1. Methods Overview

1.1. Open Cohort Stepped Wedge Design

Let M denote the total number of clusters participating in the study. At the first time point t_0, all clusters are in the control condition and have not yet received the intervention. At subsequent time points $t_1, t_2, \ldots, t_s$, clusters begin receiving the intervention. Assume cluster m $(m = 1, 2, \ldots, M)$ switches to the intervention at step $s(s = 1, 2, \ldots, S)$. However, at a step s the transition to intervention condition can happen to either just one cluster or multiple clusters at a time. Over S steps, all clusters progressively switch to the intervention.

Similarly, let the total duration of the experiment be T (where $t = t_0, t_1, \ldots, t_n$) with the starting time at t_0. The time interval lengths $\Delta t_k = t_k - t_{k-1}$ represents the time difference between two consecutive measurement points. These time intervals length Δt_k can be either equal or unequal, typically determined by the practical needs of the experiment or resource availability. For example, the time intervals lengths can be every 1 month or every 3 months or every 6 months. A switch s happens at t_s; which is immediately after t_k measurement point. The end of the experiment occurs at time point t_n, by which time all clusters have transitioned to the intervention phase, and the study moves into the final data collection stage. At this point, comparisons between the control and intervention phases across different clusters can be made to assess the overall effect of the intervention.

For example, in an experiment with 3 clusters ($M = 3$ clinics), one clinic transitions at each switch s ($S = 3$ steps), with outcome measures performed every 3 months ($\Delta t_k = 3$ months), switches happening every 3 months ($\Delta t_s =$ 3 months), for a total of 6 time points $(t_0, t_1, t_2, t_3, t_4, t_5)$, which correspond to the baseline period, 3 subsequent cluster transitions and the study's end time, respectively. Then the experiment would proceed as follows: $[t_0 - t_1)$ all clusters are in the control condition; at t_1, one randomly selected cluster will switch to the intervention phase; at t_2, another cluster will switch; and finally, at t_3, the last cluster will transition to the intervention phase. At this point, all clusters have received the intervention at t_4, the last individual enters the study, and the experiment concludes at $t_n = t_5$.

In an open cohort study, participants are allowed to enter or exit the study at any time during the experiment's duration T [1]. Let s_m be the step at which cluster m transitions to the intervention, and Δt_s represent the time interval lengths between steps, then $t_{sm} = s_m \times \Delta t_s$ represents that switch time for cluster m . Let individual i in cluster m, where $i = 1, 2, \ldots, N_m$. They will only experience:

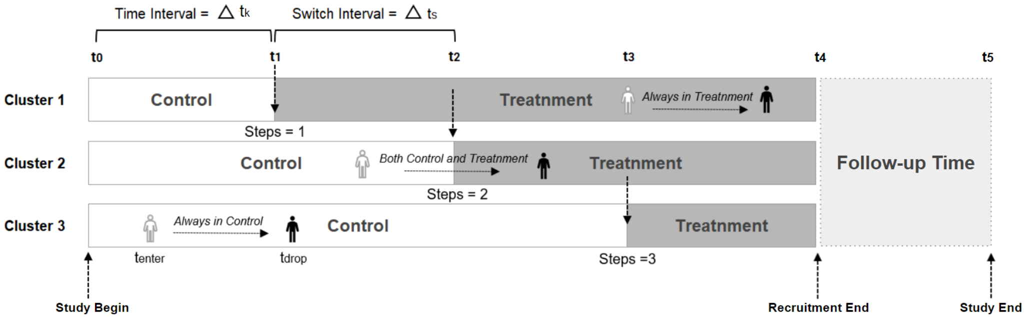

Figure 1: Visualization of Stepped Wedge Design in Open Cohort Study; M=3 Clinics

1. The control condition if the last observation time (t^{c}_{im}) or the event time $(t^{*}_{im}) < t_{sm}$.
2. The treatment condition if the entry time $(t^{e}_{im}\ ) > t_{sm}$.
3. Both conditions if they enter the study between these time points.

In *Figure 1*, the stepped wedge design is illustrated in an open cohort study format. This diagram shows how different clusters (e.g., clinics) transition from the control condition to the treatment condition over time.

1.2. Analysis of Discrete Time to Event Outcomes in SWCRT

The discrete-time hazard model, based on Allison's framework, is particularly suitable for analyzing time-to-event data in a SWCRTs, where time intervals are predefined, and events are observed in discrete periods [2]. To enhance its applicability to SWCRTs, we incorporated open cohort and random effects across clusters into the existing method, ensuring a better fit for the unique characteristics of SWCRTs. Starting from t_0 through t_n, each d_k corresponds to a consecutive interval bounded by the measurement time points $[t_{k-1}, t_k)$. Assume t^{*}_{im} and t^{c}_{im} represent the actual event time and censoring time for subject i in cluster m, respectively. The times are not directly observed but are recorded within specific intervals $d_1, d_2, \ldots d_n$. Thus, we define the we define the observed event interval d^{*}_{im} as:

$$d^{*}_{im} = \min\{k : t^{*}_{im} \leq t_k\}, \tag{1}$$

where d^{*}_{im} is the first interval such that the event time t^{*}_{im} is less than or equal to the upper boundary t_k of the interval.
Similarly, we define the observed censoring interval d^{c}_{im} as:

$$d^{c}_{im} = \max\{k : t^{c}_{im} \geq t_{k-1}\}. \tag{2}$$

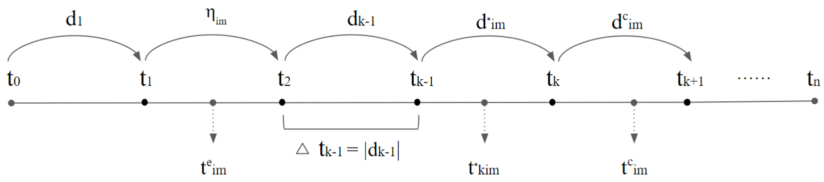

Figure 2: The Relation of Time Point and Interval in SWCRTs Setting

Here, d^c_{im} is the largest interval such that the censoring time t^c_{im} is greater than or equal to the lower boundary t_{k-1} of the interval, representing the last interval before the subject is censored or the study ends.

Participants may not all enter the trial at its onset. Therefore, we further define the time to event $\xi_{im} = d^*_{im} - \eta_{im} + 1$, where η_{im} represents the interval in which subject i from cluster m joins the trial.

Then, let the hazard rate for subject i in cluster m, denoted as $h_{im}(k)$, represent the probability that the event occurs at the k-th time interval, given that it has not occurred prior to this time, we can write the hazard rate as:

$$h_{im}(k) = \Pr(\xi_{im} = k \mid \xi_{im} \geq k, x_{im}), \tag{3}$$

where x_{im} is the vector of explanatory variables. The next step is to specify how the hazard rate depends on both time and the explanatory variables. One of the most common approaches to modeling the hazard in discrete-time survival data is through the complementary log-log function [3]. This approach is well-suited for discrete-time data as it models the probability of an event occurring in each time interval and focuses on the cumulative hazard over time [3]. The complementary log-log function is expressed as follows:

$$\log\left[-\log(1 - h_{im}(k))\right] = \alpha_k + \beta' x_{im}, \tag{4}$$

where α_k is the contribution from the k-th censoring interval, and the term β' represents the coefficient vector associated with the covariates, capturing their influence on the hazard rate across time intervals.

In SWCRTs, we can include both time-varying (e.g., biomarker level) and time-invariant covariates (e.g., gender and age at start of study). Let's assume individual i is in cluster m, where their time-invariant covariates are denoted as Z_{imr}, representing the r-th covariate among the total R time-invariant covariates. The switch time for cluster m is denoted as t_{sm}. Then, the treatment indicator $I_{im}(k)$can be defined as:

$$I_{im}(k) = \begin{cases} 1, & \text{if } t_k > t_{sm} \\ 0, & \text{if } t_k \leq t_{sm} \end{cases} \tag{5}$$

Each cluster is assumed to have its own baseline hazard, and the treatment effect also can vary across clusters due to cluster-specific characteristics. To account for these sources of variability in the outcome, we incorporate random effects at both the cluster-level and the treatment within each cluster. Specifically, τ_m represents the random effect for each cluster, where $\tau_m \sim N(0, \sigma_m^2)$, and $\tau_{m\phi}$ denotes the random effect of the treatment within each cluster, where $\tau_{m\phi} \sim N(0, \sigma_{m\phi}^2)$. Accordingly, equation (4) can be expanded as:

$$\begin{aligned} \log\left[-\log(1 - h_{im}(k))\right] = \alpha_k + \tau_m + \gamma_{m,\eta_{im}+k-1} \\ + (\phi + \tau_{m\phi}) I_{im}(k) + \sum_{r=1}^{R} \beta_r Z_{imr} \end{aligned} \tag{6}$$

Here, α_k represents the effect of the k-th censoring interval, counted from the entry interval η_{im}, and $\gamma_{m,\eta_{im}+k-1}$ denotes the cluster-specific time effect at interval $\eta_{im}+k-1$, which accounts for differences in follow-up duration among individuals, thereby avoiding the assumption that all participants have identical follow-up periods. Additionally, ϕ represents the regression coefficient for treatment effect, and β_r is the coefficient for the time-independent covariates. Hence, the hazard function $h_{im}(k)$ from function (6) can be represented as:

$$\begin{aligned} h_{im}(k) = 1 - \exp\Bigg[- \exp\Big(\alpha_k + \tau_m + \gamma_{m,\eta_{im}+k-1} \\ + (\phi + \tau_{m\phi}) I_{im}(k) + \sum_{r=1}^{R} \beta_r Z_{imr}\Big)\Bigg] \end{aligned} \tag{7}$$

The corresponding survival function can be expressed as the complement of the cumulative hazard function:

$$S_{im}(k) = \prod_{s=1}^{k} (1 - h_{im}(s)) \tag{8}$$

Now, since we already have the hazard function and the survival function, for each individual i in cluster m the likelihood depends on whether the event occurs at time interval d_k or if the data is censored. For uncensored data, the likelihood contribution is the product of hazard function at the event time interval d_k with the survival function at d_{k-1}, just before time interval d_k:

$$L_{im}(k) = h_{im}(k) S_{im}(k-1) \tag{9}$$

Otherwise, for censored data, the likelihood contribution is the survival probability up to the censoring time interval d_k:

$$L_{im}(k) = S_{im}(k) \tag{10}$$

Thus, the overall likelihood function is the product of the likelihood contributions for both types of data. Then, let Γ_{im} represent the censoring indicator, which specifies whether the event was observed during the trial period, recall from equation (1) and (2):

$$\Gamma_{im} = \begin{cases} 1 & \text{if } d^*_{im} \leq d^c_{im} \\ 0 & \text{if } d^*_{im} > d^c_{im} \end{cases} \tag{11}$$

Also, by substituting equation (8) into the equation (9) and (10), we arrive at the overall likelihood:

$$\begin{aligned} L &= \prod_{m=1}^{M} \prod_{i=1}^{N} \left[\left(h_{im}(k) S_{im}(k-1) \right)^{\Gamma_{im}} S_{im}(k)^{1-\Gamma_{im}} \right] \\ &= \prod_{m=1}^{M} \prod_{i=1}^{N} \left[\left(h_{im}(k) \prod_{s=1}^{k-1} (1 - h_{im}(s)) \right)^{\Gamma_{im}} \left(\prod_{s=1}^{k} (1 - h_{im}(s)) \right)^{1-\Gamma_{im}} \right] \\ &= \prod_{m=1}^{M} \prod_{i=1}^{N} \left[\left(\frac{h_{im}(k)}{1 - h_{im}(k)} \right)^{\Gamma_{im}} \prod_{s=1}^{k} (1 - h_{im}(s)) \right] \end{aligned} \tag{12}$$

Then we can have the log likelihood function:

$$\log L = \sum_{m=1}^{M} \sum_{i=1}^{N} \left[\Gamma_{im} \log \left(\frac{h_{im}(k)}{1 - h_{im}(k)} \right) + \sum_{s=1}^{k} \log\left(1 - h_{im}(s)\right) \right] \tag{13}$$

We can further simplify this equation by defining another indicator $y_{imk} = 1$, which indicates that individual i in cluster m experiences the event at time interval d_k; otherwise, $y_{imk} = 0$. Hence we can get the final likelihood function for SWCRTs:

$$\log L = \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{s=1}^{k}\left[y_{imk}\log\left(\frac{h_{im}(s)}{1-h_{im}(s)}\right)+\log\left(1-h_{im}(s)\right)\right]$$
$$= \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{s=1}^{k}\left[y_{imk}\log(h_{im}(s))+(1-y_{imk})\log(1-h_{im}(s))\right] \tag{14}$$

1.3. Data Simulation

The likelihood function and hazard function for the discrete model in SWCRTs have been established in **section 1.2**. To simulate trials, we must simulate data and fit that model to them. In SWCRTs, it is important to note that event times are inherently continuous. However, in discrete analysis, we can only capture discrete data due to the constrains of practical considerations. To better represent this scenario, our data simulation follows a two-step process:

1. Generate continuous event times using a Weibull distribution.
2. Convert the continuous event times into discrete intervals.

The Weibull distribution is a widely used non-parametric model, providing flexibility in shaping the pattern of generated data[4]. The behavior of the distribution is controlled by two key parameters: the shape parameter λ and the scale parameter ν, which can be changed during the simulation[4]. In the data simulation process, we define several key time points: the study start t_0, study end t_n, switch begin $t_{s=1}$, and recruitment end t_{re}. The switch begin marks the time when the first intervention switch S is implemented, while recruitment end represents the point after which no new participants are recruited into the study.

To simulate individual-level data, we first generate the entry time for individual i in cluster m, denoted as t^{e}_{im}, using a uniform distribution between t_0 and t_{re}. This represents the moment when the individual becomes at risk for the event under study.

After entering the study, the individual may either experience the event or leave the study prematurely (i.e., censoring). Therefore, we also need to simulate both the potential last observational time t_{nim} and the event occurrence time t^{*}_{im}. The dropout time from the t_0, $\hat{t}_{nim}$, is simulated using a Weibull distribution with the probability density function:

$$f(x) = \frac{a}{c}\left(\frac{x}{c}\right)^{a-1} e^{-\left(\frac{x}{c}\right)^{a}} \tag{15}$$

Where $a = \exp(\lambda)$, and $c = t_{re} - t_{s=0}$ represents the duration of the recruitment period. Consequently, the actual dropout time is calculated as

$t_{nim} = t^e_{im} + \hat{t}_{nim}$, indicating when the individual leaves the study, if censoring occurs.
Subjects will no longer be observed once the event occurs or they withdraw from the study. Therefore, the final censoring time, t^c_{im}, is determined by comparing the study end time t_n (or a predefined follow-up period, e.g., 1 year), the dropout time t_{nim}, and the event time t^*_{im}. The smallest value among these represents the actual censoring time, t^c_{im}.
Finally, we simulate the uncensored event time for each subject starting from the t_0, denoted as $\hat{t}^*_{im}$. Let $U \sim U(0, 1)$ be a uniformly distributed random variable, and let w_{im} represent the time difference between subject i entering the study and the switch time for cluster m, where $w_{im} = t_{sm} - t^e_{im}$. The outcome can be modeled under two distinct cases:
If $-\log(U) < \lambda \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m\right) w^\nu_{im}$, which indicates that there is no treatment effect involved, the event time t^*_{kim} is calculated as:

$$\hat{t}^*_{im} = \left[\frac{-\log(U)}{\lambda \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m\right)}\right]^{\frac{1}{\nu}} \tag{16}$$

Otherwise, if $-\log(U) \geq \lambda \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m\right) w^\nu_{im}$, the event time is computed as:

$$\begin{aligned}\hat{t}^*_{im} = \Bigg[& \frac{-\log(U) - \lambda \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m\right) w^\nu_{im}}{\lambda \exp(\phi) \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m + \tau_{m\phi}\right)} \\ & + \frac{\lambda \exp(\phi) \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m + \tau_{m\phi}\right) w^\nu_{im}}{\lambda \exp(\phi) \exp\left(\sum_{r=1}^{R} \beta_r Z_{imr} + \tau_m + \tau_{m\phi}\right)}\Bigg]^{\frac{1}{\nu}}\end{aligned} \tag{17}$$

Where ϕ denotes the regression coefficient for treatment effect, indicating the presence of the intervention. Similarly, the actual event time for each subject is calculated as $t^*_{im} = t^e_{im} + \hat{t}^*_{im}$.
Subsequently, all time variables—including censoring time, dropout time, and entry time—are transformed into interval segments and reformatted into a longitudinal structure to enable discrete-time analysis. The intervals are based on the time interval lengths Δt_k, where the lower bound is greater than 1. For example, if subject i enters the study 20 days after t_0, the discrete format of η_{im} is $\lfloor 20/\Delta t_n \rfloor + 1$. Hence, if the $\Delta t_n = 30$ days, the discrete value for η_{im} would be $0 + 1 = 1$.

References

[1] Skrivankova, V. W., et al. (2021). Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): Explanation and elaboration. *BMJ*, 375, n2233. https://doi.org/10.1136/bmj.n2233

[2] Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13, 61–98. https://doi.org/10.2307/270718

[3] Suresh, K., Severn, C., & Ghosh, D. (2022). Survival prediction models: An introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22. https://doi.org/10.1186/s12874-022-01679-6

[4] Gómez, Y., Gallardo, D., Marchant, C., Sánchez, L., & Bourguignon, M. (2023). An In-Depth Review of the Weibull Model with a Focus on Various Parameterizations. *Mathematics*, 12, 56. https://doi.org/10.3390/math12010056