AIMS Mathematics, 10(12): 29168–29188.
DOI: 10.3934/math.20251283
Received: 17 October 2025
Revised: 02 December 2025
Accepted: 04 December 2025
Published: 11 December 2025

https://www.aimspress.com/journal/Math

Research article

Two-stage Fill-it-up design method incorporating historical control data in binary endpoint clinical trials

Junjiang Zhong[1], Haoyun Guo[2], Nan Sun[3] and Junjie Li[1,*]

[1] School of Mathematics and Statistics, Xiamen University of Technology, Xiamen 361024, China

[2] School of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China

[3] BeOne Medicines, Shanghai 200023, China

* **Correspondence:** Email: lijunjie@xmut.edu.cn.

Abstract: Leveraging historical control data to augment randomized control data in clinical trials has become an important strategy for improving the efficiency of statistical inference, particularly in contexts with limited sample availability. However, potential heterogeneity between historical and current datasets may introduce bias in treatment effect estimation and compromise inferential validity. This study proposes a two-stage "Fill-it-up" design for clinical trials with binary endpoints to enable the rigorous integration of historical control data under controlled statistical risk. Analytical procedures for sample size determination and practical implementation steps for both stages are provided. Simulation studies demonstrate that the family-wise error rate can be effectively controlled below the pre-specified significance level, while the average sample size is substantially reduced compared with a conventional single-stage design that excludes historical controls. The efficiency gains become more pronounced as between-group heterogeneity decreases. The proposed two-stage Fill-it-up design offers a frequentist framework for safely and efficiently incorporating historical control data into binary endpoint trials. Given that additional recruitment is required when equivalence is not established, its practical application is best suited for studies where there is strong prior confidence in the quality and comparability of historical data, such as in rare disease or pediatric settings. This design provides a pragmatic approach for enhancing the efficiency and ethical sustainability of modern clinical research.

Keywords: Fill-it-up design; historical control; binary endpoint; equivalence
Mathematics Subject Classification: 62F03, 62L05, 62P10

1. Introduction

In clinical research, the integration and appropriate use of external historical data have become topics of growing methodological and regulatory interest. To promote the standardization and transparency of such practices, the U.S. Food and Drug Administration released a draft guidance in 2023 on the design and analysis of externally controlled trials [1]. This guidance document systematically outlines the core principles, design considerations, and statistical recommendations for externally controlled studies, providing an important regulatory reference for both clinical and statistical researchers. From a methodological perspective, the incorporation of external controls (e.g., historical datasets or real-world data) can improve the efficiency of drug development by enhancing statistical power and reducing the required sample size in randomized controlled trials, ultimately shortening the overall development timeline. This approach is particularly valuable in settings where patient recruitment is inherently difficult, such as studies involving rare diseases or pediatric populations. In these contexts, external historical data can serve as a valuable supplementary information source, partially mitigating the limitations imposed by small sample sizes. When properly adjusted for potential heterogeneity between data sources, this integration can provide more robust evidence for the evaluation of the efficacy and safety of treatment, thus facilitating the evaluation and development of therapeutics for special populations.

In recent years, there has been a growing methodological interest in approaches that leverage historical data to enhance the efficiency of clinical research. A critical consideration in this context is the potential heterogeneity between current and historical datasets, as failure to adequately account for such differences may compromise the validity and reliability of study inferences. To address this issue, a variety of Bayesian dynamic borrowing methods have been developed to adaptively incorporate information from external or historical controls while mitigating potential discrepancies between data sources. Representative approaches include power prior [2], modified power prior [3–5], commensurate prior [6], probability-weighted power prior [7], and robust meta-analytic-predictive (MAP) prior [8, 9].

In addition to the Bayesian approaches described above, frequentist methods are also commonly employed to evaluate whether historical data should be incorporated into current analyses [10]. The fundamental principle of these approaches lies in formally testing the consistency between current and historical datasets. If no statistically significant difference is detected, the two data sets can be combined to improve statistical efficiency; conversely, if a difference is identified, historical data are excluded to avoid bias. Li et al. [11] proposed an equivalence-based test-then-pool approach, in which the degree of similarity between datasets is quantified through the overlap area between their probability distributions, thus providing an objective criterion for determining the appropriateness of data pooling.

A central challenge in modern experimental design is how to incorporate external or historical information while maintaining statistical validity when external and internal data are not fully exchangeable. Existing approaches such as Bayesian dynamic borrowing (e.g., power priors or commensurate priors) and test-then-pool strategies provide useful tools, but they often rely on subjective tuning, introduce discrete decision boundaries, or lack rigorous guaranties for Type I error control under data heterogeneity. To address these limitations, the objective of this study is to develop a unified, model-based, and data-adaptive design framework that enables continuous information

borrowing and provides robust inference even when the compatibility between external and internal data is uncertain.

More recently, Wied et al. [12] introduced the Fill-it-up design, which evaluates the comparability of the historical and current control groups through a preliminary equivalence test. If equivalence is confirmed, the historical controls are incorporated into the ongoing randomized trial; otherwise, they are excluded, and randomization continues until the planned sample size is achieved. However, the Fill-it-up design proposed by Wied et al. [12] is limited in applicability, as it was developed specifically for continuous, normally distributed endpoints and cannot be extended directly to binary endpoints. To address this limitation, this paper proposes a two-stage Fill-it-up design tailored for binary endpoints, allowing efficient integration of historical control data while maintaining statistical rigor.

The remainder of this paper is organized as follows. Section 2 presents a motivating clinical trial example. Section 3 describes the methodology of the proposed two-stage Fill-it-up design. Section 4 examines the frequentist operating characteristics of the method, including sample size determination, Type I error control, and power evaluation. Section 5 reports the results of the simulation studies conducted to assess the design's performance in improving sample efficiency. Section 6 provides a comparison between our proposed approach and the Bayesian MAP approach. Section 7 illustrates the practical application of the design in a clinical trial setting. Finally, Sections 8 and 9 conclude the paper with a summary of key findings and a discussion of limitations.

2. A motivation example

Belimumab, a GlaxoSmithKline-developed biologic agent, specifically targets systemic lupus erythematosus (SLE). To assess its therapeutic efficacy in pediatric patients, an initial cohort of 124 subjects was screened, of whom 92 met the eligibility criteria and were randomized to a controlled clinical trial. Among them, 39 patients received placebo (control group) and 53 received belimumab at a dose of 10 mg/kg (treatment group) [13]. The primary endpoint was SLE Responder Index 4 (SRI4) response rate at Week 52, defined as $\geq$ 4-point reduction form baseline in SELENA-SLEDD score. The corresponding observed response rates were 52.83% and 43.55% for treatment and control groups, respectively. The superiority hypothesis was performed using only current control data, and the resulting p value was 0.427.

Although PLUTO was designed and analyzed using a frequentist approach without borrowing historical trial data, the SRI4 response rate from a historical trial from adult clinical studies of belimumab [14] was available at the time of the PLUTO trial. We would like to see how these historical trial data may have been used to borrow information for the control group and how it would have impacted the superiority analysis. In this study, we consider PLUTO as a current trial to compare the effect of 10 mg/kg of belimumab with placebo on the SRI4 response rate at Week 52 after randomization. The placebo data (control group) from adult clinical studies of belimumab [14] are used as historical control data in the study. To increase the effective sample size and improve the power of statistical inference, detailed data from pediatric and adult trials are summarized in Table 1. In Section 7, we demonstrate how incorporating historical control information can meaningfully improve the statistical efficiency of the superiority evaluation, thus enhancing the ability to detect true differences between treatment and control.

Table 1. Summary of response rates for historical and current trials.

Study	Reference	Group	n	Responders	Response rate
Current study (PLUTO)	[13]	Control	39	17	43.55%
		Treatment	53	28	52.83%
Historical study	[14]	Control	287	125	43.59%

3. Methodology

3.1. Notations

Let the ith endpoint for the current treatment group be denoted by y_{Ei} ($i = 1, \ldots, n_E$), and that for the current control group by y_{Ci} ($i = 1, \ldots, n_C$). In addition, suppose that there exists a historical control group with a sample size of n_H, whose ith endpoint is denoted by y_{Hi} ($i = 1, \ldots, n_H$). It is generally assumed that the historical control group and the current control group share a certain degree of similarity in the distribution of their endpoints; this similarity is further evaluated through an equivalence test.

All endpoints in the above groups are binary outcomes taking values of 0 or 1. The efficacy of the treatment, current control, and historical control groups are denoted by p_E, p_C, and p_H, respectively, and can be estimated by their corresponding sample means $\bar{y}_E, \bar{y}_C$, and $\bar{y}_H$. If the equivalence between the current and historical control groups cannot be established, additional samples (n'_E and n'_C) are enrolled in the treatment and current control groups, respectively. After supplementation, the total sample sizes for the treatment and control groups become $N_E = n_E + n'_E$ and $N_C = n_C + n'_C$, respectively. The updated mean estimates of the treatment effects based on the expanded samples are then denoted by $\bar{y}'_E$ and $\bar{y}'_C$.

3.2. Two-stage Fill-it-up design

The implementation of the proposed two-stage Fill-it-up design consists of two sequential stages. Stage I begins with a small randomized controlled trial, where the initial sample size is determined under the assumption that historical control data will be eligible for inclusion in the final analysis. The equivalence between the current and historical control groups is formally evaluated at this stage. If equivalence is not established—indicating a statistically significant difference between the two control groups—the design proceeds to Stage II, during which randomization continues and additional subjects are enrolled until the pre-specified statistical power for the clinical trial is achieved.

This design framework incorporates three key hypothesis tests. The first test assesses the equivalence between the current control and historical control groups. If equivalence pre-test (Ept) is confirmed, the two datasets are pooled and a subsequent superiority test (S_1) is conducted between the treatment group and the combined control group. In contrast, if equivalence is rejected, the historical control data are excluded, and Stage II enrollment is triggered. In this stage, an additional $n'_E + n'_C$ participants are randomized equally between the treatment and control arms. The data pooled from both stages are then used to perform the final superiority test (S_2) comparing the treatment and current control groups. A schematic overview of the two-stage Fill-it-up design and its implementation process is presented in Figure 1.

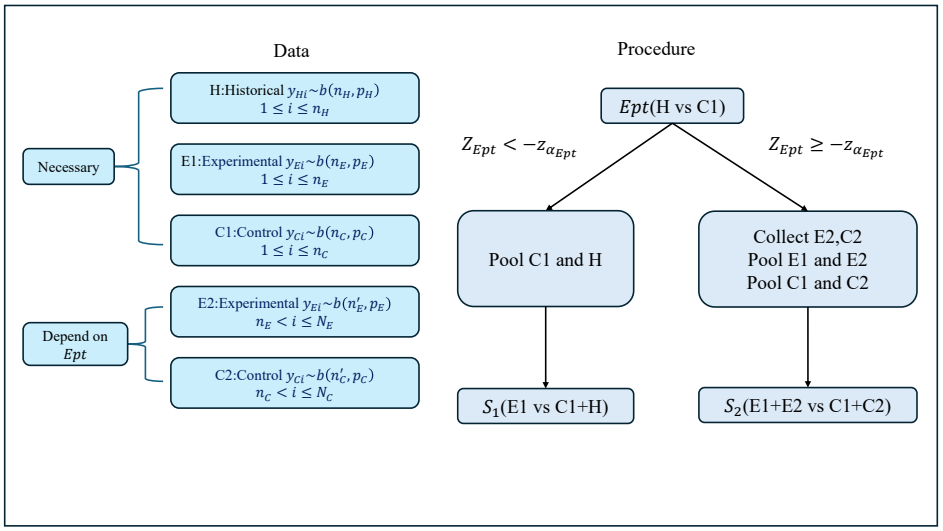

Figure 1. Flow chart of the procedure of the two-stage Fill-it-up design.

3.3. Equivalence test

We first introduce the initial hypothesis test among the three—the pre-specified equivalence test between the current control group and the historical control group. To proceed, several key parameters are defined. Let Δ denote the pre-specified equivalence margin ($\Delta > 0$), whose magnitude is determined based on clinical considerations and relevant prior studies. Let α_{Ept} represent the significance level for this pre-test, and $z_{\alpha_{Ept}}$ denote the corresponding $100(1-\alpha_{Ept})\%$ percentile of the standard normal distribution. Without loss of generality, assume that a higher value of p_i indicates greater efficacy of treatment i ($i = E, C, H$). The equivalence pre-test is therefore formulated as follows:

$$H_0^{Ept} : |p_C - p_H| \ge \Delta \text{ versus } H_1^{Ept} : |p_C - p_H| < \Delta. \tag{3.1}$$

The equivalence hypothesis (3.1) is equivalent to the following two one-sided hypotheses:

$$H_{01}^{Ept} : p_C - p_H \le -\Delta \text{ versus } H_{11}^{Ept} : p_C - p_H > -\Delta, \tag{3.2}$$

$$H_{02}^{Ept} : p_C - p_H \ge \Delta \text{ versus } H_{12}^{Ept} : p_C - p_H < \Delta. \tag{3.3}$$

Schuirmann [15] has demonstrated that a α_{Ept} level test of hypothesis (3.1) has the same decision rule as two α_{Ept} level tests of hypotheses (3.2) and (3.3). In fact, the equivalence pre-test (3.1) is an intersection-union test. An interesting feature of intersection-union tests is that no multiplicity adjustment is necessary to control the size of a test, but individual hypotheses cannot be tested at levels higher than the nominal significance level [16].

To assess hypotheses (3.2) and (3.3), the Wald-type test statistics are defined as:

$$Z_{Ept,1} = \frac{\bar{y}_C - \bar{y}_H + \Delta}{\sqrt{\widehat{\mathrm{Var}}(\bar{y}_C - \bar{y}_H)}} \text{ and } Z_{Ept,2} = \frac{\bar{y}_C - \bar{y}_H - \Delta}{\sqrt{\widehat{\mathrm{Var}}(\bar{y}_C - \bar{y}_H)}},$$

where

$$\widehat{\mathrm{Var}}(\bar{y}_C - \bar{y}_H) = \frac{\bar{y}_C(1-\bar{y}_C)}{n_C} + \frac{\bar{y}_H(1-\bar{y}_H)}{n_H}.$$

The null hypothesis H_{01}^{Ept} is rejected if $Z_{Ept,1} > z_{\alpha_{Ept}}$, and H_{02}^{Ept} is rejected if $Z_{Ept,2} < -z_{\alpha_{Ept}}$. The null hypothesis for the equivalence test H_0^{Ept} is rejected if and only if both H_{01}^{Ept} and H_{02}^{Ept} are rejected. In particular, the conditions $Z_{Ept,1} > z_{\alpha_{Ept}}$ and $Z_{Ept,2} < -z_{\alpha_{Ept}}$ are equivalent to:

$$Z_{Ept} = \frac{|\bar{y}_C - \bar{y}_H| - \Delta}{\sqrt{\widehat{\mathrm{Var}}(\bar{y}_C - \bar{y}_H)}} < -z_{\alpha_{Ept}}. \tag{3.4}$$

3.4. Superiority test with historical controls

The second hypothesis test is performed only if the preceding equivalence pre-test confirms equivalence between the current control group and the historical control group. This test evaluates the superiority of the treatment group compared to the pooled control group. The treatment group includes n_E subjects, consistent with the sample size of the first-stage randomized trial. The pooled control group consists of the current control group (n_C) and the historical control group (n_H), resulting in a combined sample size of $n_C + n_H$. The corresponding superiority hypothesis test (S_1) can be formulated as follows:

$$H_0^{S_1} : p_E - [\omega p_H + (1-\omega)p_C] = 0 \text{ versus } H_1^{S_1} : p_E - [\omega p_H + (1-\omega)p_C] > 0.$$

Here, ω denotes the weighting coefficient assigned to historical data in estimating the efficacy of the pooled control group, reflecting the relative contribution of historical information to the current analysis. The value of $\omega = \frac{n_H}{n_H+n_C}$ should be calibrated according to the reliability and relevance of historical data. In practice, it is often defined as a proportional weight based on sample sizes—specifically, the ratio of the sample size of the historical control to the total sample size of the pooled control group. This weighting scheme provides a pragmatic balance between representativeness and precision by allowing larger datasets to exert a proportionally greater influence. The corresponding test statistic for assessing superiority is given by:

$$Z_{S_1} = \frac{\bar{y}_E - \left(\frac{n_H}{n_H+n_C} \cdot \bar{y}_H + \frac{n_C}{n_H+n_C} \cdot \bar{y}_C\right)}{\sqrt{\frac{\bar{y}_E(1-\bar{y}_E)}{n_E} + \frac{n_H\bar{y}_H(1-\bar{y}_H)+n_C\bar{y}_C(1-\bar{y}_C)}{(n_H+n_C)^2}}}.$$

3.5. Superiority test without historical controls

Finally, the third test is conducted only when the previously described equivalence pre-test fails to establish equivalence between the current and historical control groups. In this case, a superiority test is performed to compare the supplemented treatment group with the supplemented control group. Let the additional sample sizes for the treatment and control groups be denoted by n'_E and n'_C, respectively. After supplementation, the total sample sizes for the treatment and control groups become $N_E = n_E + n'_E$ and $N_C = n_C + n'_C$, respectively. Both groups include only data obtained from the ongoing clinical trial, without the incorporation of historical information. The corresponding superiority hypothesis test (S_2) is formulated as follows:

$$H_0^{S_2} : p_E - p_C = 0 \text{ versus } H_1^{S_2} : p_E - p_C > 0.$$

The corresponding test statistic is given by:

$$Z_{S_2} = \frac{\bar{y}'_E - \bar{y}'_C}{\sqrt{\frac{\bar{y}'_E(1-\bar{y}'_E)}{n_E+n'_E} + \frac{\bar{y}'_C(1-\bar{y}'_C)}{n_C+n'_C}}}.$$

Let ψ_{Ept}, ψ_{S_1}, and ψ_{S_2} denote the decision functions corresponding to the three hypothesis tests described above. Thus, we have

$$\psi_{Ept} = \begin{cases} 1 & \text{if } Z_{Ept} < -z_{\alpha_{Ept}} \\ 0 & \text{if } Z_{Ept} \geq -z_{\alpha_{Ept}} \end{cases}, \quad \psi_{S_1} = \begin{cases} 1 & \text{if } Z_{S_1} > z_{\alpha_{S_1}} \\ 0 & \text{if } Z_{S_1} \leq z_{\alpha_{S_1}} \end{cases}, \quad \psi_{S_2} = \begin{cases} 1 & \text{if } Z_{S_2} > z_{\alpha_{S_2}} \\ 0 & \text{if } Z_{S_2} \leq z_{\alpha_{S_2}} \end{cases}.$$

For each test, a decision function value of 1 indicates rejection of the null hypothesis, whereas a value of 0 indicates failure to reject the null hypothesis. With respect to the aforementioned hypotheses, there are four possible testing results, which are the following:

Case A: Reject both H_0^{Ept} and $H_0^{S_1}$. Case B: Reject H_0^{Ept} and accept $H_0^{S_1}$.
Case C: Accept both H_0^{Ept} and $H_0^{S_2}$. Case D: Accept H_0^{Ept} and reject $H_0^{S_2}$.

Consequently, the overall decision function for the proposed two-stage Fill-it-up design (TSD) can be defined as follows:

$$\begin{aligned} \psi_{TSD} &= max\{\psi_{Ept} \cdot \psi_{S_1}, (1-\psi_{Ept}) \cdot \psi_{S_2}\} \\ &= \begin{cases} 1 & \text{if } \{Z_{Ept} < -z_{\alpha_{Ept}}, Z_{S_1} > z_{\alpha_{S_1}}\} \text{ or } \{Z_{Ept} \geq -z_{\alpha_{Ept}}, Z_{S_2} > z_{\alpha_{S_2}}\}, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The value of the overall decision function of 1 indicates that both null hypotheses H_0^{Ept} and $H_0^{S_1}$ are rejected (Case A), or the null hypothesis H_0^{Ept} is accepted but the null hypothesis $H_0^{S_2}$ is rejected (Case D). In particular, the choice of significance levels α_{S_1}, α_{S_2}, and α_{Ept} will be explained in the following section.

4. Key issues in two-stage Fill-it-up designs

4.1. Sample size determination

In terms of determining the sample size, a key question is how to establish the fraction γ of patients assigned in the first step. We first revisit the relationship among sample sizes in the two stages, denoted $N_E = n_E + n'_E, N_C = n_C + n'_C$, and $\gamma N_E = n_E, \gamma N_C = n_C$. According to the sample size determination framework proposed by Chow et al. [17], the following equality can be derived for the sample size required in test S_1:

$$\begin{aligned} \frac{1}{\delta^2}\left(z_{\alpha_{S_1}} + z_{\beta_{S_1}}\right)^2 &= \frac{n_E(n_H+n_C)}{(n_H+n_C)p_E(1-p_E) + n_E[\omega p_H(1-p_H) + (1-\omega)p_C(1-p_C)]} \\ &= \frac{\gamma N_E(n_H+\gamma N_C)}{(n_H+\gamma N_C)p_E(1-p_E) + \gamma N_E[\omega p_H(1-p_H) + (1-\omega)p_C(1-p_C)]} \\ &\geq \frac{4\gamma N_E(n_H+\gamma N_C)}{(n_H+\gamma N_C) + \gamma N_E}, \ \ (\text{as } p_i(1-p_i) \leq \frac{1}{4}, i = E, H, C). \end{aligned} \tag{4.1}$$

Here, $\delta > 0$ represents the treatment effect of the experimental group relative to the control group. The same rationale applies to the test S_2:

$$\frac{1}{\delta^2}\Big(z_{\alpha_{S_2}} + z_{\beta_{S_2}}\Big)^2 = \frac{N_E N_C}{N_C p_E(1-p_E) + N_E p_C(1-p_C)} \geq \frac{4N_E N_C}{N_C + N_E} \ \ (\text{as } p_i(1-p_i) \leq \frac{1}{4}, i = E, C). \quad (4.2)$$

Since the true values of p_E, p_C are unknown in real clinical trials, we therefore consider a conservative approach to determine the sample size, i.e.,

$$\frac{1}{\delta^2}\Big(z_{\alpha_{S_2}} + z_{\beta_{S_2}}\Big)^2 = \frac{4N_E N_C}{N_C + N_E}. \quad (4.3)$$

By combining (4.1) and (4.3), we have

$$\frac{N_E + N_C}{N_C N_E} \cdot \frac{\gamma N_E(n_H + \gamma N_C)}{(n_H + \gamma N_C) + \gamma N_E} \leq \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2.$$

Thus, a quadratic inequality in γ with the solutions:

$$\begin{aligned}
&\frac{1}{2}\Big(-\frac{n_H}{N_C} + \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2 - \sqrt{\Big(\frac{n_H}{N_C} - \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{1-\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2\Big)^2 + \frac{4n_H}{N_E + N_C} \cdot \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2}\Big) \leq \gamma \leq \\
&\frac{1}{2}\Big(-\frac{n_H}{N_C} + \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2 + \sqrt{\Big(\frac{n_H}{N_C} - \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2\Big)^2 + \frac{4n_H}{N_E + N_C} \cdot \Big(\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}}\Big)^2}\Big). \quad (4.4)
\end{aligned}$$

In practical applications, it is reasonable to assume that the Type I and Type II error rates for the two tests are the same, i.e., $\alpha_{S_1} = \alpha_{S_2}, \beta_{S_1} = \beta_{S_2}$. Hence,

$$\frac{z_{\alpha_{S_1}} + z_{\beta_{S_1}}}{z_{\alpha_{S_2}} + z_{\beta_{S_2}}} = 1.$$

In this study, we focus mainly on considering the simple balanced design, that is, $N_E = N_C = N$. Consequently, (4.4) reduces to

$$\frac{1}{2N}\Big(N - n_H - \sqrt{N^2 + n_H^2}\Big) \leq \gamma \leq \frac{1}{2N}\Big(N - n_H + \sqrt{N^2 + n_H^2}\Big).$$

It should be noted that $\frac{1}{2N}(N - n_H - \sqrt{N^2 + n_H^2}) < 0$. However, γ is the positive number and decreases with increasing n_H with a maximum value of $\gamma = 1$ in the case where n_H is zero. Therefore, the choices of γ are suggested in the following formula:

$$\gamma = \frac{1}{2N}\Big(N - n_H + \sqrt{N^2 + n_H^2}\Big).$$

Here, this formula is exactly the same as in Wied et al. [12]. In addition, $\gamma \geq 1/2$ since $N^2 + n_H^2 \geq n_H^2$. This means that at least half of the total sample size is necessary in the randomized trial of the first step.

4.2. Relationship between equivalence testing and superiority testing

The equivalence margin, denoted as Δ, serves as the pivotal parameter throughout the two-stage testing procedure. Its specification directly determines the outcome of the equivalence pre-test and, consequently, affects both the data integration process and the reliability of the subsequent statistical inference in the superiority assessment. In the equivalence pre-test, rejection of the null hypothesis H_0 is a necessary condition for merging the historical control group with the concurrent control group. Once equivalence is established, the superiority test S_1 is conducted based on the combined control group. The weighted efficacy of treatment compared to the pooled control group can be expressed as:

$$t = p_E - (\omega \cdot p_H + (1-\omega) \cdot p_C) = p_E - p_C + \omega(p_C - p_H) = \delta + \omega(p_C - p_H),$$

where $\delta = p_E - p_C$ denotes the treatment effect, with $\delta > 0$ assumed for simplicity.

Based on the true efficacy difference between the historical control group and the current control group, two distinct scenarios can be delineated, each necessitating appropriate control of statistical inference risk.

Scenario 1: The efficacy of the historical control group exceeds that of the current control group ($p_H > p_C$).

When $p_H > p_C$, the weighted efficacy $\omega \cdot p_H + (1-\omega) \cdot p_C$ of the pooled control group becomes greater than the true efficacy p_C of the current control. This inflation of the pooled efficacy estimate, resulting from the inclusion of superior historical data, leads to an underestimation of the true treatment effect. Consequently, the apparent difference between the treatment and the pooled control groups is attenuated, potentially obscuring a genuine treatment benefit. Such bias increases the likelihood of false-negative conclusions in the subsequent superiority test—that is, misclassifying an effective treatment as non-effective.

In this context, specifying a smaller equivalence margin (Δ) imposes a more conservative equivalence criterion, thereby reducing the probability of erroneously merging non-equivalent data (where $p_H > p_C$) as equivalent. This adjustment helps mitigate the risk of false-negative outcomes and improves the robustness of the overall inferential procedure.

Scenario 2: The efficacy of the historical control group is inferior to that of the current control group ($p_H < p_C$).

When $p_H < p_C$, the weighted efficacy $\omega \cdot p_H + (1-\omega) \cdot p_C$ of the pooled control group becomes lower than the true efficacy p_C of the current control group, as the inclusion of less effective historical data depresses the overall efficacy estimate. This pooled efficacy underestimation artificially enlarges the apparent treatment effect, potentially leading to false-positive conclusions in the subsequent superiority test—that is, classifying a non-superior treatment as superior. Such erroneous inferences can misguide clinical decision-making by promoting ineffective therapies and, in many contexts, pose greater ethical and practical risks than false negatives. To mitigate this risk, a more stringent equivalence criterion should be imposed to prevent the inclusion of non-equivalent historical data (i.e., cases where $p_H < p_C$) in the pooled analysis.

In summary, regardless of whether the historical control group demonstrates higher or lower efficacy than the current control group, adopting a narrower equivalence margin serves as a safeguard against inferential bias in the superiority test. The equivalence margin, however, cannot be too small as it might result in a lack of power for Ept. This balance between sample size efficiency and inferential reliability represents a fundamental consideration in the design and implementation of two-stage hybrid trials.

4.3. Determining the range for the equivalence margin

The equivalence margin (Δ) is a key design parameter that governs the decision to pool historical and current control data. In practical applications, Δ is typically determined based on clinical judgment regarding the maximum acceptable difference between the historical and concurrent control response rates that can be regarded as clinically negligible. Historical evidence—such as meta-analytic summaries of previous control arms—can also inform the expected range of natural variability, thereby guiding the choice of a plausible and defensible margin. Statistically, the value of Δ directly influences both Type I and Type II error rates in the two-stage Fill-it-up design. A larger Δ increases the likelihood of declaring equivalence, which enhances sample size efficiency but may inflate the Type I error rate if heterogeneous historical data are inadvertently incorporated. In contrast, a smaller Δ enforces a more conservative equivalence criterion, reducing the risk of false-positive equivalence declarations but decreasing the probability of pooling and therefore increasing the required sample size and the risk of Type II errors.

To ensure statistical robustness of the two-stage testing procedure, the equivalence margin (Δ) must simultaneously satisfy both the upper-bound constraint and the lower-bound constraint. The derivation and underlying rationale are described as follows.

The upper bound of the equivalence margin (Δ) is determined by the expected true treatment effect between the experimental and control groups. Specifically, Δ must satisfy $\Delta < \delta$, where δ denotes the theoretical upper limit of the treatment effect. This condition ensures that the equivalence threshold does not exceed the plausible treatment difference, thereby preventing overly permissive equivalence declarations.

A lower bound for the margin of equivalence can also be established by examining the test statistic Z_{Ept}:

$$\Delta \geq \sqrt{\frac{\bar{y}_H(1-\bar{y}_H)}{n_H} + \frac{\bar{y}_C(1-\bar{y}_C)}{n_C}} \cdot z_{\alpha_{Ept}}.$$

The failure of this inequality implies that the null hypothesis of equivalence can never be rejected, regardless of the sample size, thus rendering the equivalence test infeasible. In particular, the following condition must hold:

$$\sqrt{\frac{\bar{y}_H(1-\bar{y}_H)}{n_H} + \frac{\bar{y}_C(1-\bar{y}_C)}{n_C}} \leq \frac{1}{2}\sqrt{\frac{1}{n_H} + \frac{1}{n_C}}.$$

Combining these two constraints yields the final feasible range of the equivalence margin, i.e.,

$$\frac{1}{2}\sqrt{\frac{1}{n_H} + \frac{1}{n_C}} \cdot z_{\alpha_{Ept}} \leq \Delta < \delta.$$

This range ensures both the practical interpretability and statistical validity of the equivalence testing, maintaining an appropriate balance between sensitivity to true similarities and protection against false equivalence.

4.4. Family-wise error rate and power function

In statistical inference for two-stage Fill-it-up designs, strict control of the family-wise error rate (FWER), is essential to ensure that it does not exceed the pre-specified significance level (α), typically

set at 0.05 as defined in the clinical trial protocol. Formally, the FWER must satisfy the following requirements:

$$\text{FWER} = E(\psi_{TSD}) = \underbrace{E\left(\psi_{Ept} \cdot \psi_{S_1}\right)}_{:=\alpha_{Ept,S_1}} + \underbrace{E\left(\left(1-\psi_{Ept}\right) \cdot \psi_{S_2}\right)}_{:=\alpha_{Ept,S_2}} \leq \alpha. \tag{4.5}$$

Within this two-stage framework, hypothesis testing proceeds conditionally. If the null hypothesis of equivalence is rejected ($\psi_{Ept} = 1$), a superiority test (S_1) is performed to compare the treatment group and the pooled control group. In this case, the expected rejection probability corresponds to the joint event $\psi_{Ept} \cdot \psi_{S_1}$. In contrast, if equivalence cannot be established ($\psi_{Ept} = 0$), the procedure advances to a superiority test (S_2) comparing the augmented treatment and control groups, with the expected rejection probability represented by $(1 - \psi_{Ept}) \cdot \psi_{S_2}$. The formulas for the calculations of α_{Ept,S_1} and α_{Ept,S_2} can be found in Appendix A.

Next, we evaluate the statistical power of the two-stage Fill-it-up design. To quantify the power, the true difference δ in efficacy between the experimental treatment and the concurrent control group is first specified. Under the conditional triggering mechanism of the two-stage Fill-it-up design—where the outcome of the equivalence pre-test determines which subsequent superiority test is conducted—the overall power of the design can be expressed as follows:

$$1 - \beta_{TSD} = 1 - (\beta_{Ept,S_1} + \beta_{Ept,S_2}).$$

Here, β_{Ept,S_1} denotes the Type II error associated with the superiority test S_1, which compares the treatment group against the pooled control group when equivalence is confirmed in the pre-test. Conversely, β_{Ept,S_1} represents the Type II error that occurs when the equivalence pre-test fails to establish equivalence and the subsequent superiority test S_2 (comparing the supplemented treatment and supplemented control groups) fails to detect the true treatment difference. The formulas for the calculations of β_{Ept,S_1} and β_{Ept,S_2} can be found in Appendix B.

Assume $\delta = 0.1, n_H = 500$, and $\Delta = 0.085$. Here, we consider two representative scenarios in which the true efficacy of the current control group was fixed at $p_C = 0.9$ and $p_C = 0.5$, respectively. The FWER results differ substantially between the scenarios $p_C = 0.9$ and $p_C = 0.5$ due to changes in the underlying variance of the control-group outcome. In terms of real-world interpretation, the case $p_C = 0.5$ generally corresponds to moderate control success rates typical in symptomatic or partially effective treatments, whereas $p_C = 0.9$ represents near-perfect control performance, which is more common in prophylactic treatments or highly effective standard-of-care settings.

The FWER of the two-stage Fill-it-up design was first evaluated under varying pre-specified equivalence significance levels (α_{Ept}). FWER depends on the difference in efficacy between the historical and current control groups $p_H - p_C$, denoted as difference (Ept) and on the weighted efficacy of the pooled control group (denoted as difference(S_1)). The corresponding results are illustrated in Figure 2 (for $p_C = 0.9$) and Figure 3 (for $p_C = 0.5$). As shown in Figures 2 and 3, the maximum FWER consistently remains below 5%, aligning well with the pre-specified overall significance level of $\alpha = 0.05$. This finding confirms that the proposed two-stage Fill-it-up design provides robust control of the FWER, thereby meeting the fundamental inferential requirement of clinical trials, namely the avoidance of false-positive efficacy conclusions.

Furthermore, with all other parameters remaining constant, a smaller α_{Ept} is associated with a lower FWER. The rationale is that a more stringent equivalence criterion (smaller α_{Ept}) restricts the likelihood

of the equivalence pre-test incorrectly classifying non-equivalent historical and current control data as equivalent, thereby preventing their inappropriate inclusion in the pooled analysis. This reduces bias-induced inflation of the Type I error rate of the superiority test and, consequently, lowers the overall FWER.

Set $n_H = 500, \alpha_{Ept} = \alpha_{S_1} = \alpha_{S_2} = 0.05$ and $\beta_{S_1} = \beta_{S_2} = 0.20$. Figure 4 illustrates the statistical power of the two-stage Fill-it-up design under various parameter configurations. As shown in Figure 4, when the true difference in treatment effect (δ) and the equivalence margin (Δ) are appropriately specified, the test power achieved remains consistently above 80%, meeting the standard "80% power" commonly adopted in the design of clinical trial.

Furthermore, with other parameters held constant, when the true treatment difference (δ) between the experimental and current control groups is relatively small, both a smaller equivalence margin (Δ) and a more stringent equivalence pre-test significance level (α_{Ept}) are required to maintain adequate statistical power. This finding underscores the inherent trade-off between equivalence stringency and power efficiency in two-stage design optimization.

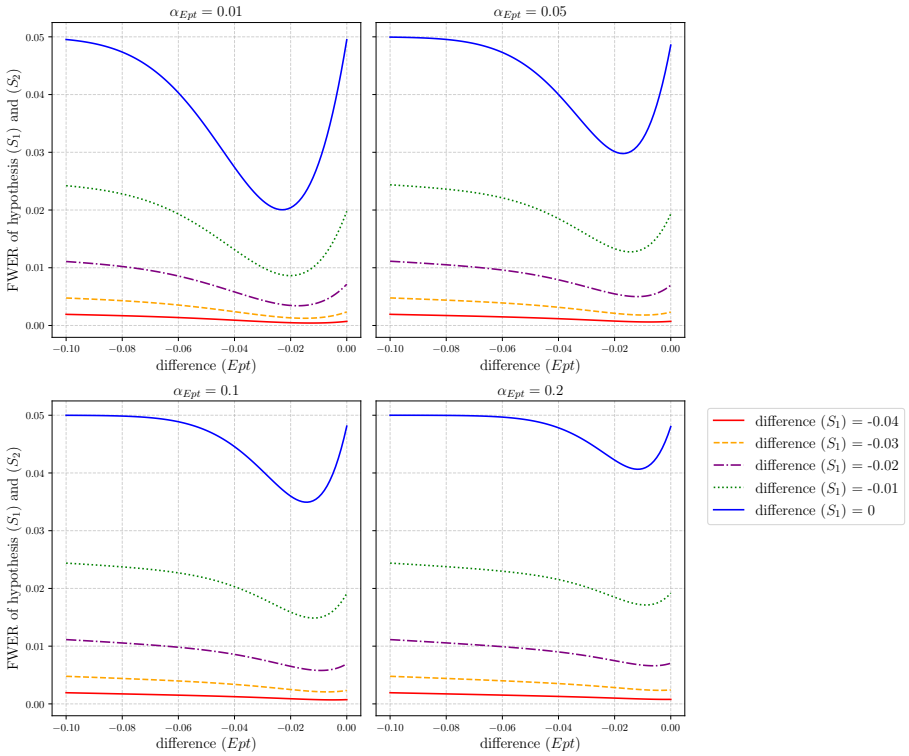

Figure 2. Family-wise error rate testing simultaneously superiority test (S_1) and (S_2) for different scenarios of the Fill-it-up design depending on the choice of the significance level of the equivalence pre-test. An effect size $\delta = 0.1$ with $p_C = 0.9, n_H = 500$, and an equivalence margin of $\Delta = 0.085$ is examined.

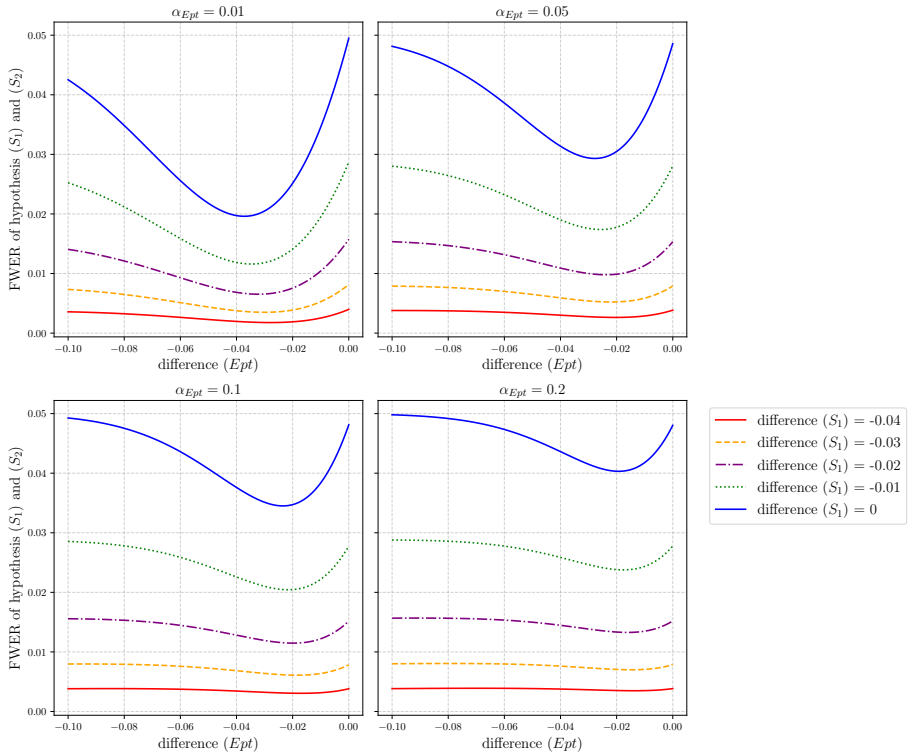

Figure 3. Family-wise error rate testing simultaneously superiority test (S_1) and (S_2) for different scenarios of the Fill-it-up design depending on the choice of the significance level of the equivalence pre-test. An effect size $\delta = 0.1$ with $p_C = 0.5, n_H = 500$, and an equivalence margin of $\Delta = 0.085$ is examined.

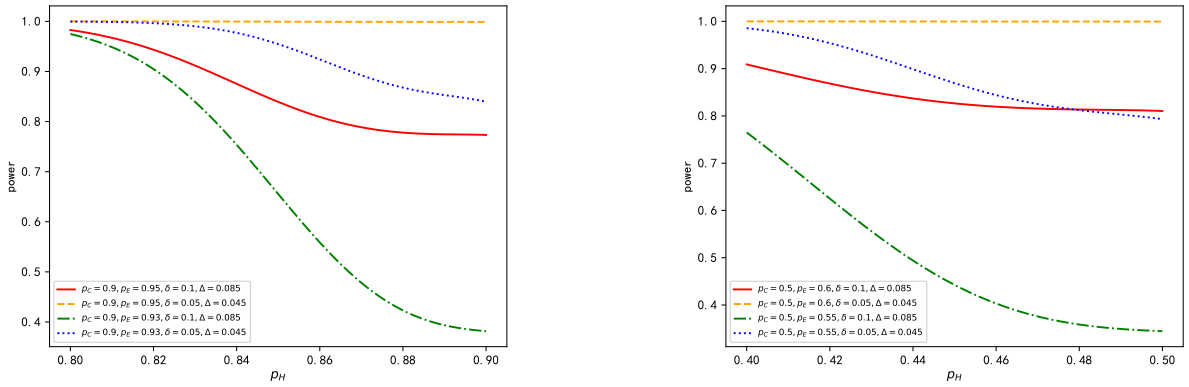

Figure 4. Power functions for different two-stage Fill-it-up design scenarios, depending on the choice difference of p_H, δ, and equivalence margin Δ. Left: Settings are $p_C = 0.9, n_H = 500, \alpha_{Ept} = \alpha_{S_1} = \alpha_{S_2} = 0.05, \beta_{S_1} = \beta_{S_2} = 0.2$. Right: Settings are $p_C = 0.5, n_H = 500, \alpha_{Ept} = \alpha_{S_1} = \alpha_{S_2} = 0.05, \beta_{S_1} = \beta_{S_2} = 0.2$.

5. Numerical simulation

The preceding sections addressed the overall Type I error rate and statistical power of the proposed two-stage Fill-it-up design. This section focuses on the determination of the required sample size and the evaluation of its performance through numerical simulations. Here, N_{Total} denotes the total sample size required when the pre-specified equivalence test fails to reject the null hypothesis—comprising both the initial Stage I sample and the supplemental Stage II sample. In contrast, γN_{Total} represents the total sample size required when the equivalence test rejects the null hypothesis, thus requiring only the Stage I sample without additional recruitment. To quantify sample size requirements and assess potential efficiency gains in different scenarios, Monte Carlo simulations were conducted. The average sample number, ASN, is given by

$$\text{ASN} = \gamma N_{Total} + (1 - \alpha_{Ept})(N_{Total} - \gamma N_{Total}).$$

The simulated ASN ($\widehat{\text{ASN}}$) was estimated based on 100, 000 replications.

The core parameters used for the initial sample size calculation of the two-stage Fill-it-up design are first specified. Table 2 summarizes these parameters, including the sample size of the historical control group (n_H), the difference in the true treatment effect between the experimental and current control groups (δ), the pre-specified significance level for the pre-test of equivalence (α_{Ept}), the two-sided Type I error rate for the superiority test, and the equivalence margin (Δ).

Table 2. Choice of parameters for two-stage Fill-it-up design in simulation study.

Parameter	Parameter description	Value
n_H	Sample size of historical control group H	500
δ	Effect size	0.1
$\beta_{S_1} = \beta_{S_2}$	Type II error rate of (S_1) and (S_2)	0.15
$\alpha_{S_1} = \alpha_{S_2}$	Significance Level of (S_1) and (S_2)	0.05
α_{Ept}	Significance Level of (Ept)	{0.025,0.05,0.1,0.2}
Δ	Equivalence Margin	0.085

Table 3 presents the corresponding sample size determinations and simulation results. As shown in Table 3, confirming the equivalence between the current and historical control groups leads to a substantial reduction in the overall sample size. For example, when the level of significance equivalence pre-test α_{Ept} is set to 0.05, the total sample size decreases from $N_{Total} = 706$ to $\gamma N_{Total} = 466$, representing a reduction of 240 subjects. Compared to the conventional single-step design that uses only the (S_2) test only with power $1 - \beta_{S_2} = 0.8$ and significance level $\alpha_{S_2} = 0.05$, the required sample size would be 606 based on a straightforward calculation, still producing a reduction of approximately 140 subjects. However, if the equivalence pre-test fails to reject the null hypothesis, an additional 100 subjects would be required compared with conducting the test (S_2) alone.

The results presented in Table 3 further indicate that, with all other parameters kept constant, increasing the pre-specified equivalence significance level (α_{Ept}) can markedly reduce the $\widehat{\text{ASN}}$ required by the design. For example, when $p_H = 0.5$ and $p_C = 0.5$, the $\widehat{\text{ASN}}$ decreases from 672 at $\alpha_{Ept} = 0.025$ to 513 at $\alpha_{Ept} = 0.20$, corresponding to a 23.7% reduction in the sample size.

Nevertheless, caution is warranted in practice: Setting an overly large α_{Ept} should be avoided, as a too-permissive equivalence criterion can lead to biased estimation of the treatment effect within the mixed control population, thus inflating the risk of Type I or Type II errors in the subsequent superiority test.

Table 3. Sample size determination and simulation results.

α_{Ept}	p_H	p_C	N_{Total}	γN_{Total}	$\widehat{\text{ASN}}$
0.025	0.50	0.50	706	466	672
0.05	0.50	0.50	706	466	616
0.10	0.50	0.50	706	466	560
0.20	0.50	0.50	706	466	513
0.025	0.45	0.50	706	466	690
0.05	0.45	0.50	706	466	662
0.10	0.45	0.50	706	466	628
0.20	0.45	0.50	706	466	584
0.025	0.80	0.80	360	212	356
0.05	0.80	0.80	360	212	320
0.10	0.80	0.80	360	212	283
0.20	0.80	0.80	360	212	250
0.025	0.75	0.80	360	212	360
0.05	0.75	0.80	360	212	347
0.10	0.75	0.80	360	212	322
0.20	0.75	0.80	360	212	292

Furthermore, the true difference in efficacy between the historical and current control groups ($p_C - p_H$) exerts a substantial influence on ASN. Across all levels of α_{Ept}, the $\widehat{\text{ASN}}$ decreases as the consistency between the two control groups improves (e.g., when p_H increases from 0.45 to 0.50, in accordance with $p_C = 0.50$). The closer the observed consistency approaches the equivalence threshold, the higher the probability of rejecting the null hypothesis in the pre-test, enabling the integration of historical control data and thus avoiding second-stage sample supplementation. Consequently, a smaller efficacy discrepancy leads to greater sample size efficiency. In contrast, when $p_C - p_H$ is large, establishing equivalence becomes more difficult, increasing the likelihood of entering the second stage of the design and resulting in a larger $\widehat{\text{ASN}}$. This finding reinforces that the degree of concordance between historical and current control data is the key determinant of sample size savings in the two-stage Fill-it-up design—the smaller the discrepancy, the more pronounced the efficiency gain.

6. Comparison with Bayesian approach

As noted earlier, numerous Bayesian approaches for incorporating historical control data have been extensively studied. To compare our two-stage Fill-it-up design (TSD) with established methods, we consider the meta-analytic predictive (MAP) prior approach [18]. In particular, the robust MAP prior has been shown to provide desirable operating characteristics, making it an appropriate comparator

for evaluating family-wise error rate and power. For consistency with previous assessments of the Fill-it-up design, we implement the MAP approach using the RBesT package in R [19].

In this section, we conduct simulation studies to compare the performance of the TSD method with that of the robust MAP prior in terms of FWER and power. Here, we fixed the sample sizes at $n_E = n_C = 100, n_H = 400$, and $\alpha_{Ept} = \alpha_{S_1} = \alpha_{S_2} = 0.05$. The simulation mainly examines how different treatment effects ($p_E - p_C$) influence operating characteristics in the scenario $p_C = p_H$. Tables 4 and 5 report the results for power and FWER, respectively. Here, the simulated results are based on 10,000 replications. The simulation results presented in Table 4 indicate that our approach generally outperforms the MAP method. For example, in Case 1, the power achieved by the TSD approach is approximately 33% higher than that of the MAP approach. With respect to FWER control, the two approaches show comparable performance, with no substantial differences observed.

Table 4. Simulation power for MAP and TSD approaches.

			Power	
p_E	p_C	p_H	MAP	TSD
0.60	0.50	0.50	0.445	0.592
0.65	0.50	0.50	0.774	0.879
0.70	0.50	0.50	0.957	0.985
0.90	0.80	0.80	0.688	0.862
0.95	0.80	0.80	0.979	0.998
0.975	0.80	0.80	0.998	1.000

Table 5. Simulation FWER for MAP and TSD approaches.

			FWER	
p_E	p_C	p_H	MAP	TSD
0.20	0.20	0.20	0.053	0.040
0.30	0.30	0.30	0.041	0.047
0.40	0.40	0.40	0.049	0.048
0.50	0.50	0.50	0.043	0.048
0.60	0.60	0.60	0.038	0.055
0.70	0.70	0.70	0.043	0.053

7. Application

This section demonstrates the advantage of the proposed two-stage Fill-it-up design in practice. The two clinical trials discussed in Section 2 are revisited. The data obtained in the two studies (see Table 1) are used to provide an estimate.

Here, we illustrate the practical implementation of the proposed two-stage Fill-it-up design using the two real-world examples. In the first stage, an equivalence pilot analysis was conducted with a pre-specified significance level of $\alpha_{Ept} = 0.20$ and an equivalence margin of $\Delta = 0.085$. The null hypothesis of non-equivalence was rejected (p value = 0.159), indicating that the historical and current

control groups could be considered comparable. Consequently, the two control datasets were pooled for the subsequent superiority assessment. In the second stage, a superiority analysis based on the combined control data yielded a p value of 0.105. As reported in Section 2, using only the current control group produced a p value of 0.427. These findings suggest that the integration of historical control information can substantially improve the statistical efficiency of the superiority evaluation, thus enhancing the ability to detect true differences between treatment and control and to demonstrate superiority.

8. Conclusions

In this paper, we propose a two-stage "Fill-it-up" design for clinical trials with binary endpoints to enable the rigorous integration of historical control data under controlled statistical risk. This adaptive framework enables sample size optimization when historical data are suitable for integration while minimizing the risk of bias arising from heterogeneous data pooling. The simulation results further confirm that, under equivalence conditions, data integration substantially reduces the average sample size. Nevertheless, the maximum required sample size under the two-stage design (corresponding to scenarios where equivalence is not achieved) may exceed that of a conventional single-stage design without historical information. In addition, greater heterogeneity between the historical and current controls increases overall sample requirements. Thus, in practical implementation, ensuring strong comparability between datasets remains essential, as excessive heterogeneity diminishes the efficiency gains of the proposed design.

Although the present study focuses on binary endpoints, the conceptual structure of the two-stage Fill-it-up design can be extended to more complex outcome types such as mixed endpoints (e.g., binary–continuous composites) or longitudinal repeated-measures data [20]. Recent methodological developments also provide mathematical tools that complement the ideas underlying the proposed two-stage design [21].

In many practical settings, multiple historical datasets are available, differing in quality, relevance, and sample size. Integrating such heterogeneous data introduces several challenges. Differences in study design, endpoints, or patient characteristics may lead to non-exchangeability, rendering naive pooling inappropriate. Small or lower-quality datasets can also exert undue influence if not properly weighted. Extending the equivalence pre-test to multiple datasets requires sequential or global multivariate testing, which increases inferential complexity. Potential solutions include data-driven weighting schemes, sequential or simultaneous equivalence tests to identify suitable datasets for borrowing, and robust estimators to limit the impact of poorly aligned sources. Developing a comprehensive multi-source extension of the two-stage Fill-it-up design is an important direction for future work and may substantially expand its applicability in real-world clinical research.

The proposed two-stage Fill-it-up design relies on normal-approximation–based Wald statistics for both the equivalence pre-test and the subsequent superiority tests. Although these approximations perform well in moderate to large samples, their accuracy may degrade in small-population settings or when the underlying event probabilities are extreme. In such cases, the equivalence test may become conservative due to inflated standard errors, thereby reducing the probability of pooling historical controls and consequently lowering the power of the final superiority test. Conversely, when event probabilities are near 0 or 1, the normal approximation may understate variability, potentially making

the equivalence test slightly liberal. These effects highlight the importance of evaluating the operating characteristics of the TSD through simulation when the sample size is limited.

Possible remedies to improve finite-sample performance include using Fisher's exact test or exact unconditional tests for the equivalence stage, or applying bootstrap-based calibrations to test statistics. Although exploring these alternatives is beyond the scope of the present study, they represent promising extensions for future work and may further enhance the applicability of TSD in small-population clinical trials.

The heterogeneity between historical and current control data can substantially affect Type I error and the power of the two-stage Fill-it-up design. Its impact can be assessed through several approaches: The equivalence pre-test provides a formal comparability check; simulation-based sensitivity analyses that vary response rates, dispersion, or sample sizes offer empirical evaluation of robustness. To mitigate heterogeneity-induced Type I error inflation or power loss, several strategies may be used. These include conservative specification of the equivalence margin Δ, adjustment of the equivalence-test significance level α_{Ept}. Integrating screening, sensitivity analysis, and robust estimation allows the design to control Type I error while retaining efficiency when historical data are adequately comparable.

Author contributions

Junjiang Zhong: Conceptualization, methodology, writing—original draft preparation, writing—review and editing and funding acquisition; Haoyun Guo: Methodology, software, writing—original draft preparation and writing—review and editing; Nan Sun: Conceptualization and writing—review and editing; Junjie Li: Writing—original draft preparation, writing—review and editing and funding acquisition. All authors have read and agreed to the final version of the manuscript.

Use of Generative-AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We are grateful to the three referees for their valuable suggestions, which have led to the significant improvement of this paper. This research was funded by the Natural Science Foundation of Xiamen, China (3502Z20227220), the Ministry of Education of China project on Humanities and Social Sciences (21YJC910011), and the Natural Science Foundation of Fujian Province of China (2023J011433, 2023J011434).

Conflict of interest

The authors declare no conflicts of interest.

References

1. *Food and Drug Administration, Considerations for the design and conduct of externally controlled trials for drug and biological products*, Food and Drug Administration, 2023. Available from: https://www.fda.gov/media/164960/download
2. M. H. Chen, J. G. Ibrahim, Power prior distributions for regression models, *Statist. Sci.*, **15** (2000), 46–60. https://doi.org/10.1214/ss/1009212673
3. Y. Duan, K. Ye, E. P. Smith, Evaluating water quality using power priors to incorporate historical information, *Environmetrics*, **17** (2006), 95–106. https://doi.org/10.1002/env.752
4. M. H. Chen, J. G. Ibrahim, P. Lam, A. Yu, Y. Zhang, Bayesian design of non-inferiority trials for medical devices using historical data, *Biometrics*, **67** (2011), 1163–1170. https://doi.org/10.1111/j.1541-0420.2011.01561.x
5. B. P. Hobbs, B. P. Carlin, S. J. Mandrekar, D. J. Sargent, Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials, *Biometrics*, **67** (2011), 1047–1056. https://doi.org/10.1111/j.1541-0420.2011.01564.x
6. B. P. Hobbs, D. J. Sargent, B. P. Carlin, Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models, *Bayesian Anal.*, **7** (2012), 639–674. https://doi.org/10.1214/12-BA722
7. M. Bennett, S. White, N. Best, A. Mander, A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: A comparison to standard methods, *Pharm. Stat.*, **20** (2021), 462–484. https://doi.org/10.1002/pst.2088
8. S. Gsteiger, B. Neuenschwander, F. Mercier, H. Schmidli, Using historical control information for the design and analysis of clinical trials with overdispersed count data, *Statist. Med.*, **32** (2013), 3609–3622. https://doi.org/10.1002/sim.5851
9. H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, B. Neuenschwander, Robust meta-analytic-predictive priors in clinical trials with historical control information, *Biometrics*, **70** (2014), 1023–1032. https://doi.org/10.1111/biom.12242
10. K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, et al., Use of historical control data for assessing treatment effects in clinical trials, *Pharm. Stat.*, **13** (2014), 41–54. https://doi.org/10.1002/pst.1589
11. W. Li, F. Liu, D. Snavely, Revisit of test-then-pool methods and some practical considerations, *Pharm. Stat.*, **19** (2020), 498–517. https://doi.org/10.1002/pst.2009
12. S. Wied, M. Posch, R. D. Hilgers, Evaluation of the fill-it-up-design to use historical control data in randomized clinical trials with two arm parallel group design, *BMC Med. Res. Methodol.*, **24** (2024), 197. https://doi.org/10.1186/s12874-024-02306-2
13. H. I. Brunner, C. Abud-Mendoza, D. O. Viola, I. C. Penades, D. Levy, J. Anton, et al., Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: Results from a randomized, placebo-controlled trial, *Ann. Rheum. Dis.*, **79** (2020), 1340–1348. https://doi.org/10.1136/annrheumdis-2020-217101

14. S. V. Navarra, R. M. Guzmán, A. E. Gallacher, S. Hall, R. A. Levy, R. E. Jimenez, et al., Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: A randomized, placebo-controlled, phase 3 trial, *Lancet*, **377** (2011), 721–731. https://doi.org/10.1016/S0140-6736(10)61354-2

15. D. J. Schuirmann, A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability, *J. Pharmacokinet. Biopharm.*, **15** (1987), 657–680. https://doi.org/10.1007/BF01068419

16. R. L. Berger, Multiparameter hypothesis testing and acceptance sampling, *Technometrics*, **24** (1982), 295–300. https://doi.org/10.2307/1267823

17. S. C. Chow, J. Shao, H. Wang, Y. Lokhnygina, *Sample size calculations in clinical research*, New York: Chapman and hall/CRC, 2017. https://doi.org/10.1201/9781315183084

18. B. Neuenschwander, G. Capkun-Niggli, M. Branson, D. Spiegelhalter, Summarizing historical information on controls in clinical trials, *Clin. Trials*, **7** (2010), 5–18. https://doi.org/10.1177/1740774509356002

19. S. Weber, Y. Li, J. W. Seaman III, T. Kakizume, H. Schmidli, Applying meta-analytic-predictive priors with the R Bayesian evidence synthesis tools, *J. Stat. Softw.*, **100** (2021), 1–32. https://doi.org/10.18637/jss.v100.i19

20. G. L. Hickey, P. Philipson, A. Jorgensen, R. Kolamunnage-Dona, Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues, *BMC Med. Res. Methodol.*, **16** (2016), 117. https://doi.org/10.1186/s12874-016-0212-5

21. A. H. Ekong, M. O. Olayiwola, A. G. Dawodu, A. I. Osinuga, Latent gaussian approach to joint modelling of longitudinal and mixture cure outcomes, *Comput. J. Math. Stat. Sci.*, **4** (2025), 72–95. https://doi.org/10.21608/cjmss.2024.303748.1061

Appendix A: Calculation of FWER

The first component of the FWER, α_{Ept,S_1}, represents the probability of rejecting both the equivalence test (Ept) and the superiority test (S_1). The second component, α_{Ept,S_2}, represents the probability of accepting the equivalence test (Ept) and rejecting the superiority test (S_2). Thus, α_{Ept,S_1} and α_{Ept,S_2}, can be expressed as follows:

$$\alpha_{Ept,S_1} = \int_{z_{\alpha_{S_1}}}^{\infty} \int_{-\infty}^{-z_{\alpha_{Ept}}} f_1(x,y)dxdy + \int_{z_{\alpha_{S_1}}}^{\infty} \int_{z_{\alpha_{Ept}}}^{\infty} f_1(x,y)dxdy, \tag{8.1}$$

$$\alpha_{Ept,S_2} = \int_{z_{\alpha_{S_2}}}^{\infty} \int_{-z_{\alpha_{Ept}}}^{z_{\alpha_{Ept}}} f_2(x,y)dxdy, \tag{8.2}$$

where $f_1(x,y)$ and $f_2(x,y)$ denote the joint probability density functions of the bivariate normal distributions with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, respectively. Here,

$$\boldsymbol{\mu_1} = \left(\frac{|p_C - p_H| - \Delta}{\sqrt{\frac{p_C(1-p_C)}{n_C} + \frac{p_H(1-p_H)}{n_H}}}, \frac{p_E - (w \cdot p_H + (1-w) \cdot p_C)}{\sqrt{\frac{p_E(1-p_E)}{n_E} + \frac{w^2 p_H(1-p_H)}{n_H} + \frac{(1-w)^2 p_C(1-p_C)}{n_C}}} \right)^T,$$

$$\boldsymbol{\mu_2} = \left(\frac{|p_C - p_H| - \Delta}{\sqrt{\frac{p_C(1-p_C)}{n_C} + \frac{p_H(1-p_H)}{n_H}}}, \frac{p_E - p_C}{\sqrt{\frac{p_E(1-p_E)}{n_E+n'_E} + \frac{p_C(1-p_C)}{n_C+n'_C}}} \right)^T,$$

and

$$\boldsymbol{\Sigma_1} = \begin{bmatrix} 1 & Cov(Z_{S_1}, Z_{Ept}) \\ Cov(Z_{S_1}, Z_{Ept}) & 1 \end{bmatrix}, \boldsymbol{\Sigma_2} = \begin{bmatrix} 1 & Cov(Z_{S_2}, Z_{Ept}) \\ Cov(Z_{S_2}, Z_{Ept}) & 1 \end{bmatrix},$$

where

$$Cov(Z_{S_1}, Z_{Ept}) = C_1 \cdot C_2 \cdot \left(- w \cdot \frac{p_H(1-p_H)}{n_H} + (1-w)\frac{p_C(1-p_C)}{n_C} \right),$$
$$Cov(Z_{S_2}, Z_{Ept}) = C_2 \cdot C_3 \cdot \frac{p_C(1-p_C)}{n_C + n'_C},$$

and

$$C_1 = \frac{1}{\sqrt{\frac{p_E(1-p_E)}{n_E} + \frac{w^2 p_H(1-p_H)}{n_H} + \frac{(1-w)^2 p_C(1-p_C)}{n_C}}}, C_2 = \frac{1}{\sqrt{\frac{p_C(1-p_C)}{n_C} + \frac{p_H(1-p_H)}{n_H}}}, C_3 = \frac{1}{\sqrt{\frac{p_E(1-p_E)}{n_E+n'_E} + \frac{p_C(1-p_C)}{n_C+n'_C}}}.$$

Appendix B: Calculation of power

The corresponding expressions for β_{Ept,S_1} and β_{Ept,S_2} are given as follows:

$$\beta_{Ept,S_1} = \int_{-\infty}^{z_{\alpha_{S_1}}} \int_{-\infty}^{-z_{\alpha_{Ept}}} f_1(x,y)dxdy + \int_{-\infty}^{z_{\alpha_{S_1}}} \int_{z_{\alpha_{Ept}}}^{\infty} f_1(x,y)dxdy, \tag{8.3}$$

$$\beta_{Ept,S_2} = \int_{-\infty}^{z_{\alpha_{S_2}}} \int_{-z_{\alpha_{Ept}}}^{z_{\alpha_{Ept}}} f_2(x,y)dxdy. \tag{8.4}$$

All computations presented in Eqs (8.1)–(8.4) were conducted using statistical software R. The `mvtnorm` package in R was particularly utilized to evaluate multivariate normal probabilities. Next, we provide a step-by-step algorithm describing how the `mvtnorm` package in R was used to compute β_{Ept,S_2}.

Step 0. Input: (i) mean: 2-dimensional mean vector $\boldsymbol{\mu_2}$; (ii) sigma: 2×2 covariance matrix $\boldsymbol{\Sigma}_2$; (iii) lower: lower vectors $(-\text{Inf}, -z_{\alpha_{Ept}})^T$ (Here, we use −Inf or Inf for unbounded limits); (iv) upper: upper integration bounds $(z_{\alpha_{S_2}}, z_{\alpha_{Ept}})^T$.

Step 1. Load the `mvtnorm` package in R: library (mvtnorm).

Step 2. Compute the multivariate normal probability β_{Ept,S_2}: prob = pmvnorm(lower = lower, upper = upper, mean = mean, sigma = sigma).

AIMS Press

© 2025 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)