

Supplementary Information

Development and Validation of VentPilot: An AI-based Recommendation System for Mechanical Ventilation

List of Contents

TRIPOD+AI Checklist.....	3
Methods S1. Additional exclusion criteria for validation cohorts	5
Methods S2. VentPilot reinforcement-learning framework.....	6
Methods S3. Intensivist feedback workflow and preference model construction	9
Methods S4. Standard-of-care guardrails.....	10
Methods S5. Set tidal volume prediction for volume-targeted modes.....	11
Methods S6. Agreement score and inverse probability-weighted outcome analyses	12
Methods S7. Ventilator setting comparisons across physiologic strata	14
Methods S8. Condition-based comparison of recommendation and setting-adjustment frequency and timing.....	15
Methods S9. Temporal consistency analyses.....	16
Methods S10. Rule-based comparator analyses.....	17
Table S1. Time-varying variables used in patient trajectories.....	18
Table S2. Standard-of-care guardrails incorporated into VentPilot.....	20
Table S3. Physiologic conditions and ventilator-setting directions used in the condition-based analysis	21
Table S4. Extended baseline clinical characteristics of derivation and validation cohorts	22
Table S5. Baseline clinical characteristics of the AI-similar and AI-dissimilar groups before inverse probability weighting.....	24
Table S6. Baseline clinical characteristics of the AI-similar and AI-dissimilar groups after inverse probability weighting.....	26
Table S7. Fitted Q-evaluation results.....	28
Table S8. Recommendation and setting-adjustment frequency and timing in the condition-based analysis	29
Table S9. Fitted Q-evaluation in ASV-exposed patients in the external validation cohort	30
Table S10. Covariate balance in ASV-exposed and AI-similar non-ASV groups	31
Table S11. Weighted outcomes in ASV-exposed and AI-similar non-ASV groups	33
Table S12. Subgroup analysis results	34
Table S13. Sensitivity analysis results	35
Table S14. Calibration of fitted Q-evaluation against empirical clinician returns	37
Figure S1. Dataset flowchart.....	38
Figure S2. Iterative intensivist-feedback loop for collecting preference labels.....	39
Figure S3. Ventilator setting comparisons across physiologic strata in the internal validation cohort	40
Figure S4. Ventilator setting comparisons across physiologic strata in the external validation cohort.....	41
Figure S5. Timing of VentPilot recommendations and clinician setting adjustments.....	42
Figure S6. Temporal consistency results	43
Figure S7. An example case of VentPilot recommendations for ventilator settings	44

TRIPOD+AI Checklist



Version: 11-January-2024

Section/Topic	Item	Development / evaluation ¹	Checklist item	Reported on page
TITLE				
<i>Title</i>	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
ABSTRACT				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	3-4
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	6
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	6
	3c	D;E	Describe any known health inequalities between sociodemographic groups	NA
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	6
METHODS				
<i>Data</i>	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	7
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	7
<i>Participants</i>	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	7
	6b	D;E	Describe the eligibility criteria for study participants	7, Methods S1
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	7-8, Methods S2
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	7-8, Methods S2
<i>Outcome</i>	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	8-9, Methods S2
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	NA
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	NA
<i>Predictors</i>	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	8, Methods S2
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	8, Methods S2
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	NA
<i>Sample size</i>	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	NA
<i>Missing data</i>	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	7-8, Methods S1-S2
<i>Analytical methods</i>	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	7-9, Methods S2-S6
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	8, Methods S2
	12c	D	Specify the type of model, rationale ² , all model-building steps, including any hyperparameter tuning, and method for internal validation	8, Methods S2-S5
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations ³	7-12
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	9-12
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	NA
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	8, Methods S2, S5
<i>Class imbalance</i>	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	NA
<i>Fairness</i>	14	D;E	Describe any approaches that were used to address model fairness and their rationale	NA
<i>Model output</i>	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	8, Methods S2, S4-S5

¹ D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

² Separately for all model building approaches.

³ TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

<i>Training versus evaluation</i>	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	6-7
<i>Ethical approval</i>	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	6
OPEN SCIENCE				
<i>Funding</i>	18a	D;E	Give the source of funding and the role of the funders for the present study	14
<i>Conflicts of interest</i>	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	NA
<i>Protocol</i>	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	NA
<i>Registration</i>	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	NA
<i>Data sharing</i>	18e	D;E	Provide details of the availability of the study data	6, 14
<i>Code sharing</i>	18f	D;E	Provide details of the availability of the analytical code ⁴	NA
PATIENT & PUBLIC INVOLVEMENT				
<i>Patient & Public Involvement</i>	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	NA
RESULTS				
<i>Participants</i>	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	8
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	8
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	8
<i>Model development</i>	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	8
<i>Model specification</i>	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary) ⁵	7
<i>Model performance</i>	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	9-11
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details ³ .	10-11
<i>Model updating</i>	24	E	Report the results from any model updating, including the updated model and subsequent performance	NA
DISCUSSION				
<i>Interpretation</i>	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	11-12
<i>Limitations</i>	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	12-13
<i>Usability of the model in the context of current care</i>	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	Methods S2, S4
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	5, 12
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	13

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

⁴ This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.

⁵ This relates to the code to implement the model to get estimates of risk for a new individual.

Methods S1. Additional exclusion criteria for validation cohorts

In addition to the derivation-cohort exclusion criteria, patients were excluded from the internal and external validation cohorts if they had a tracheostomy, received invasive mechanical ventilation for less than 36 hours, lacked essential information required to define the validation trajectory (e.g., intubation/extubation dates, ICU admission type, or ventilator mode), or had partially missing trajectory data. Partially missing trajectory data were defined as missing segments of physiologic or ventilator time-series data totaling at least 25% of the interval from intubation to extubation. The 36-hour threshold was selected to provide sufficient longitudinal observations for sequential policy evaluation and patient-level concordance assessment. Because baseline covariates in the primary concordance analysis were derived from the first 24 hours after intubation, this threshold also ensured at least 12 hours of post-baseline observation for concordance assessment.

Methods S2. VentPilot reinforcement-learning framework

Trajectory construction

For model derivation, patient trajectories were represented on a fixed 1-minute time grid spanning the intubation period to preserve temporal ordering between patient states and subsequent ventilator adjustments as closely as possible during learning. For off-policy evaluation in the internal and external validation cohorts, trajectories were represented on a fixed 15-minute time grid to reflect a more pragmatic deployment setting and to align with the average update interval of the Mayo Clinic external validation dataset. If ventilation continued beyond 28 days, trajectories were truncated at day 28 to align with the 28-day outcome window used for ventilator-free days (VFDs).

Time-varying variables listed in Table S1 were aligned to the prespecified time grid to form multivariable patient trajectories. When multiple measurements were available within a given interval, continuous variables were summarized using the interval mean. If a ventilator-setting change occurred within the interval, physiologic measurements were summarized only up to the time of the first setting change to prevent post-adjustment measurements from being incorporated into the preceding state. Ventilator settings were represented by their current values rather than interval averages. When no new measurement was recorded within an interval, the most recent available value was carried forward using a time-limited sample-and-hold approach consistent with prior studies [1,2]. Values were forward-filled only within prespecified variable-specific windows. For intermittently measured laboratory variables, the carry-forward window was 24 hours; the elapsed time since the most recent measurement was also included as a model input to represent measurement recency. Values exceeding the allowed window were treated as missing. Remaining missing values were retained as not-a-number (NaN) values and handled directly by the model.

We additionally appended patient-level variables, including age, height, weight, sex, Sequential Organ Failure Assessment (SOFA) score, and Acute Physiology and Chronic Health Evaluation II (APACHE II) score, to each trajectory. For SOFA and APACHE II, the most recent Electronic Health Record (EHR)-recorded value between ICU admission and intubation was used when available; otherwise, the first value recorded within 24 hours after intubation was used. When a post-intubation score was used, the score was treated as missing at trajectory timestamps preceding its recorded time, so that the score was not propagated backward to earlier decision points. The appended patient-level variables were used both to normalize derived ventilatory measures, such as tidal volume normalized by ideal body weight, and to provide patient-level clinical context for training the model.

MDP formulation

We formulated ventilator-setting recommendation as an episodic Markov decision process (MDP) defined on discretized patient trajectories [3]. At a given timestamp t , the patient state s_t comprised the time-varying clinical and ventilator variables listed in Table S1 together with appended patient-level variables. The action a_t was defined as a 6-dimensional vector of changes in ventilator settings between consecutive grid points. Each action dimension corresponded to the change applied to a ventilator control variable (sFiO₂, sPEEP, sPC, sPS, sTV, and sRR), with only mode-applicable components active at a given time point. Each episode $\xi = (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_T, a_T, r_T)$ extended from initiation of invasive mechanical ventilation to extubation or day 28, whichever occurred first, where r_t denotes the reward at time t . The objective was to learn a policy π that maximized the expected undiscounted cumulative reward over the trajectory:

$$J(\pi) = \mathbb{E}_{\xi \sim \pi} \left[\sum_{t=1}^T r_t \right]$$

Reward definition

The reward was defined as a weighted sum of two components: a VFDs-based component and a preference-based component derived from intensivist feedback:

$$r = w_{VFDs} \cdot r_{VFDs} + w_{pref} \cdot r_{pref}$$

For a given fixed time grid, let H denote the maximum possible number of timestamps in a trajectory (i.e., when the patient remains intubated until the 28-day horizon), and let $\Delta = 28/H$ denote the time interval between consecutive timestamps in days. The VFDs reward (r_{VFDs}) was defined as a constant per-step negative reward, with an additional terminal penalty for trajectories ending in death within the 28-day window:

$$r_{VFDs,t} = \begin{cases} -(H - t + 1)\Delta & \text{If } t \text{ is the terminal timestamp and the patient died within 28 days,} \\ -\Delta & \text{Otherwise.} \end{cases}$$

Note that the cumulative VFDs reward over a realized trajectory, $\sum_{t=1}^T r_{VFDs,t}$, where T denotes the realized trajectory length, is equal to the corresponding patient’s VFDs minus 28.

The preference reward (r_{pref}) was defined as the learned preference model output shifted and scaled by Δ :

$$r_{pref,t} = (\hat{R}(s_t, a_t) - 1)\Delta$$

where $\hat{R}$ denotes the learned preference model with output range $[0, 1]$, which was trained on board-certified intensivist ratings of action appropriateness for a subset of state-action pairs (Methods S3). The factor Δ was introduced to align the per-step scale of the preference reward with that of the VFDs reward. We set $w_{VFDs} = 1$ and $w_{pref} = 0.1$ as the default.

Model development

VentPilot was developed in three stages. First, we learned an Implicit Q-Learning (IQL)-based critic on the derivation dataset $\mathcal{D}_{drv} = \{(s_t, a_t, r_t, s_{t+1})\}$ [4]. Specifically, we estimated a state-value function $V_\psi(s)$ and a state-action value function $Q_\theta(s, a)$ using an implicit backup based on expectile regression:

$$\begin{aligned} L_V(\psi) &= \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}_{drv}} \left[L_2^\tau \left(Q_\theta(s_t, a_t) - V_\psi(s_t) \right) \right] \\ L_Q(\theta) &= \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}_{drv}} \left[\left(Q_\theta(s_t, a_t) - (r_t + V_\psi(s_{t+1})) \right)^2 \right] \end{aligned}$$

where the expectile loss L_2^τ was defined as

$$L_2^\tau(u) = |\tau - \mathbf{1}(u < 0)| \cdot u^2$$

The state-value and state-action value functions were parameterized using an FT-Transformer, a transformer architecture designed for tabular data [5].

Because both the VFDs-based reward and the preference reward are bounded, the attainable remaining cumulative reward at timestamp t is also bounded. Let $R_{VFDs,t}$ and $R_{pref,t}$ denote the return-to-go for the VFDs and preference components, respectively. Then,

$$\begin{aligned} R_{VFDs,t} &\in [-(H - t + 1)\Delta, 0] \\ R_{pref,t} &\in [-(H - t + 1)\Delta, 0] \end{aligned}$$

Accordingly, for the overall reward return-to-go $R_t = w_{VFDs} \cdot R_{VFDs,t} + w_{pref} \cdot R_{pref,t}$, the attainable range is

$$R_t \in [-(w_{VFDs} + w_{pref})(H - t + 1)\Delta, 0]$$

We therefore clipped both value estimates onto this attainable range during training:

$$\begin{aligned} L_V^{clip}(\psi) &= \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}_{drv}} \left[L_2^\tau \left(\Pi_t(Q_\theta(s_t, a_t)) - V_\psi(s_t) \right) \right] \\ L_Q^{clip}(\theta) &= \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}_{drv}} \left[\left(Q_\theta(s_t, a_t) - \Pi_t(r_t + V_\psi(s_{t+1})) \right)^2 \right] \end{aligned}$$

where

$$\Pi_t(x) = \min(0, \max(-(w_{VFDs} + w_{pref})(H - t + 1)\Delta, x))$$

This time-indexed return-range clipping leverages the bounded 28-day horizon to keep value estimates within clinically attainable ranges during learning.

Second, we extracted a clinically interpretable deployment policy from the learned value functions by fitting separate optimal decision trees for each ventilator control variable [6–8]. For policy extraction, we used a control-specific balanced sampling distribution $\mathcal{D}_{drv,k}$, obtained by reweighting sampling probabilities in $\mathcal{D}_{drv}$ so that change ($a_{t,k} \neq 0$) and no-change ($a_{t,k} = 0$) actions were sampled with equal probability, to prevent the more frequent no-change observations from dominating the tree-fitting objective. The tree policy π_{ϕ_k} for control variable k was fitted by minimizing a regularized weighted regression objective:

$$L_{\pi,k}(\phi_k) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}_{drv,k}} \left[\exp \left((Q_\theta(s_t, a_t) - V_\psi(s_t)) / \beta \right) \cdot (\pi_{\phi_k}(s_t) - a_{t,k})^2 \right] + \lambda |\pi_{\phi_k}|$$

where $a_{t,k}$ denotes the logged action for ventilator control variable k and $|\pi_{\phi_k}|$ denotes the tree size, defined as the number of branch nodes. Under this objective, logged actions with larger estimated advantage ($Q_\theta(s_t, a_t) - V_\psi(s_t)$) under the learned value functions contributed more strongly to policy fitting [9]. Because the advantage was computed from the full multidimensional action vector, each control-specific tree was fitted using information about the joint value of the observed ventilator-setting combination rather than an independently estimated value for that control variable alone. Hyperparameters for value learning and policy extraction, including τ , β , and λ , were selected using 5-fold cross-validation on the derivation cohort.

Third, we post-processed the extracted trees to enforce prespecified standard-of-care guardrails (details in Methods S4) [10–13]. After post-processing, the control-specific trees were combined to define the overall vector-valued deployment policy π_ϕ , with only mode-applicable components active at each decision point. These steps yielded the final guideline-compliant deployment policy. This deployment policy was finalized before subsequent evaluation in the validation cohorts. At deployment, VentPilot directly applied π_ϕ to the current patient state to generate ventilator-setting recommendations, with set tidal volume recommendations for volume-targeted modes obtained as described in Methods S5.

Model evaluation

We used off-policy evaluation to estimate the value of the developed policy. Specifically, we used fitted Q-evaluation (FQE) on the internal and external validation trajectories [14]. For a fixed target policy π and validation transition set $\mathcal{D}_{val} = \{(s_t, a_t, r_t, s_{t+1})\}$, an evaluation critic $Q_{\tilde{\theta}}(s, a)$ was fit by iterative Bellman regression:

$$L_Q^{val}(\tilde{\theta}) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}_{val}} \left[\left(Q_{\tilde{\theta}}(s_t, a_t) - \Pi_t \left(r_t + Q_{\tilde{\theta}}(s_{t+1}, \pi_\phi(s_{t+1})) \right) \right)^2 \right]$$

The estimated policy value was then computed as

$$\hat{J}(\pi_\phi) = \frac{1}{N} \sum_{i=1}^N Q_{\tilde{\theta}}(s_0^{(i)}, \pi_\phi(s_0^{(i)}))$$

where $s_0^{(i)}$ denotes the initial state of trajectory i and N is the number of validation trajectories. FQE was performed separately for the overall reward and for each reward component. The estimated overall return was therefore not constrained to equal the weighted sum of the separately estimated component returns. For the observed clinician behavior policy, return was estimated as the empirical mean of observed episode returns in the corresponding validation cohort. Uncertainty was quantified by trajectory-level nonparametric bootstrapping.

As a calibration analysis, we additionally fit the evaluation critic to the observed clinician behavior policy, using the observed subsequent clinician action in the Bellman target. The resulting overall return estimate was compared with the empirical mean of observed episode returns.

Methods S3. Intensivist feedback workflow and preference model construction

To incorporate clinician-aligned supervision, we constructed a preference model using feedback from a committee of three board-certified intensivists at SNUH, each with at least 12 years of clinical experience (Figure S2). Candidate cases for review were selected using a behavior-cloning (BC) model trained on the derivation cohort. Specifically, the BC model generated a predicted action for each state, and we identified outlier state-action pairs where the model’s predicted actions differed substantially from the observed actual action for at least one ventilator control variable. In total, 1,186 candidate state–action pairs were reviewed by the committee.

Each candidate outlier was first independently reviewed by all committee members, who assigned a preliminary appropriateness grade on a 4-point scale and recorded a brief rationale. During this stage, reviewers were not shown one another’s preliminary grades. The grading scale denoted appropriate actions (grade 1), acceptable but unnecessary actions (grade 2), actions that were suboptimal but not entirely incorrect (grade 3), and inappropriate actions (grade 4). Reviewers were instructed to assess the action in the context of the contemporaneous clinical state and ventilator setting trajectory available at that decision point. Reviewers were not provided with VentPilot recommendations, preference-model outputs, or downstream clinical outcomes during grading. Of the 1,186 reviewed cases, 354 (29.8%) were excluded as outside the scope of ventilator-setting assessment, including cases associated with suction procedures, bronchoscopy, sedative administration, or chest radiography.

Cases with discordant preliminary grades were adjudicated in a consensus meeting. During adjudication, committee members reviewed the same state-action pair together, discussed the rationale for their initial labels, and reconciled disagreements through iterative discussion. A final label was assigned only after consensus was reached. Outliers graded as 3 or 4 after adjudication were treated as non-preferred samples and were separated from the original candidate set. This review cycle was repeated until all outlier samples had either been excluded as out-of-scope events or assigned a final adjudicated grade. The remaining 832 cases received final labels and contributed labeled samples for preference-model development.

For preference model training, the adjudicated grades were mapped to a preference target as follows: grades 1–2 $\rightarrow$ 1 and grades 3–4 $\rightarrow$ 0. This mapping defined a preference signal in which values closer to 1 indicate more desirable actions and values closer to 0 indicate less desirable actions. We then trained a preference model using gradient-boosted decision trees (XGBoost) to predict the mapped preference target from state-action features [15]. Inputs consisted of the state and action variables at the decision timestamp, and the target was the adjudicated preference target. Hyperparameters of the preference model were selected using 5-fold cross-validation within the adjudicated labeled dataset. The learned preference model provided the per-step preference reward used during VentPilot training (r_{pref}).

Methods S4. Standard-of-care guardrails

Despite substantial heterogeneity in real-world ventilator management, several minimal safety requirements are broadly accepted in clinical practice. We summarized these requirements in Table S2 and treated them as minimum standard-of-care guardrails for the final deployment policy [10–13].

After extracting the tree-based policy, we post-processed the learned tree structures to enforce these guardrails. Specifically, we reviewed the learned branches and downstream recommendations against the rule set in Table S2 and modified local branch conditions and/or terminal recommendations where the reachable recommendations could violate the guardrails. These edits were designed to be minimal and local, so that recommendations were brought into compliance while preserving as much of the learned tree structure as possible.

Overall, guardrail-based post-processing affected 26 of 159 terminal leaves (16.4%). In a post hoc characterization performed after policy finalization, these modifications changed less than 1% of setting recommendations across evaluation timestamps in either validation cohort (0.36% in the internal cohort and 0.64% in the external cohort).

Methods S5. Set tidal volume prediction for volume-targeted modes

We trained our AI model using data from SNUH, where pressure-targeted ventilation modes are predominantly used. As a result, it is insufficient to train a model to predict set tidal volumes with this data alone. To enable tidal volume prediction for the Mayo Clinic external validation data, we applied the sPC model trained from the SNUH data and converted the sPC recommendations into sTV recommendations using the dynamic lung compliance formula [16]. The sTV recommendation was calculated by multiplying the sPC recommendation by the 4-hour average dynamic lung compliance and rounding the resulting value to the nearest 10 mL.

Methods S6. Agreement score and inverse probability-weighted outcome analyses

Agreement score and AI-similar group definition

Agreement between observed ventilator settings and VentPilot recommendations was assessed at chart-update time points, defined as timestamps when at least one time-varying variable in Table S1 received a newly recorded EHR value. We used these event-triggered time points to focus the agreement analysis on moments when clinicians were more likely to reassess ventilator settings.

For each control variable i (sFiO₂, sPEEP, sPC/sPS for pressure-targeted modes, sTV for volume-targeted modes, and sRR), we computed the discrepancy between the action recommended by VentPilot and the actual observed action at a chart-update timestamp t as: $\Delta_{i,t} = a_{i,t}^{\text{delivered}} - a_{i,t}^{\text{recommended}}$. Discrepancies were standardized using a variable-specific scaling factor s_i defined as the standard deviation of $\Delta_{i,t}$ estimated from the derivation cohort. For each patient and control variable, we then computed a standardized root mean square deviation over time:

$$RMSD_i = \sqrt{\frac{1}{B} \sum_{t=1}^B (\Delta_{i,t}/s_i)^2}$$

where B denotes the number of chart-update timestamps for that patient. When a control variable was not applicable for a given timestamp (e.g., tidal volume in pressure-targeted modes), it was excluded from the RMSD calculation for that timestamp.

The patient-level discrepancy score was defined as the mean of $RMSD_i$ across the set of applicable control variables: $D = \frac{1}{6} \sum_{i=1}^6 RMSD_i$, where lower values indicate higher agreement. In the primary analysis, concordance was assessed using all eligible chart-update timestamps from 24 hours after intubation until extubation or day 28, whichever occurred first. Patient-level baseline covariates used for propensity-score estimation were derived from data collected during the first 24 hours after intubation. Patients whose discrepancy scores ranked in the lowest 50% within each validation cohort were classified as AI-similar, and all others were classified as AI-dissimilar.

As sensitivity analyses, we (i) recalculated concordance on the fixed 15-minute validation grid rather than at chart-update timestamps; (ii) performed a fixed early-window analysis in which baseline covariates were recalculated using data from 0–12 hours after intubation and concordance was assessed using eligible chart-update timestamps during the subsequent 12–36-hour period; and (iii) evaluated stricter discrepancy-score thresholds by comparing the lowest 40% with the highest 40% and the lowest 30% with the highest 30% within each validation cohort, excluding intermediate patients from the threshold-based analyses. In the Mayo Clinic cohort, we additionally recalculated concordance using only pressure-targeted ventilation periods, thereby excluding converted sTV recommendations from the agreement score.

The fixed early-window analysis was designed to mitigate potential time-varying confounding by restricting concordance assessment to an early prespecified period. Because validation eligibility required at least 36 hours of invasive mechanical ventilation, all patients included in the fixed early-window analysis contributed the same 24-hour concordance-assessment window. For this analysis, concordance scores, AI-similarity classifications, propensity scores, and stabilized inverse probability weights were recalculated; all other analytic specifications were unchanged from the primary analysis.

Inverse probability weighting

Propensity scores were used to estimate the probability of AI-similarity based on baseline covariates that could confound the agreement–outcome association [17]. Propensity scores were estimated using multivariable logistic regression including prespecified baseline covariates: ICU admission type, age, sex, body mass index, P/F ratio, SOFA score, APACHE II score, primary diagnosis category, major comorbidities, and additional baseline respiratory mechanics measures (full variable list in Table S4). Stabilized inverse probability weights targeting the average treatment effect (ATE) were computed as: $sw_i = P(A = 1)/e_i$ for AI-similar patients and $sw_i = P(A = 0)/(1 - e_i)$ for AI-dissimilar patients, where e_i is the propensity score and A indicates AI-similarity. Covariate balance was assessed using standardized mean differences (SMDs) before and after weighting; SMD < 0.1 was considered indicative of adequate balance [18]. To reduce the influence of extreme weights, inverse probability weights were truncated at the 99th percentile [19].

Considered patient outcomes

For this analysis, the primary outcome was ventilator-free days within 28 days, defined as days alive and free from invasive mechanical ventilation during the first 28 days after intubation; patients who died within this period were assigned zero ventilator-free days [20]. Secondary outcomes included 28-day mortality, ICU mortality, in-hospital mortality, ICU and hospital length of stay, and duration of mechanical ventilation among ICU survivors. For continuous outcomes, P values were obtained using survey-weighted Wilcoxon rank-sum tests, and weighted mean or median differences, as appropriate, were reported. For binary outcomes, P values were obtained using survey-weighted chi-square tests, and absolute risk differences were reported. Percentile 95% confidence intervals (CIs) for these differences were estimated using 1,000 nonparametric bootstrap samples. Within each bootstrap sample, the propensity-score model was refitted, stabilized inverse probability weights were recalculated and truncated at the 99th percentile, and the weighted outcome difference was re-estimated.

Methods S7. Ventilator setting comparisons across physiologic strata

To descriptively compare VentPilot recommendations with observed settings across prespecified physiologic strata, we evaluated eligible timestamps in the validation cohorts. We defined the following strata: $\text{SpO}_2 < 92\%$, $\text{SpO}_2 \geq 98\%$, P/F ratio < 100 mmHg, P/F ratio ≥ 300 mmHg, $\text{pH} < 7.3$, $\text{pH} \geq 7.45$, TV/IBW < 4 mL/kg, and TV/IBW ≥ 8 mL/kg. At each eligible timestamp, we paired the contemporaneous observed setting with the corresponding VentPilot recommendation for each control variable. For each stratum and control variable, we plotted the empirical distributions of observed and VentPilot-recommended settings and summarized the difference using the mean difference and Cohen's d, calculated using the pooled standard deviation. Because sPC and sPS apply only to pressure-targeted modes and sTV to volume-targeted modes, the number of eligible decision points varied across panels.

Methods S8. Condition-based comparison of recommendation and setting-adjustment frequency and timing

To compare the frequency and timing of VentPilot recommendations with observed clinician ventilator-setting adjustments, we performed a condition-based analysis in the internal and external validation cohorts. The physiologic conditions and corresponding ventilator-setting response directions are listed in Table S3.

For each patient and condition type, a condition episode was defined as a contiguous period during which the corresponding physiologic condition persisted for at least 5 minutes. For each episode, we identified the first VentPilot recommendation in the condition-specific direction and the first observed clinician adjustment of the relevant ventilator setting in the same direction. Recommendation frequency was defined as the proportion of episodes with a VentPilot recommendation in the condition-specific direction, and setting-adjustment frequency as the proportion with a clinician adjustment of the relevant ventilator setting in the same direction.

Time to recommendation was defined as the elapsed time from condition onset to the first VentPilot recommendation in the condition-specific direction. Time to setting adjustment was defined analogously using the first observed clinician adjustment in the same direction. Timing comparisons were restricted to episodes in which both events occurred. Frequencies are reported as number (%), and times are summarized as median (IQR) in hours. The time difference was defined as time to recommendation minus time to setting adjustment. Two-sided McNemar tests were used to compare event frequencies, and two-sided Wilcoxon signed-rank tests were used to compare paired times.

Methods S9. Temporal consistency analyses

To assess temporal consistency, we quantified variation in the frequency of setting-change events according to time of day (6-hour bins) and day of week in the SNUH cohort. Clinician events were defined as timestamps with any observed adjustments in ventilator control variables, and VentPilot events were defined as timestamps with any recommended change in at least one control variable. Observed counts were expressed as a percentage of the expected count under a uniform-in-time assumption (100% indicates no temporal variation). Confidence intervals were computed using Wilson score intervals for the proportion of events falling in each bin and transformed to the observed/expected scale, and two-sided binomial tests assessed deviation from uniformity. These analyses were performed only in the SNUH cohort because timestamps in the Mayo Clinic dataset were randomly shifted for de-identification.

Methods S10. Rule-based comparator analyses

As a comparator analysis against an existing rule-based automated ventilation strategy, we performed comparator analyses against adaptive support ventilation (ASV) in the Mayo Clinic external validation cohort [21]. For off-policy evaluation, we identified patients with ASV exposure during the intubation period, defined as ASV use for at least 25% of the invasive mechanical ventilation period, and compared the VentPilot policy return with the empirical clinician behavior policy return within this subgroup using the same fitted Q-evaluation framework as in the primary analysis. For inverse probability-weighted analyses, patients with ASV exposure were compared with patients managed without ASV whose observed care was classified as AI-similar using the same agreement-score framework as in the primary analysis. Patients with ASV use for more than 0% but less than 25% of the invasive mechanical ventilation period were not included in this comparator analysis. Propensity score estimation, weight construction and truncation, covariate balance assessment, and outcome definitions followed the primary inverse probability-weighted analysis.

Table S1. Time-varying variables used in patient trajectories

Name	Abbreviation	Unit
Saturation of peripheral oxygen	SpO ₂	%
Mean blood pressure	MBP	mmHg
Heart rate	HR	bpm
Potential of hydrogen in arterial blood	pH	-
Partial pressure of oxygen in arterial blood	PaO ₂	mmHg
Partial pressure of carbon dioxide in arterial blood	PaCO ₂	mmHg
Saturation of oxygen in arterial blood	SaO ₂	%
Bicarbonate in arterial blood	HCO ₃ ⁻	mmol/L
Lactic acid	lactate	mmol/L
Peak inspiratory pressure	PIP	cmH ₂ O
Tidal volume	V _T	mL
Minute ventilation	MV	L/min
Fraction of inspired oxygen	FiO ₂	%
Positive end-expiratory pressure	PEEP	cmH ₂ O
Respiratory rate	RR	/min
End-tidal CO ₂	EtCO ₂	mmHg
Airway occlusion pressure (drop in Paw during the first 100ms of inspiratory effort)	P0.1	cmH ₂ O
Set fraction of inspired oxygen	sFiO ₂	%
Set positive end-expiratory pressure	sPEEP	cmH ₂ O
Set pressure control above PEEP	sPC	cmH ₂ O
Set pressure support above PEEP	sPS	cmH ₂ O
Set respiratory rate	sRR	/min
Set tidal volume	sTV	mL
Inspiration: expiration time ratio	I:E ratio	-
Spontaneous inspiration time	T _{insp}	sec
Inspiratory cycle off	Cycleoff	%

Inspiratory rise time	T _{insp.rise}	% or sec
Trigger sensitivity (flow, pressure)	Sensitivity	L/min or cmH ₂ O
Ventilator mode	-	-

Table S2. Standard-of-care guardrails incorporated into VentPilot

Guideline	Affected control variables
Maintain PaO ₂ between 55 and 110 mmHg [10]	sFiO ₂
Maintain tidal volume between 4 and 8 mL/kg of ideal body weight [11–13]	sPC, sTV
Maintain plateau pressure below 30 cmH ₂ O [11,13]	sPEEP, sPC, sTV

Table S3. Physiologic conditions and ventilator-setting directions used in the condition-based analysis

Control variable	Response direction	Physiologic condition
sFiO ₂	Increase	SpO ₂ < 92% OR PaO ₂ < 60mmHg [11]
sFiO ₂	Decrease	SpO ₂ > 96% AND PaO ₂ > 120mmHg
sPEEP	Increase	SpO ₂ < 92% AND P/F ratio < 200mmHg AND Low PEEP ^a [24]
sPEEP	Decrease	SpO ₂ > 96% AND P/F ratio > 200mmHg AND High PEEP ^a [24]
sPC, sTV	Increase	pH < 7.3 AND TV/IBW < 6mL/kg [11,24]
sPC, sTV	Decrease	pH > 7.45 AND TV/IBW > 6mL/kg [11,24]
sRR	Increase	pH < 7.3 AND sRR < 30/min [24]
sRR	Decrease	pH > 7.45 AND sRR > 12/min [24]
sPS	Increase	TV/IBW < 6mL/kg AND RR > 25/min [11]
sPS	Decrease	TV/IBW > 6mL/kg AND RR < 17/min [25]

Abbreviation: TV/IBW, tidal volume / ideal body weight.

^a: Low PEEP and High PEEP were defined based on the ARDSNet low-PEEP and high-PEEP FiO₂–PEEP tables [24]. “Low PEEP” indicates that the observed set PEEP was at or below the table-recommended PEEP for a given FiO₂ (using the low-PEEP table), whereas “High PEEP” indicates that the observed set PEEP was at or above the table-recommended PEEP for the FiO₂ (using the high-PEEP table).

Table S4. Extended baseline clinical characteristics of derivation and validation cohorts

Variable	Derivation cohort (n = 3002)	Internal validation cohort (n = 283)	Mayo Clinic external validation cohort (n = 1011)	Overall (n = 4296)
Type of ICU admission				
Medical	452 (15.1)	135 (47.7)	600 (59.3)	1187 (27.6)
Surgical	2550 (84.9)	148 (52.3)	411 (40.7)	3109 (72.4)
Age, y	59.9 ± 15.8	67.2 ± 14.2	63.2 ± 15.1	61.2 ± 15.7
Sex				
Male	1654 (55.1)	177 (62.5)	566 (56.0)	2397 (55.8)
Female	1348 (44.9)	106 (37.5)	445 (44.0)	1899 (44.2)
Body mass index, kg/m ²	23.8 ± 4.3	22.9 ± 4.3	30.4 ± 7.7	25.3 ± 6.0
Primary diagnosis				
Cardiovascular	554 (18.5)	21 (7.4)	220 (21.8)	795 (18.5)
Gastrointestinal	105 (3.5)	17 (6.0)	76 (7.5)	198 (4.6)
Hepatobiliary	123 (4.1)	13 (4.6)	50 (4.9)	186 (4.3)
Neoplastic	1552 (51.7)	97 (34.3)	24 (2.4)	1673 (38.9)
Neurologic	140 (4.7)	45 (15.9)	90 (8.9)	275 (6.4)
Renal	14 (0.5)	8 (2.8)	29 (2.9)	51 (1.2)
Respiratory	85 (2.8)	14 (4.9)	131 (13.0)	230 (5.4)
Sepsis, including pneumonia	96 (3.2)	28 (9.9)	149 (14.7)	273 (6.3)
Other	333 (11.1)	40 (14.1)	242 (23.9)	615 (14.3)
Comorbidities				
Cardiovascular disease	745 (24.8)	108 (38.2)	725 (71.7)	1578 (36.7)
Diabetes	487 (16.2)	79 (27.9)	346 (34.2)	912 (21.2)
COPD	129 (4.3)	28 (9.9)	245 (24.2)	402 (9.4)
Other chronic lung disease ^a	288 (9.6)	69 (24.4)	182 (18.0)	539 (12.5)
Chronic kidney disease	164 (5.5)	30 (10.6)	361 (35.7)	555 (12.9)
Malignancy	2039 (67.9)	185 (65.4)	280 (27.7)	2504 (58.3)
P/F ratio, mmHg	383 (276-488)	268 (196-387)	247 (175-332)	328 (221-440)
SOFA score	8 (6-12)	12 (9-15)	10 (8-12)	10 (7-12)
APACHE II score	25 (21-29)	31 (26-35)	25 (22-28)	25 (22-29)
Oxygenation Index	1.6 (1.2-2.7)	3.0 (1.7-4.6)	3.8 (2.5-5.8)	2.3 (1.4-4.1)

Ventilatory ratio ^b	1.2 (1.0-1.4)	1.3 (1.1-1.5)	1.2 (1.0-1.5)	1.2 (1.0-1.5)
Driving pressure, cmH ₂ O	9 (7-11)	12 (10-15)	14 (10-17)	10 (8-13)
Elastance, cmH ₂ O/(mL/kg) ^c	1.2 (1.0-1.5)	1.7 (1.3-2.1)	2.1 (1.7-2.7)	1.4 (1.1-1.9)
Mechanical power, J/min/kg ^c	0.2 (0.1-0.2)	0.3 (0.2-0.3)	0.3 (0.2-0.4)	0.2 (0.1-0.3)

Abbreviations: APACHE, Acute Physiology and Chronic Health Evaluation; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; SNUH, Seoul National University Hospital; SOFA, Sequential Organ Failure Assessment.

Values are reported as n (%) for categorical variables and median (IQR) or mean $\pm$ SD for continuous variables.

^a: Includes interstitial lung disease or other chronic lung diseases, excluding COPD.

^b: Defined as (minute ventilation $\times$ PaCO₂)/(ideal body weight $\times$ 100 $\times$ 37.5).

^c: Values were normalized with ideal body weight.

Table S5. Baseline clinical characteristics of the AI-similar and AI-dissimilar groups before inverse probability weighting

Variable	SNUH internal validation cohort			Mayo Clinic external validation cohort		
	AI-similar (n=142)	AI-dissimilar (n=141)	SMD	AI-similar (n=506)	AI-dissimilar (n=505)	SMD
Type of ICU admission						
Medical	58 (40.8)	77 (54.6)	-0.277	273 (54.0)	327 (64.8)	-0.221
Surgical	84 (59.2)	64 (45.4)	0.277	233 (46.0)	178 (35.2)	0.221
Age, y	67.4 ± 15.1	67.0 ± 13.2	0.030	64.1 ± 14.8	62.3 ± 15.4	0.123
Sex						
Male	90 (63.4)	87 (61.7)	0.035	275 (54.3)	291 (57.6)	-0.066
Female	52 (36.6)	54 (38.3)	-0.035	231 (45.7)	214 (42.4)	0.066
Body mass index, kg/m ²	22.5 ± 4.2	23.4 ± 4.4	-0.208	30.1 ± 7.1	30.8 ± 8.4	-0.085
Primary diagnosis						
Cardiovascular	12 (8.5)	9 (6.4)	0.079	114 (22.5)	106 (21.0)	0.037
Gastrointestinal	9 (6.3)	8 (5.7)	0.028	44 (8.7)	32 (6.3)	0.089
Hepatobiliary	6 (4.2)	7 (5.0)	-0.035	26 (5.1)	24 (4.8)	0.018
Neoplastic	45 (31.7)	52 (36.9)	-0.109	14 (2.8)	10 (2.0)	0.052
Neurologic	31 (21.8)	14 (9.9)	0.329	58 (11.5)	32 (6.3)	0.181
Renal	2 (1.4)	6 (4.3)	-0.172	8 (1.6)	21 (4.2)	-0.155
Respiratory	7 (4.9)	7 (5.0)	-0.002	48 (9.5)	83 (16.4)	-0.208
Sepsis, including pneumonia	9 (6.3)	19 (13.5)	-0.240	62 (12.3)	87 (17.2)	-0.141
Other	21 (14.8)	19 (13.5)	0.038	132 (26.1)	110 (21.8)	0.101
Comorbidities						
Cardiovascular disease	59 (41.5)	49 (34.8)	0.140	363 (71.7)	362 (71.7)	0.001
Diabetes	37 (26.1)	42 (29.8)	-0.083	163 (32.2)	183 (36.2)	-0.085
COPD	13 (9.2)	15 (10.6)	-0.050	108 (21.3)	137 (27.1)	-0.135
Other chronic lung disease ^a	29 (20.4)	40 (28.4)	-0.185	64 (12.6)	118 (23.4)	-0.281

Chronic kidney disease	13 (9.2)	17 (12.1)	-0.094	169 (33.4)	192 (38.0)	-0.096
Malignancy	87 (61.3)	98 (69.5)	-0.173	133 (26.3)	147 (29.1)	-0.063
P/F ratio, mmHg	328 (237-416)	222 (170-305)	0.686	260 (195-334)	222 (160-329)	0.234
SOFA score	11 (9-14)	14 (11-16)	-0.615	10 (8-12)	11 (9-12)	-0.345
APACHE II score	30 (26-34)	32 (27-37)	-0.193	25 (22-27)	26 (22-28)	-0.169
Oxygenation Index	2.2 (1.4-3.4)	4.1 (2.4-5.7)	-0.708	3.3 (2.3-4.6)	4.3 (2.9-6.6)	-0.494
Ventilatory ratio ^b	1.2 (0.9-1.4)	1.4 (1.2-1.7)	-0.762	1.1 (1.0-1.4)	1.3 (1.1-1.6)	-0.488
Driving pressure, cmH ₂ O	11 (9-14)	13 (11-15)	-0.441	13 (10-16)	15 (12-18)	-0.480
Elastance, cmH ₂ O/(mL/kg) ^c	1.5 (1.3-2.0)	1.8 (1.4-2.3)	-0.349	2.0 (1.6-2.5)	2.3 (1.8-2.9)	-0.336
Mechanical power, J/min/kg ^c	0.2 (0.1-0.3)	0.3 (0.2-0.4)	-0.723	0.3 (0.2-0.3)	0.3 (0.3-0.4)	-0.785

Abbreviations: AI, artificial intelligence; APACHE, Acute Physiology and Chronic Health Evaluation; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; SMD, standardized mean difference; SNUH, Seoul National University Hospital; SOFA, Sequential Organ Failure Assessment.

Values are reported as n (%) for categorical variables and median (IQR) or mean $\pm$ SD for continuous variables.

^a: Includes interstitial lung disease or other chronic lung diseases, excluding COPD.

^b: Defined as (minute ventilation $\times$ PaCO₂)/(ideal body weight $\times$ 100 $\times$ 37.5).

^c: Values were normalized with ideal body weight.

Table S6. Baseline clinical characteristics of the AI-similar and AI-dissimilar groups after inverse probability weighting

Variable	SNUH internal validation cohort			Mayo Clinic external validation cohort		
	AI-similar (n=137) ^a	AI-dissimilar (n=138) ^a	SMD	AI-similar (n=512) ^a	AI-dissimilar (n=505) ^a	SMD
Type of ICU admission						
Medical	68 (49.4)	66 (47.5)	0.039	311 (60.7)	297 (58.8)	0.039
Surgical	69 (50.6)	73 (52.5)	-0.039	201 (39.3)	208 (41.2)	-0.039
Age, y	67.3 ± 14.4	67.3 ± 13.3	0.001	63.1 ± 14.7	63.1 ± 15.3	0.005
Sex						
Male	86 (62.4)	85 (61.9)	0.010	282 (55.0)	280 (55.4)	-0.009
Female	52 (37.6)	53 (38.1)	-0.010	231 (45.0)	224 (44.6)	0.009
Body mass index, kg/m ²	22.8 ± 4.2	22.9 ± 4.2	-0.021	30.5 ± 7.6	30.4 ± 8.2	0.018
Primary diagnosis						
Cardiovascular	11 (8.1)	11 (7.7)	0.016	112 (22.0)	110 (21.8)	0.004
Gastrointestinal	7 (5.4)	7 (5.0)	0.018	43 (8.4)	40 (7.9)	0.016
Hepatobiliary	5 (3.9)	6 (4.4)	-0.024	25 (4.9)	26 (5.1)	-0.010
Neoplastic	48 (34.9)	48 (34.9)	0.001	15 (2.9)	15 (3.1)	-0.006
Neurologic	23 (16.6)	24 (17.6)	-0.026	43 (8.3)	43 (8.5)	-0.007
Renal	2 (1.5)	4 (2.7)	-0.082	11 (2.2)	14 (2.7)	-0.032
Respiratory	8 (6.0)	6 (4.3)	0.073	65 (12.7)	67 (13.3)	-0.020
Sepsis, including pneumonia	12 (8.6)	14 (9.9)	-0.044	75 (14.7)	73 (14.4)	0.010
Other	20 (15.0)	19 (13.5)	0.040	123 (23.9)	117 (23.2)	0.017
Comorbidities						
Cardiovascular disease	52 (37.7)	54 (39.4)	-0.034	362 (70.7)	361 (71.4)	-0.016
Diabetes	39 (28.2)	44 (31.9)	-0.080	185 (36.1)	174 (34.5)	0.034
COPD	14 (10.1)	13 (9.3)	0.024	118 (23.1)	121 (24.0)	-0.022
Other chronic lung disease ^b	34 (24.9)	34 (24.4)	0.012	102 (19.8)	91 (18.0)	0.048

Chronic kidney disease	13 (9.7)	14 (10.3)	-0.018	186 (36.3)	178 (35.2)	0.022
Malignancy	91 (66.4)	90 (65.1)	0.027	147 (28.7)	144 (28.5)	0.005
P/F ratio, mmHg	275 (208-392)	268 (190-365)	0.086	241 (175-317)	248 (169-348)	-0.016
SOFA score	12 (9-15)	12 (9-15)	-0.067	10 (8-12)	10 (8-12)	0.033
APACHE II score	31 (26-34)	31 (27-35)	-0.030	25 (23-28)	25 (22-28)	-0.003
Oxygenation Index	2.7 (1.7-4.0)	3.0 (1.5-4.5)	-0.089	3.8 (2.5-5.8)	3.8 (2.5-5.8)	0.008
Ventilatory ratio ^c	1.3 (1.0-1.5)	1.3 (1.1-1.6)	-0.075	1.2 (1.0-1.5)	1.2 (1.0-1.5)	0.031
Driving pressure, cmH ₂ O	12 (9-15)	12 (10-15)	-0.018	14 (10-17)	14 (10-17)	-0.006
Elastance, cmH ₂ O/(mL/kg) ^d	1.6 (1.3-2.2)	1.7 (1.3-2.2)	0.009	2.2 (1.7-2.8)	2.1 (1.6-2.7)	0.056
Mechanical power, J/min/kg ^d	0.3 (0.2-0.3)	0.3 (0.2-0.3)	-0.091	0.3 (0.2-0.4)	0.3 (0.2-0.4)	0.049

Abbreviations: AI, artificial intelligence; APACHE, Acute Physiology and Chronic Health Evaluation; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; SMD, standardized mean difference; SNUH, Seoul National University Hospital; SOFA, Sequential Organ Failure Assessment.

Values are presented as weighted n (%) for categorical variables and weighted mean $\pm$ SD or weighted median (IQR) for continuous variables.

^a: The n values indicate weighted sample sizes, calculated as the sum of the final stabilized inverse probability weights after truncation at the 99th percentile and rounded to the nearest integer.

^b: Includes interstitial lung disease or other chronic lung diseases, excluding COPD.

^c: Defined as (minute ventilation $\times$ PaCO₂)/(ideal body weight $\times$ 100 $\times$ 37.5).

^d: Values were normalized with ideal body weight.

Table S7. Fitted Q-evaluation results

Reward type	AI	Without intensivist feedback	Without high-resolution train data	Without intensivist feedback and high-resolution train data	Clinician
SNUH internal validation cohort					
Overall	-12.6 (-13.7 to -11.5)	-13.0 (-14.2 to -11.8)	-12.9 (-14.3 to -11.7)	-13.3 (-14.9 to -12.3)	-14.2 (-15.3 to -13.0)
VFDs, days ^a	16.2 (15.4 to 17.1)	15.7 (14.7 to 16.8)	15.9 (14.4 to 17.0)	15.5 (13.9 to 16.7)	14.8 (13.4 to 16.1)
Preference	-8.6 (-10.2 to -6.9)	-10.3 (-11.9 to -9.0)	-8.9 (-10.9 to -6.5)	-10.6 (-11.9 to -9.1)	-10.1 (-11.4 to -8.8)
Mayo Clinic external validation cohort					
Overall	-11.0 (-11.9 to -10.2)	-11.6 (-12.4 to -10.8)	-11.5 (-12.4 to -10.7)	-12.1 (-12.6 to -11.5)	-12.3 (-12.9 to -11.5)
VFDs, days ^a	17.7 (16.9 to 18.5)	17.5 (16.4 to 18.1)	17.1 (16.2 to 17.7)	16.6 (15.9 to 16.9)	16.5 (15.7 to 17.1)
Preference	-7.6 (-8.1 to -7.0)	-8.0 (-8.5 to -7.6)	-7.9 (-8.6 to -7.4)	-8.2 (-8.7 to -7.8)	-8.4 (-9.0 to -7.4)

Abbreviations: SNUH, Seoul National University Hospital; VFDs, ventilator-free days.

The clinician behavior policy return was estimated as the empirical mean of observed episode returns in the validation cohorts, whereas returns for VentPilot and ablated policies were estimated using fitted Q-evaluation. Uncertainty for all reported estimates was quantified by trajectory-level nonparametric bootstrapping.

^a: For the VFDs component, returns were shifted by +28 to match estimated VFDs. Other rows are reported on the original return scale. This linear shift does not affect between-policy comparisons.

Table S8. Recommendation and setting-adjustment frequency and timing in the condition-based analysis

Hospital	Control variable	Response direction	Episodes	Response frequency (AI)	Response frequency (Clinician)	P value	Both responded	Time to recommend (AI)	Time to adjustment (Clinician)	Time difference	P value
SNUH	sFiO ₂	Increase	1911	1340 (70.1)	740 (38.7)	<0.001	527	0.0 (0.0 to 0.0)	0.1 (0.0 to 0.7)	-0.0 (-0.4 to 0.0)	<0.001
SNUH	sFiO ₂	Decrease	11770	10973 (93.2)	3384 (28.8)	<0.001	3097	0.0 (0.0 to 0.0)	0.5 (0.0 to 1.2)	-0.5 (-1.1 to 0.0)	<0.001
SNUH	sPEEP	Increase	485	322 (66.4)	53 (10.9)	<0.001	38	0.0 (0.0 to 0.0)	0.1 (0.0 to 0.4)	-0.1 (-0.4 to 0.0)	<0.001
SNUH	sPEEP	Decrease	2543	1101 (43.3)	223 (8.8)	<0.001	102	0.0 (0.0 to 0.0)	1.0 (0.3 to 2.3)	-1.0 (-2.3 to -0.2)	<0.001
SNUH	sPC, sTV	Increase	1498	744 (49.7)	218 (14.6)	<0.001	147	0.0 (0.0 to 0.0)	0.4 (0.0 to 1.4)	-0.3 (-1.0 to 0.0)	<0.001
SNUH	sPC, sTV	Decrease	5454	4890 (89.7)	910 (16.7)	<0.001	856	0.0 (0.0 to 0.0)	1.0 (0.2 to 2.7)	-1.0 (-2.7 to -0.2)	<0.001
SNUH	sRR	Increase	2248	1011 (45.0)	379 (16.9)	<0.001	202	0.0 (0.0 to 0.0)	0.1 (0.0 to 1.0)	-0.1 (-1.0 to 0.0)	<0.001
SNUH	sRR	Decrease	4760	1149 (24.1)	915 (19.2)	<0.001	284	0.0 (0.0 to 0.0)	1.0 (0.1 to 2.0)	-1.0 (-2.0 to -0.1)	<0.001
SNUH	sPS	Increase	1079	368 (34.1)	87 (8.1)	<0.001	48	0.0 (0.0 to 0.0)	0.3 (0.2 to 1.0)	-0.3 (-1.0 to -0.1)	<0.001
SNUH	sPS	Decrease	8496	3636 (42.8)	893 (10.5)	<0.001	556	0.0 (0.0 to 0.0)	1.0 (0.0 to 2.0)	-1.0 (-2.0 to 0.0)	<0.001
MAYO	sFiO ₂	Increase	2431	717 (29.5)	475 (19.5)	<0.001	133	0.0 (0.0 to 0.0)	0.5 (0.2 to 1.5)	-0.5 (-1.2 to -0.2)	<0.001
MAYO	sFiO ₂	Decrease	1137	538 (47.3)	312 (27.4)	<0.001	135	0.0 (0.0 to 0.0)	3.0 (1.0 to 7.9)	-2.5 (-6.0 to -0.8)	<0.001
MAYO	sPEEP	Increase	402	113 (28.1)	47 (11.7)	<0.001	9	0.0 (0.0 to 0.2)	1.0 (0.8 to 3.8)	-0.8 (-1.0 to -0.5)	0.004
MAYO	sPEEP	Decrease	421	195 (46.3)	59 (14.0)	<0.001	27	0.0 (0.0 to 0.3)	7.0 (2.7 to 14.1)	-4.7 (-13.0 to -2.6)	<0.001
MAYO	sPC, sTV	Increase	104	53 (51.0)	18 (17.3)	<0.001	12	0.0 (0.0 to 1.8)	2.4 (1.1 to 6.9)	-1.0 (-3.8 to -0.3)	0.002
MAYO	sPC, sTV	Decrease	803	799 (99.5)	83 (10.3)	<0.001	83	0.0 (0.0 to 0.0)	3.2 (1.0 to 7.5)	-3.2 (-7.5 to -1.0)	<0.001
MAYO	sRR	Increase	245	115 (46.9)	50 (20.4)	<0.001	30	0.0 (0.0 to 0.0)	0.6 (0.1 to 2.0)	-0.5 (-1.2 to -0.1)	<0.001
MAYO	sRR	Decrease	1056	244 (23.1)	152 (14.4)	<0.001	71	0.0 (0.0 to 0.0)	1.0 (0.0 to 2.9)	-1.0 (-2.1 to 0.0)	<0.001
MAYO	sPS	Increase	953	109 (11.4)	67 (7.0)	<0.001	16	0.0 (0.0 to 0.4)	1.0 (0.3 to 3.2)	-0.5 (-1.3 to -0.2)	0.002
MAYO	sPS	Decrease	2589	1336 (51.6)	182 (7.0)	<0.001	150	0.0 (0.0 to 0.0)	1.0 (0.1 to 3.0)	-1.0 (-3.0 to -0.1)	<0.001

Abbreviations: AI, artificial intelligence; SNUH, Seoul National University Hospital; MAYO, Mayo Clinic.

Condition episodes were defined using the physiologic conditions in Table S3. Only conditions persisting for at least 5 minutes were included, and each episode was defined as a contiguous period during which the condition persisted. The VentPilot recommendation and clinician setting-adjustment columns report the number (%) of episodes with an event in the condition-specific response direction. Timing analyses were restricted to episodes in which both events occurred. Time to recommendation, time to setting adjustment, and their difference are summarized as median (IQR) in hours, with the time difference defined as time to recommendation minus time to setting adjustment. P values for event frequencies and paired times were obtained using two-sided McNemar tests and two-sided Wilcoxon signed-rank tests, respectively.

Table S9. Fitted Q-evaluation in ASV-exposed patients in the external validation cohort

Reward type	AI	Clinician with ASV	Difference (95% CI)
Overall	-13.0 (-14.6 to -11.5)	-14.9 (-16.7 to -12.8)	1.9 (0.2 to 2.9)
VFDs, days ^a	15.3 (13.9 to 16.8)	14.3 (12.3 to 16.0)	1.0 (0.4 to 1.7)
Preference	-10.6 (-13.7 to -8.0)	-12.0 (-13.4 to -10.5)	1.4 (-0.3 to 2.9)

Abbreviations: AI, artificial intelligence; ASV, adaptive support ventilation; VFDs, ventilator-free days.

ASV exposure was defined as ASV use for at least 25% of the invasive mechanical ventilation period. The clinician behavior policy return was estimated as the empirical mean of observed trajectory returns within this subgroup, whereas the VentPilot policy return was estimated using fitted Q-evaluation. ^a: For the VFDs component, returns were shifted by +28 to match estimated VFDs. Other rows are reported on the original return scale. This linear shift does not affect between-policy comparisons.

Table S10. Covariate balance in ASV-exposed and AI-similar non-ASV groups

Variable	Before weighting			After weighting		
	AI-similar & non-ASV (n=359)	ASV-exposed (n=114)	SMD	AI-similar & non-ASV (n=356) ^a	ASV-exposed (n=110) ^a	SMD
Type of ICU admission						
Medical	179 (49.9)	84 (73.7)	-0.504	196 (54.9)	63 (57.2)	-0.048
Surgical	180 (50.1)	30 (26.3)	0.504	161 (45.1)	47 (42.7)	0.048
Age, y	64.4 ± 14.5	62.2 ± 14.8	0.147	64.0 ± 14.7	63.6 ± 14.9	0.022
Sex						
Male	207 (57.7)	61 (53.5)	0.083	202 (56.8)	62 (56.6)	0.004
Female	152 (42.3)	53 (46.5)	-0.083	154 (43.2)	48 (43.4)	-0.004
Body mass index, kg/m ²	31.0 ± 7.7	30.1 ± 7.6	0.116	30.8 ± 7.8	30.3 ± 7.9	0.065
Primary diagnosis						
Cardiovascular	85 (23.7)	18 (15.8)	0.199	79 (22.1)	21 (18.8)	0.083
Gastrointestinal	29 (8.1)	13 (11.4)	-0.112	31 (8.7)	8 (7.5)	0.044
Hepatobiliary	19 (5.3)	6 (5.3)	0.001	19 (5.4)	6 (5.1)	0.011
Neoplastic	7 (1.9)	3 (2.6)	-0.045	7 (2.0)	2 (1.9)	0.014
Neurologic	38 (10.6)	14 (12.3)	-0.053	40 (11.3)	15 (13.6)	-0.070
Renal	6 (1.7)	4 (3.5)	-0.115	7 (1.9)	2 (1.6)	0.022
Respiratory	38 (10.6)	15 (13.2)	-0.079	40 (11.1)	15 (13.7)	-0.076
Sepsis, including pneumonia	46 (12.8)	13 (11.4)	0.043	45 (12.6)	12 (11.2)	0.042
Other	91 (25.3)	28 (24.6)	0.018	89 (24.9)	29 (26.8)	-0.041
Comorbidities						
Cardiovascular disease	264 (73.5)	87 (76.3)	-0.064	262 (73.7)	82 (74.7)	-0.024
Diabetes	132 (36.8)	39 (34.2)	0.053	128 (36.0)	40 (36.3)	-0.008
COPD	80 (22.3)	21 (18.4)	0.096	78 (22.0)	26 (23.3)	-0.030
Other chronic lung disease ^b	61 (17.0)	19 (16.7)	0.009	60 (16.9)	18 (16.8)	0.002

Chronic kidney disease	119 (33.1)	48 (42.1)	-0.185	123 (34.6)	41 (37.5)	-0.061
Malignancy	88 (24.5)	25 (21.9)	0.061	85 (23.8)	28 (25.3)	-0.034
P/F ratio, mmHg	267 (204-342)	276 (171-348)	-0.002	267 (204-342)	276 (182-352)	-0.020
SOFA score	10 (8-12)	10 (7-12)	-0.061	10 (8-12)	9 (7-11)	0.091
APACHE II score	25 (22-28)	25 (22-28)	-0.061	25 (22-28)	25 (23-28)	-0.065
Oxygenation Index	3.3 (2.3-4.6)	3.0 (2.2-5.7)	-0.116	3.3 (2.3-4.6)	2.9 (2.2-4.4)	0.026
Ventilatory ratio ^c	1.2 (1.0-1.4)	1.2 (1.0-1.4)	-0.048	1.2 (1.0-1.4)	1.2 (1.0-1.4)	0.033
Driving pressure, cmH ₂ O	13 (10-16)	13 (11-17)	-0.082	13 (11-16)	13 (11-17)	-0.031
Elastance, cmH ₂ O/(mL/kg) ^d	2.0 (1.5-2.5)	2.0 (1.7-2.5)	-0.011	2.0 (1.5-2.5)	2.1 (1.7-2.6)	-0.044
Mechanical power, J/min/kg ^d	0.3 (0.2-0.3)	0.3 (0.2-0.4)	-0.162	0.3 (0.2-0.3)	0.3 (0.2-0.3)	0.021

Abbreviations: AI, artificial intelligence; ASV, Adaptive Support Ventilation; APACHE, Acute Physiology and Chronic Health Evaluation; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; SMD, standardized mean difference; SOFA, Sequential Organ Failure Assessment.

Before weighting, values are presented as n (%) for categorical variables and mean $\pm$ SD or median (IQR) for continuous variables. After weighting, categorical variables are reported as weighted n (%) and continuous variables as weighted mean $\pm$ SD or weighted medians (IQR). ASV-exposed patients were defined as those with ASV use for at least 25% of the invasive mechanical ventilation period. Non-ASV patients had no documented ASV use during the invasive mechanical ventilation period.

^a: The n values indicate weighted sample sizes, calculated as the sum of the final stabilized inverse probability weights after truncation at the 99th percentile and rounded to the nearest integer.

^b: Includes interstitial lung disease or other chronic lung diseases, excluding COPD.

^c: Defined as (minute ventilation $\times$ PaCO₂)/(ideal body weight $\times$ 100 $\times$ 37.5).

^d: Values were normalized with ideal body weight.

Table S11. Weighted outcomes in ASV-exposed and AI-similar non-ASV groups

Outcome	AI-similar & non-ASV (n = 356) ^a	ASV-exposed (n = 110) ^a	Absolute difference (95% CI) ^b
Primary outcome - ventilator-free days (within 28 days)			
Mean	19.0 ± 10.8	15.1 ± 11.3	3.9 (1.6 to 6.3)
Median	24.9 (20.7 to 25.9)	21.2 (0.0 to 25.1)	3.7 (1.5 to 6.0)
Secondary outcomes			
Ventilation duration among ICU survivors, days	2.6 (1.9 to 3.6)	3.7 (2.7 to 6.7)	-1.0 (-2.0 to -0.5)
Ventilation duration among in-hospital survivors, days	2.6 (1.9 to 3.6)	3.6 (2.7 to 6.9)	-1.0 (-2.2 to -0.3)
28-day mortality, n (%)	86 (24.1)	38 (34.5)	-10.4 (-20.4 to -0.5)
ICU mortality, n (%)	51 (14.5)	23 (20.6)	-6.1 (-14.5 to 2.3)
In-hospital mortality, n (%)	87 (24.3)	34 (30.7)	-6.4 (-16.1 to 3.4)
ICU length of stay, days			
Total	4.5 (3.0 to 6.0)	5.7 (4.2 to 8.7)	-1.2 (-1.8 to -0.4)
ICU survivors	4.6 (3.1 to 6.0)	5.9 (4.0 to 9.5)	-1.3 (-2.3 to -0.2)
ICU non-survivors	3.7 (2.7 to 6.1)	5.7 (4.3 to 6.1)	-2.0 (-2.7 to -0.0)
Hospital length of stay, days			
Total	13.8 (8.2 to 22.9)	12.8 (6.9 to 22.7)	0.9 (-2.8 to 4.7)
Hospital survivors	16.6 (10.0 to 28.7)	16.4 (9.8 to 24.9)	0.2 (-3.6 to 4.0)
Hospital non-survivors	6.9 (3.7 to 13.6)	6.9 (5.2 to 11.2)	0.0 (-2.9 to 1.8)

Abbreviations: AI, artificial intelligence; ASV, Adaptive Support Ventilation; CI, confidence interval; ICU, intensive care unit.

Values after weighting are presented as weighted n (%) for binary outcomes and weighted mean ± SD or weighted median (IQR) for continuous outcomes. The primary outcome is presented as both weighted mean ± SD and weighted median (IQR); other continuous outcomes are presented as weighted median (IQR).

^a: The n values indicate weighted sample sizes, calculated as the sum of the final stabilized inverse probability weights after truncation at the 99th percentile and rounded to the nearest integer.

^b: For continuous outcomes, absolute differences are in days. For binary outcomes, absolute differences are risk differences expressed in percentage points.

Table S12. Subgroup analysis results

Subgroup	No. of patients (%)	Difference in fitted Q-evaluation return	Difference in mean VFDs after IPW
SNUH internal validation cohort			
Age, y			
< 65	92 (32.5)	1.8 (0.7 to 2.7)	4.4 (1.6 to 8.8)
≥ 65	191 (67.5)	1.5 (1.1 to 1.9)	4.9 (0.4 to 10.5)
Sex			
Male	177 (62.5)	1.5 (0.3 to 2.7)	4.5 (0.1 to 8.5)
Female	106 (37.5)	1.7 (0.6 to 2.9)	6.8 (2.0 to 10.6)
P/F ratio			
< 200	75 (26.5)	2.2 (0.5 to 3.9)	6.7 (1.3 to 11.4)
≥ 200	208 (73.5)	1.5 (1.1 to 1.8)	3.9 (0.6 to 7.3)
Mayo Clinic external validation cohort			
Age, y			
< 65	486 (48.1)	1.0 (0.2 to 2.1)	3.8 (0.9 to 5.9)
≥ 65	525 (51.9)	1.5 (1.0 to 1.9)	3.5 (1.7 to 5.8)
Sex			
Male	566 (56.0)	1.3 (0.5 to 1.9)	3.4 (0.9 to 5.2)
Female	445 (44.0)	1.2 (0.4 to 1.8)	3.8 (1.4 to 5.7)
P/F ratio			
< 200	347 (34.3)	1.7 (1.2 to 2.1)	4.5 (2.1 to 6.2)
≥ 200	664 (65.7)	0.8 (0.5 to 1.1)	3.1 (1.3 to 5.1)
Primary diagnosis & comorbidities			
Respiratory	131 (13.0)	1.3 (0.2 to 2.9)	7.8 (3.9 to 11.5)
Sepsis, including pneumonia	149 (14.7)	1.6 (0.3 to 2.5)	3.0 (0.7 to 7.8)
Cardiovascular disease	725 (71.7)	1.4 (0.1 to 2.2)	3.3 (1.6 to 4.9)
COPD	245 (24.2)	1.5 (0.5 to 2.9)	3.6 (0.7 to 6.5)

Abbreviations: AI, artificial intelligence; COPD, chronic obstructive pulmonary disease; IPW, inverse probability weighting; SNUH, Seoul National University Hospital; VFDs, ventilator-free days.

The fitted Q-evaluation column reports the estimated return of the AI policy minus the return of the clinician policy, and the VFDs column reports the inverse-probability-weighted difference in mean ventilator-free days within 28 days (AI-similar minus AI-dissimilar). Values are shown with 95% confidence intervals. Subgroup analyses were performed across age, sex, P/F ratio. Additional diagnosis- and comorbidity-defined subgroup analyses were performed only in the Mayo Clinic external validation cohort because subgroup-specific re-estimation was considered less stable in the smaller SNUH internal validation cohort. Respiratory diagnosis and COPD were selected because of their direct relevance to ventilator management, sepsis including pneumonia because it represents a common critical care syndrome in mechanically ventilated ICU patients, and cardiovascular disease because it is a prevalent comorbidity that may influence hemodynamics and ventilatory tolerance. Within each subgroup, fitted Q-evaluation and inverse probability-weighted analyses were repeated using the same analytic framework as in the primary analyses, with propensity scores re-estimated within each subgroup.

Table S13. Sensitivity analysis results

Analysis	Difference in OPE return	Difference in mean VFDs after IPW
SNUH internal validation cohort		
Primary analysis	1.6 (0.8 to 2.5)	5.0 (2.4 to 7.6)
Time grid for OPE		
5-min fixed grid	1.3 (0.7 to 1.8)	-
60-min fixed grid	1.6 (0.8 to 2.2)	-
Alternative OPE estimator		
Weighted importance sampling	1.7 (0.4 to 3.1)	-
Time points for agreement assessment in IPW analysis		
15-min fixed grid	-	4.8 (2.5 to 6.7)
Concordance assessment window for IPW analysis		
Fixed 24h window, 12 to 36 h after intubation	-	4.1 (1.0 to 6.7)
AI-similarity threshold for IPW analysis		
40%/40% split	-	7.9 (4.9 to 10.6)
30%/30% split	-	9.4 (6.2 to 12.5)
Preference reward weighting		
0.03	0.9 (0.3 to 1.5)	4.8 (2.3 to 8.6)
0.3	1.7 (0.9 to 2.2)	5.1 (2.7 to 8.5)
Policy-extraction sampling ratio (change:no-change)		
40:60	1.7 (0.9 to 2.5)	4.2 (1.9 to 6.5)
60:40	1.6 (0.9 to 2.2)	5.3 (2.7 to 8.0)
Mayo Clinic external validation cohort		
Primary analysis	1.3 (0.6 to 2.0)	3.1 (1.7 to 4.6)
Time grid for OPE		
5-min fixed grid	1.0 (0.3 to 1.7)	-
60-min fixed grid	1.0 (0.1 to 1.8)	-
Alternative OPE estimator		
Weighted importance sampling	1.2 (0.3 to 2.0)	-
Time points for agreement assessment in IPW analysis		
15-min fixed grid	-	3.2 (1.8 to 4.6)

Concordance assessment window for IPW analysis		
Fixed 24h window, 12 to 36 h after intubation	-	2.8 (1.1 to 4.2)
AI-similarity threshold for IPW analysis		
40%/40% split	-	4.1 (2.6 to 5.7)
30%/30% split	-	6.2 (4.5 to 7.9)
Preference reward weighting		
0.03	1.2 (0.6 to 1.8)	3.1 (1.9 to 4.8)
0.3	1.2 (0.5 to 2.0)	3.0 (1.5 to 4.6)
Policy-extraction sampling ratio (change:no-change)		
40:60	0.8 (0.2 to 2.4)	2.9 (1.4 to 4.3)
60:40	1.4 (0.5 to 2.3)	3.0 (1.4 to 5.1)
Ventilation mode for concordance assessment in IPW analysis		
Pressure-targeted modes only	-	2.7 (1.3 to 4.1)

Abbreviations: AI, artificial intelligence; IPW, inverse probability weighting; OPE, off-policy evaluation; SNUH, Seoul National University Hospital; VFDs, ventilator-free days.

Values in the off-policy evaluation column represent the estimated return of the VentPilot policy minus that of the clinician behavior policy. Unless otherwise specified, off-policy evaluation estimates were obtained using fitted Q-evaluation. The weighted importance sampling (WIS) row provides an alternative off-policy evaluation estimate for the primary comparison between VentPilot and the clinician behavior policy using the same overall reward definition [14]. Whereas fitted Q-evaluation estimates policy value by fitting value functions to observed transitions, WIS estimates policy value by reweighting observed trajectories according to their alignment with the target policy. Because the extracted tree policy was deterministic, WIS used a stochastic approximation induced from the learned trees, with target-policy probabilities derived from advantage-weighted empirical action frequencies within the corresponding terminal leaves. Multivariable action probabilities were factorized as the product of component-wise probabilities across mode-applicable control variables to avoid sparse estimation of the high-dimensional joint action distribution. To aid interpretation of the WIS estimate, we report the effective sample size (ESS), which was 148.7 of 283 trajectories (52.5%) in the SNUH cohort and 429.2 of 1,011 trajectories (42.5%) in the Mayo Clinic cohort [26]. Larger ESS values indicate that the WIS estimate is supported by a broader effective set of weighted trajectories and is therefore less influenced by a small number of highly weighted trajectories. Values in the VFD column represent the weighted mean difference in ventilator-free days within 28 days between AI-similar and AI-dissimilar patients.

Primary analyses used a 50:50 change-to-no-change sampling ratio during policy extraction, a 15-minute fixed validation grid for off-policy evaluation, chart-update timestamps defined by newly recorded electronic medical record data for concordance assessment, concordance calculated from 24 hours after intubation until extubation or day 28, a 50%/50% threshold for AI-similar versus AI-dissimilar classification, and a default preference-reward weight of 0.1. Sensitivity analyses evaluated 5-minute and 60-minute fixed grids for off-policy evaluation, weighted importance sampling as an alternative off-policy estimator, concordance reassessment on the 15-minute fixed grid, a fixed early-window analysis using baseline covariates from 0–12 hours and concordance from 12–36 hours after intubation, stricter discrepancy-score thresholds comparing the lowest 40% with the highest 40% and the lowest 30% with the highest 30% of patients within each validation cohort, alternative preference-reward weights, alternative change-to-no-change sampling ratios during policy extraction, and concordance restricted to pressure-targeted ventilation periods in the Mayo Clinic cohort.

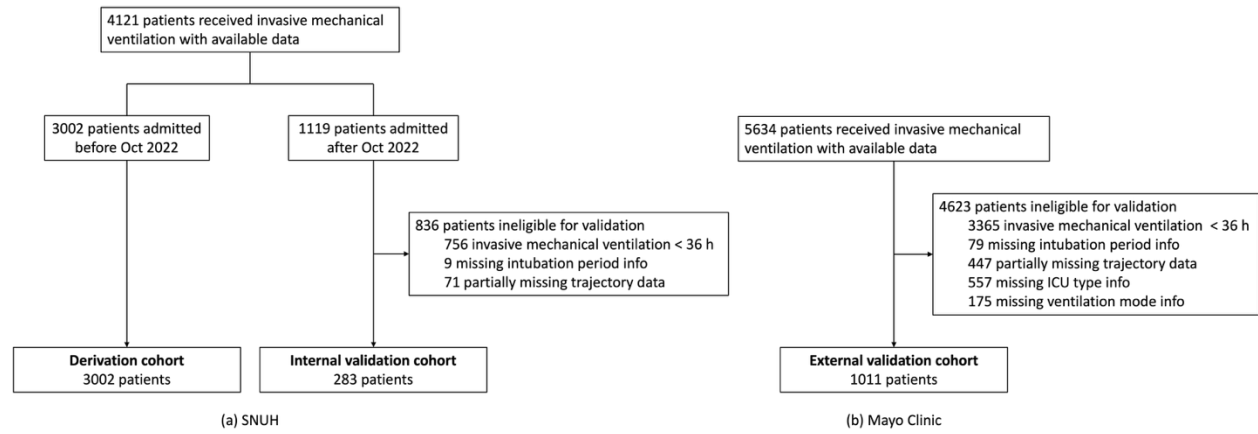
Table S14. Calibration of fitted Q-evaluation against empirical clinician returns

Cohort	Empirical clinician return	FQE-estimated clinician return	Difference (95% CI)
SNUH	-14.2 (-15.3 to -13.0)	-14.4 (-15.6 to -13.4)	-0.2 (-1.4 to 1.1)
Mayo Clinic	-12.3 (-12.9 to -11.5)	-12.6 (-13.9 to -11.5)	-0.3 (-0.7 to 0.2)

Abbreviations: FQE, fitted Q-evaluation; SNUH, Seoul National University Hospital.

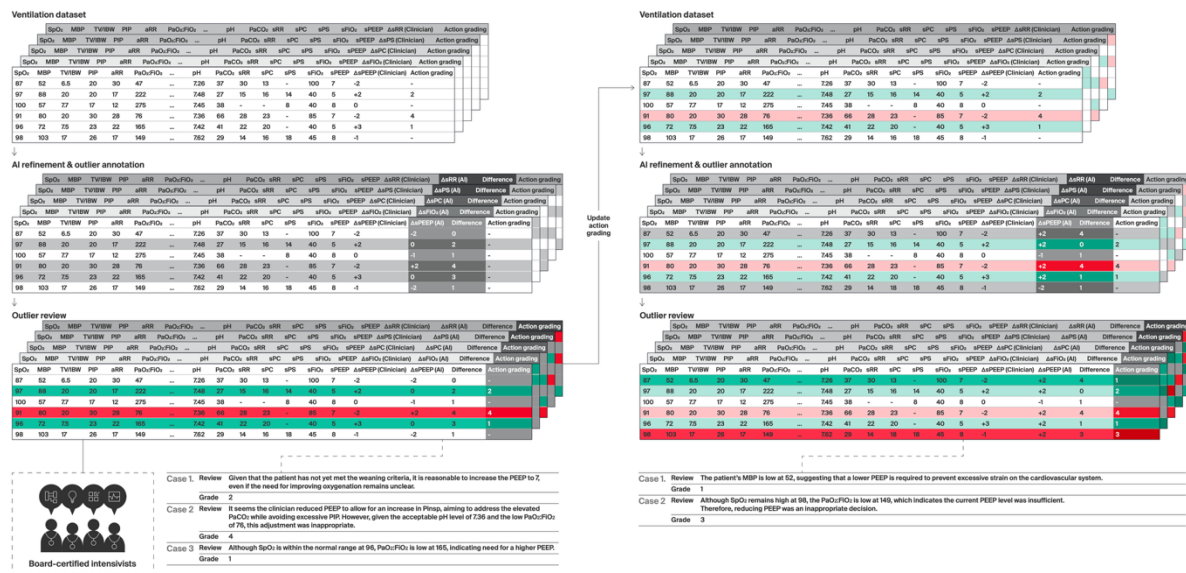
The empirical clinician return was calculated as the mean observed episode return. For the FQE estimate, the evaluation critic was fit using the observed subsequent clinician action in the Bellman target. Differences are FQE-estimated minus empirical returns. Confidence intervals were obtained using trajectory-level nonparametric bootstrapping.

Figure S1. Dataset flowchart



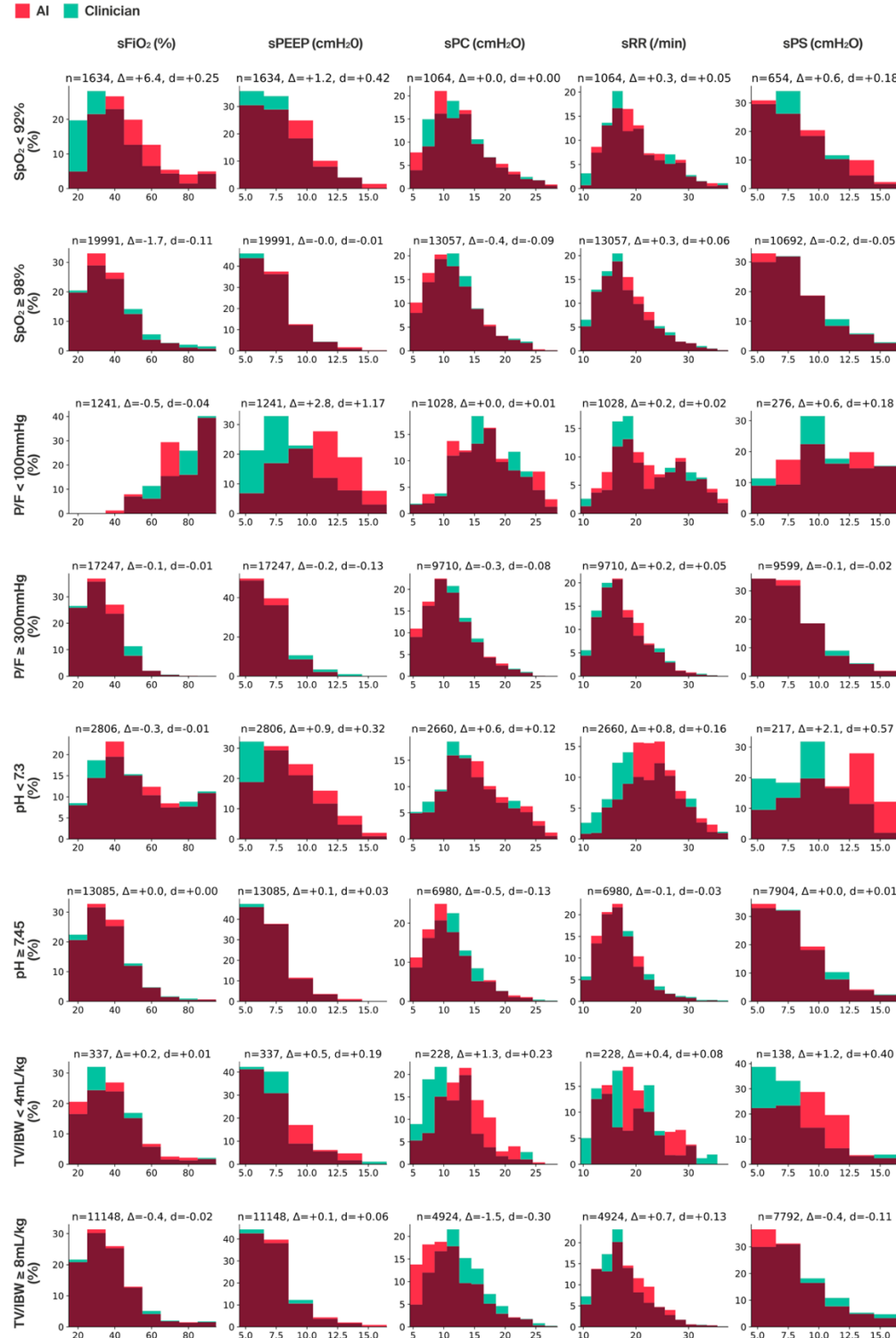
Abbreviations: ICU, intensive care unit.

Figure S2. Iterative intensivist-feedback loop for collecting preference labels



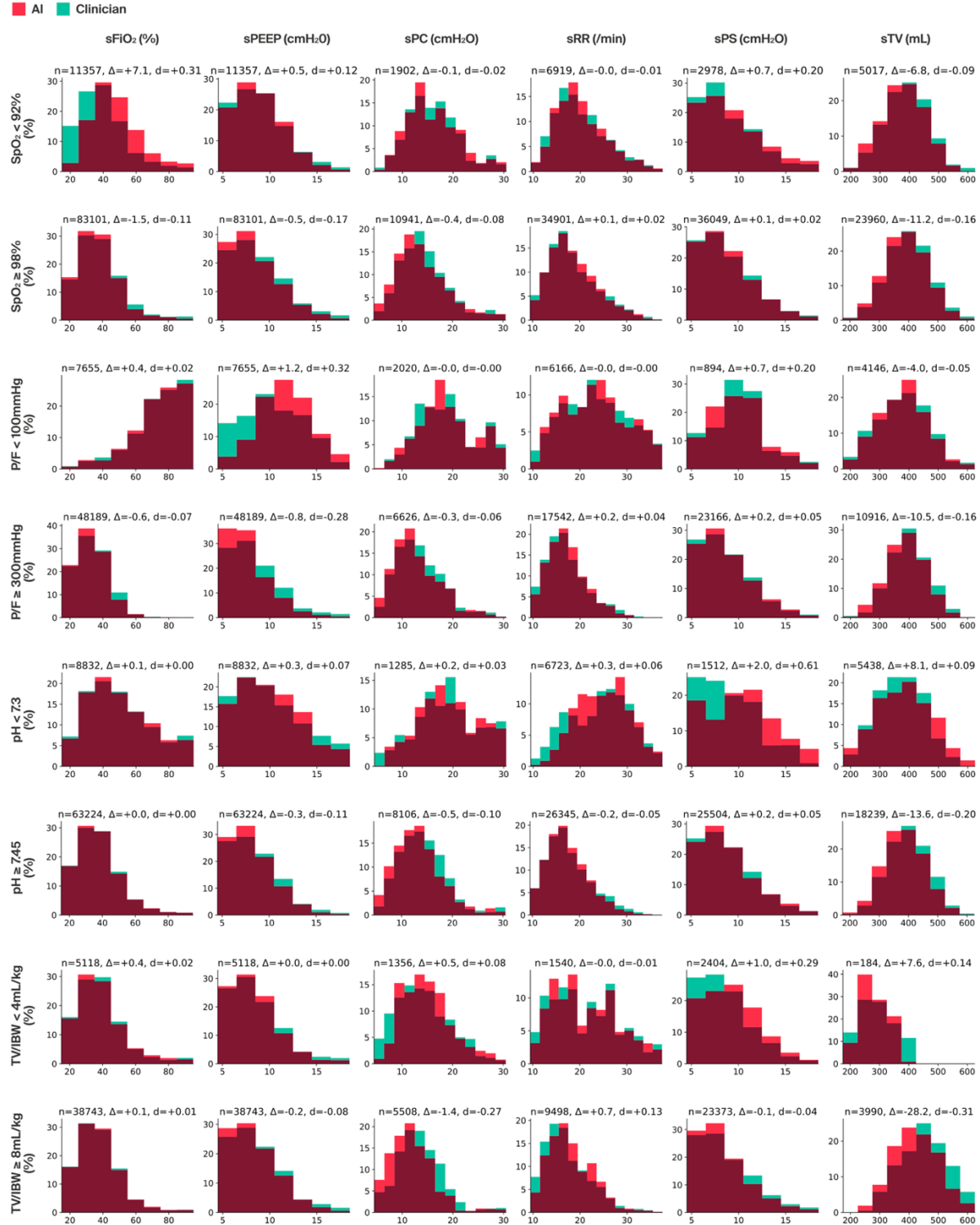
A behavior-cloning model trained on the derivation cohort was used to identify discrepancy (outlier) events where model-recommended ventilator setpoints differed substantially from observed clinician settings. For each outlier event, a committee of board-certified intensivists reviewed the full clinical context and graded the observed clinician action on a 1–4 scale. The newly labeled cases were then added to the review dataset, enabling iterative refinement of the labeled set used to train the preference model.

Figure S3. Ventilator setting comparisons across physiologic strata in the internal validation cohort



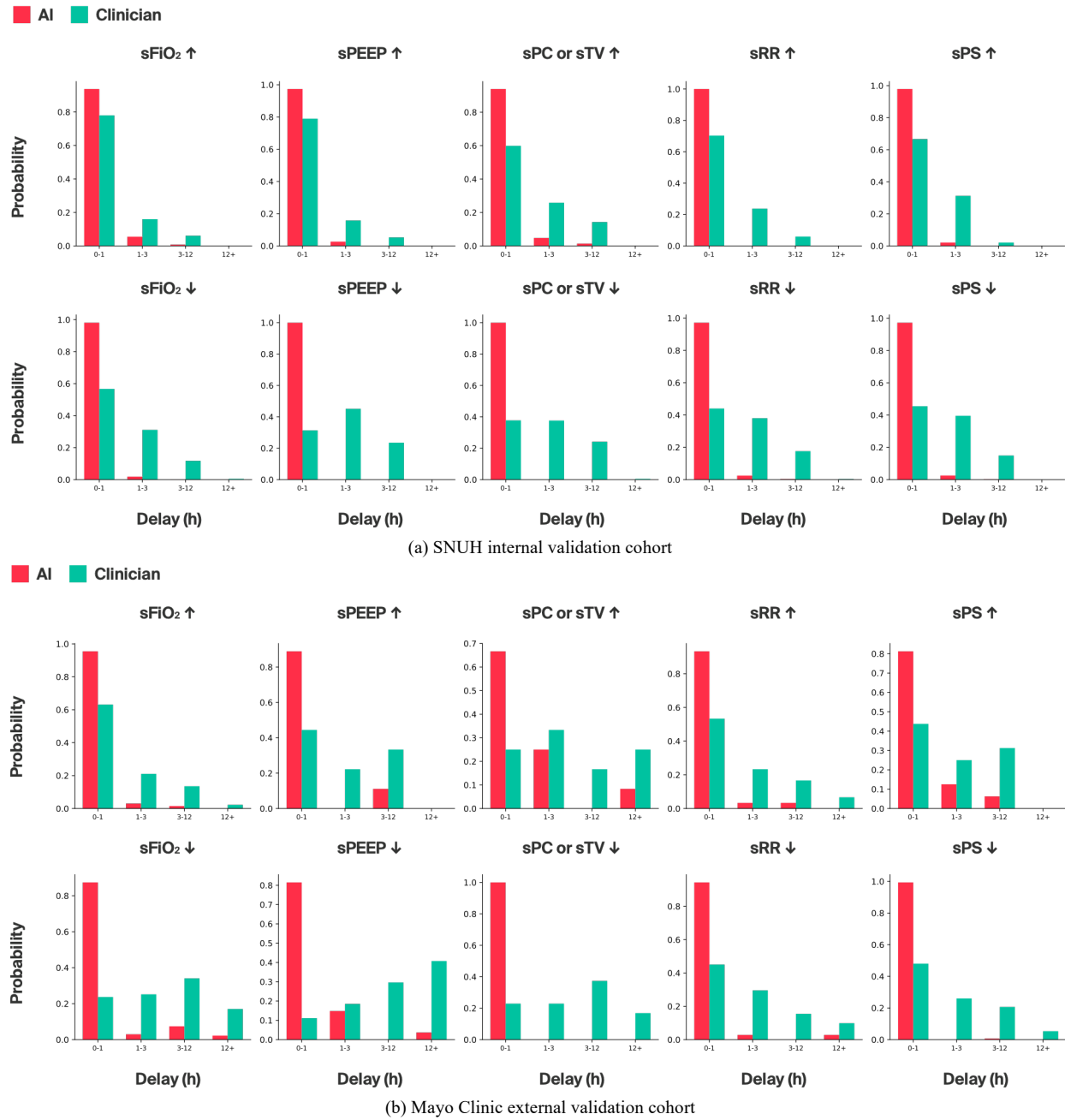
Overlaid distributions of observed settings and VentPilot recommendations across clinically motivated physiologic strata in the SNUH internal validation cohort. Rows denote SpO₂ < 92%, SpO₂ ≥ 98%, P/F ratio < 100 mmHg, P/F ratio ≥ 300 mmHg, pH < 7.3, pH ≥ 7.45, TV/IBW < 4 mL/kg, and TV/IBW ≥ 8 mL/kg. Columns denote sFiO₂, sPEEP, sPC, sRR, and sPS. Each panel reports the number of eligible decision points (n), the mean difference (Δ), and Cohen's d. Positive values indicate higher settings recommended by VentPilot, whereas negative values indicate lower settings.

Figure S4. Ventilator setting comparisons across physiologic strata in the external validation cohort



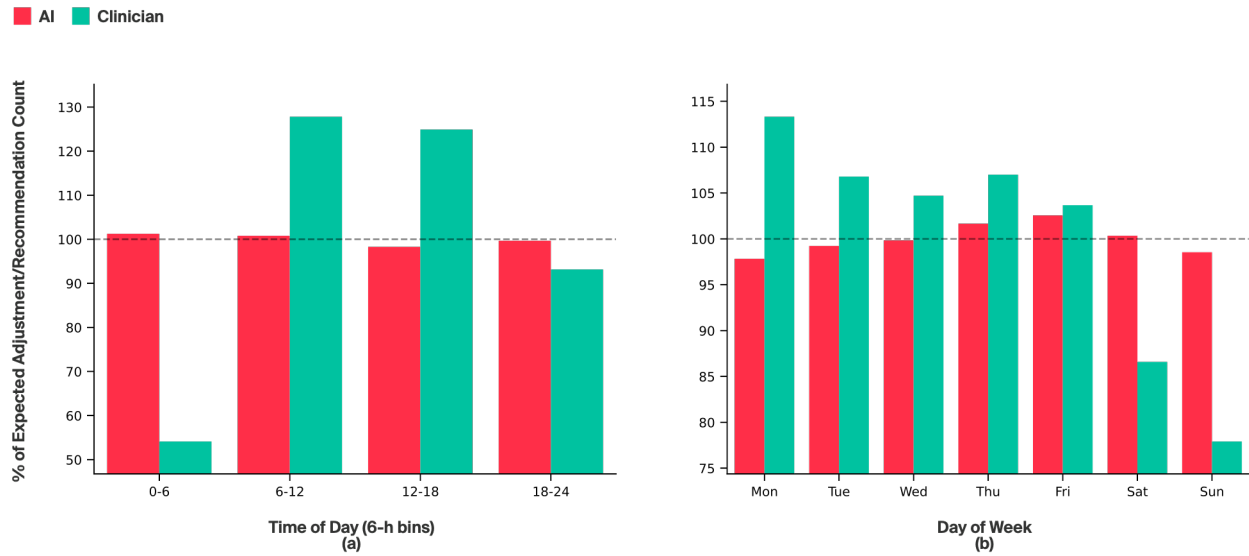
Overlaid distributions of observed settings and VentPilot recommendations across clinically motivated physiologic strata in the Mayo Clinic external validation cohort. Rows denote SpO₂ <92%, SpO₂ ≥98%, P/F ratio <100 mmHg, P/F ratio ≥300 mmHg, pH <7.3, pH ≥7.45, TV/IBW <4 mL/kg, and TV/IBW ≥8 mL/kg. Columns denote sFiO₂, sPEEP, sPC, sRR, sPS, and sTV. Each panel reports the number of eligible decision points (n), the mean difference (Δ), and Cohen's d. Positive values indicate higher settings recommended by VentPilot, whereas negative values indicate lower settings.

Figure S5. Timing of VentPilot recommendations and clinician setting adjustments



Distribution of time from condition onset to the first VentPilot recommendation in the condition-specific response direction and the first observed clinician adjustment of the relevant ventilator setting in the same direction, among condition episodes in which both events occurred. Bars indicate the proportion of episodes with times of 0–1, 1–3, 3–12, or >12 hours. Physiologic conditions and corresponding directions are defined in Table S3, and only conditions persisting for at least 5 minutes were included.

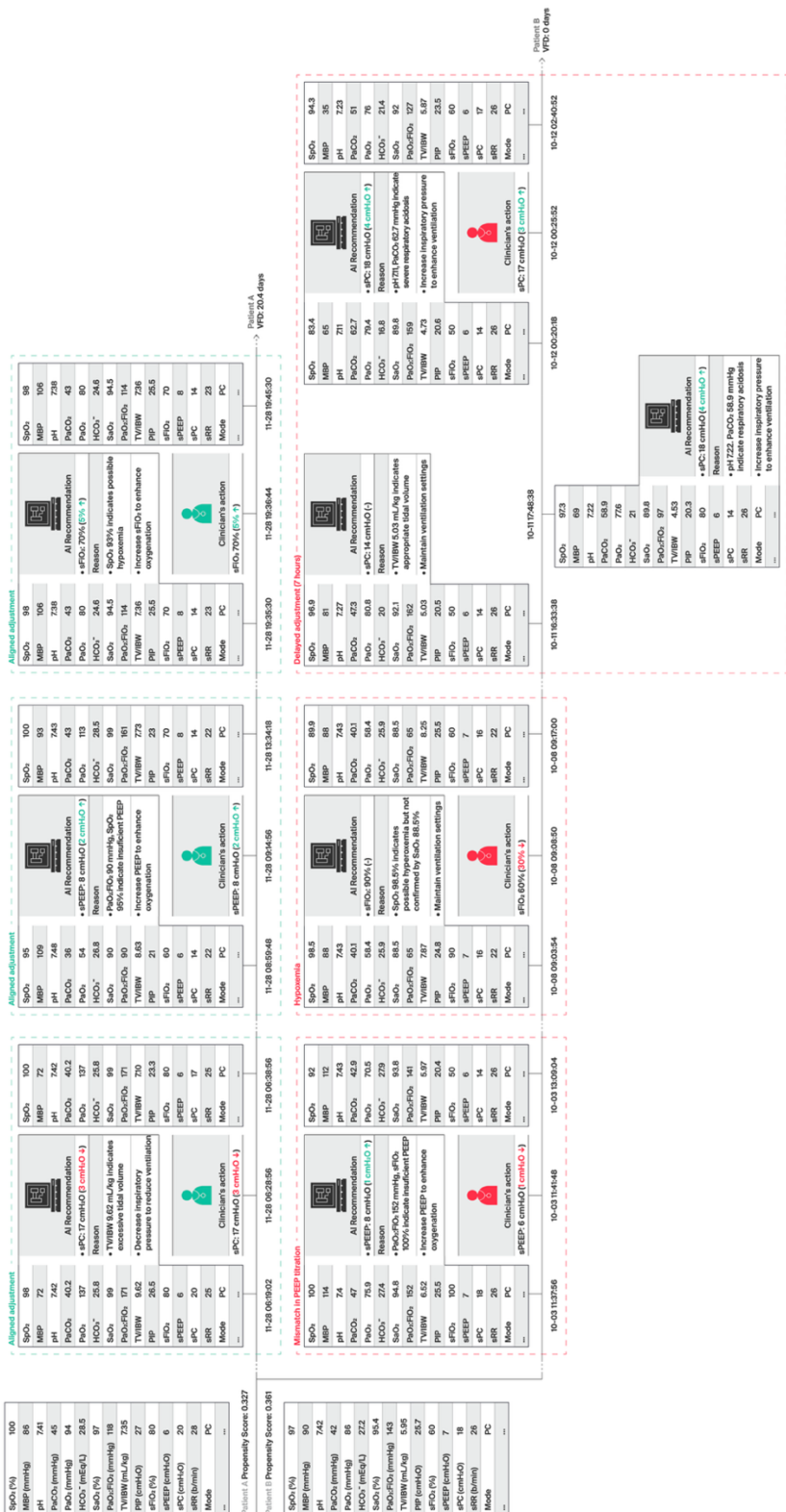
Figure S6. Temporal consistency results



(a) Time-of-day variation (SNUH). Frequency of VentPilot recommendation events and clinician setting-adjustment events across 6-hour time-of-day bins (00:00–05:59, 06:00–11:59, 12:00–17:59, 18:00–23:59). Bars show observed event counts expressed as a percentage of the expected count under a uniform-in-time assumption (100% dashed line). Clinician events are timestamps with any observed change in ventilator control variables; VentPilot events are timestamps with any recommended change in at least one control variable.

(b) Day-of-week variation (SNUH). Frequency of events across days of the week shown on the same observed/expected scale.

Figure S7. An example case of VentPilot recommendations for ventilator settings



For patient A, the clinician's actual settings were closely aligned with the AI's recommendations. In contrast, for patient B, the clinician's adjustments differed from the AI's recommendations. For instance, in the first event, the AI recommended increasing the current set PEEP, as the P/F ratio and set FiO₂ indicated the current PEEP level is insufficient. However, the clinician reduced the set PEEP, followed by a sudden drop in P/F ratio and SpO₂. The second and third timestamps highlight additional instances where the clinician's actions either contradicted or significantly lagged behind the AI's recommendations.

Although patients A and B had similar initial conditions, with propensity scores of 0.327 and 0.361, respectively, patient A—whose treatment was similar to the AI's suggestions—had 20.4 VFDs, while patient B, whose treatment differed from the AI's recommendations, had 0 VFDs.

References

1. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med. Nature Publishing Group*; 2018;24:1716–20. <https://doi.org/10.1038/s41591-018-0213-5>
2. Hug CW (Caleb W. Detecting hazardous intensive care patient episodes using real-time mortality models [Internet] [Thesis]. Massachusetts Institute of Technology; 2009 [cited 2024 May 2]. <https://dspace.mit.edu/handle/1721.1/53290>. Accessed 2 May 2024
3. Puterman ML. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.; 1994. <https://doi.org/10.1002/9780470316887>
4. Kostrikov I, Nair A, Levine S. Offline Reinforcement Learning with Implicit Q-Learning [Internet]. arXiv; 2021 [cited 2025 Dec 31]. <https://doi.org/10.48550/arXiv.2110.06169>
5. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting Deep Learning Models for Tabular Data. *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2021 [cited 2026 Apr 9]. p. 18932–43. <https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html>. Accessed 9 Apr 2026
6. Bertsimas D, Dunn J. Optimal classification trees. *Mach Learn*. 2017;106:1039–82. <https://doi.org/10.1007/s10994-017-5633-9>
7. Lin J, Zhong C, Hu D, Rudin C, Seltzer M. Generalized and Scalable Optimal Sparse Decision Trees. *Proceedings of the 37th International Conference on Machine Learning* [Internet]. PMLR; 2020 [cited 2024 May 2]. p. 6150–60. <https://proceedings.mlr.press/v119/lin20g.html>. Accessed 2 May 2024
8. McTavish H, Zhong C, Achermann R, Karimalis I, Chen J, Rudin C, et al. Fast Sparse Decision Tree Optimization via Reference Ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence* [Internet]. 2022 [cited 2024 May 2]. p. 9604–13. <https://doi.org/10.1609/aaai.v36i9.21194>
9. Peng XB, Kumar A, Zhang G, Levine S. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning [Internet]. arXiv; 2019 [cited 2026 Apr 9]. <https://doi.org/10.48550/arXiv.1910.00177>
10. Semler MW, Casey JD, Lloyd BD, Hastings PG, Hays MA, Stollings JL, et al. Oxygen-Saturation Targets for Critically Ill Adults Receiving Mechanical Ventilation. *New England Journal of Medicine. Massachusetts Medical Society*; 2022;387:1759–69. <https://doi.org/10.1056/NEJMoa2208415>
11. Fan E, Del Sorbo L, Goligher EC, Hodgson CL, Munshi L, Walkey AJ, et al. An Official American Thoracic Society/European Society of Intensive Care Medicine/Society of Critical Care Medicine Clinical Practice Guideline: Mechanical Ventilation in Adult Patients with Acute Respiratory Distress Syndrome. *Am J Respir Crit Care Med. American Thoracic Society - AJRCCM*; 2017;195:1253–63. <https://doi.org/10.1164/rccm.201703-0548ST>
12. Grasselli G, Calfee CS, Camporota L, Poole D, Amato MBP, Antonelli M, et al. ESICM guidelines on acute respiratory distress syndrome: definition, phenotyping and respiratory support strategies. *Intensive Care Med*. 2023;49:727–59. <https://doi.org/10.1007/s00134-023-07050-7>
13. Qadir N, Sahetya S, Munshi L, Summers C, Abrams D, Beitler J, et al. An Update on Management of Adult Patients with Acute Respiratory Distress Syndrome: An Official American Thoracic Society Clinical Practice Guideline. *Am J Respir Crit Care Med*. 2024;209:24–36. <https://doi.org/10.1164/rccm.202311-2011ST>
14. Voloshin C, Le HM, Jiang N, Yue Y. Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning [Internet]. arXiv; 2021 [cited 2025 Dec 31]. <https://doi.org/10.48550/arXiv.1911.06854>
15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Internet]. 2016 [cited 2026 Jan 1]. p. 785–94. <https://doi.org/10.1145/2939672.2939785>
16. Marshall R. The physical properties of the lungs in relation to the subdivisions of lung volume. *Clin Sci*. 1957;16:507–15.

17. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34:3661–79. <https://doi.org/10.1002/sim.6607>
18. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46:399–424. <https://doi.org/10.1080/00273171.2011.568786>
19. Cole SR, Hernán MA. Constructing Inverse Probability Weights for Marginal Structural Models. *Am J Epidemiol*. 2008;168:656–64. <https://doi.org/10.1093/aje/kwn164>
20. Yehya N, Harhay MO, Curley MAQ, Schoenfeld DA, Reeder RW. Reappraisal of Ventilator-Free Days in Critical Care Research. *American Journal of Respiratory and Critical Care Medicine*. 2019;200:828. <https://doi.org/10.1164/rccm.201810-2050CP>
21. Kirakli C, Naz I, Ediboglu O, Tatar D, Budak A, Tellioglu E. A randomized controlled trial comparing the ventilation duration between adaptive support ventilation and pressure assist/control ventilation in medical patients in the ICU. *Chest*. 2015;147:1503–9. <https://doi.org/10.1378/chest.14-2599>
22. Chu DK, Kim LH-Y, Young PJ, Zamiri N, Almenawer SA, Jaeschke R, et al. Mortality and morbidity in acutely ill adults treated with liberal versus conservative oxygen therapy (IOTA): a systematic review and meta-analysis. *Lancet*. 2018;391:1693–705. [https://doi.org/10.1016/S0140-6736\(18\)30479-3](https://doi.org/10.1016/S0140-6736(18)30479-3)
23. Boles J-M, Bion J, Connors A, Herridge M, Marsh B, Melot C, et al. Weaning from mechanical ventilation. *Eur Respir J*. 2007;29:1033–56. <https://doi.org/10.1183/09031936.00010206>
24. Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, et al. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N Engl J Med*. 2000;342:1301–8. <https://doi.org/10.1056/NEJM200005043421801>
25. Pletsch-Assuncao R, Caleffi Pereira M, Ferreira JG, Cardenas LZ, de Albuquerque ALP, de Carvalho CRR, et al. Accuracy of Invasive and Noninvasive Parameters for Diagnosing Ventilatory Overassistance During Pressure Support Ventilation. *Crit Care Med*. 2018;46:411–7. <https://doi.org/10.1097/CCM.0000000000002871>
26. Martino L, Elvira V, Louzada F. Effective Sample Size for Importance Sampling based on discrepancy measures. *Signal Processing*. 2017;131:386–401. <https://doi.org/10.1016/j.sigpro.2016.08.025>