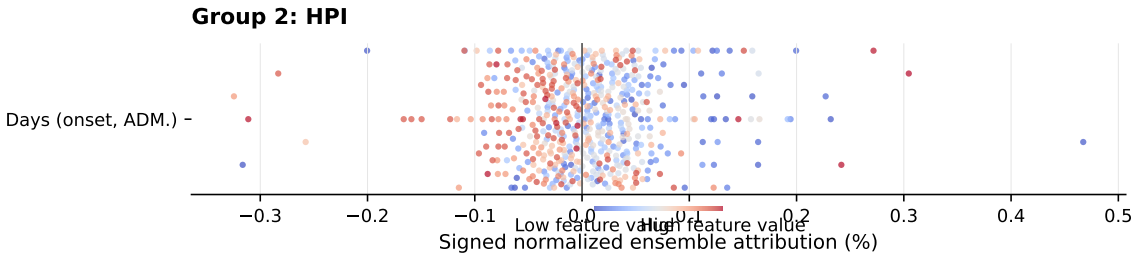
Supplementary Figures 1-16 display signed normalized ensemble SHAP attributions for the most important features within each clinical category. Each point represents one pooled evaluation observation after performance-weighted normalized attributions were averaged across the ten trials in its fixed evaluation batch. Positive values increase the normalized ensemble output relative to its baseline, whereas negative values decrease it. Color denotes the within-batch percentile rank of the observed feature value.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



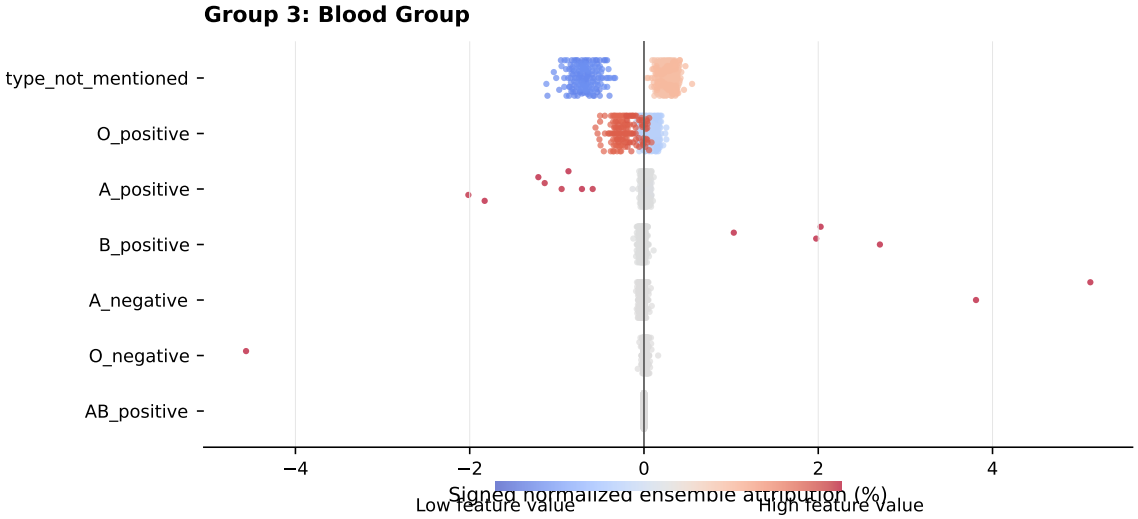
Supplementary Figure 1: Signed normalized ensemble SHAP attributions for Group 1 (Demographics). The plot shows the 9 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



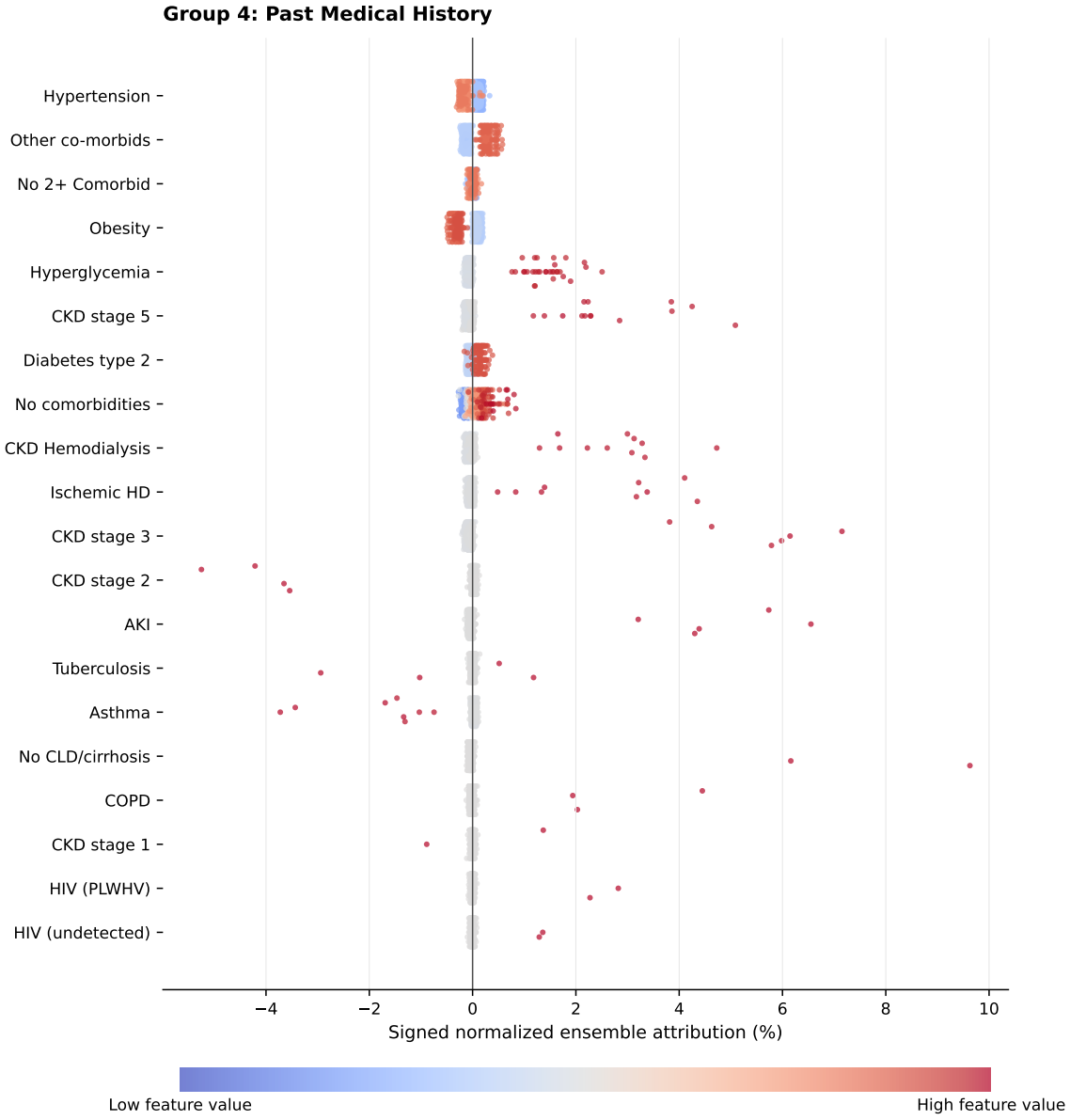
Supplementary Figure 2: Signed normalized ensemble SHAP attributions for Group 2 (History of Present Illness (HPI)). The plot shows the 1 feature with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



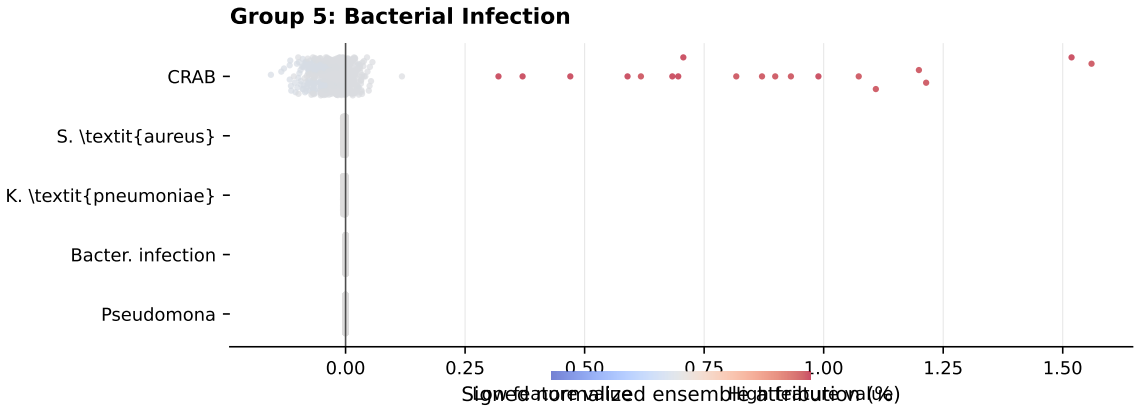
Supplementary Figure 3: Signed normalized ensemble SHAP attributions for Group 3 (Blood Group). The plot shows the 7 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



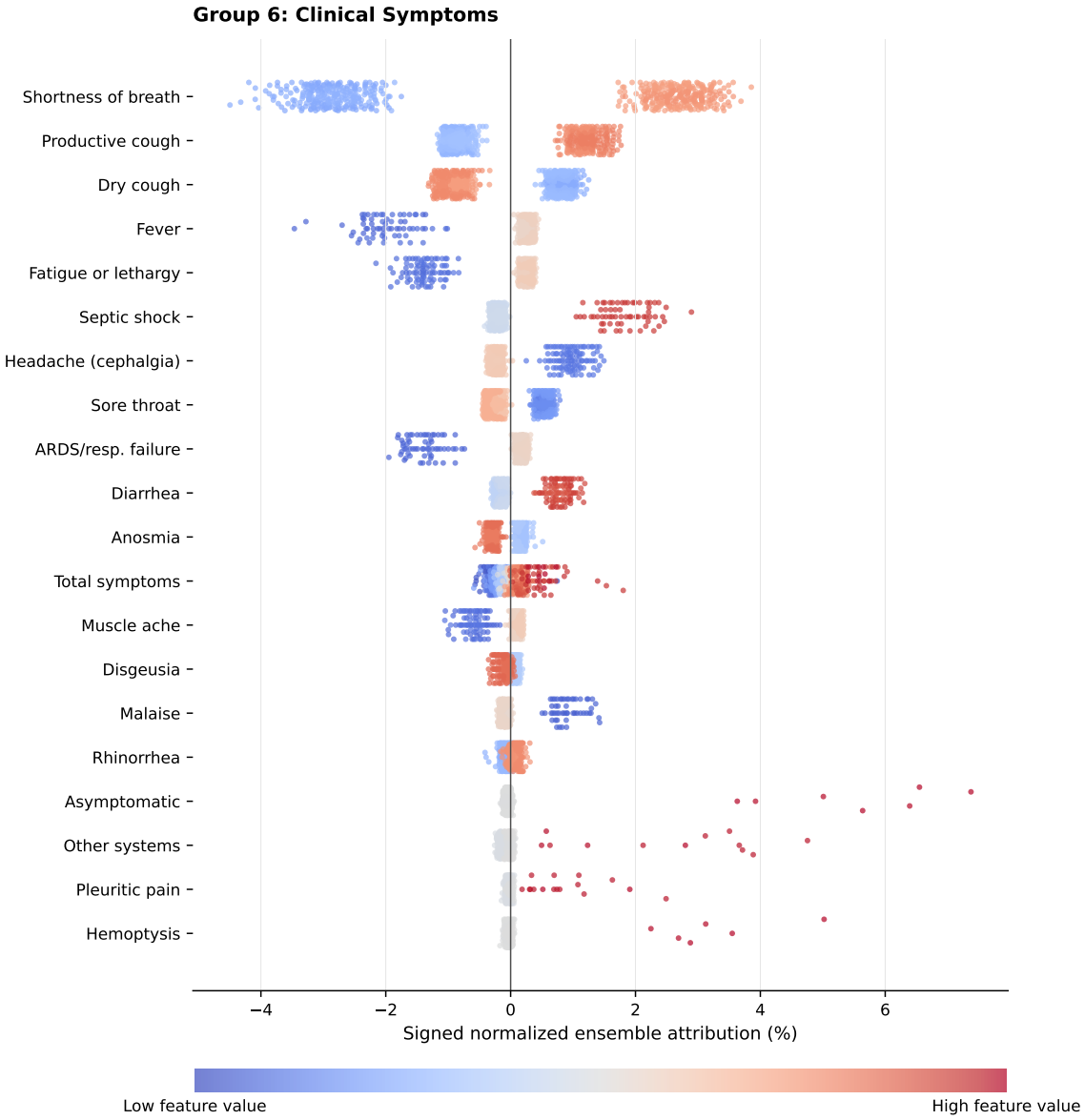
Supplementary Figure 4: Signed normalized ensemble SHAP attributions for Group 4 (Past Medical History). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



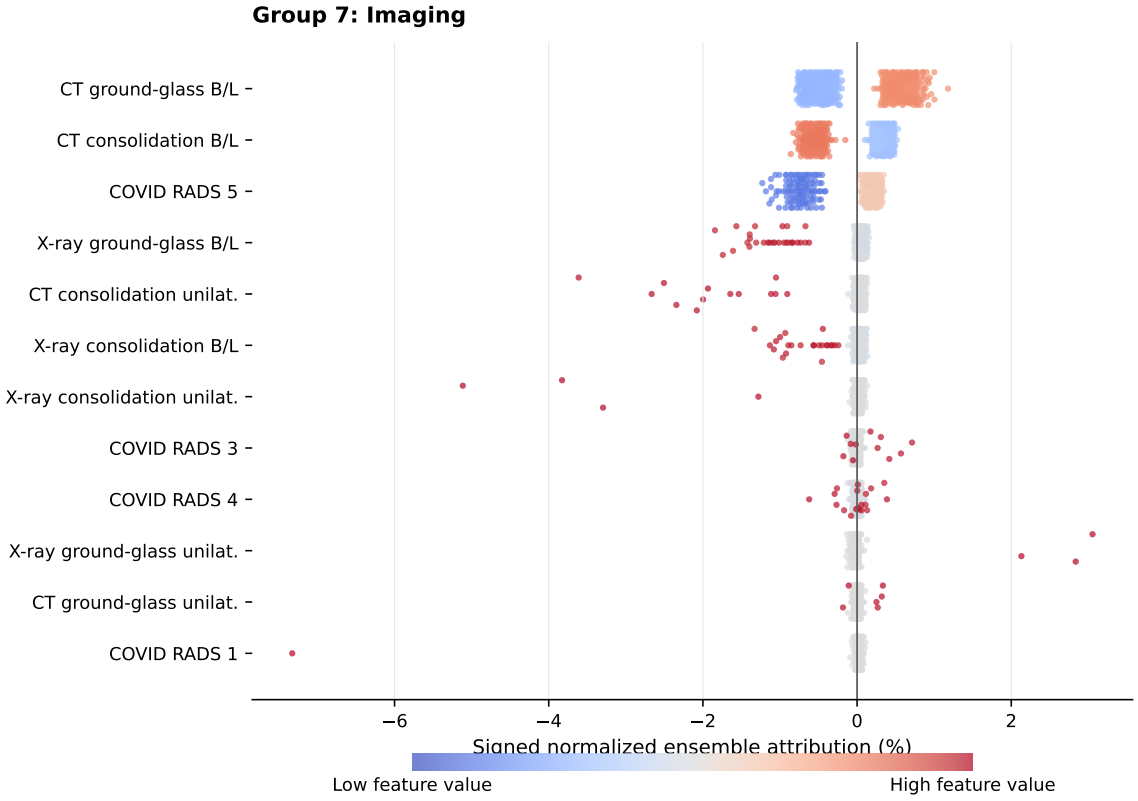
Supplementary Figure 5: Signed normalized ensemble SHAP attributions for Group 5 (Bacterial Infection). The plot shows the 5 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



Supplementary Figure 6: Signed normalized ensemble SHAP attributions for Group 6 (Clinical Symptoms). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



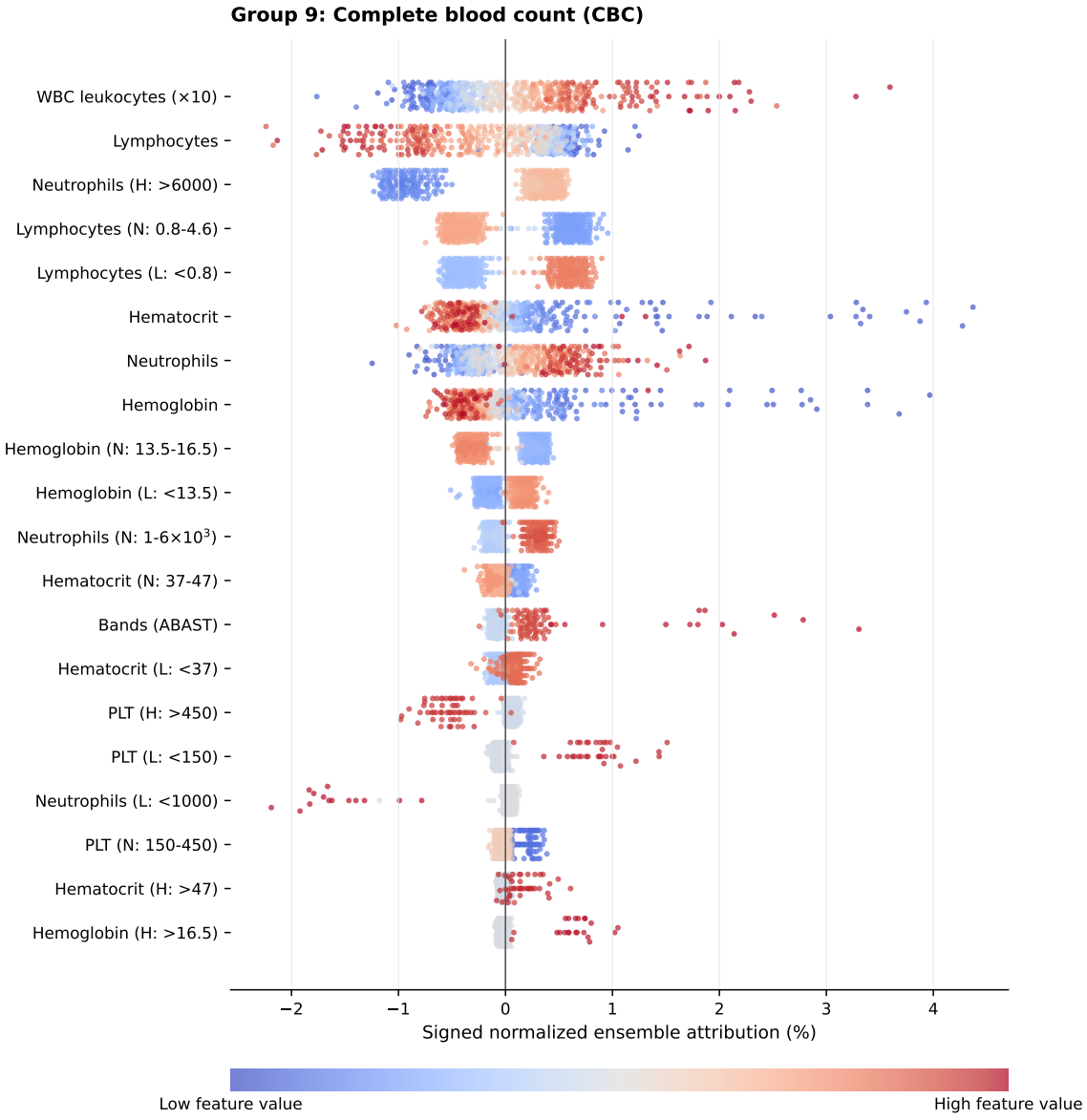
Supplementary Figure 7: Signed normalized ensemble SHAP attributions for Group 7 (Imaging). The plot shows the 12 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



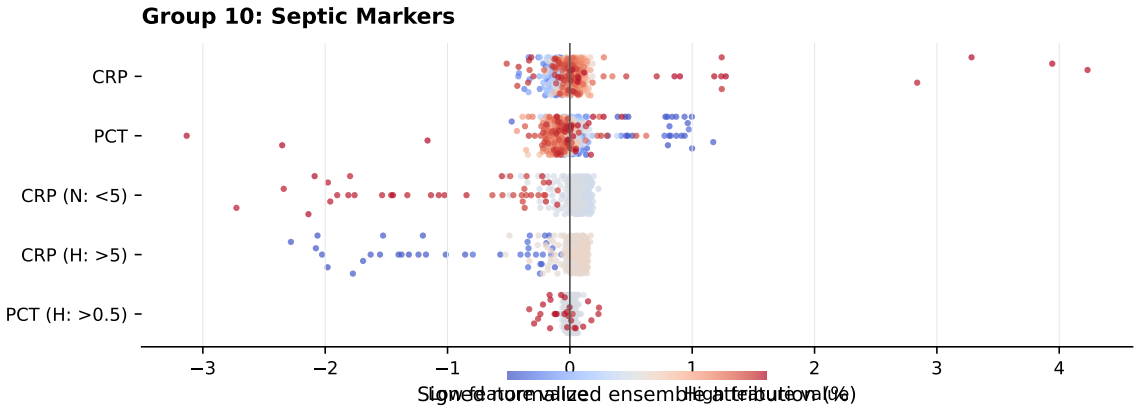
Supplementary Figure 8: Signed normalized ensemble SHAP attributions for Group 8 (Liver Function Tests). The plot shows the 16 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



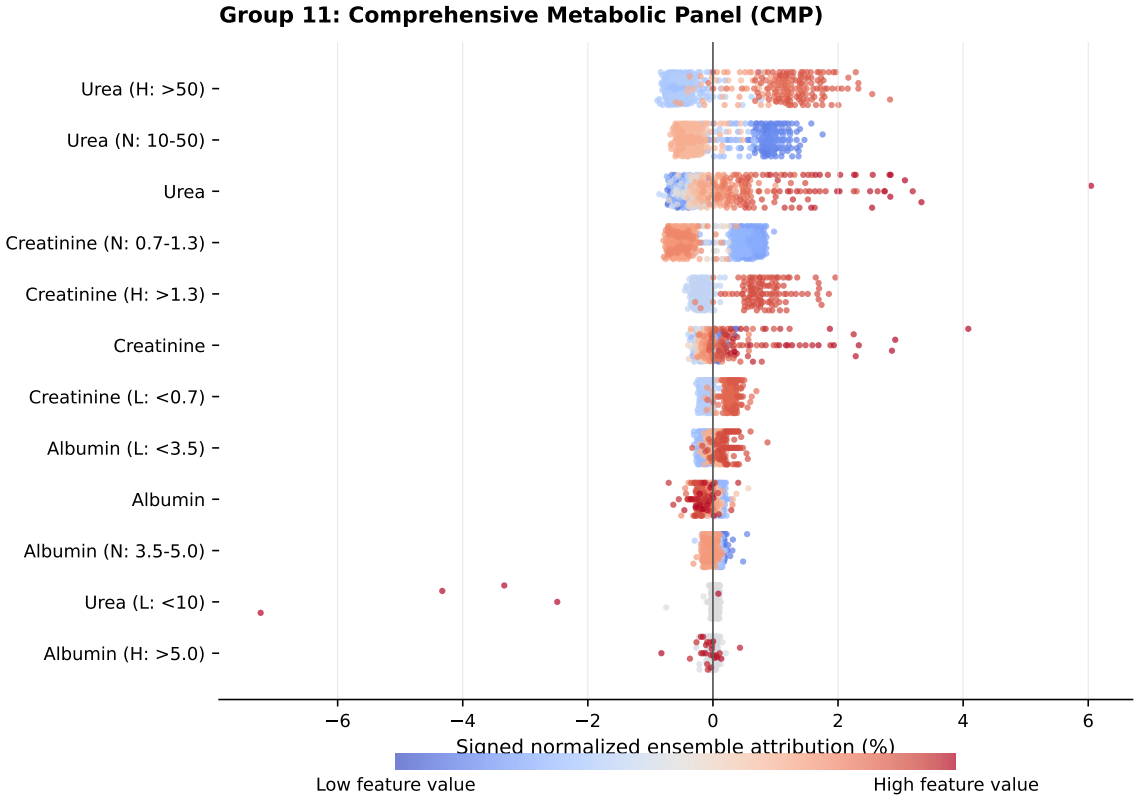
Supplementary Figure 9: Signed normalized ensemble SHAP attributions for Group 9 (Complete Blood Count (CBC)). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



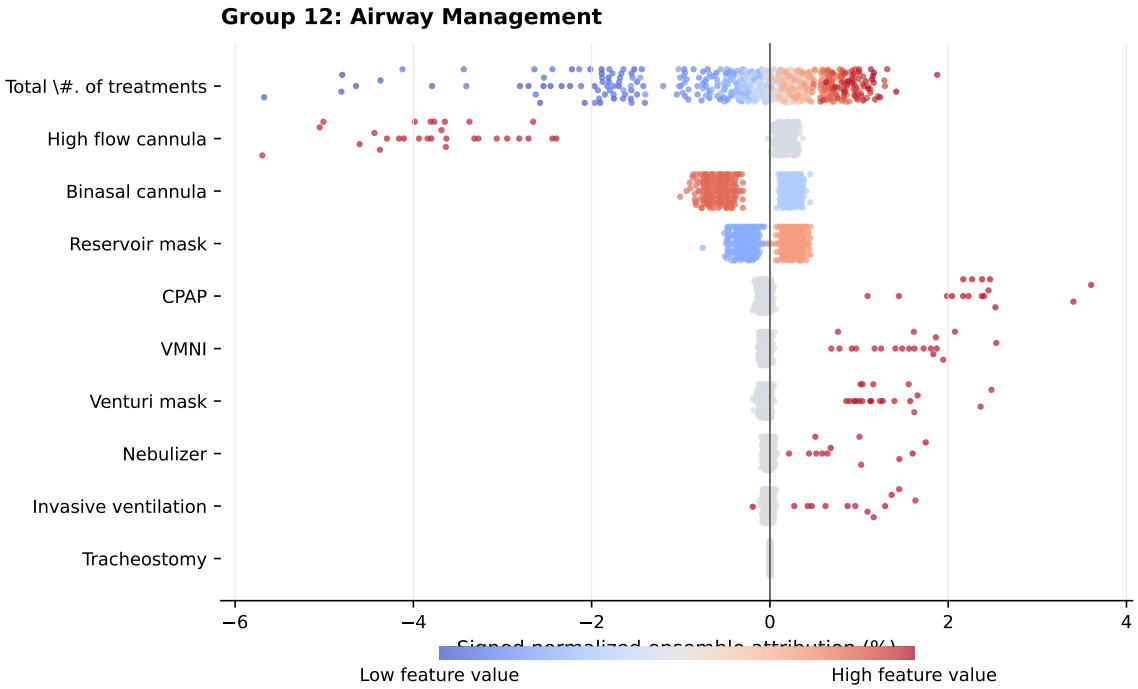
Supplementary Figure 10: Signed normalized ensemble SHAP attributions for Group 10 (Septic Markers). The plot shows the 5 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



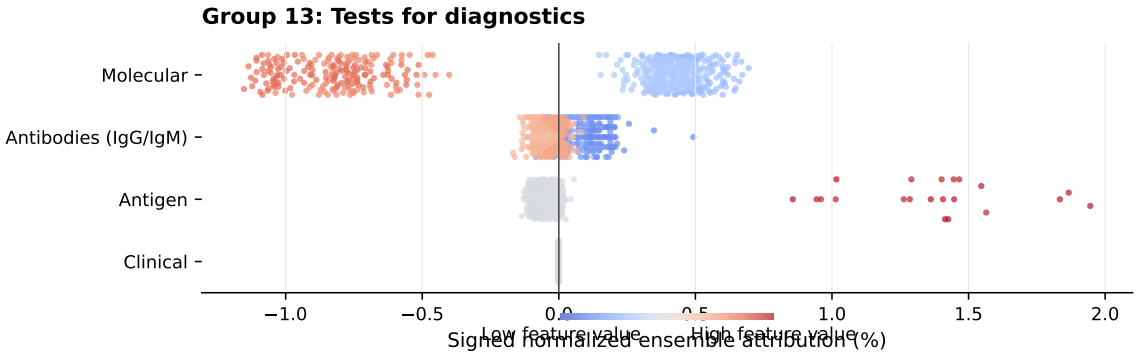
Supplementary Figure 11: Signed normalized ensemble SHAP attributions for Group 11 (Comprehensive Metabolic Panel (CMP)). The plot shows the 12 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



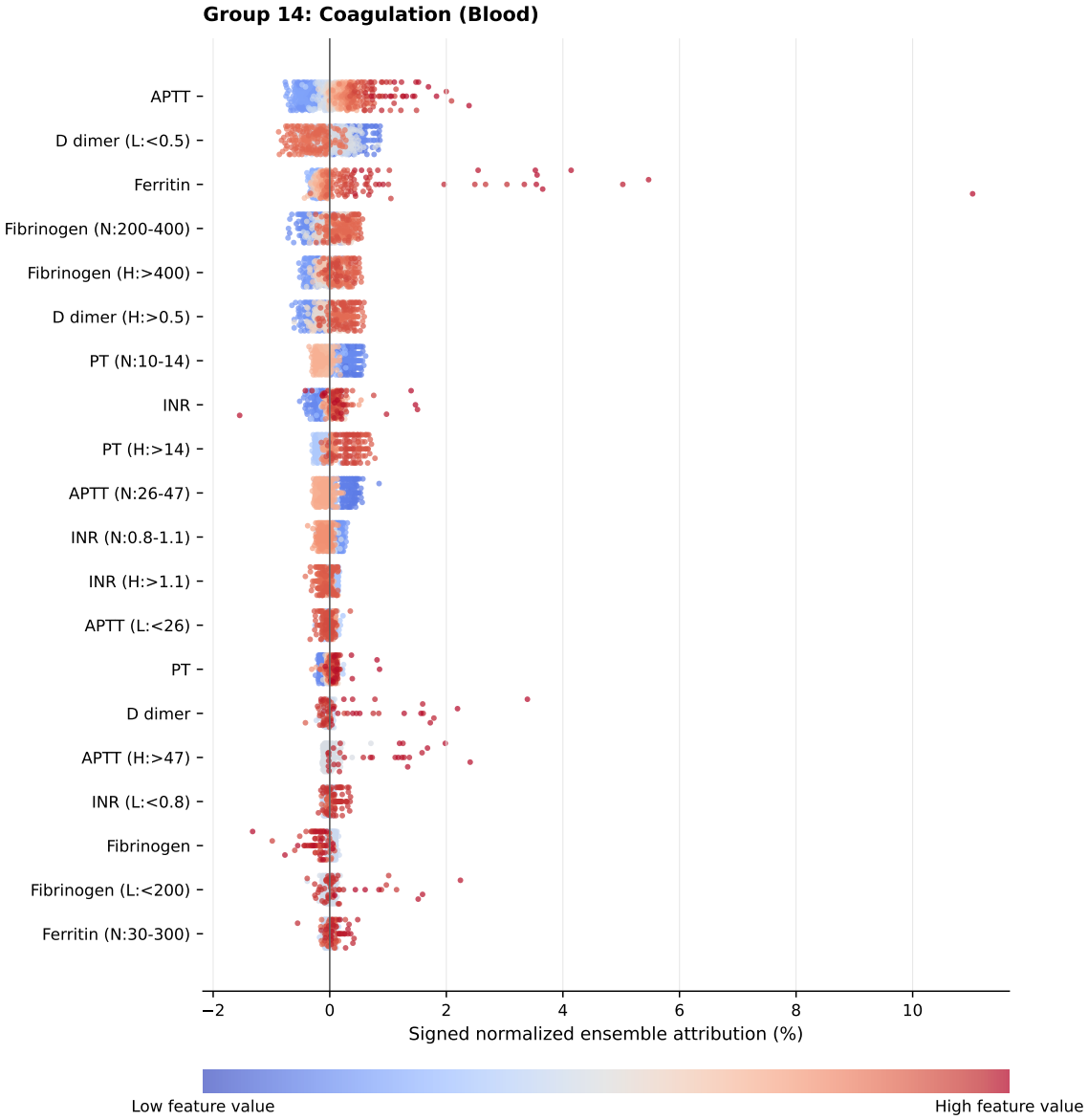
Supplementary Figure 12: Signed normalized ensemble SHAP attributions for Group 12 (Airway Management). The plot shows the 10 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



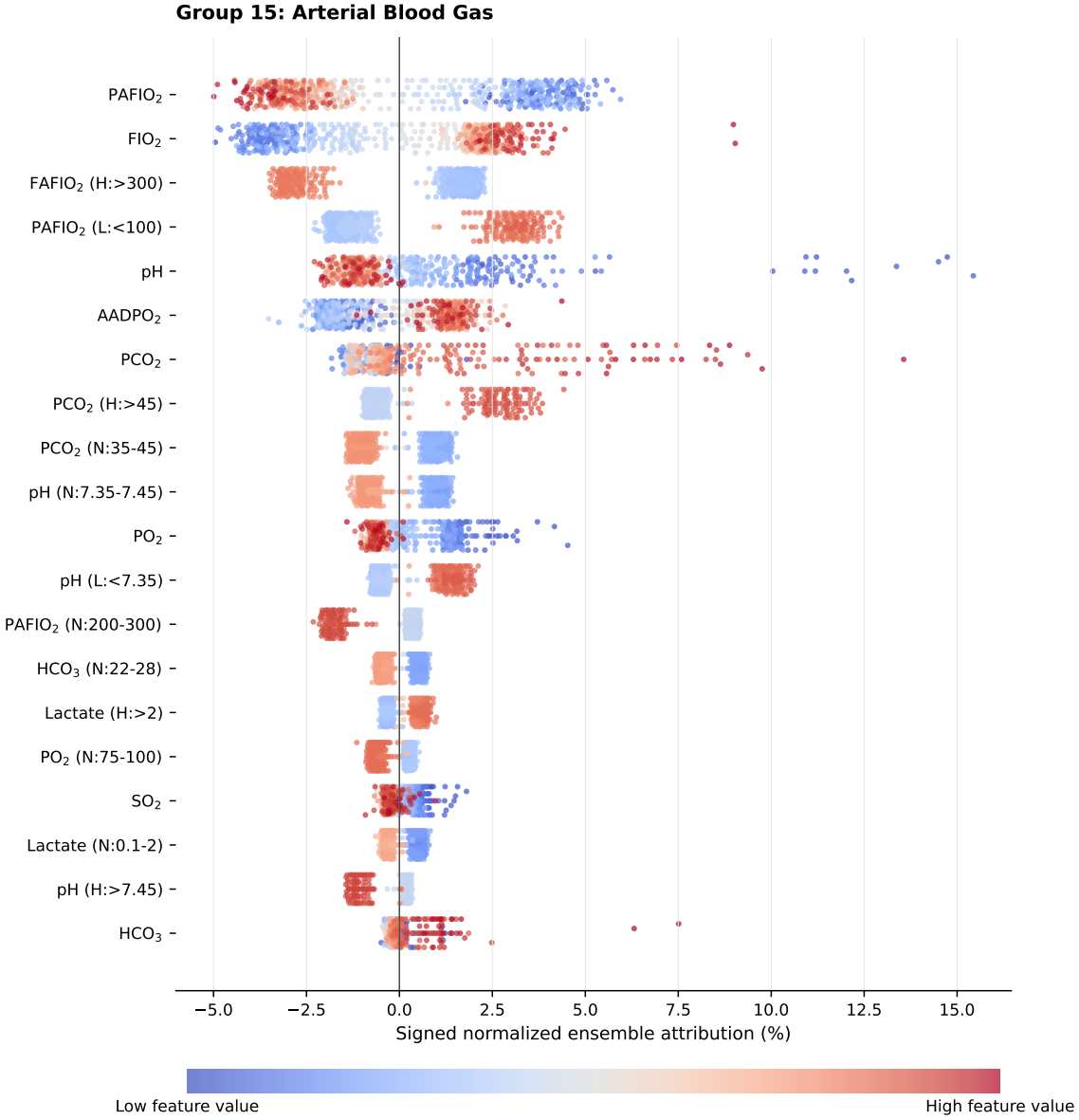
Supplementary Figure 13: Signed normalized ensemble SHAP attributions for Group 13 (Diagnostic Tests). The plot shows the 4 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



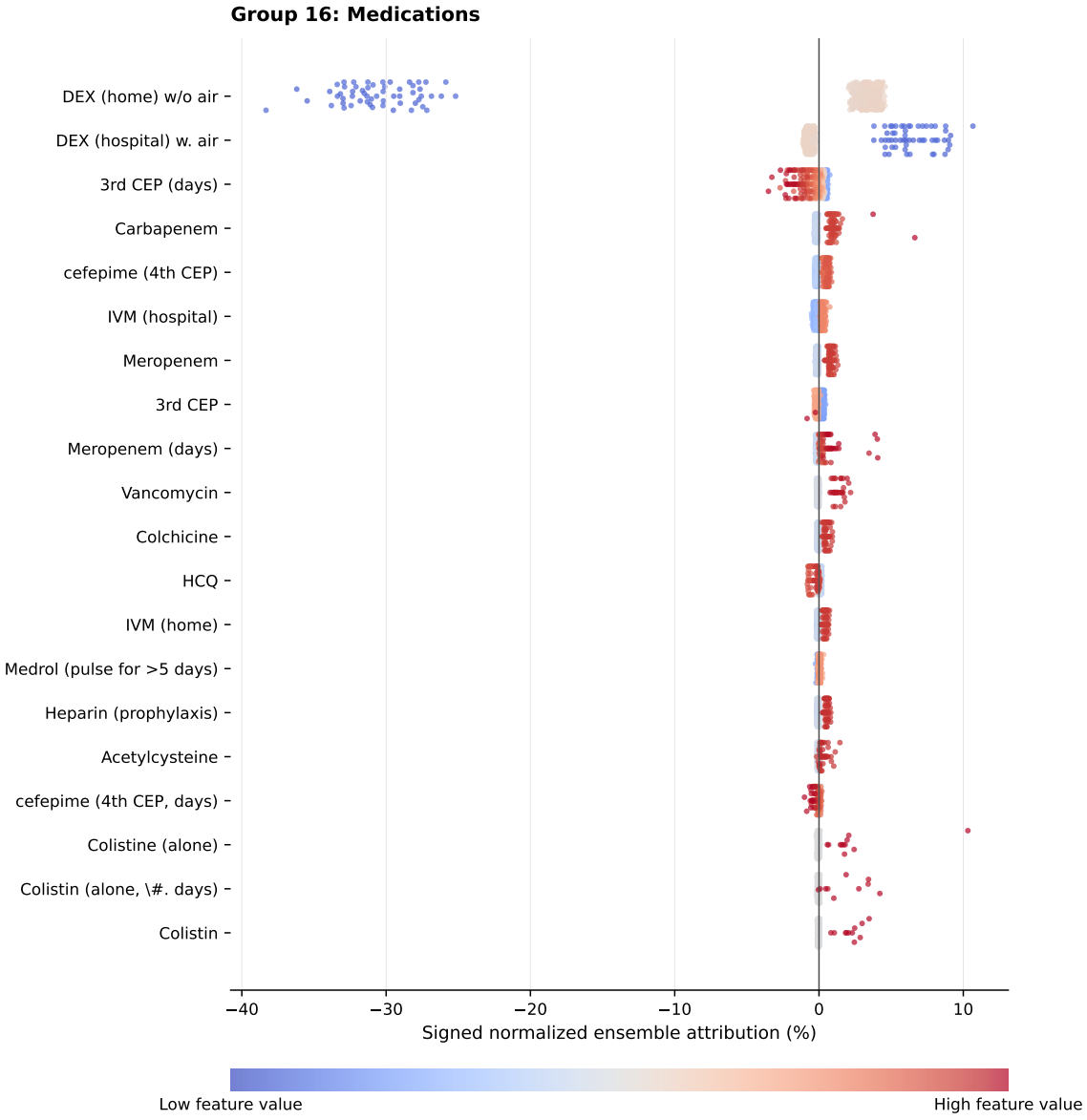
Supplementary Figure 14: Signed normalized ensemble SHAP attributions for Group 14 (Coagulation). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



Supplementary Figure 15: Signed normalized ensemble SHAP attributions for Group 15 (Arterial Blood Gas). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



Supplementary Figure 16: Signed normalized ensemble SHAP attributions for Group 16 (Medications). The plot shows the 20 features with the greatest mean absolute importance in this category. Each point represents one evaluation observation after performance-weighted normalized attributions were averaged across the ten training trials in its fixed evaluation batch. Points to the right of zero increase the normalized ensemble output relative to its baseline; points to the left decrease it. Blue and red indicate lower and higher observed feature values, respectively. Numerical importance and alignment summaries are reported in main-manuscript Table 3.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Tables

Condensed Subgroup Results

The complete results for all 3,113 subgroup–feature comparisons, including both Benjamini–Hochberg adjustments, are provided as Supplementary Data S1 in machine-readable CSV format. Tables 1 and 2 summarize inferential eligibility and the 64 comparisons with nominal one-sided $p < 0.05$, respectively. No comparison remained below 0.05 after adjustment within subgroup or across all eligible comparisons.

Supplementary Table 1: Summary of inferential eligibility and multiplicity adjustment by subgroup.

Subgroup	n_G	Total	Eligible	Small n	Zero var.	KS > 0.10	Defining	Nominal $p < 0.05$	Min. p	Min. q_{within}	Min. q_{global}
Age ≥ 50 years	109	283	257	0	19	4	3	8	0.017	0.499	0.499
Male	374	283	262	0	19	2	0	0	0.059	0.499	0.499
Female	177	283	262	0	19	2	0	12	0.010	0.498	0.498
Blood type A	10	283	0	283	0	0	0	0	–	–	
Blood type B	4	283	0	283	0	0	0	0	–	–	
Blood type O	143	283	257	0	19	2	5	7	0.004	0.496	0.497
Asthma or COPD	11	283	0	283	0	0	0	0	–	–	
Metabolic or cardiovascular disease	303	283	258	0	19	5	1	4	0.023	0.497	0.499
Kidney disease	30	283	209	0	19	53	2	18	< 0.001	0.153	0.499
Chronic viral or liver disease	4	283	0	283	0	0	0	0	–	–	
At least two documented comorbidities	214	283	258	0	19	6	0	15	0.002	0.467	0.499
Total	–	3,113	1,763	1,132	133	74	11	64	< 0.001	0.153	0.499

Notes. Total is the number of encoded features evaluated descriptively in each subgroup. Eligible denotes comparisons included in inferential testing. Small n denotes subgroups containing fewer than 20 evaluation observations; Zero var. denotes zero patient-level attribution variance; KS > 0.10 denotes inadequate post-reweighting balance; Defining denotes features used to define the subgroup. Min. q_{within} is the smallest Benjamini–Hochberg-adjusted value within the subgroup, and Min. q_{global} is the smallest adjusted value under correction across all 1,763 eligible comparisons.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC BY-NC-ND 4.0 International license .

Supplementary Table 2: Subgroup-feature comparisons with nominal one-sided $p < 0.05$.

Subgroup	Feature	Ref.	Adj.	Δ	Δ%	Nominal p	q_{within}	q_{global}
Age ≥ 50 years	ALT (Norm: 0-42)	0.200	0.215	-0.015	-7.3%	0.017	0.499	0.499
Age ≥ 50 years	Invasive ventilation	0.085	0.128	-0.042	-49.9%	0.022	0.499	0.499
Age ≥ 50 years	A_negative	0.061	0.161	-0.100	-163.0%	0.030	0.499	0.499
Age ≥ 50 years	X-ray ground-glass B/L	0.134	0.183	-0.050	-37.4%	0.033	0.499	0.499
Age ≥ 50 years	Vancomycin	0.172	0.234	-0.062	-36.4%	0.038	0.499	0.499
Age ≥ 50 years	Tigecycline	0.100	0.180	-0.080	-79.4%	0.044	0.499	0.499
Age ≥ 50 years	Hemoglobin (L: <13.5)	0.210	0.220	-0.010	-5.0%	0.044	0.499	0.499
Age ≥ 50 years	Tocilizumab (2 doses)	0.099	0.163	-0.064	-64.2%	0.044	0.499	0.499
Female	Meropenem	0.239	0.192	0.047	19.7%	0.010	0.498	0.499
Female	Acetylcysteine	0.128	0.162	-0.034	-26.8%	0.017	0.498	0.499
Female	Ferritin (N:30-300)	0.111	0.129	-0.018	-16.3%	0.025	0.498	0.499
Female	Obesity	0.181	0.201	-0.020	-11.1%	0.028	0.498	0.499
Female	Ertapenem	0.000	0.000	0.000	29.3%	0.032	0.498	0.499
Female	CKD stage 5	0.160	0.239	-0.079	-49.1%	0.036	0.498	0.499
Female	PT (L:<10)	0.090	0.075	0.015	16.7%	0.037	0.498	0.499
Female	Hematocrit (H: >47)	0.083	0.069	0.013	16.1%	0.037	0.498	0.499
Female	Carbapenem	0.332	0.274	0.058	17.4%	0.040	0.498	0.499
Female	INR	0.212	0.175	0.037	17.4%	0.044	0.498	0.499
Female	Hyperglycemia	0.169	0.126	0.043	25.6%	0.048	0.498	0.499
Female	HCQ	0.158	0.137	0.021	13.3%	0.050	0.498	0.499
Blood type O	Age 30+	0.245	0.403	-0.158	-64.4%	0.004	0.496	0.499
Blood type O	High flow cannula	0.383	0.233	0.150	39.2%	0.016	0.496	0.499
Blood type O	Age 18+	0.412	0.589	-0.177	-42.8%	0.027	0.496	0.499

Continued on next page

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 2 – continued

Subgroup	Feature	Ref.	Adj.	Δ	Δ%	Nominal p	q_{within}	q_{global}
Blood type O	Albumin (H: >5.0)	0.075	0.090	-0.015	-19.5%	0.031	0.496	0.499
Blood type O	Total symptoms 2+	0.081	0.189	-0.108	-133.6%	0.034	0.496	0.499
Blood type O	Medrol (pulse for >5 days)	0.152	0.142	0.010	6.9%	0.035	0.496	0.499
Blood type O	Malaise	0.187	0.231	-0.044	-23.3%	0.040	0.496	0.499
Metabolic or cardiovascular disease	Pregnancy in 3rd Trim	0.000	0.000	0.000	12.1%	0.023	0.497	0.499
Metabolic or cardiovascular disease	Other co-morbids	0.206	0.220	-0.015	-7.1%	0.036	0.497	0.499
Metabolic or cardiovascular disease	HCO_3 (H:>28)	0.286	0.252	0.035	12.2%	0.043	0.497	0.499
Metabolic or cardiovascular disease	Hyperglycemia	0.169	0.211	-0.043	-25.3%	0.045	0.497	0.499
Kidney disease	COPD	0.070	0.248	-0.178	-254.2%	< 0.001	0.153	0.499
Kidney disease	X-ray consolidation unilat.	0.082	0.295	-0.213	-261.0%	0.001	0.153	0.499
Kidney disease	B_positive	0.064	0.216	-0.152	-239.6%	0.008	0.378	0.499
Kidney disease	$PaO_2/FiO_2 < 100$	1.980	1.595	0.385	19.4%	0.008	0.378	0.499
Kidney disease	Ferritin (L:<30)	0.091	0.199	-0.108	-118.5%	0.013	0.378	0.499
Kidney disease	AB_positive	0.000	0.000	-0.000	-43.8%	0.013	0.378	0.499
Kidney disease	Dry cough	0.868	0.793	0.075	8.6%	0.015	0.378	0.499
Kidney disease	Nonalcoholic fatty liver disease	0.000	0.000	-0.000	-89.2%	0.015	0.378	0.499
Kidney disease	Neutrophils (L: <1000)	0.107	0.212	-0.106	-99.3%	0.016	0.378	0.499
Kidney disease	Invasive ventilation	0.085	0.159	-0.073	-86.2%	0.020	0.424	0.499
Kidney disease	$AADPO_2$	1.529	1.333	0.196	12.8%	0.024	0.441	0.499
Kidney disease	PO_2 (H:>100)	0.092	0.132	-0.040	-43.9%	0.026	0.441	0.499

Continued on next page

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 2 – continued

Subgroup	Feature	Ref.	Adj.	Δ	Δ%	Nominal p	q_{within}	q_{global}
Kidney disease	A_positive	0.078	0.151	-0.072	-92.3%	0.027	0.441	0.499
Kidney disease	PCO_2 (N:35-45)	1.002	0.926	0.076	7.6%	0.034	0.483	0.499
Kidney disease	pH (N:7.35-7.45)	0.914	0.844	0.070	7.6%	0.039	0.483	0.499
Kidney disease	Fever	0.472	0.666	-0.194	-41.2%	0.041	0.483	0.499
Kidney disease	Total symptoms 2+	0.081	0.259	-0.178	-219.9%	0.044	0.483	0.499
Kidney disease	PT	0.144	0.203	-0.059	-40.6%	0.044	0.483	0.499
At least two documented comorbidities	Hyperglycemia	0.169	0.254	-0.085	-50.6%	0.002	0.467	0.499
At least two documented comorbidities	Heparin (prophylaxis)	0.139	0.170	-0.031	-22.2%	0.009	0.495	0.499
At least two documented comorbidities	Fatigue or lethargy	0.439	0.367	0.072	16.4%	0.013	0.495	0.499
At least two documented comorbidities	Septic shock	0.416	0.511	-0.095	-22.8%	0.017	0.495	0.499
At least two documented comorbidities	INR (L:<0.8)	0.118	0.103	0.015	12.5%	0.018	0.495	0.499
At least two documented comorbidities	CKD Hemodialysis	0.134	0.212	-0.078	-58.5%	0.018	0.495	0.499
At least two documented comorbidities	HCO_3 (H:>28)	0.286	0.240	0.047	16.2%	0.021	0.495	0.499
At least two documented comorbidities	Other systems	0.149	0.229	-0.081	-54.4%	0.028	0.495	0.499
At least two documented comorbidities	Age 60+	1.148	1.110	0.038	3.3%	0.038	0.495	0.499

Continued on next page

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 2 – continued

Subgroup	Feature	Ref.	Adj.	Δ	Δ%	Nominal p	q_{within}	q_{global}
At least two documented comorbidities	Age 40+	0.203	0.179	0.024	11.6%	0.042	0.495	0.499
At least two documented comorbidities	Caspofungin (days)	0.000	0.000	-0.000	-58.4%	0.044	0.495	0.499
At least two documented comorbidities	Dry cough	0.868	0.843	0.025	2.9%	0.044	0.495	0.499
At least two documented comorbidities	ALP	0.217	0.203	0.014	6.5%	0.045	0.495	0.499
At least two documented comorbidities	CT consolidation unilat.	0.117	0.164	-0.047	-40.3%	0.047	0.495	0.499
At least two documented comorbidities	CKD stage 3	0.125	0.218	-0.093	-74.4%	0.048	0.495	0.499

Notes. Ref., pooled-reference importance; Adj., distribution-adjusted subgroup importance. Δ is pooled-reference importance minus adjusted-subgroup importance. Positive values indicate greater attribution magnitude in the pooled evaluation reference, whereas negative values indicate greater magnitude in the adjusted subgroup. Nominal p is the one-sided tail probability evaluated in the direction of the observed difference. q_{within} applies the Benjamini–Hochberg procedure among eligible features within each subgroup; q_{global} applies it across all 1,763 eligible subgroup–feature comparisons. None of the displayed comparisons had $q < 0.05$ under either adjustment. Complete results, including the 1,350 rows excluded from inference and their exclusion reasons, are supplied in Supplementary Data S1.

medRxiv preprint doi: https://doi.org/10.64898/2026.09.07.26362459; this version posted September 8, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .