Intensive Care Med (2026) 52:1472–1483
https://doi.org/10.1007/s00134-026-08441-2

ORIGINAL

Reproducible clinical archetypes in acute respiratory failure: a multi-cohort trajectory analysis

Dominic C. Marshall[1,2*], Ashleigh D. Green[1], Matthieu Komorowski[3], Brijesh V. Patel[1], Sonali Parbhoo[4] and David B. Antcliffe[1]

© 2026 The Author(s)

Abstract

Purpose: Acute hypoxemic respiratory failure (AHRF) is common and biologically heterogeneous. Static phenotyping at a single time point does not capture illness evolution and risks stage-mixing; reproducible clinical course archetypes may address this. We aimed to identify, externally validate, and predict trajectory classes (TCs) of persistent AHRF.

Methods: We analyzed MIMIC-IV (derivation; $n = 3938$) and two external validation cohorts (UK/Netherlands; $n = 6480$) comprising adults with $PaO_2/FiO_2 < 300$ mmHg and PEEP ≥ 5 cmH_2O for ≥ 72 h. Daily mean PaO_2/FiO_2 to day 14 and time to ICU discharge/death were jointly modelled using a competing-risk latent class mixed model. Early TC prediction used a 12-variable XGBoost model. We explored prevalence of ARDS and hyperinflammatory subphenotypes between TCs.

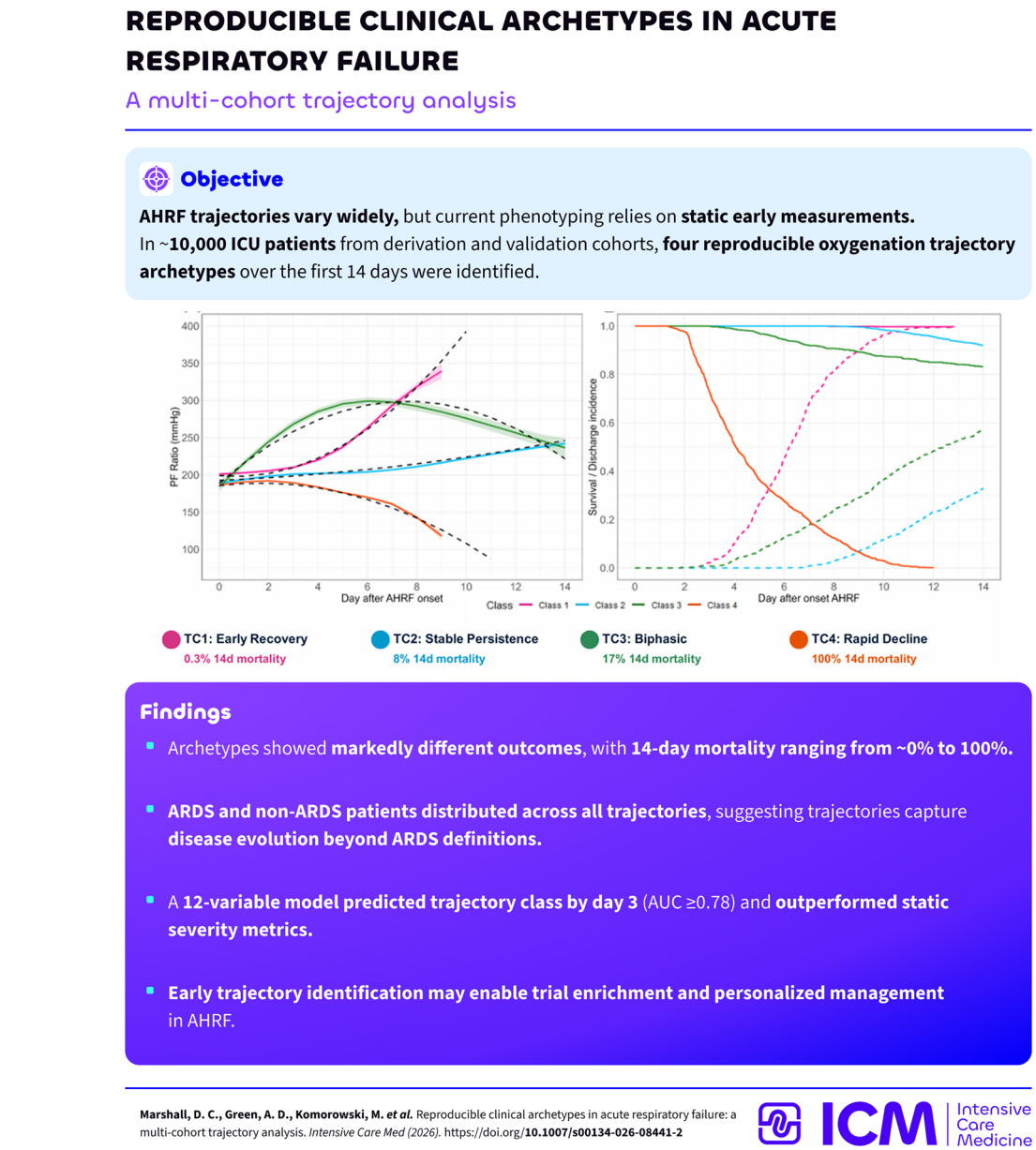
Results: A four-class model provided optimal fit: (TC1) early recovery (0.3% 14-day mortality); (TC2) stable persistence (8% 14-day mortality); (TC3) biphasic improvement–deterioration (17% 14-day mortality); and (TC4) rapid decline (100% 14-day mortality). These archetypes generalized to external cohorts with high assignment certainty. TCs demonstrated distinct patterns in other clinical biomarker trajectories. TC4 was enriched for the hyperinflammatory subphenotype (41–53%), while TC2 was most common in patients with ARDS (50%). Early TC prediction models achieved mean AUCs ≥ 0.78 (0.70–0.86) by day 3 in external validation.

Conclusions: Four reproducible oxygenation archetypes capture the 14-day course of persistent respiratory failure. By providing early prognostic value distinct from static baseline severity, these trajectories have the potential to guide therapeutic strategies, reduce patient heterogeneity in trials, and direct biological phenotyping.

*Correspondence: dominic.marshall12@imperial.ac.uk
[1] Division of Anaesthetics, Pain Medicine and Intensive Care, Department of Surgery and Cancer, Imperial College London, London, UK
Full author information is available at the end of the article

Springer

Graphic abstract:



Keywords: Acute hypoxemic respiratory failure, Trajectories, Subphenotyping

Introduction

Acute hypoxemic respiratory failure (AHRF) is a dynamic clinical syndrome, yet our current precision medicine frameworks remain largely static. Substantial progress has been made in identifying biological subphenotypes in Acute Respiratory Distress Syndrome (ARDS), most notably the reproducible "hyper-" and "hypo-inflammatory" classes [1–5]. Yet these frameworks typically classify patients at a single timepoint and are anchored to ARDS-specific samples. Recent work has begun to explore the temporal stability of these traits [6], but cross-sectional phenotyping does not capture how illness evolves over days. As hypoxemia persists, the initial insult may resolve or be compounded by ICU-acquired inflammatory and infectious "second hits," and by evolving organ dysfunction

during persistent critical illness [7, 8]. Static phenotyping, therefore, risks stage-mixing, in which patients with similar early physiology follow divergent biological and clinical paths. A robust method that captures longitudinal disease course is needed to support deep phenotyping and trial enrichment.

Physiological trajectories offer a pragmatic approach. The PaO_2/FiO_2 ratio (P/F), while imperfect [9], remains the most widely operationalized severity metric for hypoxemic respiratory failure. Prior oxygenation-trajectory studies have often focused on short observation windows, typically the first 3 to 4 days, and have been largely confined to ARDS cohorts or dichotomies, such as persistent versus resolving hypoxemia [10–12]. Few studies explicitly account for the competing outcomes (ICU discharge or death) that truncate longitudinal oxygenation follow-up. If they are ignored, survivor bias may obscure rapidly recovering trajectories and underestimate the severity of rapidly deteriorating courses. Whether oxygenation trajectories represent transportable clinical archetypes (distinct, reproducible patterns of longitudinal disease evolution) that generalize across healthcare systems and eras, or instead reflect local practice variation, remains uncertain.

Establishing reproducible trajectory archetypes would not only be of value for prognosis but also serve as a tool to proactively guide clinical management, such as optimizing the timing of ventilator weaning, anticipating secondary complications, or escalating to salvage therapies. Beyond the bedside, this dynamic approach may provide a vital scaffold for translational research. Deep biological phenotyping, particularly via intensive approaches, such as bronchoscopy, is costly and typically restricted to small cohorts [13]. Using longitudinal trajectory archetypes to guide and anchor biospecimen collection, researchers could precisely target their invasive sampling to specific clinical courses, thereby reducing heterogeneity and increasing the biological contrast.

We hypothesised that AHRF comprises a finite set of clinically meaningful oxygenation trajectory archetypes that differ in oxygenation course and outcome. (14) Our primary aim was to identify and externally validate these trajectories using a competing-risk latent class mixed model that jointly models daily P/F ratios, ICU discharge, and death. Our secondary objective was to determine whether class membership can be predicted early (as early as possible within the first 7 days) using routinely collected clinical variables. Preliminary findings from this study have been presented as an abstract [14].

Take-home message:

We identified and externally validated four reproducible, dynamic oxygenation trajectories in persistent acute hypoxemic respiratory failure: early recovery, stable persistence, biphasic improvement-then-deterioration, and rapid decline. Predictable within three days, these archetypes provide early prognostic value beyond static baseline severity, offering a novel framework for trial enrichment and directing biological sampling.

Methods

Data sources

MIMIC-IV v1.0 (2008–2019) was the derivation cohort [15], selected for its high temporal resolution of physiological data and open-access nature, facilitating methodological reproducibility. For external validation, we curated a specific cohort from three hospitals within Imperial College Healthcare NHS Trust (ICHT, London, 2009–2024) and utilised the Amsterdam UMCdb (AUMC, 2003–2016) [16].

Cohort selection

We focused on persistent AHRF because previous work shows that P/F measured on ICU day 3 dichotomises patients into rapidly resolving versus persistent groups with distinct outcomes [11]. Transient hypoxemia (e.g., post-operative oversedation, fluid overload) resolving within 3 days carries low mortality, has a low proportion of patients who will benefit from new treatment strategies, and reflects different pathobiology from persistent AHRF.

Inclusion criteria: Adults who, for ≥ 72 h (or for ≥ 48 h if they died on day 3), satisfied both (i) P/F < 300 mmHg and (ii) positive end-expiratory pressure (PEEP) ≥ 5 cmH_2O. Exclusion criteria: Age < 18 y or receiving extracorporeal membrane oxygenation.

Patients who died on day 3 were retained to avoid survival bias against the most severe cases. The PEEP threshold ensured inclusion of individuals receiving active ventilatory support.

We also identified a cohort with ARDS in MIMIC-IV from the persistent AHRF cohort (random sample $n = 2100$) to allow a comparison of trajectories between AHRF in the context of ARDS or from other aetiologies (non-ARDS AHRF) (Methods S1, Figure S1).

Measurement and outcomes

Our primary longitudinal variable was the daily mean P/F ratio which was extracted from the onset of AHRF until day 14. Where P/F ratio data were missing, these were imputed from oxygen saturations where possible using previously described methods [17] (Methods S2). Time to ICU discharge or death was recorded from the

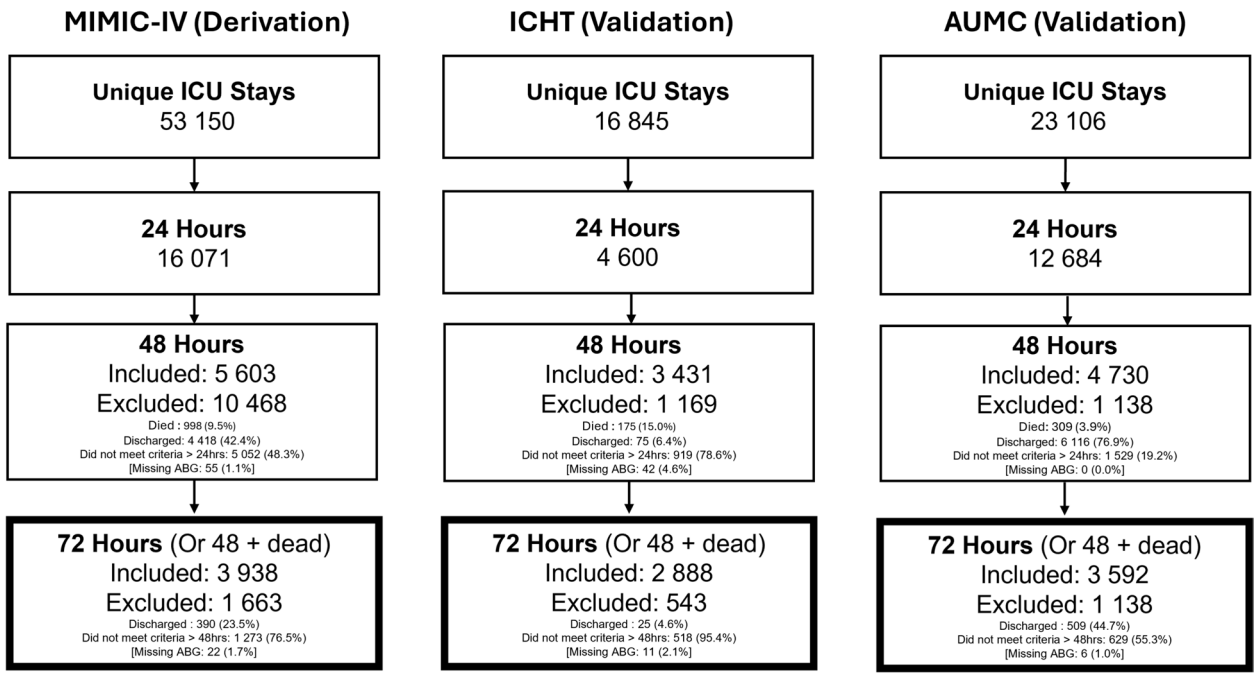


Fig. 1 CONSORT diagram of cohort selection. Selection of patients with persistent AHRF across the derivation (MIMIC-IV) and external validation (ICHT, AUMC) cohorts. Persistent AHRF was defined as P/F < 300 mmHg with PEEP ≥ 5 cmH_2O for ≥ 72 h (or ≥ 48 h if death occurred on day 3). A breakdown of reasons for exclusion at each stage is provided as n (%). The subset of patients excluded due to missing arterial blood gas (ABG) data is reported as a percentage of all criteria failures at that specific time point

day of onset of AHRF. To ensure the robustness of daily P/F, we performed a sensitivity analysis comparing the daily mean, daily median, and first-available P/F values across all cohorts (Methods S10).

Trajectory modelling

Our primary modelling approach was a competing-risk latent class mixed model (CRLCMM) [18]. This approach was selected because it simultaneously captures unobserved heterogeneity in longitudinal P/F trajectories and accounts for patient-level random effects, enabling probabilistic assignment to subgroups that differ in their clinical evolution. We incorporated a competing-risk component for death and discharge, as these events are significant predictors of missing P/F measurements and violate standard missing-at-random assumptions (Methods S3, Figure S2) [19]. Time effects were modelled using a quadratic function, which provided the optimal balance of statistical fit and clinical interpretability. The final number of trajectory classes (TCs) was determined using standard fit statistics (including Bayesian Information Criterion and entropy), class stability via bootstrap consensus-clustering, and a minimum class size requirement of > 10% of the cohort (Methods S4). Additional exploratory baseline models used to supply starting values are detailed in Methods S5. To examine other clinical variables not included in the model, we constructed daily trajectories stratified by TC.

External validation

After fitting the final CRLCMM in MIMIC-IV, we froze all model parameters and applied the model unchanged to the ICHT and AUMC cohorts. In both validation datasets we computed posterior class probabilities for each patient with the frozen MIMIC model and assigned patients to the class of highest probability. Trajectory fidelity was examined by comparing the model-predicted class median P/F curve with the observed class median P/F ratios over days 0–14. This was repeated for predicted discharge and mortality. In each cohort and class, we performed a bias analysis by comparing predicted vs. observed P/F measurements computing the mean signed difference on each day. Further validation analyses are described in Methods S6.

Finally, as further validation a de-novo four-class CRLCMM was fitted directly in external validation cohorts and TCs were visually compared with the MIMIC-IV-derived classes.

Table 1 Baseline characteristics of the derivation and external validation cohorts at AHRF onset

Variable	MIMIC-IV	ICHT	AUMC
n	3938	2888	3592
Demographics			
Age, years[a]	64 [53–74]	61 [50–71]	60–69 (50–59 – 70–79)
Male sex, *n* (%)	2408 (61.1%)	1888 (65.4%)	2251 (62.6%)
Outcomes			
ICU mortality, *n* (%)	1270 (32.2%)	972 (33.7%)	941 (26.2%)
ICU length of stay, days	10 [6–16]	15 [8–27]	12 [7–22]
Vital signs			
Heart rate, bpm	88 [77–101]	85 [72–98]	86 [73–100]
Mean BP, mmHg	74 [69–81]	79 [74–86]	78 [73–85]
Temperature, °C	37.1 [36.7–37.6]	36.7 [36.3–37.3]	36.2 [33.7–36.9]
Respiratory rate, bpm	21 [18–24]	19 [17–23]	19 [15–24]
Organ support			
Invasive mechanical ventilation, *n* (%)[b]	3183 (80.8%)	1982 (68.6%)	3077 (85.6%)
Vasopressor use, *n* (%)	1750 (44.4%)	1653 (57.2%)	2728 (75.9%)
Norepinephrine equiv, µg/kg/min[c]	0.18 [0.08–0.58]	0.12 [0.07–0.24]	0.09 [0.05–0.13]
Respiratory parameters			
PaO_2/FiO_2, mmHg	188 [148–234]	195 [141–248]	225 [185–262]
$PaCO_2$, mmHg	41 [36–46]	40 [36–45]	41 [37–46]
PEEP, cm H_2O	8 [6–11]	8 [6–10]	11 [9–14]
Minute ventilation, L/min	9.8 [8.3–11.6]	9.2 [7.7–10.9]	9.1 [7.7–10.8]
Driving pressure, cm H_2O	15 [12–18]	15 [12–18]	17 [13–21]
Laboratory values			
pH	7.36 [7.31–7.41]	7.39 [7.34–7.43]	7.36 [7.31–7.41]
Lactate, mmol/L	1.9 [1.2–3.2]	1.4 [1.0–2.2]	1.8 [1.2–3.1]
Bicarbonate, mmol/L	22.6 [19.8–25.7]	24.2 [21.6–26.8]	21.9 [19.9–24.8]
Creatinine, mg/dL	1.27 [0.84–2.09]	0.94 [0.69–1.62]	1.10 [0.79–1.67]
WBC, $\times 10^9$/L	12.6 [9.1–17.1]	11.6 [8.3–15.9]	12.2 [8.5–17.1]
Platelet count, $\times 10^9$/L	177 [117–246]	173 [109–258]	177 [113–259]
Hyperinflammatory, *n* (%)	1258 (31.9%)	802 (27.8%)	1103 (30.7%)

Continuous variables are reported as median [interquartile range] and categorical variables as *n* (%), unless otherwise stated

[a] Age in AmsterdamUMCdb is available only as age ranges

[b] Patients not receiving invasive mechanical ventilation received non-invasive ventilation

[c] Norepinephrine equivalent is the median dose among patients receiving vasopressors

Early trajectory class prediction and added-value analysis at day 3

We used 12 routinely charted variables, 10 dynamic measurements (Methods S7) plus age and sex to predict CRLCMM assigned TCs. Models were trained in MIMIC-IV using XGBoost utilizing rolling daily feature matrices. The prediction window was limited to day 0–7; beyond this point, the trajectory classes became clinically distinguishable, diminishing the utility of an early prediction tool. Performance and calibration were computed for each day and TC. External validation applied the frozen MIMIC models unchanged to the ICHT and AUMC cohorts after variable harmonisation; for comparison, site-specific models were also retrained on 80% of each external cohort and evaluated on the remaining 20% (Methods S7).

We performed a leakage-free auxiliary day 3 added-value analysis in MIMIC-IV comparing discrimination for 14-day mortality using day 3 P/F, day 3 P/F plus day 3 SOFA, and day 3 predicted trajectory probabilities (Methods S8).

Exploratory analyses: ARDS and inflammatory subphenotypes

In our cohort labelled as ARDS or non-ARDS AHRF, we evaluated the confidence of TC assignment (posterior probability), and compared predicted trajectories

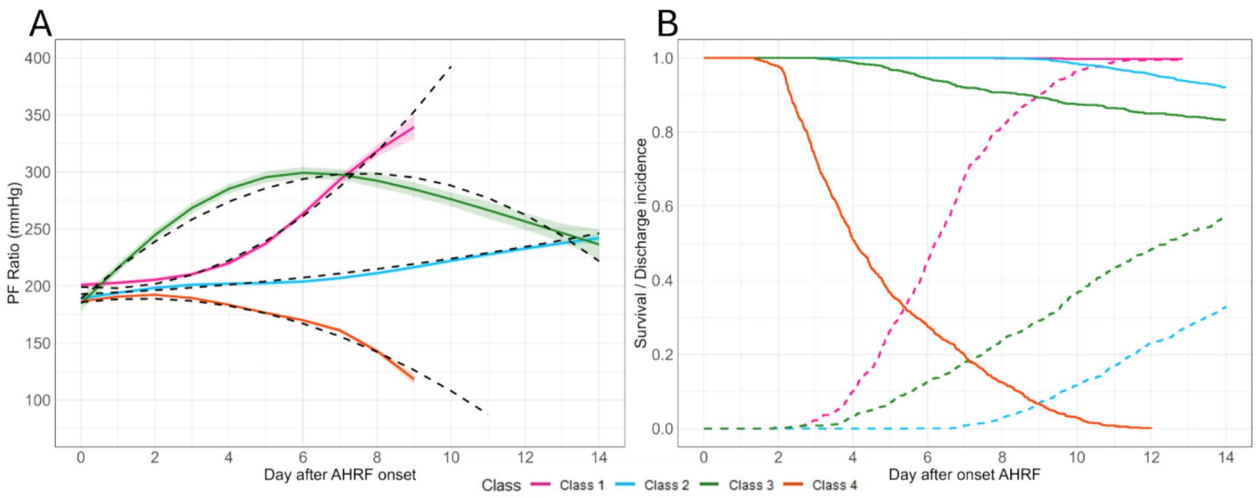


Fig. 2 Trajectory classes in the MIMIC-IV cohort. **A** Observed P/F ratio by trajectory class (solid colored lines show the daily median; shaded ribbons show 95% confidence interval) and CRLCMM model-estimated class median trajectory (black dotted line). **B** Model-estimated cumulative incidence of ICU survival (solid lines) and ICU discharge (dashed lines) by class. TC1 (pink) Early Recovery; TC2 (blue) Stable Persistence; TC3 (green) Biphasic-Improvement-then-Deterioration; TC4 (orange) Rapid Decline

with observed trajectories, as well as the distribution of patients across classes.

To explore how previously described inflammatory subphenotypes relate to longitudinal P/F trajectories, we applied an open-source AI clinical classifier to routinely collected data to assign hyper- vs hypoinflammatory subphenotypes [6]. Subphenotypes were assigned on the day of AHRF onset using a probability threshold of 0.5 [6].

Statistical software and reproducibility

All modelling was performed in R (version 4.4.0) (Methods S9). Code is available in the GitHub repository (https://github.com/Dom-Marshall/AHRF-Trajectories-code).

Results

Patient characteristics in the development and validation cohort

3938 of 53,150 unique patient stays from MIMIC-IV met the inclusion criteria (Fig. 1). The median age was 64 (IQR 53–74), 2408 (61%) were male, and the ICU mortality rate was 32% (Table 1). On the day of onset of AHRF, the median P/F ratio was 188 mmHg (IQR 148–234). Between onset and day 14, 36,253 days of mean P/F measurements were available (17.6% imputed). The median length of stay was 10 days (IQR 6–16).

In the validation dataset (ICHT) 2888 of 16,845 unique patient stays met our criteria with a median age of 61 (IQR, 50–71), 1888 males (65%) and an ICU mortality rate of 34% (Table 1). On the day of AHRF onset the median P/F ratio was 195 mm Hg (141–248). Between onset and day 14 of AHRF, 32,466 days of mean P/F measurements were used (0% imputed) with a median length of stay of 15 days (IQR 8–27). In the second external validation cohort (AUMC), 3592 of 23,106 unique stays met criteria with median P/F 225 (IQR 185–262) on the day of AHRF onset (Table 1).

AHRF trajectories identified by competing risk latent class mixed model

A four-TC competing-risk latent class mixed model (CRLCMM) with a quadratic term for fixed and mixed effects for P/F ratios and competing-risk for death or discharge was selected due to overall optimal model fit (Figure S3). Four TCs had a major improvement in entropy compared with three (0.67 vs 0.54), a median posterior probability of class membership of 0.85 (0.80 in a sensitivity analysis restricted to non-imputed P/F measurements), a minimum class size of 19% and clinically distinct P/F and survival trajectories. Five TCs had similar fit statistics but a minimum class size of 5%. Four TCs were also supported by Bootstrap hierarchical clustering of CRLCMM consensus class modelling and additional clustering approaches (Supplementary Results 1–2, Figures S4–S10, Table S1). Observed mean daily P/F ratios closely matched those predicted by the model (Fig. 2). Cohort-level trajectory shapes and TC assignment were robust to the choice of daily P/F summary metric (mean, median, or first-available). The frequency of arterial blood gas (ABG) sampling by day and cohort

Table 2 Baseline characteristics at AHRF onset by trajectory class in MIMIC-IV

Variable	Trajectory class 1	Trajectory class 2	Trajectory class 3	Trajectory class 4
n (%)	764 (19.4%)	1668 (42.4%)	754 (19.1%)	752 (19.1%)
Demographics				
Age, years	63 [52–73]	64 [52–74]	64 [53–74]	68 [56–77]
Male sex, *n* (%)	468 (61.3%)	1044 (62.6%)	440 (58.4%)	456 (60.6%)
White Ethnicity, *n* (%)	499 (65.3%)	1107 (66.4%)	466 (61.8%)	452 (60.1%)
Outcomes				
ICU mortality, *n* (%)	2 (0.3%)	337 (20.2%)	179 (23.7%)	752 (100.0%)
ICU length of stay, days	6 [5–8]	15 [12–22]	10 [7–14]	4 [3–6]
Mortality—Day 7, *n* (%)	0 (0.0%)	0 (0.0%)	60 (8.0%)	604 (80.3%)
ICU discharge—Day 7, *n* (%)	521 (68.2%)	14 (0.8%)	134 (17.8%)	0 (0.0%)
Mortality—Day 14, *n* (%)	2 (0.3%)	132 (7.9%)	126 (16.7%)	752 (100.0%)
ICU discharge—Day 14, *n* (%)	762 (99.7%)	550 (33.0%)	431 (57.2%)	0 (0.0%)
Organ support				
Invasive mechanical ventilation, *n* (%)	623 (81.5%)	1340 (80.3%)	605 (80.2%)	615 (81.8%)
Norepinephrine equiv, µg/kg/min	0.10 [0.05–0.26]	0.16 [0.08–0.50]	0.20 [0.08–3.05]	0.24 [0.10–3.17]
Respiratory parameters				
Respiratory rate, breaths/min	20 [18–23]	21 [18–24]	21 [18–25]	22 [19–25]
PaO_2/FiO_2, mmHg	204 [159–245]	185 [146–232]	184 [148–232]	184 [145–233]
$PaCO_2$, mmHg	41 [37–46]	42 [37–47]	40 [36–45]	40 [35–46]
PEEP 0–24 h, cm H_2O	8 [5–10]	9 [6–12]	8 [6–11]	8 [6–11]
Driving pressure, cm H_2O	14 [11–18]	15 [12–18]	15 [12–18]	16 [12–20]
Minute ventilation, L/min	9.3 [8.1–11.1]	9.9 [8.4–11.7]	9.8 [8.1–11.5]	10.2 [8.3–12.0]
Laboratory values				
pH	7.38 [7.33–7.42]	7.36 [7.31–7.41]	7.36 [7.31–7.41]	7.34 [7.28–7.39]
Lactate, mmol/L	1.6 [1.1–2.3]	1.8 [1.2–3.1]	2.0 [1.4–3.3]	2.5 [1.5–4.1]
Bicarbonate, mmol/L	23.7 [21.3–26.8]	22.8 [20.2–25.8]	22.3 [19.2–25.5]	21.2 [18.0–24.4]
Creatinine, mg/dL	1.09 [0.78–1.60]	1.23 [0.80–1.98]	1.35 [0.87–2.30]	1.59 [1.00–2.56]
WBC, $\times 10^9$/L	12.1 [9.4–15.8]	12.4 [8.9–17.0]	13.1 [9.2–17.8]	13.5 [9.3–19.1]
Platelet count, $\times 10^9$/L	192 [131–256]	178 [121–250]	173 [114–240]	161 [97–239]
Other				
Charlson comorbidity index	5 (3–7)	6 (4–7)	6 (4–8)	6 (5–9)
Total SOFA	6 [4–8]	7 [4–9]	7 [5–10]	8 [6–11]
Hyperinflammatory *n*, (%)	147 (19.2%)	495 (29.7%)	267 (35.4%)	348 (46.3%)

Continuous variables are reported as median [interquartile range] and categorical variables as *n* (%), unless otherwise stated. Norepinephrine equivalent is the median dose among patients receiving vasopressors. Day 7 and day 14 mortality and ICU discharge are cumulative incidences measured from AHRF onset

is reported in Supplementary Results S3 (Figures S11–S19, Table S2). Trajectory Class 2 (TC2) was the largest group ($n = 1668$). The other three TCs were similar in size, 764, 754 and 752 patients for TC1, TC3, and TC4, respectively (Table 2). The four TCs differed in both P/F evolution and outcome (, Figure S20, Table 2). TC1 demonstrated a moderate rate of P/F ratio improvement with early discharge—521/764 (68%) and 762/764 (≈100%) by day 7 and 14, respectively (Early Recovery). TC2 demonstrated a relatively flat P/F ratio trajectory with comparatively lower rates of death and discharge, 132/1668 (8%) and 550/1668 (33%) by day 14, respectively (Stable Persistence). TC3 demonstrated a biphasic P/F ratio trajectory with the fastest initial improvement in P/F ratio, however, around day 7–8 this deteriorated (Biphasic Improvement-then-Deterioration). This group also demonstrated higher mortality and discharge rates than TC2 (126/754, 17% and 431/754, 57%, death and discharge by day 14, respectively). TC4 demonstrated a rapid decline in P/F and 100% (752/752) mortality by day 14 (Rapid Decline).

At baseline, TC1 (Early Recovery) had the highest P/F ratio, lowest PEEP requirement, lowest vasopressor use, lactate and creatinine (Table 2). TC2 and TC3 were broadly similar in terms of baseline co-morbidities, physiology and laboratory parameters. TC4 (Rapid Decline) demonstrated the least favorable baseline

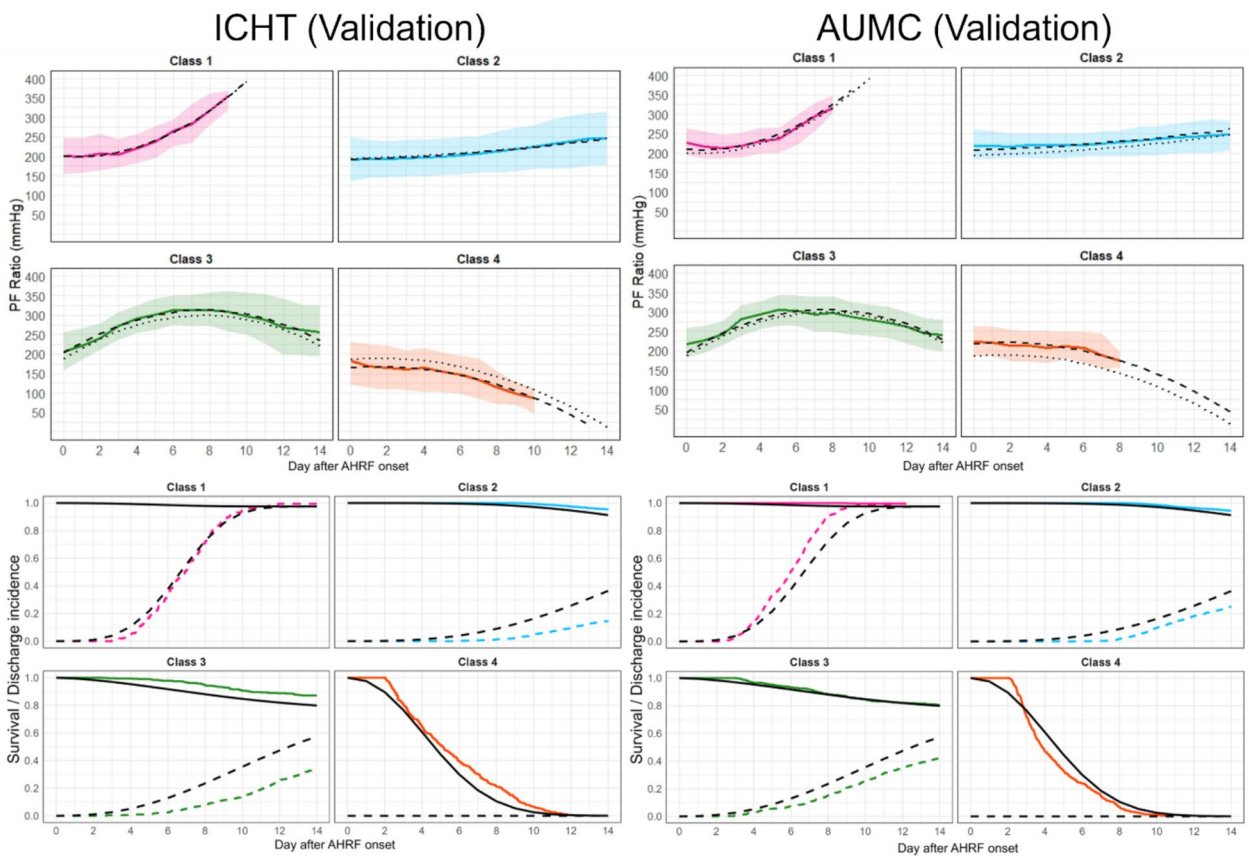


Fig. 3 Top—Predicted and observed P/F trajectories for the validation data (ICHT and AUMC). Observed P/F ratio by trajectory class (solid colored = median class P/F on each day; shaded ribbons = interquartile range), uncorrected model prediction (black dotted), and bias-corrected prediction (black dashed). Bias correction was computed as the mean signed difference (observed minus predicted) across days and applied as a class-specific offset to the predicted curve. Bottom—Observed (colored) vs predicted (black) ICU survival (solid) and discharge (dashed) incidence by trajectory class

characteristics, including modestly higher rates of some comorbidities, notably cancer (TC4 19% vs the overall cohort 13%), liver disease, and ischemic heart disease; however, their median Charlson Comorbidity Index was the same as TC2–3 (Table 2, Table S3).

Evolution in physiological variables by trajectory class

As well as each TC being characterized by P/F ratio trajectory and outcome, distinct patterns of physiological evolution were seen in other variables not included in the trajectory models (Figure S21). TC1, Early Recovery, received the lowest PEEP despite its improvement in P/F ratio being initially slower than TC3. TC4 demonstrated the least favorable trajectories across cardiovascular, respiratory and renal parameters.

External validation

Across both external cohorts, the frozen MIMIC CRL-CMM assigned TCs with high certainty (ICHT median posterior probability = 0.95; AUMC = 0.84; Figure S22). In each cohort, observed P/F trajectories and outcome curves closely matched the model-predicted TC profiles (Fig. 3). In the ICHT cohort, using a standard measure of prediction accuracy (root mean squared error), our model reduced overall P/F prediction error compared with a daily-median baseline (overall RMSE—8.1%; Table S4). In AUMC, a mean-bias analysis (predicted—observed P/F) showed minimal offset for TC1 and TC3 but a small systematic positive bias for TC2 and TC4 (+ 14.4 and + 32.5 mmHg, respectively), consistent with the higher cohort-level median P/F in AUMC versus MIMIC-IV (225 vs 188 mmHg). Non-P/F organ-failure trajectories displayed similar TC-wise patterns in both cohorts (Figure S21). Baseline characteristics by TC are reported for ICHT and AUMC in Table S5–S6.

CRLCMM was re-run independently in the ICHT cohort with four classes demonstrating optimal fit statistics and with the predicted P/F trajectories and outcomes being highly similar (Figures S23–24) to those predicted by the MIMIC-IV model. When refitted in the AUMC cohort, the core trajectory shapes were similarly recovered; however, reflecting AUMC's lower overall mortality, the model preferentially split larger recovery classes at $K=4$ before isolating the less-frequent, high-mortality TC4 archetype at $K=5$ (Figures S25–26). To confirm the temporal stability of our models, we performed an era sensitivity analysis within the ICHT cohort. Comparison of patients treated ≤ 2015 ($n=422$) to those treated from 2016 to 2024 ($n=2466$) demonstrated no material differences in class-assignment confidence, TC proportions, or empirical trajectory shapes (Figures S27–29).

Prediction of trajectory classes

The binary one-vs-rest classifiers using full observation feature matrices were most accurate in derivation and validation cohorts. Discrimination and calibration generally improved up to day 7 (Figure S30–37 and Table S7). On day 3 after AHRF onset the MIMIC-derived model achieved mean AUCs across TCs of 0.80 (0.82/0.73/0.80/0.86 for TC1–4) in MIMIC-IV, 0.79 (0.70/0.73/0.86/0.86 for TC1–4) in ICHT, and 0.78 (0.84/0.73/0.78/0.75 for TC1–4) in AUMC; specificity remained high (0.78–0.86) with modest calibration error (Expected Calibration Error 0.06–0.10; Brier 0.12–0.17). Complete performance metrics, including site-specific retraining results (e.g., AUMC TC4 AUC improving to 0.83), are detailed in Table S8. Model feature importance is demonstrated in Figure S38–39: Day 3 ventilatory variables were most significant in predicting TC1 and TC2, whereas TC3 was defined by PF ratio. In contrast, TC4 was driven by day 3 hemodynamic and metabolic failure.

In a leakage-free auxiliary analysis at day 3 in MIMIC-IV, day 3 P/F alone had AUC 0.55 for 14-day mortality and day 3 P/F plus day 3 SOFA had AUC 0.63. Day 3 predicted trajectory probabilities alone achieved AUC 0.74 (DeLong $p<0.001$ versus baseline (Day 3 SOFA and P/F)), supporting that early trajectory assignment captures information beyond static severity (Table S9).

P/F trajectories in patients with ARDS

In the subset of MIMIC-IV labelled as ARDS or non-ARDS AHRF, among patients labelled as ARDS ($n=1041$) versus non-ARDS AHRF ($n=930$), the mean posterior probability of TC assignment was similar (0.81 in ARDS vs 0.78 in non-ARDS AHRF). Observed TC-specific P/F trajectories were also comparable between the two groups (Figure S40). ARDS patients were distributed across all four TCs (Figure S41). Within the ARDS cohort, 50% of ARDS patients were assigned to TC2 (Stable Persistence), compared with 33% of non-ARDS AHRF patients (Figure S42). Mortality was similar between ARDS and non-ARDS AHRF across TCs except in TC3 (Biphasic Improvement-then-Deterioration), where mortality was 10% in ARDS versus 26% in non-ARDS AHRF.

Prevalence of inflammatory subphenotypes across P/F AHRF trajectory classes

Using a clinical AI classifier [6], the prevalence of the hyperinflammatory subphenotype at onset of AHRF was 32% in MIMIC-IV, 28% in ICHT, and 31% in AUMC (Table 1). Across cohorts TC1 had the fewest patients classified as hyperinflammatory; (MIMIC-IV 19%, ICHT 18%, AUMC 18%), whilst TC4 had the highest (ICHT 53%, MIMIC-IV 46%, AUMC 41%) (Table 2, Table S5–6).

Discussion

Using a competing-risk latent class mixed model, we identified and validated four reproducible trajectory classes of persistent AHRF for the first 14 days after onset. These trajectory archetypes: Early Recovery, Stable Persistence, Biphasic Improvement-then-Deterioration, and Rapid Decline were replicated across geographically and temporally distinct populations capturing the dynamic evolution of respiratory failure often missed by static severity scores. Although derived solely from P/F ratios and competing outcomes, each class exhibited distinctive patterns of multiorgan dysfunction and differences in the prevalence of previously described inflammatory subphenotypes. These trajectory classes are best understood as dynamic severity archetypes: reproducible patterns of oxygenation evolution and competing outcomes that encode the direction and shape of illness, rather than risk at a single time point. Importantly, class membership could be predicted within 3 days of AHRF onset using routinely collected data, adding prognostic information that a single day 3 P/F ratio does not capture. Within a population already selected for persistent hypoxaemia, a single severity measurement is a poor discriminator of subsequent course (AUC 0.55 for 14-day mortality); trajectory prediction is demonstrably superior (AUC 0.74) because it captures the rate and direction of physiological change.

The potential clinical utility of these archetypes derives from their prognostic signal: they allow clinicians to anticipate the likely disease course rather than reacting solely to static severity. For example, early identification of a recovery trajectory (TC1) could support prompt evaluation for ventilator liberation. Conversely, a predicted biphasic trajectory (TC3) cautions against premature reassurance during initial oxygenation improvement;

awareness that deterioration typically follows around days 7–8 should prompt vigilance for secondary complications rather than de-escalation. A predicted rapid decline (TC4) carries near-uniform 14-day mortality, providing early prognostic context for goals-of-care discussions or escalation planning, such as extracorporeal support referral. Importantly, this study does not establish whether specific interventions modify trajectory membership. The most immediate research application is as a day 3 stratification variable for clinical trials, where predicted trajectory may offer more relevant enrichment than static baseline severity because it predicts the course an intervention aims to modify [20].

These archetypes may also reduce heterogeneity when anchoring biological sampling to well-defined clinical courses. Although data-driven subphenotyping has revealed clinically meaningful heterogeneity in other critical illness syndromes, including cross-sectional subphenotypes and temperature trajectories in sepsis [21–23], in respiratory failure the focus has remained predominantly on ARDS and cross-sectional early time points [1, 10, 24]. Prior AHRF work has largely used simplified longitudinal definitions, such as a day 3 dichotomy of persistent versus resolving hypoxemia, with the persistent group demonstrating higher mortality and increased inflammatory markers [11]. In contrast, our study examines persistent AHRF at scale, includes external validation, accounts for informative dropout via ICU discharge and death, and links trajectory archetypes to established inflammatory subphenotypes. Our four-class model adds critical granularity, identifying a biphasic group that initially resembles the recovery class but suffers late deterioration, and distinguishes stable persistence from rapid decline. Static assessment on Day 3 would misclassify these patients as 'resolving,' potentially confounding trials aimed at early de-escalation or identifying patients at risk of late-stage complications.

A defining feature of these archetypes is their stability across heterogeneous cohorts. Across derivation and validation cohorts, the study spanned different healthcare systems in the United States, the United Kingdom, and the Netherlands, distinct eras (2003–2024), and differing practice patterns in oxygenation targets and ventilator weaning. That the same four classes were reproducible with high assignment certainty across these settings suggests they represent transportable clinical archetypes rather than artifacts of local care protocols.

While the underlying biological drivers of each trajectory remain to be defined, the contrasts in other clinical biomarkers and inflammatory subphenotype prevalence suggest distinct pathological processes in each trajectory. We hypothesise that the Early Recovery trajectory reflects prompt resolution of the inciting lung injury. The Biphasic trajectory's characteristic deterioration around days 7–8 may reflect secondary insults, such as ventilator-induced lung injury or nosocomial infection, but this remains speculative and requires prospective microbiological and ventilator-management data. Additionally, while higher baseline driving pressure was associated with more severe trajectory classes, our observational design cannot determine whether this reflects baseline injury severity, a modifiable contributor to ongoing lung damage, or both. The Rapid Decline archetype warrants scrutiny. While its true prevalence may be underrepresented due to our inherent exclusion of deaths before day 3, given its near-uniform mortality, we considered whether this trajectory primarily reflects early treatment limitations. Lacking uniform timing for withdrawal of life-sustaining therapy across cohorts, we assessed this indirectly via organ-support patterns. While most ICU deaths ultimately involve some degree of limitation, several features argue against this archetype being solely a proxy for baseline frailty or early de-escalation. First, organ-support trajectories indicate that, on average, patients in this class received high ventilatory and vasoactive support up to death. Second, compared with the cohort overall, these patients had similar Charlson comorbidity indices and only a modestly higher prevalence of metastatic cancer. Importantly, this group was consistently enriched for the hyperinflammatory subphenotype, suggesting that rapid decline could reflect severe systemic inflammation and refractory shock. Prior work indicates that mortality in hyperinflammatory subphenotypes is frequently driven by circulatory failure rather than respiratory failure alone [25]. Together, these observations suggest that Rapid Decline is unlikely to be explained solely by baseline frailty or early treatment limitation and may instead identify a subgroup with severe systemic inflammation and refractory multiorgan dysfunction.

Beyond their prognostic signal, these trajectory labels address a structural limitation of cross-sectional classification. By jointly modelling oxygenation trajectories with competing risks of ICU discharge and death, this approach reduces the informative dropout and survivor bias inherent to single time-point comparisons. These dynamic labels are designed to complement, not replace, cross-sectional subphenotypes (e.g., hyper- vs hypo-inflammatory). While the latter robustly capture early inflammatory states, they do not encode temporal evolution [26]. Directly linking clinical courses to biology holds significant promise for identifying mechanistic underpinnings, a strategy that has already yielded early results in COVID-19 [27]. Further, patients meeting ARDS criteria were distributed across all four trajectory classes, with ARDS and non-ARDS patients following comparable oxygenation courses with similar assignment certainty. This substantial overlap suggests that longitudinal

oxygenation trajectories transcend traditional syndromic definitions. Because a patient's point-in-time ARDS status poorly predicts their longitudinal course, relying solely on the ARDS label for trial enrichment or prognostication may be insufficient or even unhelpful. This finding aligns with emerging consensus that broad syndromic labels can obscure clinically meaningful heterogeneity [28]. The similarity between ARDS and non-ARDS AHRF is also reinforced by the PHIND study, which prospectively demonstrated that hyper- and hypoinflammatory subphenotypes exist at similar prevalence in both groups [29]. Consequently, trajectory-based classification may outperform traditional ARDS definitions for prognostic enrichment, allowing trials to capture patients with shared clinical trajectories who might otherwise be excluded or grouped by a single-point assessment.

Strengths of this study include the use of large, heterogeneous datasets across three countries, a modelling approach accounting for informative dropout, and a rigorously validated early-prediction framework. Limitations include a retrospective design with unmeasured confounding (e.g., goals-of-care) and practice variability across settings. Despite heterogeneity in baseline P/F trajectories between cohorts, the model maintained high certainty in class assignments, though site-specific recalibration may improve transportability. Our 72-h inclusion criterion focuses the analysis on persistent AHRF and excludes patients who die or rapidly recover before day 2/3; these archetypes, therefore, do not capture the full early spectrum of respiratory failure, in which disease biology and therapeutic window may differ. A proportion of daily P/F values in the MIMIC-IV derivation cohort were imputed via SpO_2/FiO_2 conversion. However, in a sensitivity analysis excluding imputed P/F measurements, class-assignment confidence was unchanged, and the trajectory classes were reproduced in both external validation cohorts where no imputed P/F values were used. A notable challenge is the lower prediction performance observed for TC2 (Stable Persistence). Because its trajectory is flat, it lacks the distinct early signals necessary for confident classification. Additionally, as the largest cohort (42.4%), this group is heterogeneous and may act as a catch-all for smaller TCs that generate predictive noise. Given this inherent uncertainty, future enrichment strategies targeting TC2 may be more effective if they rely on continuous predicted probabilities rather than hard class assignments. Finally, this methodological approach cannot establish causality or show that trajectory-guided management changes outcomes; prospective, biologically enriched studies are needed.

In conclusion, we have defined four reproducible oxygenation archetypes in persistent AHRF that capture heterogeneity in disease course, multiorgan dysfunction, and outcome. Importantly, these can be forecast by day 3 with the potential to inform prognostication and trial enrichment. Prospective studies should now test whether using these archetypes to guide biological sampling enhances mechanistic discovery, and whether trajectory-targeted treatments ultimately improve patient outcome.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1007/s00134-026-08441-2.

Author details
[1] Division of Anaesthetics, Pain Medicine and Intensive Care, Department of Surgery and Cancer, Imperial College London, London, UK. [2] Cleveland Clinic London, London, UK. [3] STRIVE, London, UK. [4] School of Electrical and Electronic Engineering, Imperial College London, London, UK.

Acknowledgements
DCM is supported by an MRC clinical research training fellowship (award MR/Y000404/1) and the Mittal Fund at Cleveland Clinic Philanthropy. ADG is supported by the MRC MultiSci Doctoral Training Partnership (MR/W00710X/1). DBA is supported by an NIHR DSE award (NIHR303768) and the NIHR Imperial Biomedical Research Centre. The views expressed in this article are those of the authors and not necessarily those of the MRC, National Health Service, NIHR, or Department of Health and Social Care. The funders of the study had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript, or the decision to submit for publication.

Author contributions
DCM and ADG had access to all of the data in the study and took responsibility for the integrity of the data and accuracy of the data analysis. DCM, MK, DBA and SP were responsible for the study concept and design. All authors (DCM, ADG, MK, BVP, SP, DBA) were responsible for interpretation of the manuscript, drafting of the manuscript and critical revision of the manuscript.

Data availability
The MIMIC-IV database is publicly available to credentialled researchers who complete the required human-subjects training and agree to the data-use terms (https://physionet.org/content/mimiciv/3.1/). The AmsterdamUMCdb is publicly available to researchers upon registration and acceptance of the data-use agreement; access can be requested via the project portal (https://amsterdammedicaldatascience.nl/). De-identified patient-level data from Imperial College Healthcare NHS Trust (ICHT) are not publicly available because of institutional governance and privacy restrictions. Access may be granted to qualified researchers upon reasonable request and subject to (i) a prespecified protocol, (ii) relevant regulatory approvals, and (iii) a data-sharing agreement with ICHT. Enquiries should be directed to the study team.

Declarations

Conflicts of interest
The authors declare no conflicts of interest.

Ethics
The MIMIC IV database has received ethical approval from the institutional review boards (IRB) at BIDMC and MIT, and, because the database does not contain protected health information, a waiver of the requirement for informed consent was included in the IRB approval. Use of routinely collected intensive-care data from Imperial College Healthcare NHS Trust (ICHT) was approved by the UK Health Research Authority and the London-Brighton & Sussex Research Ethics Committee (REC reference 25/HRA/0032; IRAS project ID 348389; protocol 6438939). All ICHT records were fully de-identified, and the REC likewise waived the requirement for individual informed consent. AmsterdamUMCdb (Amsterdam University Medical Centers ICU database) was created using a risk-based de-identification strategy and underwent external

privacy and ethics audits. Access is governed by an end-user license and mandatory research-integrity training; therefore, individual informed consent is not required for use of this anonymized dataset. Preliminary findings for this study were presented at the American Thoracic Society Congress 2026.

Use of artificial intelligence/large language model

During manuscript preparation, the authors employed ChatGPT (OpenAI, accessed October 2025) exclusively for proofreading (spelling, grammar, and consistency of terminology). The AI system was not used to generate scientific content, interpret data, or draft any original sections.

Open Access

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 19 February 2026 Accepted: 10 April 2026
Published: 17 May 2026

References

1. Calfee CS, Delucchi K, Parsons PE et al (2014) Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med 2:611–620. https://doi.org/10.1016/S2213-2600(14)70097-9
2. Sinha P, Calfee CS (2019) Phenotypes in acute respiratory distress syndrome: moving towards precision medicine. Curr Opin Crit Care 25:12–20
3. Sinha P, Churpek MM, Calfee CS (2020) Machine learning classifier models can identify acute respiratory distress syndrome phenotypes using readily available clinical data. Am J Respir Crit Care Med 202:996–1004
4. Reddy K, Calfee CS, McAuley DF (2021) Acute respiratory distress syndrome subphenotypes beyond the syndrome: a step toward treatable traits? Am J Respir Crit Care Med 203:1449–1451. https://doi.org/10.1164/rccm.202101-0218ED
5. Grasselli G, Calfee CS, Camporota L et al (2023) ESICM guidelines on acute respiratory distress syndrome: definition, phenotyping and respiratory support strategies. Intensive Care Med 49:727–759. https://doi.org/10.1007/s00134-023-07050-7
6. Pensier J, Fosset M, Paschold BS et al (2025) Temporal stability of phenotypes of acute respiratory distress syndrome: clinical implications for early corticosteroid therapy and mortality. Intensive Care Med 51:1784–1796. https://doi.org/10.1007/s00134-025-08089-4
7. Iwashyna TJ, Hodgson CL, Pilcher D et al (2016) Timing of onset and burden of persistent critical illness in Australia and New Zealand: a retrospective, population-based, observational study. Lancet Respir Med 4:566–573. https://doi.org/10.1016/S2213-2600(16)30098-4
8. Shaw M, Viglianti EM, McPeake J et al (2020) Timing of onset, burden, and postdischarge mortality of persistent critical illness in Scotland, 2005–2014: a retrospective, population-based, observational study. Crit Care Explor 2:e0102. https://doi.org/10.1097/CCE.0000000000000102
9. Nasa P, Bos LD, Estenssoro E et al (2025) Defining and subphenotyping ARDS: insights from an international Delphi expert panel. Lancet Respir Med 13:638–650. https://doi.org/10.1016/S2213-2600(25)00115-8
10. Chen H, Yu Q, Xie J et al (2022) Longitudinal phenotypes in patients with acute respiratory distress syndrome: a multi-database study. Crit Care 26:340. https://doi.org/10.1186/s13054-022-04211-w
11. Sathe NA, Zelnick LR, Mikacenic C et al (2021) Identification of persistent and resolving subphenotypes of acute hypoxemic respiratory failure in two independent cohorts. Crit Care 25:336. https://doi.org/10.1186/s13054-021-03755-7
12. Bai Y, Chen S, Yang H et al (2025) Dynamic oxygenation subgroup bringing new insights in ARDS: more predictive of outcomes and response to PEEP than static PaO_2/FiO_2. Thorax 80:525–534
13. Sinha P, Meyer NJ, Calfee CS (2023) Biological phenotyping in sepsis and acute respiratory distress syndrome. Annu Rev Med 74:457–471. https://doi.org/10.1146/annurev-med-043021-014005
14. Marshall DC, Green AD, Komorowski M et al (2026) Longitudinal, outcome-aware acute hypoxic respiratory failure phenotypes: discovery, external validation, and early prediction. ATS Congress
15. Johnson AEW, Bulgarelli L, Shen L et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10:1. https://doi.org/10.1038/s41597-022-01899-x
16. Thoral PJ, Peppink JM, Driessen RH et al (2021) Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. Crit Care Med 49:e563–e577. https://doi.org/10.1097/CCM.0000000000004916
17. Rice TW, Wheeler AP, Bernard GR et al (2007) Comparison of the SpO_2/FIO_2 ratio and the PaO_2/FIO_2 ratio in patients with acute lung injury or ARDS. Chest 132:410–417. https://doi.org/10.1378/chest.07-0617
18. Proust-Lima C, Philipps V, Liquet B (2017) Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. J Stat Softw 78:1–56
19. van de Schoot R, Sijbrandij M, Winter SD et al (2017) The GRoLTS-checklist: guidelines for reporting on latent trajectory studies. Struct Equ Modeling 24:451–467
20. Ware LB, Matthay MA, Mebazaa A (2020) Designing an ARDS trial for 2020 and beyond: focus on enrichment strategies. Intensive Care Med 46:2153–2156
21. Bhavani SV, Carey KA, Gilbert ER et al (2019) Identifying novel sepsis subphenotypes using temperature trajectories. Am J Respir Crit Care Med 200:327–335
22. Seymour CW, Kennedy JN, Wang S et al (2019) Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 321:2003–2017
23. Rojas JC, Lyons PG, Chhikara K et al (2025) A common longitudinal intensive care unit data format (CLIF) for critical illness research. Intensive Care Med 51:556–569
24. Bos LDJ, Sjoding M, Sinha P et al (2021) Longitudinal respiratory subphenotypes in patients with COVID-19-related acute respiratory distress syndrome: results from three observational cohorts. Lancet Respir Med 9:1374–1387. https://doi.org/10.1016/S2213-2600(21)00365-9
25. Evrard B, Sinha P, Delucchi K et al (2024) Causes and attributable fraction of death from ARDS in inflammatory phenotypes of sepsis. Crit Care 28:219. https://doi.org/10.1186/s13054-024-04943-x
26. Maddali MV, Churpek M, Pham T et al (2022) Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet Respir Med 10:367–377
27. Diray-Arce J, Fourati S, Jayavelu ND et al (2023) Multi-omic longitudinal study reveals immune correlates of clinical course among hospitalized COVID-19 patients. Cell Rep Med 4:101079. https://doi.org/10.1016/j.xcrm.2023.101079
28. Maslove DM, Tang B, Shankar-Hari M et al (2022) Redefining critical illness. Nat Med 28:1141–1148
29. Reddy K, Sinha P, Antcliffe DB et al (2026) Bedside identification of subphenotypes in acute respiratory failure (PHIND): a multicentre, observational cohort study. Lancet Respir Med. https://doi.org/10.1016/S2213-2600(26)00040-8