

Statistical Analysis Plan

TRIAL FULL TITLE	Strategy to Avoid Excessive Oxygen using Autonomous Oxygen Titration Intervention
SAP VERSION	1.0
SAP VERSION DATE	08/20/2025
TRIAL STATISTICIAN	John D. Rice, PhD
Protocol Version (SAP associated with)	1.0
TRIAL PRINCIPAL INVESTIGATOR	Adit A. Ginde, MD, MPH
SAP AUTHOR	John D. Rice, PhD

1 SAP Signatures

I give my approval for the attached SAP entitled SAVE-O2 AI dated 08/28/25.

Statistician (Author)

Name: John D. Rice, PhD

Signature: _____

Date: _____

Principal Investigator

Name: Adit A. Ginde, MD, MPH

Signature: _____

Date: _____

2 Table of Contents

1	SAP Signatures	2
2	Table of Contents	3
3	Abbreviations and Definitions	4
4	Study Objectives and Endpoints	6
4.1	Study Objectives	6
4.2	Endpoints	6
4.2.1	Primary Endpoint	6
4.2.2	Secondary Endpoints	7
4.2.3	Exploratory Endpoints	7
4.2.4	Covariates	7
5	Study Methods	7
5.1	General Study Design and Plan	7
5.2	Inclusion-Exclusion Criteria and General Study Population	8
5.2.1	Inclusion Criteria	8
5.2.2	Exclusion Criteria	8
5.3	Randomization	8
6	Power and Sample Size	9
7	General Analysis Considerations	10
7.1	Covariates	10
7.2	Differential treatment effects and subgroup analyses	10
7.3	Missing Data	10
7.4	Interim Analyses and Data Monitoring	11
8	Statistical Analysis Methods	12
8.1	Primary Endpoint	12
8.1.1	Definitions	12
8.1.2	Valid data	12
8.1.3	Statistical model specification	12
8.1.4	Hypothesis testing for primary intervention effect estimate	13
8.1.5	Missing data	13
8.1.6	Subgroup analysis and visualization	13
8.1.7	Sensitivity analyses	14
8.2	Secondary Endpoints	15
8.3	Exploratory endpoints	15
	Works Cited	16

3 Abbreviations and Definitions

AC = Alternating Current Power

AE = Adverse Events

AIS = Abbreviated Injury Scale

APACHE = Acute Physiology and Chronic Health Evaluation

ARDS = Acute Respiratory Distress Syndrome

BMP = Basic Metabolic Panel

CA = Competent Authority

CBC = Complete Blood Count

CCC = Clinical Coordinating Center

CFR = Code of Federal Regulations

CI = Confidence Interval

CMP = Clinical Monitoring Plan

COMIRB = Colorado Multiple Institutional Review Board

COPD = Chronic Obstructive Pulmonary Disease

COVID-19 = Coronavirus Disease of 2019

DCC = Data Coordinating Center

DoD = Department of Defense

DSMB = Data and Safety Monitoring Board

DSMP = Data and Safety Monitoring Plan

eConsent = Electronic Consent

eCRF = Electronic Case Report Form

ED = Emergency Department

EHR = Electronic Health Record

FDA = U.S. Food and Drug Administration

FHIR = Fast Healthcare Interoperability Resource

FiO₂ = Fraction of Inspired Oxygen

GCP = Good Clinical Practice

HFD28 = Hospital Free Days to Day 28

HFNO = High Flow Nasal Oxygen

ICC = Interclass Correlation Coefficient

ICH = International Conference on Harmonisation

ICU = Intensive Care Unit

IDE = Investigational Device Exemption

IEC = Independent Ethics Committee

IMV = Invasive Mechanical Ventilation

IRB = Institutional Review Board

ISS = Injury Severity Score

LAR = Legal Authorized Representative

LFT = Lipid Fat

LPM = Liters Per Minute

MST = Monk Skin Tone
MRI = Magnetic Resonance Imaging
MTEC = Medical Technology Enterprise Consortium
NCT = National Clinical Trial
NIV = Non-Invasive Ventilation
O2 = Oxygen
OHRP = Office for Human Research Protections
OR = Operating Room
ORHO = Office of Human Research Oversight
PaO2 = Partial Pressure of Oxygen in Arterial Blood
PCLC = Physiological Closed-Loop Control
PDF = Portable Document Format
PHI = Protected Health Information
PI = Principal Investigator
PT/PTT/INR = Prothrombin Time/Partial Thromboplastin Time/International Normalized Ratio
QC = Quality Control
SAE = Serious Adverse Event
SAP = Statistical Analysis Plan
SAVE-O2 = Strategy to Avoid Excessive Oxygen
sIRB = Single IRB
SOFD = Supplemental Oxygen Free Days
SOP = Standard Operating Procedure
SpO2 = Oxygen saturation via pulse oximetry
TBSA = Total Body Surface Area
TCCC = Tactical Combat Casualty Care
UAP = Unanticipated Problem
U.S. = United States
VUMC = Vanderbilt University Medical Center
WOCBP = Women of Childbearing Potential

4 Study Objectives and Endpoints

4.1 Study Objectives

We propose the following three objectives:

- (1) Determine the effectiveness of an autonomous oxygen titration system to improve time spent in normoxemia and reduce time spent in hypoxemia and hyperoxemia among acutely injured and ill patients receiving supplemental oxygen. We will compare patient-hours spent in normoxemia (SpO₂ 90-96%), hypoxemia (SpO₂ <88%), and hyperoxemia (SpO₂ >96%) among patients randomized to autonomous vs manual oxygen titration.
- (2) Determine the impact of an autonomous oxygen titration system on overall utilization of supplemental oxygen. We will compare the total volume of supplemental oxygen administered to patients randomized to autonomous vs manual oxygen titration during the 72-hour intervention period.
- (3) Evaluate the effect of an autonomous oxygen titration system on patient-centered clinical outcomes, including hospital-free days to day 28, ICU length of stay, and mortality, by comparing these outcomes between patients randomized to autonomous oxygen titration versus usual care.

4.2 Endpoints

4.2.1 Primary Endpoint

The primary endpoint is the proportion of time spent within the targeted normoxemia range, defined as an oxygen saturation (SpO₂) of 90–96%, as measured by two separate continuous non-invasive pulse oximetry devices, during the first 72 hours after randomization. This proportion is calculated separately for each device stream as the number of valid minutes in the normoxemic range divided by the total number of valid minutes for that device. Minutes with SpO₂ above 96% are also counted as normoxemic if the participant is on room air as this reflects clinically appropriate oxygenation in the absence of supplemental oxygen. The status of *room air* will be defined based on available oxygen delivery information, using a consistent operational rule across sites.

Each minute is treated as a discrete unit of observation. A minute is considered to have a missing SpO₂ value for a given device if no measurement is recorded or if the recorded value falls outside the physiologically plausible range of 60% to 100%. If multiple valid SpO₂ values are recorded by a device within a given minute, they will be averaged to yield a single minute-level value. Only minutes with at least one valid reading are included in the calculation for that device. A participant contributes one device-level proportion per device with any valid data, and these proportions serve as the units of analysis in the primary model.

If a patient reaches a concluding event prior to 72 hours, the denominator will be restricted to the time with observed SpO₂ values from randomization until the event. Concluding events include, but are not limited to: discharge or transfer from the enrolling hospital; escalation to high-flow nasal oxygen, non-invasive or invasive mechanical ventilation; death; sustained transition to room air with removal of SpO₂ monitoring; or sustained discontinuation of supplemental oxygen. Sustained discontinuation is defined as at least 6 continuous hours with no documented oxygen flow and no subsequent resumption of supplemental oxygen for the remainder of the 72-hour window, after which all subsequent time will be censored regardless of whether SpO₂ monitoring continues. In addition, if a treating clinician prospectively documents an SpO₂ target range outside 90–96% (e.g., for patients with COPD), that period will be excluded from both the numerator and denominator of the primary outcome.

4.2.2 Secondary Endpoints

- (1) **Amount of supplemental oxygen administered**, defined as total estimated oxygen volume during the first 72 hours after randomization.
- (2) **Proportion of time spent in hypoxemia** (SpO₂ <88%) during the first 72 hours after randomization. Denominator will be time on observation as defined for the primary outcome.
- (3) **Proportion of time spent in hyperoxemia** (SpO₂ >96%) during the first 72 hours after randomization. Denominator will be time on observation as defined for the primary outcome.
- (4) **Time to Room Air**, defined as the time from randomization to the first period documented as room air up to 28 days after randomization, with discharge or death considered as competing risks [1].

4.2.3 Exploratory Endpoints

- (1) **Supplemental O₂ free days (SOFD)** will be calculated as 28 minus the number of calendar days from the first through the last day on which a patient receives supplemental oxygen. Patients who die before day 28 will receive a score of -1. Patients discharged before day 28 will be assumed to remain on the oxygen status recorded at discharge for the remainder of the period.
- (2) **All-cause mortality to day 28**, defined as a dichotomous vital status (survived or died) at hospital discharge or day 28, whichever is first.
- (3) **Respiratory failure**, defined as receipt of high flow nasal oxygen, non-invasive ventilation, or invasive mechanical ventilation, to day 28 or hospital discharge, whichever is first, excluding time in the operating room.
- (4) **Hospital-free days to day 28 (HFD28)** defined as the number of days alive and outside the hospital between randomization and 28 days later, censored at hospital discharge. This outcome is a composite of in-hospital mortality and length of hospital stay and has a range of 0 days (worst outcome) to 28 days (best outcome). Patients discharged to any location prior to day 28 are assumed to survive to day 28 and are assigned a number of hospital-free days equal to the difference between 28 and their hospital length of stay. Patients dying in-hospital prior to day 28 will be assigned HFD28 of -1.
- (5) **Proportion of time receiving no supplemental oxygen** (room air) during the first 72 hours after randomization.

4.2.4 Covariates

As this study is randomized at the patient level, we will only adjust for the randomization stratification factors of site and patient type (trauma/burn/surgical vs medical). We do not plan to include additional covariates unless substantial baseline imbalance is observed. However, if adjustment is pursued to improve statistical precision, the following variables are candidates for inclusion a priori: skin pigmentation (Monk scale) and baseline supplemental oxygen volume.

5 Study Methods

5.1 General Study Design and Plan

This study will be a **multicenter randomized controlled trial** to determine the effectiveness of using a closed loop/autonomous oxygen system (O₂matic PRO100) to maintain normoxemia (oxygen saturation 90-96%, target SpO₂ 93%) during the first 72 hours of acute injury or illness, compared to standard provider-driven methods (manual titration). Given the nature of the intervention, we will not blind the primary clinical team, patient, or implementation research team. However, the outcomes

assessment is primarily electronic and will be blinded to the randomized group.

5.2 Inclusion-Exclusion Criteria and General Study Population

The target population will focus on major trauma, acute care surgery, burn, and acute respiratory illness patients who are hospitalized and receiving supplemental oxygen, but not invasive mechanical ventilation. We will randomize up to 300 patients at four level 1 trauma centers within 36 hours of hospitalization and randomize to receive autonomous versus manual oxygen titration for up to 72 hours.

5.2.1 Inclusion Criteria

In order to be eligible to participate in this study, an individual must meet all of the following criteria:

- Age 18 years or older
- Hospitalized or will be hospitalized from Emergency Department for major trauma, burn, acute care surgery, or acute respiratory illness
- Able to be randomized within 36 hours of hospital arrival
- Receiving supplemental oxygen 1-10 lpm for documented or presumed hypoxemia (must be higher than baseline for those on chronic oxygen therapy)
- Signed and dated informed consent from patient or legally authorized representative (LAR)

5.2.2 Exclusion Criteria

An individual who meets any of the following criteria will be excluded from participation in this study:

- Anticipated hospital discharge within 24 hours
- Imminent plans to discontinue supplemental oxygen
- Imminent plans to administer high flow nasal oxygen, non-invasive ventilation, invasive mechanical ventilation
- Clinical team unwilling or unable to follow the prescribed oxygen titration method in either randomized group
- Known prisoner
- Known pregnancy
- Known contraindicated conditions for use of the PRO100 device: carbon monoxide poisoning, incapable of handling airway secretions, increased methemoglobin, cyanide poisoning, cluster headaches, undrained pneumothorax, sick cell crisis, paraquat poisoning or a history of bleomycin poisoning, patients for whom the SpO2 signal is not stable

5.3 Randomization

Randomization assignments are 1:1 to intervention or usual care at the patient level and treatments are assigned at randomization. Randomization will be implemented using an electronic system (REDCap) and the variable permuted block technique; treatment assignments in this two-group trial will utilize a block size between four and eight. Employing the permuted block technique ensures balance across the intervention and control groups at regular enrollment intervals within each site, thus reducing the possibility of unequal distribution across arms for which the participant is eligible. Additionally, randomization will be stratified by site and patient subgroup (trauma/burn/surgical vs medical) to optimize balance across these key characteristics.

6 Power and Sample Size

The primary outcome being a proportion between 0 and 1 will be analyzed as continuous. Using a two-sample t test, we compute the power for $n = 150$ patients per arm, shown in the table below. We assume a mean in the control group of $\mu_1 = 0.6$; the mean proportion of time in the target range among patients assigned to standard of care by L'Her et al. (2021) [2] was found to be 0.621. The effect size (difference in mean between the two treatment arms, $\delta = \mu_2 - \mu_1$) was varied between 0.1 and 0.15; this is substantially smaller than the effect size of 0.319 found by L'Her et al. (2021) [2]. An estimated standard deviation for this outcome of $\sigma \approx 0.183$ can also be computed [2]; we therefore examine a range of standard deviations between 0.2 and 0.5, to account for possibly increased variability in the outcome distribution.

μ_1	μ_2	$\delta = \mu_2 - \mu_1$	σ	Power
0.6	0.7	0.1	0.2	99.1%
0.6	0.75	0.15	0.2	100.0%
0.6	0.7	0.1	0.3	82.1%
0.6	0.75	0.15	0.3	99.1%
0.6	0.7	0.1	0.4	57.9%
0.6	0.75	0.15	0.4	89.9%
0.6	0.7	0.1	0.5	40.8%
0.6	0.75	0.15	0.5	73.6%

From this table, we see that there is >99% power for an effect size of ≥ 0.1 as long as the standard deviation of the outcome is ≤ 0.3 .

Due to the importance of subgroup analysis to the research goals, we have also computed power for the two-sample t test as applied to subgroups of various sizes. The methods are identical to those described above for the primary analysis, except that the sample size available is reduced to the size of the subgroup in question.

Subgroup size	Sample size per arm	μ_1	μ_2	$\delta = \mu_2 - \mu_1$	σ	Power
150	75	0.6	0.7	0.1	0.2	86.0%
150	75	0.6	0.75	0.15	0.2	99.5%
120	60	0.6	0.7	0.1	0.2	77.5%
120	60	0.6	0.75	0.15	0.2	98.3%
90	45	0.6	0.7	0.1	0.2	65.0%
90	45	0.6	0.75	0.15	0.2	94.0%
60	30	0.6	0.7	0.1	0.2	47.8%
60	30	0.6	0.75	0.15	0.2	81.5%
30	15	0.6	0.7	0.1	0.2	26.2%
30	15	0.6	0.75	0.15	0.2	50.9%

This table shows that we have good (>80%) power even for subgroups as small as 60 (20% of the total sample size) when the effect size is 0.15 or larger. To detect a treatment effect with 80% power at a significance level of 0.05, the minimum detectable difference in normoxemia proportion ranges from 0.092 for subgroups of 150 participants (75 per arm) to 0.212 for subgroups of 30 participants (15 per arm), assuming a standard deviation of 0.2.

7 General Analysis Considerations

Baseline descriptive statistics will be calculated for variables including but not limited to age, sex, race/ethnicity, skin pigmentation, type of patient (trauma/burn/surgical vs medical), comorbidities, home oxygen use, and supplemental oxygen flow rate at baseline. This will include mean and SD for continuous variables and frequency and percent for categorical variables. Statistics will be presented for the sample overall and stratified by treatment arm. Comparisons between treatment arm will be made to ensure balance has been achieved by randomization.

As an overall approach, we will analyze primary and secondary endpoints using a regression modeling framework, with specific distributions chosen depending on the type of outcome (e.g., binary, count, ordinal, time-to-event). The primary analysis will be as-randomized (i.e., intention to treat), so the primary exposure variable in these regression models will be a binary indicator of randomized treatment arm (equal to 1 for experimental and 0 for control). A threshold of $\alpha = 0.05$ will be used, with p-values less than α considered statistically significant and will be reported for the primary endpoint. Other secondary and exploratory endpoints will be reported as effect estimates with 95% confidence intervals. There will be no adjustment for multiple comparisons, given that there is a single primary endpoint and other endpoints are considered secondary/exploratory.

7.1 Covariates

As this trial is randomized at the individual-patient level, the only covariates that will be adjusted for are the randomization stratification factors site and patient subgroup (reason for admission trauma/burn/surgical or medical). Additional covariates may be included if they are identified a priori as likely to improve precision, such as skin pigmentation or baseline supplemental oxygen volume.

7.2 Differential treatment effects and subgroup analyses

Subgroup analyses will be conducted by including interactions in regression models between the treatment condition and the specific covariate of interest. Pre-specified subgroup analyses will include:

- race/ethnicity (non-Hispanic white, non-Hispanic black, Hispanic),
- skin pigmentation (3 categories by Monk Skin Tone Scale),
- baseline supplemental oxygen (2-3 categories, to be determined based on empirical distribution),
- home oxygen use (y/n),
- reason for hospitalization (trauma/burn/surgical vs medical)
- chronic pulmonary disease (y/n)

Exploratory subgroups include Fitzpatrick skin type (categorical), and individual topography angle (ITA) measured by spectrophotometry. These may be included if measurement quality and coverage permit reliable analysis. Due to exploratory nature of ITA and the Nix device, we anticipate using Monk categories for primary subgroup comparisons and Fitzpatrick as secondary.

A significant coefficient estimate for the interaction term(s) will suggest that the effect of treatment varies according to the value of that covariate. Separate estimates of the treatment effect will then be reported for each level of the interacting covariate. Interaction p-values less than 0.05 will be considered statistically significant; however, for underpowered subgroups, interaction p-values up to 0.20 may be interpreted as suggestive of a potential differential effect. For continuous covariates, significant interactions will be summarized graphically, along with treatment effect estimates at selected values of the covariate.

7.3 Missing Data

Continuous SpO₂ data for the primary endpoint will be collected every minute from up to two devices per participant: the PRO100 and the hospital's standard monitor. Each minute is treated as a discrete observation, and the analysis is based on the total number of observed minutes per device.

Missingness is expected due to temporary loss of signal, clinical workflow, or device-specific issues. All valid SpO₂ measurements will be included, and data from each device will be analyzed separately. If neither device provides a valid measurement at a given minute, that minute will be excluded from the outcome calculation.

All participants with any valid SpO₂ data from at least one device will be included in the analysis. The quasi-binomial model accounts for differing amounts of observed data by modeling the proportion of time in range. Participants with less data will contribute less information to the estimate but are not excluded a priori.

Room air status will be determined based on available oxygen delivery information, including flow rate, device type, and ventilation status. A consistent rule will be applied during data processing to identify periods of room air across sites. Minutes with insufficient information to determine oxygen status will be excluded from calculations requiring that information.

Missing covariate data are not expected to be common. If present, categorical variables will include a “missing” category; continuous variables will be left unadjusted unless otherwise noted.

7.4 Interim Analyses and Data Monitoring

A single formal interim analysis will be conducted for safety after approximately 150 participants have been accrued. The purpose of this analysis is to evaluate whether the intervention group shows evidence of increased hypoxemia compared to the control group.

The primary hypothesis tested at interim will be:

- H₀ (null): The mean proportion of time spent in hypoxemia is the same in both treatment arms.
- H₁ (alternative): The intervention arm has a higher mean proportion of time in hypoxemia than the control arm.

This will be assessed using a one-sided test with $\alpha = 0.01$. The primary endpoint for the interim analysis will be evaluated using the same analytic strategy as specified for the final primary analysis. Specifically, a quasi-binomial regression model will be used to compare the mean proportion of time in hypoxemia between treatment arms, with an identity link and cluster-robust standard errors to account for correlation between measurements from the two devices within each participant. The model will include terms for treatment arm and the stratification variables (site and patient subgroup).

Participants with any valid SpO₂ data from at least one device will be included in the interim analysis. Valid data will be defined using the same criteria as for the primary endpoint: minute-level SpO₂ values between 60–100%, collected during the 72-hour intervention window.

Because data may be delayed due to the 28-day follow-up window, site upload schedules, or device availability, the interim analysis will be conducted only once data from at least 90% of enrolled participants have available data for the primary endpoint. Participants whose primary endpoint data have not yet been received will be excluded from the interim analysis but retained for final analysis.

The following information will be presented to the DSMB:

- Summary of baseline characteristics by arm
- Distribution of the primary endpoint (normoxemia and hypoxemia proportions) by arm
- Point estimate and confidence interval for the treatment effect on hypoxemia
- One-sided p-value for the interim test
- Cumulative reportable adverse events and serious adverse events by arm
- Any additional subgroup or stratified summaries requested by the DSMB

The DSMB will review these data in light of the pre-specified threshold and the magnitude of any

observed differences. Rejection of the null hypothesis at $p < 0.01$ will not automatically trigger study pause or termination. Rather, the DSMB will consider both statistical and clinical significance, including the size and consistency of the observed difference, as well as other safety data. The interim test is expected to be underpowered to detect small or moderate differences in hypoxemia. Therefore, results will be interpreted with caution, and emphasis will be placed on the magnitude and precision of the estimated treatment effect in addition to statistical significance.

No interim analysis for efficacy will be conducted, and the trial will not be stopped for efficacy or futility. The primary rationale for this is that adequate accrual by key subgroups (e.g., skin pigmentation category) is required to full evaluation of the device.

8 Statistical Analysis Methods

There is a single primary endpoint, and no adjustment will be made for multiple comparisons with respect to secondary endpoints and analyses; appropriate caution will therefore be used in interpreting the results of hypothesis testing for these analyses. Data will be summarized descriptively by treatment arm, using mean and median as measures of central tendency, and standard deviation (SD) and interquartile range (IQR) as measures of spread; visual summaries (e.g., histograms) will also be examined. Regression models will include terms for the treatment arm as randomized (intention to treat approach) and stratification factors used for randomization.

8.1 Primary Endpoint

8.1.1 Definitions

The primary outcome is the proportion of time spent in normoxemia, defined as SpO₂ between 90% and 96%, during the 72 hours following randomization. This outcome is calculated using minute-level data from up to two pulse oximetry devices per participant: (1) the PRO100 (running in observation mode for control arm participants), and (2) the hospital's standard monitor. Each minute is treated as a discrete measurement. Normoxemia is calculated as the number of minutes with valid SpO₂ in the 90–96% range, or SpO₂ above 96% while the patient is confirmed to be on room air, divided by the number of minutes with valid SpO₂ measurements overall. Minutes with SpO₂ above 96% on room air will be considered clinically acceptable and counted as normoxemia to avoid penalizing appropriate weaning from supplemental oxygen.

8.1.2 Valid data

SpO₂ values are considered valid if they fall within a clinically plausible range of 60% to 100%. All participants with at least one minute of valid SpO₂ data from either device will be included in the analysis. For each device stream (PRO100 and hospital monitor), we compute a separate proportion of time in normoxemia per participant, defined as the number of valid minutes in normoxemia range (or above 96% on room air) divided by total valid minutes with observed SpO₂ from that device. These device-specific averages are the unit of analysis. If both devices yield sufficient data, the participant contributes two independent outcome values to the model. No averaging or merging of minute-level values is performed. Cluster-robust standard errors account for correlation between device streams within participants. To support interpretation of device-level outcomes, we will also summarize agreement between PRO100 and hospital monitor readings using patient-level metrics such as concordance in normoxemia classification and overall disagreement rates. In a sensitivity analysis, we will instead use the full 72-hour observation window as the denominator, with unobserved minutes excluded from the numerator but retained in the denominator.

8.1.3 Statistical model specification

The treatment effect, defined as the difference in the mean proportion of time spent in normoxemia between study arms, will be estimated using a quasi-binomial generalized linear model (GLM) with an identity link function. This model is appropriate for bounded, continuous outcomes between 0 and 1 that may exhibit overdispersion relative to the binomial distribution. The identity link models the outcome directly on the probability scale, allowing the treatment coefficient to be interpreted as an absolute risk difference in mean time spent in normoxemia.

The model will include a fixed effect for treatment arm (coded as 1 for intervention and 0 for control), along with fixed effects for the randomization stratification factors: site and patient subgroup (trauma/burn/surgical vs. medical).

Each participant may contribute one or two outcome values, each representing the proportion of time in normoxemia from a single device stream, depending on data availability. These device-level averages are modeled under a working independence assumption, which refers to the structure used in the estimation algorithm rather than an assumption that the observations are truly uncorrelated. To account for correlation between observations from the same participant, including any minute-to-minute dependence within a device stream or between devices, we will use cluster-robust standard errors at the participant level. This approach provides valid inference even if the working independence assumption is violated. The variance function will follow the quasi-binomial family, with variance proportional to $\mu(1-\mu)$.

8.1.4 Hypothesis testing for primary intervention effect estimate

The treatment effect estimate from our primary regression model, $\hat{\beta}_1$, represents the average difference in the proportion of time spent in normoxemia between intervention and control groups. This estimate incorporates data from both SpO₂ devices when available and accounts for within-subject correlation via clustered standard errors. Because each SpO₂ device stream is modeled as a separate observation, the treatment effect estimate reflects an average across devices, accounting for which devices provided usable data. Participants with valid data from both devices contribute two observations, while those with data from only one device contribute one. As a result, the estimate is implicitly weighted by the availability of valid data across the two device types. The analysis does not assume equal contribution from each device, but rather uses all available information while accounting for within-subject correlation.

The formal statistical null hypothesis to be tested is $H_0: \beta_1 = 0$, that is, there is no difference in mean proportion of time in the target range is the same between control and intervention groups; the alternative hypothesis is $H_1: \beta_1 \neq 0$. A two-sided p-value < 0.05 will be considered statistically significant. All effect estimates will be reported with 95% confidence intervals and interpreted with attention to both magnitude and clinical relevance.

8.1.5 Missing data

Missing outcome data are anticipated due to interruptions in device monitoring, variability in data transmission, or other operational factors. All participants with any valid SpO₂ data from at least one device will be included in the analysis. The model naturally adjusts for differing amounts of data through the quasi-binomial structure. No imputation will be performed for missing outcome data. The primary model uses a quasi-likelihood framework for estimation, and inference will rely on cluster-robust standard errors, which remain valid under standard regularity conditions even in the presence of unmodeled heteroskedasticity or within-participant correlation [3]. The analysis assumes that missingness is missing at random (MAR), conditional on observed covariates. Minute-level observations with missing data from both devices will be excluded from the outcome calculation, but valid data from either device will be retained when available.

8.1.6 Subgroup analysis and visualization

To aid interpretation of potential heterogeneity in treatment effects, we will present subgroup-specific estimates of the primary endpoint in a forest plot. Each plot will display the estimated treatment effect and 95% confidence interval within levels of key pre-specified subgroups. These graphical summaries are intended to complement the formal interaction models described in Section 7.2 by providing a concise visual overview of treatment effect consistency across subpopulations.

8.1.7 Sensitivity analyses

Several sensitivity analyses will be conducted to evaluate the robustness of the primary treatment effect estimate to modeling assumptions and data structure.

Although the quasi-binomial model accounts for overdispersion in bounded outcomes, we will examine whether the assumed mean–variance relationship affects the estimated treatment effect. Specifically, we will assess whether deviations from the functional form $Var(Y) = \phi\mu(1 - \mu)$ result in material differences in inference. The dispersion parameter ϕ will be estimated from the data, and sensitivity to this specification will be evaluated by comparing models using alternative variance structures, if appropriate.

We will explore potential heterogeneity in the treatment effect by device type. To do so, we will fit an expanded model including main effects for device (hospital vs. PRO100) and an interaction between treatment and device. The model is given by:

$$E(Y_{ij}|X_i, Z_i, b_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 I(j = 2) + \beta_4 X_i \cdot I(j = 2),$$

where:

- Y_{ij} is the normoxemia proportion calculated from device j for participant i .
- X_i is the treatment indicator.
- Z_i includes stratification variables (site and patient type).
- $I(j = 2)$ indicates the PRO100 device.

This model will use the same quasi-binomial structure with an identity link, working independence correlation structure, and cluster-robust standard errors to account for repeated observations per participant. This specification will allow us to assess whether the treatment effect differs between device streams, or whether one device consistently reports higher or lower oxygen saturation.

In addition to the interaction model assessing effect modification by device type, we will perform complementary sensitivity analyses using each device stream independently. First, we will re-estimate the treatment effect using only SpO₂ measurements obtained from the PRO100 device (regardless of study arm), excluding minutes for which no valid PRO100 reading is available. Second, we will repeat the analysis using only SpO₂ data from the hospital's standard pulse oximeter. Each of these models will use the same quasi-binomial regression framework with an identity link and include fixed effects for treatment arm, site, and patient subgroup, with standard errors clustered at the participant level. These analyses will allow us to evaluate the extent to which the estimated treatment effect is sensitive to the choice of device and potential device-specific measurement differences. Finally, we will assess whether the primary findings are sensitive to differences in the amount of valid SpO₂ data available across participants. To do so, we will re-fit the primary model under varying inclusion criteria. In particular, we will evaluate the impact of requiring higher or lower thresholds for data completeness (e.g., restricting to participants with ≥ 6 , ≥ 12 , or ≥ 24 hours of valid data). Although the primary analysis now includes all participants with any valid SpO₂ measurements, these sensitivity analyses will help determine whether the estimated treatment effect is influenced by the extent of observed monitoring time. This will include graphical comparisons and evaluation of changes in the treatment coefficient and its standard error under different inclusion criteria. These results will be interpreted descriptively to assess robustness but will not replace the primary analysis.

Results from all sensitivity analyses will be compared to those of the primary analysis. Substantial discrepancies will be interpreted in context but will not replace the primary model as the basis for inference.

8.2 Secondary Endpoints

For secondary outcomes, we will report effect estimates and 95% confidence intervals without formal hypothesis testing or p-values.

- (1) **Amount of supplemental oxygen administered** during the first 72 hours post-randomization will be compared between treatment arms using a quasi-Poisson generalized linear model with log link and terms for treatment arm and randomization stratification factors. An offset will be included in the model to normalize for time spent in the hospital. The model will be

$$\log E(Y_i|X_i, Z_i) = \log T_i + \beta_0 + \beta_1 X_i + \beta_2 Z_i$$

with terms as described above for the primary analysis, and the additional offset term $\log T_i$ where T_i is the patient's length of stay or 72 hours, whichever is less. This approach will account for both differing follow-up times and for the right-skewed nature of this outcome variable.

- (2) **Proportion of time spent in hypoxemia** will be analyzed as described above for the primary outcome.
- (3) **Proportion of time spent in hyperoxemia** will be analyzed as described above for the primary outcome.
- (4) **Time to Room Air** will be analyzed in the competing risks framework [1], where death and discharge are competing risks with the event of first room air. Note that the definition of “room air” used for this time-to-event outcome differs from that used in the primary endpoint analysis. For the secondary outcome, room air is defined as the first recorded episode of zero supplemental oxygen flow (i.e., flow rate = 0 L/min), regardless of whether supplemental oxygen is resumed later. In contrast, the primary endpoint censors observation time only after a sustained discontinuation of supplemental oxygen. The analysis will be based on a Cox proportional hazards regression model. Time zero will be defined as time of randomization. Cumulative incidence plots will be used to examine the unadjusted distribution of time to the event of interest and make comparisons between treatment conditions. The regression model for these analyses will be

$$\log h_i(t) = \log h_0(t) + \beta_1 X_i + \beta_2 Z_i$$

- $h_i(t)$ is the hazard for the event of interest (reaching room air) at time t for patient i
- $h_0(t)$ is the baseline hazard
- X_i is the treatment indicator for patient i , equal to 0 for control and 1 for intervention
- β_1 is the treatment effect, the log hazard ratio between intervention and control arms

8.3 Exploratory endpoints

For exploratory outcomes, results will also be presented descriptively with point estimates and 95% confidence intervals only, without p-values, given their hypothesis-generating purpose.

- (1) **Supplemental O2 free days (SOFD)** will be analyzed using a proportional odds model to account for the unique features of this outcome. The model for these analyses will be

$$\log \frac{P(Y_i \leq l)}{P(Y_i > l)} = \delta_l + \beta_1 X_i + \beta_2 Z_i, l = 1, \dots, L - 1$$

A separate δ_l intercept is estimated for each of the thresholds between consecutive levels of SOFD, $-1, 0, 1, \dots, 28$. The treatment effect is β_1 and represents a log cumulative odds ratio between the randomized treatment arms.

- (2) **All-cause mortality to day 28** will be analyzed in the competing risks framework, as for time to room air described above. In this case, death is the event of interest and discharge alive is the competing risk, and censoring will be applied to patients alive and not discharged by day 28.

- (3) **Respiratory failure**, will be analyzed in the competing risks framework, as for mortality above, with the event of respiratory failure competing with death and discharge alive.
- (4) **Hospital-free days to day 28 (HFD28)** will be analyzed using a proportional odds model as described above for SOFD.
- (5) **Duration of time receiving no supplemental oxygen** (room air) during the first 72 hours after randomization will be analyzed in the same manner as the primary outcome; specifically, we will compare the mean proportion of observed patient-time without supplemental oxygen between treatment arms using a linear regression model.

Works Cited

- [1] H. Putter, M. Fiocco and R. B. Geskus, "Tutorial in biostatistics: Competing risks and multi-state models," *STATISTICS IN MEDICINE*, vol. 26, p. 2389–2430, 2007.
- [2] E. L'Her, S. Jaber, D. Verzilli, C. Jacob, B. Huiban, E. Futier, T. Kerforne, V. Pateu, P.-A. Bouchard, M. Consigny and F. Lellouche, "Automated closed-loop versus standard manual oxygen administration after major abdominal or thoracic surgery: an international multicentre randomised controlled study," *European Respiratory Journal*, vol. 57, p. 2000182, 2021.
- [3] H. White and H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817-838, 1980.
- [4] G. Verbeke and G. Molenberghs, *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics, Springer-Verlag, 2000.
- [5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data* (2nd ed.), Wiley-Interscience, 2002.