ICU Background Early Awareness for Critical deterioratiON (BEACON)

Statistical Analysis Plan

Gunnar Rätsch
ETH Zürich

Martin Faltys
Inselspital, Bern

Quinten Johnson
ETH Zürich

Manuel Burger
ETH Zürich

Jérôme Pasquier
ETH Zürich

SAP version 1.1
May 14th, 2025

Document scope

This document is developed in direct alignment with the proposed protocol for the express purpose of elaborating on the statistical analysis planned for the quantitative analysis of endpoints.

Protocol version

v1.1, dated 14.05.2025

BASEC Number

2025 - 00623

Version history

Version	Change History	Date
1.0	First version	25.03.2025
1.1	Revised version to add justification for sample size calculation with cluster crossover design and clarification on cluster bootstrapping	14.05.2025

SAP Authors / Contributors

Ver.	Contributor	Function
1.0, 1.1	Prof. Dr. Gunnar Rätsch	Professor in Biomedical Informatics
1.0, 1.1	Dr. med. Martin Faltys	Senior Physician, Inselspital
1.0, 1.1	Quinten Johnson	Project Manager
1.0, 1.1	Manuel Burger	MSc Data Science
1.0, 1.1	Dr. Jérôme Pasquier	Statistician

Roles and responsibilities

Sponsor: Prof. Dr. sc. nat. Gunnar Rätsch, on behalf of ETH Zürich

Date:_____ Signature:_____

Principal Investigator: Dr. med. Martin Faltys, Inselspital

Date:_____ Signature:_____

Study Statistician: Dr. Jérôme Pasquier

Date:_____ Signature:_____

Contents

Terms and Definitions

Term	Definition
ETH	Eidgenössische Technische Hochschule Zurich, Switzerland
Inselspital	University Hospital Bern
ICU	Intensive Care Unit / Department of Intensive Medicine
Clinician	A healthcare professional involved in diagnosing and treating patients (ICU physicians or nurses).
auROC	Area Under the Receiver Operating Characteristic (for binary classifier performance).
CIP	Clinical Investigation Plan

1 Background and Rationale

1.1 Prior research: Predictive models for organ function deterioration

Critically ill patients require treatment in an intensive care unit (ICU), where continuous monitoring of organ function parameters enables recognition of physiological deterioration and rapid onset of appropriate interventions. Clinicians monitor the patient's state, by analyzing the available information and, through further examinations of the patient. The intensive care physician has to constantly assess and decide which patient requires the most urgent attention, based on their medical conditions and the likelihood of acute deterioration.

This decision-making process is challenging because, while the physician focuses on one patient, other patients in the unit may not receive attention. This could potentially lead to unrecognized changes in the condition of the currently unattended patient. Our previous data[1,2] shows that BEACON scores (developed and validated with retrospective data by our group) can recognize early signs of potential organ failure events. We thus hypothesize that informing the physician about such possible early signs of deterioration, could improve the situational awareness of the ICU clinicians and possibly help to recognize patient's state deteriorations earlier (e.g., when the physician is otherwise currently not at the bedside). Improved situational awareness of the physician is hypothesized to improve performance of the ICU unit and possibly be associated with improved patient outcomes.

BEACON (Background Early Awareness for Critical deterioratiON) scores are used alongside standard ICU monitoring and standard clinical management. It is designed in-house by our group to increase unit performance without disruption of physicians. Physicians are expected to receive less than one notification of early signs of deterioration per shift (more if the physician is responsible for many patients). Overall, the BEACON scores may offer a significant benefit by assisting physicians' situational awareness and possibly enabling timely medical interventions with negligible risks. The expected effect on outcomes of individual patients is expected to be small. We therefore expect to see effects on clinical outcomes only on an ICU unit level and only when observing thousands of patients.

This analysis is designed as a cluster-randomized crossover analysis on unit performance where sub-units of the Inselspital ICU are randomly allocated to either the study arm (BEACON in addition to unit standard) or the control group (unit standard). We expect that the information provided by BEACON scores allows for better unit performance. No harm are foreseen for the patients admitted to the ICU sub-unit with or without additional BEACON scores.

Importantly, the BEACON set will have no impact on ICU standards: all ICU patients are treated 24/7 by a board-certified ICU physician according to the unit's standard operating procedures. This is an early feasibility analysis of the BEACON scores that will undergo next iterations / further improvements and validation settings after this study and before potential clinical use.

The (ICU) medical field as a whole might profit from computerized assistance[3,4] and clinicians support artificial intelligence (AI)-based tools development and use[5,6,7]. The field nevertheless experiences a so-called "translational gap"[8] between an abundance of retrospective studies and only minimal prospective clinical validation studies. This is recognized as one of the most pressing issues of the field. This early feasibility analysis can pioneer such efforts and (independently of the individual success of BEACON)

significantly contribute to the general knowledge required to implement and use such endeavor and help to improve ICU unit performance.

1.2 Primary Objective and Hypothesis

Primary Objective: The primary objective is to determine whether BEACON improves ICU sub-unit performance, i.e. reduces total mortality and/or mitigates severity of organ failures in an adult university ICU setting.

Primary Hypothesis: We hypothesize that the availability of BEACON scores will reduce 28-day all-cause mortality and/or mitigate organ-deterioration severity, as captured by our hierarchical "win-ratio" endpoint (described in Section 5.2),

While softer process measures (e.g., physician satisfaction, response times) can provide insight into workflow changes, the fundamental aim in the ICU is to reduce mortality and organ failure. These hard clinical outcomes are not only the ultimate indicators of patient benefit, but also the most objective and measurable endpoints. Therefore, this study focuses on these outcomes as they directly reflect improved unit performance.

1.3 Secondary objectives

Secondary objectives of the study are to estimate the impact of the availability of BEACON scores in the ICU sub-unit on the following outcomes:

1. Time Free of Circulatory Endpoint and Alive at 28 Days: total number of days within the first 28 days post-ICU-admission during which the patient remains alive and free from circulatory failure. The latter is defined as (1) elevated arterial lactate of $\geqslant 2$ mmol/L, and (2) either MAP $\leqslant 65$ mmHg, or the patient is receiving vasopressors or inotropes.
2. Time Free of Respiratory Endpoint and Alive at 28 Days: total number of days within the first 28 days post-ICU-admission during which the patient remains alive and free from respiratory failure (defined as P/F index < 150 mmHg, where arterial oxygen partial pressure is P and fractional inspired oxygen is F).
3. All-Cause Mortality at 28 Days: Mortality from any cause (including within the ICU and hospital settings) during the first 28 days following ICU-admission.
4. Mean total SOFA Score over the ICU stay: Mean value of the total SOFA score calculated on a daily basis within the first 28 days of ICU admission.

1.4 Patient disease definitions

Below are relevant definitions for organ dysfunction/failure events in this study:

- **Circulatory Failure:** Defined as mean arterial pressure (MAP) $\leqslant 65$ mmHg and/or the use of vasopressors/inotropes, plus lactate $\geqslant 2.0$ mmol/L.
- **Respiratory Failure:** Defined as a partial pressure of arterial oxygen (PaO_2) to fraction of inspired oxygen (FiO_2) ratio (PaO_2/FiO_2 ratio) of < 150 mmHg.
- **Other Organ Dysfunction:** Assessed via the Sequential Organ Failure Assessment (SOFA) score for renal, coagulation, and liver systems, with each organ system scored 0–4.

(Additional details on definitions and scoring criteria are in the main protocol's Appendix II.)

2 Study methods

2.1 Study design

Single-center, stratified, cluster-randomized[9,10] (the ICU sub-units will be randomized) multiple-period crossover analysis with outcome-assessor blinding. The analysis will be conducted in the adult intensive care units (ICUs) of Inselspital, Bern, comprising two distinct ICU units: the Blue and the Yellow ICU Unit. Each unit contains two sub-units, which will serve as the clustering units for randomization. (NB: The assignment to the Blue/Yellow unit is non-random and based on the expected length of stay of the patient. The assignment to the subunits (within the Yellow/Blue unit) can be considered random and only depend on utilization, but not on patient characteristics). The study follows a two-phase design. In the initial phase, sub-units within each ICU unit will be randomly allocated in a 1:1 ratio to either the additional BEACON scoring group or control group. Patients admitted to sub-units allocated to the BEACON group will receive standard care plus BEACON scores, while those admitted to control group will receive standard care only. After predefined cluster periods, allocations will be swapped between the BEACON and control groups. A wash-out period of two weeks will be observed between swaps to minimize carryover effects. During the wash-out period, no new study patients will be enrolled. Once the swap is complete, only newly admitted patients will be enrolled under the sub-units's new allocation. Patients still occupying sub-units from the preceding period will be censored from the study at the time of the swap and excluded from outcome analysis beyond that point. We will apply the same approach for patients who will be transferred from one unit to a different unit during the ICU admission for logistical/organizational reasons. Importantly, individual patients will only be exposed to one analysis condition—either the BEACON or control—based on the room allocation at the time of their admission. Only the sub-units themselves undergo crossover, ensuring temporal balance while maintaining patient-level exposure to a single condition. At least three swaps will be conducted – one before, during, and after the interim analysis – unless stopping criteria are met during the interim analysis, in which case only one swap will occur prior to the interim analysis (Section 2.4). This will ensure that each ICU sub-unit undergoes a minimum of four exposure periods, comprising two control periods and two BEACON periods. The initial number of periods will be determined prior to the start of the study and may be increased based on the findings of the interim analysis.

NB: We do not expect differences between the sub-units within the yellow/blue units. Nurses are on a rotation schedule across subunits and the physicians are assigned quasi-random to the subunits. Also, the patients are essentially assigned randomly, but subunit utilization as well as infections can influence the assignment of patients to specific sub-units or beds within the ICU. The cross-over design is aimed to reduce the effect of any remaining differences between the sub-units.

2.2 Randomization

Randomization will occur at the sub-unit level within each ICU unit (Blue and Yellow). Within each unit, sub-units will be initially randomized in a 1:1 ratio to the study group or control group using a computer-generated randomization schedule. To maintain allocation concealment, the initial randomization will be performed by the study's data management team, independent of the clinical staff. The chief nurse, responsible for patient sub-unit assignments based on clinical and logistical considerations, will not be involved in the study.

Following the initial randomization, sub-unit allocations will be swapped between the study and control groups at predefined intervals as described in the study protocol. A two-week wash-out period will be observed between sub-unit swaps, during which no new study patients will be enrolled. After the wash-out period, only newly admitted patients will be eligible for enrollment under the sub-unit's new allocation. Patients remaining in the sub-unit from the preceding period will be censored from the study at the time of the swap, ensuring that individual patients do not experience both conditions.

This is a cluster randomized crossover analysis where at least 16 cluster-periods are defined by four ICU sub-units and at least four exposure periods.

2.3 Sample size

A formal sample-size calculation is presented in Section 3.1 below. In summary, the study targets an effect size of 1.16 (Win Ratio) with $\alpha = 5\%$ (one-sided) and power = 80% and we plan with an effective sample size of $N = 1962$. The effective sample size requirements may increase up to $N = 2800$ as a result of interim analyses.

Based on typical ICU admission rates, the study is expected to observe the maximum of $N = 2800$ patients over approximately 20 months (including an assumed 3% dropout).

2.4 Statistical interim analysis and stopping guidance

An interim analysis will be conducted to allow for a potential sample size re-estimation. Specifically, once a proportion of the initially planned sample size have been enrolled and have complete primary endpoint data, the observed win ratio will be compared to the initially assumed value of 1.16. If the interim analysis indicates that the observed win ratio is materially lower than expected (but not so low as to suggest futility), the total sample size may be increased up to a maximum of 2800 patients. This approach is motivated by the practical method proposed by Mehta and Pocock, which provides for type I error control and preserves overall study power when the interim effect size is lower than anticipated[11].

The adaptive sample size increase will proceed only if the interim estimate falls within a predefined "promising zone," as described by Mehta and Pocock, whereby the study appears unlikely to meet its original power target unless additional patients are recruited. Conversely, if the interim results are consistent with the originally planned assumptions (i.e., the observed win ratio is near 1.16 or higher), no adaptation will be made.

If any unforeseen organizational concerns arise, the sponsor can pause or stop the study prematurely at any time.

2.5 Timing and outcome assessments

- In-hospital/ICU Data Collection: Continuous from ICU admission until discharge or death.
- Primary Endpoint: Evaluated up to day 28 post-ICU admission (mortality plus organ dysfunction during ICU stay).
- Secondary Endpoints: Measured throughout ICU stay and at day 28 (see Section 5.3).
- Longer Follow-up: If needed, hospital readmissions and mortality up to 28 days are captured from the Swiss national death registry and the hospital EHR (Section 9.2 of the CIP).

A final statistical analysis will be performed upon completion of the study (i.e., after the last subject completes the 28-day follow-up). Analyses will be conducted by the study statistician, who will remain blinded to group allocation until final data lock.

3 Statistical Principles

3.1 Study power

This study is designed to have 80% power at a one-sided significance level of $\alpha = 0.05$ to detect a win ratio greater than 1 for the study group compared with the control group. The anticipated effect size and proportion of ties are based on:

- A baseline 28-day mortality of approximately 20–30% in the emergency admission ICU population,
- An expected 15–20% relative reduction in mortality, along with improved organ-failure outcomes for the study group.

From previous observations and analyses on retrospective data, we estimate the proportion of ties to be 15% and assume a favorable outcome increase of 6.4% in the study group.

A tie occurs when given a pair of patients, one from the study group and one from the control group, it is not possible to determine which one has a better outcome. The probability of a tie is denoted as p_{tie}.

The increase in the proportion of favorable outcomes is defined as the *the proportion in favor of treatment* and is denoted as Δ.

The proportion in favor of treatment Δ is closely related to the win ratio Ψ:

$$\Psi = \frac{1 - p_{\text{tie}} + \Delta}{1 - p_{\text{tie}} - \Delta}.$$

Based on the estimated proportion of ties and the assumed win ratio (derived from the proportion in favor of treatment), the sample size is calculated as[12]:

$$N \approx \frac{\sigma^2 \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\log^2(\Psi)},$$

where:

- $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the critical values from the standard normal distribution for the specified significance level and power,
- σ^2 is given by
$$\sigma^2 = \frac{4\left(1 + p_{\text{tie}}\right)}{3\,k\,(1-k)\left(1 - p_{\text{tie}}\right)},$$
- k denotes the proportion of patients allocated to each group (the formula uses $k\,(1-k)$, so it is symmetric for the two groups).

Note: The PASS Sample Size Software from NCSS and the R package WRestimates both implement this method for power calculation.

Because this study follows a cluster-randomized crossover (CRXO) design, a correction for the sample size may be warranted as described by Giraudeau[9] and Arnup[10]. The sample size with the correction factor is given by,

$$N \approx \frac{\sigma^2 \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\log^2(\Psi)} \underbrace{\left[1 + (m-1)\,\rho - m\,\eta\right]}_{\text{CRXO correction factor}},$$

where m is the cluster size, ρ the within-period correlation, and η between-period correlation.

In this study, we assume that the between-period correlation η is approximately equal to the within-period correlation ρ. Under this assumption, the CRXO correction factor is close to one, yielding minimal change in the required sample size relative to a standard (non-CRXO) design.

For a win ratio of 1.16, accounting for 15% with win ratio ties and 6.4% with improved outcomes in the study group, the estimated total sample size for this study is 1962 patients.

3.2 Significance

The primary analysis will be conducted using a one-sided test at a significance level of 0.05, as we are specifically testing the hypothesis that the addition of a clinician awareness tool to the standard of care (as in the study group) is superior to the standard of care. This one-sided approach is justified by the research objective, which focuses exclusively on detecting an improvement in outcomes within the study group. Secondary outcomes and subgroup analyses will be interpreted with caution and considered exploratory unless otherwise specified.

3.3 Analysis population

The analysis population will include emergency ICU patients without a declined general research consent. A separate per-protocol analysis may be conducted to exclude major protocol deviations.

4 Study Population

4.1 Study eligibility

The ICU of the Inselspital Bern consists of two units, which will be considered for this study and its subsequent analysis. Each unit is subdivided into two sub-units.

To analyze the ICU unit performance, we will take the data of all emergency patients admitted to the respective ICU into account, except if they fulfill any of these criteria:

- age < 18 years
- presence of documented refused general consent form (at database closure);
- admission for the sole purpose of dying/organ donation or evaluation of such;

4.2 Recruitment

No active recruitment procedures will take place.

4.3 Withdrawal/Follow-up

Data from patients who refused the GC by the end of data collection will not be used and deleted. These patients will be replaced as to not compromise the power of the study.

4.4 Baseline patient characteristics

Relevant patient characteristics to be collected include:

- Demographics (age, sex, weight, height)
- Severity of illness score (e.g. APACHE)

4.5 Potential confounding covariates

Any known factors that might influence organ failure severity (e.g., baseline chronic organ dysfunction) or mortality (e.g., malignancy) will be documented. Randomization of ICU sub-units and cross-over is expected to balance these, but formal adjusted models may be considered which account for attributes, such as APACHE score on admission.

5 Analysis

5.1 Timing of final analysis

A final analysis will be performed once all enrolled patients have reached the 28-day follow-up or died/discharged before Day 28. (Section 2.4).

5.2 Primary Outcome definitions

The primary endpoint is determined using a stratified clustered hierarchical win ratio. Patient data is clustered by ICU sub-unit and stratified by APACHE score. See sections 2.1 and 5.4 for more details on the study design and the statistical analysis. For each stratum, each patient in the study group is compared with all patients in the control group within that same stratum. The hierarchical levels are:

1. **All-cause mortality within 28 days of ICU admission (binary):** Derived from the Inselspital or Swiss death registries.
2. **Number of organ systems failed** (respiratory, cardiovascular, coagulation, liver, renal) with SOFA scores newly increasing to ≥ 3 points during the ICU stay up to day 28.
3. **Sum of highest circulatory and respiratory SOFA scores** during the ICU stay (up to day 28), independently assessed and subsequently summed.

If one patient "wins" at the first level (i.e., is alive while the other is dead), the comparison is decided. If a tie occurs (both alive or both dead at 28 days), the next level is compared, and so forth. This approach prioritizes the most clinically serious outcome (mortality), while capturing additional organ dysfunction at different levels.

5.3 Secondary Outcome definitions

1. **Time Free of Circulatory Endpoint and Alive at 28 Days:** Total number of days within the first 28 days post-ICU admission where the patient is alive and free of circulatory failure. Circulatory failure is defined as lactate $\geqslant 2$ mmol/L plus MAP $\leqslant 65$ mmHg or receipt of vasopressors/inotropes.
2. **Time Free of Respiratory Endpoint and Alive at 28 Days:** Total number of days within the first 28 days post-ICU admission where the patient is alive and free from respiratory failure (PaO_2/FiO_2 ratio < 150 mmHg).
3. **All-Cause Hospital Mortality at 28 Days:** Mortality from any cause within 28 days following ICU admission.
4. **Mean Total SOFA Score** Over the ICU stay (up to 28 days)

5.4 Analysis methods

Primary Analysis

The primary endpoint will be evaluated by determining the win ratio (WR), where a WR greater than 1 indicates a favorable effect of the clinician awareness tool on ICU unit performance. The APACHE score, which is an important prognostic factor, will be used to form strata, and a stratified win ratio will be calculated. Hypothesis testing and confidence interval estimation for the WR will be conducted by using bootstrap methods[13,14,15], taking the randomization clusters into account. More details on the win ratio estimation are given in appendix A.

Secondary Analyses

- **Time-to-event analyses:** For "time free of failure endpoints," Kaplan-Meier curves or restricted mean survival time methods will be used.
- **All-cause hospital mortality at 28 Days:** A standard chi-square or Fisher exact test will be applied, and Kaplan-Meier survival curves may be displayed.
- **Mean total SOFA score:** The mean SOFA score will be compared between groups using a two-sample t-test or a an adjusted linear regression model.

Analysis of the other outcomes of interest

The other outcomes of interest listed in the protocol (section 2.1.3) will be analysed using the appropriate statistical methods as follows: Surival outcomes such as number of days alive and without invasive mechanical ventilation or vasopressor support will be analysed with Kaplan-Meier curves or Cox regression, continuous outcomes such as length of stay in ICU, and length of hospital stay will be analyzed using linear regression or non-parametric methods (e.g., Mann-Whitney U test), depending on the normality of

distributions. Categorical outcomes such as within-ICU mortality will be analyzed using logistic regression or Fisher's exact tests, as appropriate. Comparative performance in predefined subgroups will be assessed using interaction tests within regression models. Qualitative outcomes relating to acceptability, usability, adoption, appropriateness, feasibility, penetration, sustainability, ICU sub-unit efficacy, and physician perception will be analyzed through descriptive statistics (e.g., frequencies, proportions, medians) for quantitative measures, and thematic analysis for qualitative interview data.

5.5 Adjustment for covariates

The primary outcome is adjusted for the APACHE score by incorporating strata.

Although randomization of ICU sub-units and cross-over is expected to balance both measured and unmeasured confounders, multivariable regression models may be employed to adjust secondary outcomes for covariates such as baseline APACHE score, comorbidities, or age. This prespecified sensitivity analysis aims to enhance precision and address any potential chance imbalances.

5.6 Subgroup analyses

Subgroup analyses (e.g., by age groups, baseline disease severity, presence/absence of sepsis, etc.) will be prespecified to examine the consistency of the effect within the study group compared to the control group. These analyses will be exploratory. Formal interaction tests between the subgroup factor and BEACON arm may be performed.

5.7 Missing data

We will document all missing or incomplete data. If any key endpoints have missing data, the missing data mechanism will be assessed (e.g., missing completely at random, MCAR; missing at random, MAR; missing not at random, MNAR). If missingness is low, complete-case analysis may be sufficient. If non-negligible, multiple imputation methods or appropriate sensitivity analyses (e.g., best-worst scenario) will be performed.

5.8 Additional analyses

Additional (post-hoc) analyses may be conducted to explore potential trends and hypotheses that arise during the study. Such analyses will be clearly identified as exploratory.

5.9 Deviations from the SAP

Any deviations from the prespecified analyses in this SAP must be documented and justified in the final study report or publication, explaining the reasons and potential impact on results.

5.10 Statistical software

All analyses will be performed using **R** (e.g., version 4.0 or later) and/or **Python** (3.8 or later) for data cleaning, modeling, and graphical representations.

6 Summary Tables, Figures and Listings

6.1 List of tables

Table #	Description
1	Baseline characteristics of participants
2	Primary outcome (Win-ratio) analysis
3	Secondary endpoints descriptive statistics
4	Subgroup analysis results

6.2 List of figures

Figure #	Description
1	CONSORT flow diagram
2	Kaplan-Meier curves (for time-dependent secondary outcomes)
3	Forest plot for subgroup analysis

References

[1] Hyland, S. L. et al. "Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning". In: *Nature Medicine* 26.3 (2020), pp. 364–373. DOI: 10.1038/s41591-020-0789-4.

[2] Matthias, H. et al. "Early prediction of respiratory failure in the intensive care unit". In: *arXiv [cs.LG]* (2021).

[3] Li, Y. et al. "Advances in the application of AI robots in critical care: Scoping review". en. In: *J. Med. Internet Res.* 26 (2024), e54095.

[4] Leong, T.-Y., Aronsky, D., and Shabot, M. M. "Computer-based decision support for critical and emergency care". en. In: *J. Biomed. Inform.* 41.3 (2008), pp. 409–412.

[5] Meijden, S. L. van der et al. "Intensive care unit physicians' perspectives on artificial intelligence-based clinical decision support tools: Preimplementation survey study". en. In: *JMIR Hum. Factors* 10 (2023), e39114.

[6] Poncette, A.-S. et al. "Improvements in patient monitoring in the intensive care unit: Survey study". en. In: *J. Med. Internet Res.* 22.6 (2020), e19091.

[7] Bergauer, L. et al. "Avatar-based patient monitoring improves information transfer, diagnostic confidence and reduces perceived workload in intensive care units: computer-based, multicentre comparison study". en. In: *Sci. Rep.* 13.1 (2023), p. 5908.

[8] Van De, S. D. et al. "Moving from bytes bedside: systematic review use artificial intelligence intensive care unit". In: *Intensive Care Med* 47.7 (2021), pp. 750–760.

[9] Giraudeau, B., Ravaud, P., and Donner, A. "Sample Size Calculation for Cluster Randomized Cross-over Trials". In: *Statistics in Medicine* 27.27 (2008), pp. 5578–5585. DOI: 10.1002/sim.3383.

[10] Arnup, S. J. et al. "Understanding the Cluster Randomised Crossover Design: A Graphical Illustration of the Components of Variation and a Sample Size Tutorial". In: *Trials* 18.1 (2017), p. 381. DOI: 10.1186/s13063-017-2113-2.

[11] Mehta, C. R. and Pocock, S. J. "Adaptive Increase in Sample Size When Interim Results Are Promising: A Practical Guide with Examples". In: *Statistics in Medicine* 30.28 (2011), pp. 3267–3284. DOI: 10.1002/sim.4102.

[12] Yu, R. X. and Ganju, J. "Sample Size Formula for a Win Ratio Endpoint". In: *Statistics in Medicine* 41.6 (2022), pp. 950–963. DOI: 10.1002/sim.9297.

[13] Dong, G. et al. "The Stratified Win Ratio". In: *Journal of Biopharmaceutical Statistics* 28.4 (2018), pp. 778–796. DOI: 10.1080/10543406.2017.1397007.

[14] Pocock, S. J. et al. "The Win Ratio: A New Approach to the Analysis of Composite Endpoints in Clinical Trials Based on Clinical Priorities". In: *European Heart Journal* 33.2 (2012), pp. 176–182. DOI: 10.1093/eurheartj/ehr352.

[15] Ng, E. S.-W., Grieve, R., and Carpenter, J. R. "Two-Stage Nonparametric Bootstrap Sampling with Shrinkage Correction for Clustered Data". In: *The Stata Journal* 13.1 (2013), pp. 141–164. DOI: 10.1177/1536867X1301300111.

[16] Dong, G. et al. "The Stratified Win Statistics (Win Ratio, Win Odds, and Net Benefit)". In: *Pharmaceutical Statistics* 22.4 (2023), pp. 748–756. DOI: 10.1002/pst.2293.

[17] Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and Their Application*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1997.

A Definition of the Win Ratio Estimator

The APACHE score will define the strata, and a post-stratified win ratio will be computed. Due to the study design, the variance estimator will account for the randomization clusters. The following section outlines the stratified win ratio estimator and the method for calculating the bootstrap distribution.

A.1 Stratified Win Ratio Estimator

We consider S strata, and define:

- $X^{(s)}$ and $Y^{(s)}$ represent the random outcomes for two patients, one from the BEACON group and the other from the control group, respectively, within the stratum s, where $s = 1, \ldots, S$.
- $X^{(s)} \succ Y^{(s)}$ as the event occurring if $X^{(s)}$ wins over $Y^{(s)}$.
- $\phi_1(X, Y)$ and $\phi_2(X, Y)$ as the indicator functions for the events $X \succ Y$ and $Y \succ X$, respectively:

$$\phi_1(X, Y) \;=\; 1_{\{X \succ Y\}} \quad \text{and} \quad \phi_2(X, Y) \;=\; 1_{\{Y \succ X\}}.$$

- $\tau_k^{(s)} = E[\phi_k(X^{(s)}, Y^{(s)})]$, $k = 1, 2$, as the probabilities of the events $X^{(s)} \succ Y^{(s)}$ and $Y^{(s)} \succ X^{(s)}$, respectively.

Given two stratified samples

$$X_1^{(1)}, \ldots, X_{m^{(1)}}^{(1)}, \ldots, X_1^{(S)}, \ldots, X_{m^{(S)}}^{(S)} \quad \text{and} \quad Y_1^{(1)}, \ldots, Y_{n^{(1)}}^{(1)}, \ldots, Y_1^{(S)}, \ldots, Y_{n^{(S)}}^{(S)},$$

where $N^{(s)} = m^{(s)} + n^{(s)}$ is the sample size in stratum s, an unbiased estimator of $\tau_k^{(s)}$ is

$$\hat{\tau}_k^{(s)} \;=\; \frac{1}{m^{(s)}\, n^{(s)}} \sum_{i=1}^{m^{(s)}} \sum_{j=1}^{n^{(s)}} \phi_k\big(X_i^{(s)}, Y_j^{(s)}\big), \quad k = 1, 2,$$

The stratified estimators $\hat{\tau}_k^{\text{strat}}$ are then defined by[13,16]

$$\hat{\tau}_k^{\text{strat}} \;=\; \frac{1}{W} \sum_{s=1}^{S} w^{(s)}\, \hat{\tau_k}^{(s)}, \quad \text{where} \quad w^{(s)} \;=\; \frac{m^{(s)}\, n^{(s)}}{N^{(s)}} \quad \text{and} \quad W \;=\; \sum_{s=1}^{S} w^{(s)}.$$

Finally, the stratified win ratio is estimated by

$$\hat{\Psi}^{\text{strat}} \;=\; \frac{\hat{\tau}_1^{\text{strat}}}{\hat{\tau}_2^{\text{strat}}}.$$

In essence, $\hat{\tau}_1^{\text{strat}}$ measures how often a study group patient "wins" over a control patient across all strata, while $\hat{\tau}_2^{\text{strat}}$ measures how often the control patient "wins." The resulting ratio $\hat{\Psi}^{\text{strat}}$ quantifies the overall comparative benefit of the BEACON clinician awareness tool, prioritizing mortality first and then organ dysfunction.

Note that

$$\hat{\Psi}^{\text{strat}} \;=\; \frac{\sum_{s=1}^{S} \dfrac{\nu_1^{(s)}}{N^{(s)}}}{\sum_{s=1}^{S} \dfrac{\nu_2^{(s)}}{N^{(s)}}}, \quad \text{where} \quad \nu_k^{(s)} = \sum_{i=1}^{m^{(s)}} \sum_{j=1}^{n^{(s)}} \phi_k\big(X_i^{(s)}, Y_j^{(s)}\big), \quad k = 1, 2.$$

This expression corresponds to a Mantel-Haenszel-like weighted estimator. The weighting scheme is inspired by its similarity to the stratified odds estimator proposed by Dong et al.[13].

A.2 Bootstrap Distribution

A two-stage bootstrap approach will be employed. For each bootstrap replicate, we will begin by sampling four sub-units with replacement. Within each selected sub-unit, we will then randomly sample periods—at least four per sub-unit—with replacement, ensuring that entire cluster-periods are preserved in the replicates[17,15]. This method maintains the integrity of both the cluster and crossover structure. The win ratio will be calculated for each replicate, and the variance of the estimator will be obtained from the resulting empirical distribution.