MEGA ROX individualised oxygen treatment effects

Statistical analysis plan

Draft version 1.3 | 10 November 2026
Chief investigator Professor Paul Young
Trial registration ACTRN12620000391976

1 Background and rationale

Patients may differ in their response to conservative versus liberal oxygen therapy. Baseline risk can influence absolute treatment benefit, while combinations of clinical characteristics may identify further variation in treatment effects. A previous machine-learning model was developed in the Pragmatic Investigation of Optimal Oxygen Targets (PILOT) trial and externally evaluated in the Intensive Care Unit Randomised Trial Comparing Two Approaches to Oxygen Therapy (ICU-ROX) [1]. The Mega Randomised Registry Trial Comparing Conservative versus Liberal Oxygenation Targets (MEGA-ROX) provides a substantially larger randomised population in which to develop new treatment-effect models and externally evaluate reduced-variable versions of our previously published model [2].

We will follow the general risk-modelling and effect-modelling approach used in the Low Oxygen Intervention for Cardiac Arrest Injury Limitation (LOGICAL) individualised oxygen-target analysis plan [3]. We will assess model discrimination, calibration and treatment effects in temporally later samples, using validation_1 for model selection and validation_2 for final evaluation after refitting the selected models using the combined training and validation_1 data [4].

2 Objectives and study populations

We will undertake five principal analyses under three objectives, all using MEGA-ROX patients. ANZ denotes Australia and New Zealand; APD denotes the Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database. APD linkage provides additional baseline predictors for the ANZ model. The final two analyses will externally evaluate reduced-variable versions of our JAMA model, developed using the original study data with predictors available in MEGA-ROX. All evaluation patients below are from MEGA-ROX.

Table 1 Analyses and populations

Analysis	Patients	Predictors	Training and temporal validation
1 Global model	40,003	Baseline MEGA-ROX study variables	Training 20,001; validation_1 10,001; validation_2 10,001. Refit: 30,002.
2a ANZ core model	25,720	Baseline MEGA-ROX study variables	Global split retained. Refit on ANZ training and validation_1 patients.
2b ANZ model with APD	Same 25,720	Study variables plus eligible baseline APD variables	Same patients, split and refitting sample as 2a. Assess added APD value.
3a Reduced JAMA models overall	27,172	Available baseline predictors, including the additional measurements; not the full published predictor set	All eligible patients. Models fixed using original study data; no MEGA-ROX training split.
3b Reduced JAMA models in ANZ	13,576	Same reduced-variable models as analysis 3a	Eligible ANZ patients from 3a; no MEGA-ROX training split.

The 27,172 patients are the subset of the global population in whom the seven additional baseline measurements were collected. This does not mean that every predictor required by the published JAMA model is available. The 13,576 patients in analysis 3b are included in both analysis 3a and the ANZ population; all ANZ patients are included in the global population. These overlapping counts are not additive. Global training, validation_1 and validation_2 counts follow the temporal 50:25:25 split in Section 5. The 30,002-patient refitting sample combines training and validation_1 and is not an additional patient group. ANZ and additional-measurement subsets retain the global assignments, so their proportions need not be 50:25:25. We will report final sample sizes, recruitment periods, APD linkage and predictor availability.

3 Outcome and treatment effect

The primary outcome will be in-hospital all-cause mortality up to 90 days after randomisation, using the parent trial definition [2]. Patients discharged alive from the index hospital before day 90 will be classified as survivors for this endpoint. We expect the primary outcome to be available for all 40,003 patients and do not plan to impute primary outcomes.

We will analyse patients according to their randomised treatment, regardless of adherence to the assigned strategy. The predicted individualised treatment effect will be the mortality risk with conservative oxygen minus the mortality risk with liberal oxygen. Negative values favour conservative oxygen and positive values favour liberal oxygen. We will report absolute effects in percentage points. These predictions describe average effects among patients with similar characteristics.

The published JAMA model predicts 28-day mortality. For the reduced-variable models, we will evaluate the association of their predictions with treatment effects on the MEGA-ROX primary outcome and, as a supporting analysis, derive 28-day in-hospital mortality from the trial death and discharge data. Where complete post-discharge vital status is available, we will report all-cause day-28 mortality separately.

4 Baseline predictors

4 1 MEGA ROX study variables

The global model and ANZ core model will use age, sex, primary admission diagnosis and surgical status recorded using the Intensive Care National Audit & Research Centre (ICNARC) Coding Method (ICM), baseline hypoxic ischaemic encephalopathy (HIE), sepsis, other acute brain pathology and suspected or confirmed COVID-19, baseline PaO_2 and FIO_2, and intervals between hospital admission, ICU admission and randomisation. We will use the cleaned and locked study database as supplied.

We will also derive two separate categorical candidate predictors from the baseline primary admission code using the 2018 five-part ICM, including its surgical Type [5]. Each will classify patients as having the highest, middle or lowest expected risk of early anaemia or low cardiac output (Tables 2 and 3). The groups will use only information available before randomisation and will be fixed before validation analyses. These are clinical hypotheses, not measured physiological states or validated risk scores. Missing or unmapped codes will be treated as unknown.

Table 2 Clinical groups for expected early anaemia risk

Category	Clinical basis	Broad clinical examples
Highest	A strong direct anaemia mechanism, or surgery for a selected heart or aortic condition.	Major bleeding; selected solid-organ injuries; marrow failure, acute leukaemia or haemolysis; selected heart or aortic surgery.
Middle	A plausible but less specific association, including other surgical conditions.	Chronic renal failure; cirrhosis; malignancy; septic shock; coagulopathy; regional bleeding; single long-bone fracture; other surgery.
Lowest	Little additional diagnostic evidence of systemic anaemia and no surgical upgrade.	Most non-surgical isolated airway, neurological or rhythm diagnoses without a bleeding, haematological or relevant chronic-disease mechanism.

Table 3 Clinical groups for expected low cardiac output risk

Category	Clinical basis	Broad clinical examples
Highest	A diagnosis directly denoting ventricular failure or a low-flow state, or an acute lesion with a strong low-flow mechanism.	Cardiogenic shock; ventricular failure; hypovolaemic shock; tamponade; acute mechanical cardiac complications; ruptured major aneurysm.
Middle	Low output is plausible but depends on severity, treatment or operative context.	Myocardial infarction; cardiomyopathy; valve disease; selected heart or thoracic aortic surgery; pulmonary embolism; important arrhythmias; septic shock; major bleeding without an explicit low-flow diagnosis.
Lowest	Little additional evidence of low output from the recorded diagnosis.	Most isolated respiratory, neurological, renal and other diagnoses without a cardiac, shock, major bleeding or selected systemic mechanism.

4 2 Additional ANZ variables

The ANZ model with APD data will additionally use premorbid frailty, chronic health conditions, admission source, preadmission function or residence, treatment limitations documented before randomisation, and weight, height or body mass index (BMI). Chronic health variables will include respiratory, cardiovascular, renal and liver disease and immunosuppression by disease or therapy. Patients without a successful APD match will retain their study data, with additional variables handled as missing. We will also examine successfully linked patients separately.

Only variables representing information available before randomisation will be eligible. We will exclude post-randomisation physiology and treatments, first-24-hour severity scores whose measurement window extends beyond randomisation, and subsequent outcomes.

4 3 Additional data for the JAMA model

Temperature, heart rate, mean arterial pressure, respiratory rate, PEEP, $PaCO_2$ and creatinine were collected at baseline in 27,172 MEGA-ROX patients, including 13,576 ANZ patients, to support evaluation of the JAMA model. Other predictors required by the published model were not collected, so we cannot validate the full published model. These measurements define the populations for objective 3. They will not enter the primary global or ANZ models because they were not collected throughout those populations; a separate supporting model will use them as described in Section 11.

5 Temporal training and validation samples

We will order all 40,003 patients chronologically by randomisation date and time and assign the earliest 50% to training, the next 25% to validation_1 and the latest 25% to validation_2. The training sample will contain the first 20,001 patients; validation_1 the next 10,001; and validation_2 the final 10,001. Ties in recorded randomisation date and time will be resolved by a fixed patient-identifier order specified before outcomes are examined. The ANZ analyses and additional-measurement subset will retain the global patient assignments. Both ANZ models will therefore use identical patients within each sample.

Initial predictor processing, risk-model fitting, candidate treatment-effect model fitting and hyperparameter tuning will use the training sample only. All candidate models will be fixed before validation_1 outcomes are examined. We will use validation_1 to select the best candidate for each predictor set. We will then combine training and validation_1 into the combined development sample (the earliest 75%; 30,002 global patients) and refit the selected model, as specified in Section 8.2. We will complete selection and final refitting and lock all new or adapted models before examining validation_2 outcomes or externally evaluating the reduced JAMA models. Validation_2 will be reserved for final evaluation and will not be used for model selection, tuning or refitting.

6 Missing predictors and preprocessing

Within each initial training sample, we will ordinarily exclude candidate variables with more than 40% missing data and assess correlations to identify redundant predictors. We may retain a clinically important potential effect modifier with greater missingness, using imputation and a missingness indicator where appropriate. Any such exception will be specified on clinical grounds before model fitting and documented. We will initially impute missing numeric values using bagged trees fitted in the training data and apply the same imputation models to validation_1 patients. After model selection, we will retain the chosen predictor set and refit imputation models and other learned preprocessing parameters using the combined development sample before applying them unchanged to validation_2. Categorical variables will include an unknown category where appropriate. The predictor sets and preprocessing for the reduced JAMA models will be fixed using the original study data before evaluation. Systematically unavailable predictors from the published model will not be handled as ordinary sporadically missing measurements.

We will retain the two clinical risk classifications as prespecified. We will group other sparse ICM diagnoses using their hierarchy and clinical interpretation and allow suitable transformations for continuous predictors. We will make these choices within the initial training data only and retain them during final refitting. Learned numerical parameters required by the chosen transformations will be re-estimated using the combined development sample. During cross-validation, we will fit imputation, transformations, and other predictor processing in the relevant training fold before applying them to its assessment fold.

7 Risk modelling

For each of the three new predictor sets in Table 1, we will develop an elastic-net logistic model of mortality using both treatment groups combined. This allows LASSO or less aggressive shrinkage to be selected during initial training. Candidate terms will include baseline main terms and pairwise interactions between baseline predictors. We will choose tuning parameters using cross-validated AUROC in the training sample, with Brier score used to resolve ties. After treatment-effect model selection, we will refit the risk model using the combined development sample, retaining its candidate terms and selected tuning parameters.

We will assess AUROC, Brier score, calibration intercept and slope, and a calibration plot in validation_1 using the risk model fitted in the training sample and in validation_2 using the risk model refitted in the combined development sample. Validation_2 will provide the final assessment. We will report randomised treatment effects across tertiles of predicted risk and test the interaction between treatment and the continuous prognostic linear predictor. A supporting analysis will use risk quintiles to provide finer description.

Predicted mortality risk will also be a candidate input to the treatment-effect models. During cross-validation, we will generate this risk prediction using only the relevant training data, excluding the patient whose prediction is being generated. We will also use out-of-fold risk predictions when fitting candidate models in the full training sample and when refitting the selected model in the combined development sample, using five-fold cross-validation with preprocessing fitted within each fold. For validation_1 patients, we will use the risk model fitted in the full training sample; for validation_2 patients, we will use the risk model refitted in the full combined development sample.

8 Effect modelling

8 1 Candidate algorithms

We will compare eight learner families from the LOGICAL plan: Rboost, causal forest, Tian transformation with elastic net, X-learner with Bayesian additive regression trees, Rlasso, and T-, S- and X-learners using conditional random forests [3]. For LASSO-based models, we will consider baseline main terms and pairwise predictor interactions.

8 2 Model training selection and refitting

Within the training sample, we will obtain out-of-fold predictions for each candidate algorithm using five-fold cross-validation. Hyperparameters will be tuned within each training fold using five-fold cross-validation. We will then tune each candidate using five-fold cross-validation in the full training sample and fit it using all training patients. We will average predictions from five fits with different random seeds to improve stability. All algorithms will use the same assessment folds and will be compared in the same validation_1 patients.

We will select the algorithm with the highest validation_1 Qini coefficient, using the common definition in Appendix 1. Ties will be resolved by lower calibration error, defined as the mean absolute difference between predicted and observed risk differences across predicted-effect tertiles, followed by higher C-for-benefit. After selection, we will refit the selected algorithm, its component models and learned preprocessing using the combined training and validation_1 sample, retaining the selected predictor set, preprocessing choices and hyperparameters. We will use the same five-fit averaging procedure and lock the complete refitted model before evaluating validation_2. Validation_1 selection metrics will retain the predictions made before refitting.

8 3 Randomisation and weighting

MEGA-ROX used allocation ratios of 1.05:1, 1:1 or 1:1.05 at its allocation updates. We will use each patient's known assignment probability when a learner or evaluation method requires allocation adjustment. These probabilities are close to 0.5. We will use design weights to standardise arm comparisons to an equal treatment mixture.

We will specify separately the weights used for model fitting and the rule used to combine the two X-learner effect estimates. We will use propensity-based combination where supported and verify the intended weighting in the chosen BART implementation before validation_1 evaluation. The initial fitting, final refitting and evaluation procedures will use the actual allocation probabilities rather than final observed arm proportions.

9 Validation of the new models

9 1 Treatment effect heterogeneity

We will report validation_1 and validation_2 results separately. Validation_1 results will use predictions made by the training-fitted models before refitting; validation_2 results will use the final models refitted in the combined development sample. In each sample, we will describe the predicted treatment-effect distribution and divide patients into lower, middle and upper tertiles using their predictions. Tertiles will describe ranking; treatment preference will depend on whether predictions are below or above zero. We will retain tied scores together.

The primary inferential assessment of each final refitted treatment-effect model will be the treatment-by-predicted-effect-tertile interaction for 90-day in-hospital mortality in validation_2. We will use a logistic model with treatment, predicted-effect tertile and their interaction, without additional prognostic covariate adjustment. We will use known-probability design weights, ICU-robust standard errors and a joint Wald test of the interaction terms. This tests heterogeneity on the odds-ratio scale.

For each tertile, we will report patient and death counts by randomised group, weighted mortality risks, the conservative-minus-liberal risk difference and the conservative-to-liberal risk ratio, with 95% confidence intervals. We will compare mean predicted effects with observed randomised risk differences to assess calibration. We will report Qini coefficients and C-for-benefit with 95% confidence intervals to describe treatment-benefit discrimination. Qini in validation_1 will be the primary metric for algorithm selection; the interaction test in validation_2 will be the primary inferential assessment. Validation_1 performance will be descriptive because it informs algorithm selection.

9 2 Added value of APD variables

We will compare the final refitted ANZ core model with the final refitted ANZ model containing additional APD variables in the same validation_2 patients and report corresponding pre-refit validation_1 comparisons descriptively. We will report paired differences in discrimination, calibration and estimated outcomes under model-directed treatment. This will distinguish the contribution of additional predictors from differences in patient populations.

9 3 Model directed treatment

We will estimate mortality if each validation_2 patient received the treatment predicted to be preferable by the final refitted model, assigning conservative oxygen for a negative predicted effect and liberal oxygen otherwise. We will use augmented inverse-probability weighting, combining observed randomised outcomes, known allocation probabilities and outcome predictions fitted in the combined development sample. We will compare this estimate with always-conservative, always-liberal and an equal random mixture of the two strategies. Corresponding validation_1 estimates will use the original training-fitted treatment-effect and outcome models and will be reported descriptively. These will be estimates from trial data rather than evidence from prospective use of the model.

10 External validation of reduced JAMA models

As investigators in the original JAMA study, we will use the original study data to develop one or more reduced-variable versions of the published treatment-effect model, restricted to predictors available in MEGA-ROX. We cannot validate the full published model because MEGA-ROX did not collect every required predictor, even among patients with the seven additional baseline measurements. We will document the reduced predictor sets, preprocessing and all departures from the published model.

We will specify the reduced versions and fix their preprocessing and fitted parameters using the original study data before external evaluation in MEGA-ROX. We will then evaluate each version in all eligible patients within the 27,172-patient population and its 13,576-patient ANZ subset, irrespective of their training, validation_1 or validation_2 assignment for the new models. We will not use MEGA-ROX outcomes to select or refit these treatment-effect models, and we will report all reduced versions evaluated. Any adaptation using MEGA-ROX outcomes will instead follow the temporal training, model-selection, refitting and final-validation procedures described in Section 11.

We will report predicted-effect distributions, randomised treatment effects by prediction tertile, discrimination and calibration for each reduced model, overall and in ANZ. We will use 90-day in-hospital mortality and the supporting day-28 outcome, identifying differences in outcome ascertainment. The 90-day analysis will assess transport to a different outcome horizon; an original 28-day prediction will not be interpreted as a calibrated 90-day effect. Results will describe the reduced-variable models and will not establish the performance of the full published model.

For estimates of mortality under model-directed treatment, any outcome-regression models used in augmented inverse-probability weighting will use cross-validation. These are evaluation models only and will not refit or alter the fixed treatment-effect predictions. Final comparisons with the primary refitted new models will use shared MEGA-ROX validation_2 patients; corresponding comparisons with the training-fitted new models in validation_1 will be descriptive. The overall and ANZ evaluations overlap and will not be described as independent replications.

11 Model interpretation and sensitivity analyses

We will summarise the baseline characteristics of predicted-effect tertiles and interpret the final refitted models. Variable importance will describe the change in predictions when a field is replaced by its reference value from the combined development sample. We will show partial dependence for the four most important continuous predictors and illustrate how the leading predictors contribute to predictions for 100 validation_2 patients, following the general LOGICAL presentation [3].

Sensitivity analyses will examine random reassignment of patients within the combined training and validation_1 samples in a 2:1 ratio, retaining the original validation_2 sample; exclusion of the predicted-risk input; median imputation; complete predictor records; and successful APD linkage. We will apply the same initial training, validation_1 model-selection and combined-sample refitting procedures and retain validation_2 exclusively for evaluation. We will examine performance in held-out ICUs or countries as supporting transport analyses. The primary analyses will retain the prespecified temporal 50:25:25 split, with final refitting using the combined training and validation_1 sample.

As a prespecified supporting analysis, we will develop an additional model using the core study variables and seven extra baseline measurements in the 27,172-patient population. We will retain these patients' training, validation_1 and validation_2 assignments from the global temporal split, without repartitioning; the subset proportions need not be 50:25:25. We will use the same initial training-only fitting and tuning, selection in validation_1, refitting using the combined training and validation_1 sample, and final evaluation in validation_2. This analysis will proceed irrespective of the primary validation_2 results. Any adaptation of the JAMA model using MEGA-ROX outcomes will follow this same sequence of initial training, selection, refitting and final evaluation and will be reported as new-model development.

12 Uncertainty multiplicity and reporting

We will obtain 95% confidence intervals for Qini coefficients, C-for-benefit, calibration measures and paired model comparisons using 1,000 bootstrap samples resampled by ICU within each evaluation sample, keeping the model used to generate each set of predictions fixed. These intervals will describe uncertainty conditional on the fitted models.

We will retain two primary validation interaction tests in validation_2: the global new model and the ANZ model with APD variables. We will report nominal and Holm-adjusted P values, with a two-sided familywise threshold of 0.05. The ANZ core model will be the comparator for added APD value. Validation_1 results will be descriptive and will not add primary tests. Reduced JAMA models, the additional-variable model and other supporting analyses will be exploratory, with nominal P values and pointwise confidence intervals.

We will report missingness, final cohort and sample sizes, randomisation-date ranges for training, validation_1, combined refitting and validation_2, retained predictors, model specifications and analysis deviations. The three objectives may be reported in separate manuscripts, with their shared patients and analyses identified. Planned tables and figures are listed in Appendix 2.

13 Software and governance

We will use R with compatible package versions and record the R version, package versions, random seeds, temporal cut points, sample assignments and relevant source-code commits. We will preserve the training-fitted candidate models and their validation_1 predictions, and the final refitted models and their validation_2 predictions. Analysis code will be made available in accordance with trial and registry permissions.

We will use deidentified data under the parent trial's consent, ethics and data-access arrangements, including ANZICS CORE requirements and applicable Māori data governance. We will document SAP finalisation and any previous access to treatment-specific outcomes.

References

1. Buell KG, Spicer AB, Casey JD, et al. Individualized Treatment Effects of Oxygen Targets in Mechanically Ventilated Critically Ill Adults. JAMA. 2024;331:1195-1204. doi:10.1001/jama.2024.2933.

2. Young PJ, Arabi YM, Bagshaw SM, et al. Protocol and statistical analysis plan for the mega randomised registry trial research program comparing conservative versus liberal oxygenation targets in adults receiving unplanned invasive mechanical ventilation in the ICU (Mega-ROX). Crit Care Resusc. 2022;24:137-149. doi:10.51893/2022.2.OA4.

3. Johnson NJ, Udy A, Drayton BA, et al. Individualised Oxygen Targets after Cardiac Arrest A secondary analysis of the LOGICAL Trial. Protocol and statistical analysis plan version 1.1. 7 July 2026.

4. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity PATH Statement. Ann Intern Med. 2020;172:35-45. doi:10.7326/M18-3667.

5. Intensive Care National Audit & Research Centre. ICNARC Coding Method (ICM) Guide. Last updated April 2018. Available at: https://account.icnarc.org/DataServices/Attachments/Download/81e9e5dc-5742-e811-80ec-1402ec3fcd79.

Appendix 1 Common Qini definition

We will rank patients from greatest predicted conservative benefit to least. For each ranked fraction, the gain will be the sum of liberal deaths divided by their liberal assignment probabilities, minus the sum of conservative deaths divided by their conservative assignment probabilities, divided by the full evaluation sample size. The endpoint will estimate the overall mortality benefit of conservative oxygen.

We will subtract the straight line from zero to that endpoint and calculate the area between the curves using trapezoidal integration. We will include tied scores proportionally. This design-adjusted Qini coefficient will be used consistently for training cross-validation, validation_1 selection before refitting and validation_2 evaluation of the final refitted models; calibration will be assessed separately by comparing predicted and observed risk differences.

Appendix 2 Planned tables and figures

We will produce the following outputs for the relevant analyses in Table 1. Baseline summaries will show observed denominators and missingness, without significance tests comparing baseline characteristics.

Table	Content
1 Populations and baseline characteristics	Training, validation_1, combined refitting and validation_2 sample sizes and randomisation-date ranges for the new models; external-validation populations for the reduced JAMA models and supporting analyses; treatment allocation and predictor missingness.
2 Prognostic performance	AUROC, Brier score, calibration intercept and slope; treatment effects across predicted-risk groups. Training-fitted risk models in validation_1 and refitted risk models in validation_2.
3 Model training selection and refitting	Candidate algorithms, cross-validated training performance, validation_1 Qini and tie-breaker metrics before refitting, and the selected algorithm, specification and combined-sample refitting details for each predictor set.
4 Treatment effects by predicted ITE	Patient and death counts by randomised group; weighted mortality risks; RD, RR and 95% CIs; mean predicted effect; interaction P values. Separate validation_1 results before refitting and validation_2 results from the refitted models, with primary interaction tests in validation_2.
5 Performance and model comparisons	Qini, C-for-benefit and calibration; ANZ core versus APD comparison; model-directed mortality estimates; descriptive validation_1 results before refitting and final results from the refitted models in validation_2; reduced JAMA model overall and ANZ results; additional-variable model results.

The figures will show patient flow through the three temporal samples and combination of training and validation_1 for final refitting, predicted-effect distributions, observed treatment effects by tertile, Qini curves, calibration, variable importance and partial dependence. Final new-model evaluation figures will use the refitted models in validation_2; validation_1 selection results before refitting will be identified separately. Supplementary outputs will show sensitivity analyses, predictor definitions and the implementation of each reduced JAMA model and its differences from the published model.