Published in final edited form as:

Clin Trials. 2023 June ; 20(3): 223–234. doi:10.1177/17407745231158906.

Dynamic use of historical controls in clinical trials for rare disease research: a re-evaluation of the MILES trial

Nusrat Harun, PhD[1], **Nishant Gupta, MD**[2], **Francis X McCormack, MD**[2], **Maurizio Macaluso, MD, DrPH**[1,3]

[1]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[2]Division of Pulmonary Critical Care and Sleep Medicine, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

[3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Abstract

Background: Randomized controlled trials offer the best design for eliminating bias in estimating treatment effects but can be slow and costly in rare disease research. Additionally, an equal randomization approach may not be optimal in studies in which prior evidence of superiority of one or more treatments exist. Supplementing prospectively enrolled, concurrent controls with historical controls can reduce recruitment requirements and provide patients a higher likelihood of enrolling in a new and possibly superior treatment arm. Appropriate methods need to be employed to ensure comparability of concurrent and historical controls to minimize bias and variability in the treatment effect estimates and reduce the chances of drawing incorrect conclusions regarding treatment benefit.

Methods: MILES was a phase III placebo-controlled trial employing 1:1 randomization that led to FDA approval of sirolimus for treating patients with lymphangioleiomyomatosis. We re-analyzed the MILES trial data to learn whether substituting concurrent controls with controls from a historical registry could have accelerated subject enrollment while leading to similar study conclusions. We used propensity score matching to identify exchangeable historical controls from a registry balancing the baseline characteristics of the two control groups. This allowed more new patients to be assigned to the sirolimus arm. We used trial data and simulations to estimate key outcomes under an array of alternative designs.

Corresponding Author: Maurizio Macaluso, MD, DrPH, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229-3026, USA Maurizio.Macaluso@cchmc.org.

Declaration of conflicting interests
The Authors declare that there is no conflict of interest.

MILES ClinicalTrials.gov Number: NCT00414648

Research ethics and patient consent
The Cincinnati Children's Hospital Institutional Review Board (IRB) reviewed this research study: IRB ID 2022–0017. The proposed activities described in this study was determined to be not research involving human subjects as defined by DHHS and FDA regulations. IRB review and approval by this organization was not required.

Research Data
The MILES dataset the NHLBI LAM Registry data can be made available upon approval by Dr. Francis McCormack and Dr. Nishant Gupta. The R code would be available from the first author upon request.

Results: Borrowing information from historical controls would have allowed the trial to enroll fewer concurrent controls while leading to the same conclusion reached in the trial. Simulations showed similar statistical performance for borrowing as for the actual trial design without producing type I error inflation and preserving power for the same study size when concurrent and historical controls are comparable.

Conclusions: Substituting concurrent controls with propensity score-matched historical controls can allow more prospectively enrolled patients to be assigned to the active treatment and enable the trial to be conducted with smaller overall sample size, while maintaining covariate balance and study power and minimizing bias in response estimation. This approach does not fully eliminate the concern that introducing non-randomized historical controls in a trial may lead to bias in estimating treatment effects, and should be carefully considered on a case-by-case basis. Borrowing historical controls is best suited when conducting randomized controlled trials with conventional designs is challenging, as in rare disease research. High-quality data on covariates and outcomes must be available for candidate historical controls to ensure the validity of these designs. Additional precautions are needed to maintain blinding of the treatment assignment and to ensure comparability in the assessment of treatment safety.

Keywords

Randomized controlled trials; historical controls; propensity score matching; rare diseases

Background

Randomized controlled trials (RCTs) are the accepted gold standard for determining treatment efficacy and safety. Validity hinges on the random allocation of eligible subjects to the intervention and control arms. If no effective therapies are available, as is the case for many rare diseases, control subjects are assigned to a sham treatment (placebo). In rare disease research, accrual is often limited by the lack of eligible patients. Also, enrollment can be slow if patients are less motivated to participate for fear of not receiving active treatment.

Under certain conditions, existing data on patients who may serve as non-randomized historical controls may supplement information collected from prospectively enrolled, concurrent controls. Increasing the allocation of prospectively enrolled patients to the experimental arm while supplementing the control arm with historical controls is an appealing strategy to augment clinical trials.[1–3] Such a strategy can increase the probability that patients receive a promising new treatment, increase statistical power and shorten the time required to complete the trial.

Historical control data may be available from previous single-arm or randomized trials or from non-randomized studies, including disease registries and well-designed natural history studies, which may provide high-quality information. Using “real-world data” to determine treatment effectiveness would be consistent with the policies of the US Food and Drug Administration (FDA), which is open to using real-world evidence in its decision-making processes.[4,5]

Valid statistical approaches have been developed for incorporating information from historical controls[6–9] that satisfy certain "acceptability criteria".[10] A comprehensive account of different methods can be found in recent review articles.[11,12] Recent literature has focused on the criteria for choosing historical controls and appropriate statistical methods for data analysis.[13–20]

A major concern about using non-randomized historical controls is the potential for bias in estimating treatment effects. Studies may differ according to eligibility criteria and other patient characteristics, leading to covariate imbalance between studies. Thus, combining studies to compare treatments may lead to confounded effect estimates. Effect modification may also lead to differences in effect estimates between studies.

Propensity scores have been widely used to eliminate confounding and reduce bias in treatment effect estimation from observational studies.[21,22] In this approach, the conditional probability of being assigned to a certain group given a set of observed covariates is estimated under the assumption of no unmeasured confounding. This method can be used to select historical controls and replace concurrent controls in an RCT, so long as the relationship of the known confounders and effect modifiers with the response is specified correctly in the propensity score model.[23,24]

In this paper, we evaluate the impact of alternative designs that could have been applied to a completed randomized placebo-controlled trial of sirolimus for the treatment of lymphangioleiomyomatosis. We assess how using a historical registry and borrowing information from propensity score-matched historical controls would have influenced enrollment and estimation of treatment efficacy, and discuss benefits and limitations of our approach.

Methods

We demonstrate the use of propensity score matching to identify exchangeable historical controls in an RCT and replace concurrent controls while assigning more newly enrolled patients to the treatment arm according to different matching schemes. We re-analyze archived trial and registry data to illustrate the proposed designs and use simulations to investigate the operating characteristics and performance metrics of the alternative designs.

Problem setup and matching method

We suppose that the i-th subject is available from the RCT, $S_i = 1$ or the historical study, $S_i = 0$. Each subject in the RCT would have a treatment assignment denoted by $T_i = 1$ for treatment or $T_i = 0$ for control, whereas the historical study subjects would have $T_i = 0$. We also assume that the key patient-level covariates ($\boldsymbol{X}_i$) are measured in both studies. We use the potential outcome framework in the context of obtaining control subjects from different studies to estimate treatment effect. Assuming the distribution $[Y \mid T = 0, \boldsymbol{X}]$ in the two studies are similar, the RCT and external control data can be pooled together to improve the estimation of the treatment effect in the RCT by increasing sample size. The treatment effect will be unbiased given the known covariates under the assumption of no unmeasured confounders.

The propensity of the i-th subject of taking part in the study of interest can be calculated as the conditional probability given the covariates $\boldsymbol{X}_i$: $e(\boldsymbol{X}_i) = P(S_i = s / \boldsymbol{X}_i)$, $s = 0, 1$. The nearest-neighbor matching algorithm is commonly used for propensity score matching. Specifically, the k-th subject from the RCT is matched to the l-th subject from the registry if the estimated propensity scores $\hat{e}(\boldsymbol{X}_k)$ and $\hat{e}(\boldsymbol{X}_l)$ do not differ by more than a pre-determined caliper width, η.

Lymphangioleiomyomatosis studies

Lymphangioleiomyomatosis (LAM) is a progressive rare disease predominantly seen in women which is associated with cystic destruction of the lung.[25] The Multicenter International Lymphangioleiomyomatosis Efficacy of Sirolimus (MILES) study[26] was a phase III, double blind, placebo-controlled trial that led to FDA approval of sirolimus in 2015.[27,28] A total of 89 patients were enrolled across USA, Canada, and Japan. The primary outcome for the trial was the rate of change in the forced expiratory volume in 1 second (FEV1) to measure lung function decline. Employing equal randomization, patients were assigned to receive treatment (or placebo) for 12 months followed by a 12-month observation period off therapy. FEV1 was measured at baseline and at 3, 6, 9, 12, 18, and 24 month follow-up visits.

The National Heart, Lung and Blood Institute LAM Registry (registry here forth) was a comprehensive and rigorous natural history study established in the late 1990's.[29,30] It enrolled 246 women in the USA for a period of 3 years and followed them for up to 5 years through 2003. Longitudinal data, including FEV1, were collected at enrollment and every 12 months. The MILES trial was limited to patients with FEV1 ≤70% of the predicted baseline FEV1 value using age and height as predictors,[31] whereas the registry included patients of all severity levels. Overall, 108 patients from the registry would have met the MILES trial eligibility criteria and would be available as historical controls. Treatment strategies for LAM did not change appreciably between the registry timeframe and the MILES trial enrollment period.

Substituting MILES controls with registry controls using propensity score matching.—We calculate propensity scores to match MILES patients assigned to the placebo arm as they enroll with registry controls. The concurrently enrolled patient is assigned to the treatment arm if a match is found. The algorithm can be summarized as follows:

- **I.** Screen patients for eligibility;
- **II.** Evaluate treatment assignment in MILES trial as observed;
  - **1.** if treatment allocation is to the intervention arm;
    - **i.** patient receives sirolimus,
    - **ii.** use trial outcomes.
  - **2.** else if treatment allocation is to the control arm;
    - **i.** if there is a matched registry control;

a. patient receives sirolimus,

b. impute trial outcomes, and

c. use matched registry patient as control.

ii. else if no matched registry control is found;

a. patient receives placebo,

b. use trial outcomes.

The MILES trial enrolled patients from USA, Canada, and Japan, while the historical registry enrolled patients only from the USA. We investigated two different scenarios: (1) we searched for suitable historical controls for all concurrent controls; (2) we did not attempt to replace Japanese concurrent controls with the historical controls retaining their original treatment assignments. We graphically illustrate the steps described above using flowcharts in Figures 1(a) and 1(b).

We fit the propensity score model using logistic regression, including the following baseline covariates: age at enrollment (X_1), race (X_2), menopausal status (X_3), disease subtype (X_4), history of angiomyolipomas (X_5), history of pneumothorax (X_6), need for supplemental oxygen (X_7), time since diagnosis to enrollment (X_8), and baseline FEV1 values (X_9). We included an interaction term for menopausal status and baseline FEV1 based on subject matter knowledge from prior publications.[30,32] The baseline covariate distribution and FEV1 slopes for the two control groups are presented in the online supplement (Tables S1 and S2) to justify our choices. Additionally, we included interaction terms for baseline FEV1 with age, disease subtype, and need for supplemental oxygen after examining the data. More details are in the online appendix (A1). We included site (X_{10}) in the propensity model when Japanese concurrent controls were substituted by historical controls.

To impute the FEV1 at the t-th follow up time for the i-th placebo arm patient who was re-assigned to the sirolimus arm, we use a predictive model derived from the actual trial data:

$$\begin{aligned} FEV1_{it} = {} & \beta_0 + \beta_1 {*} X_{1i} + \beta_2 {*} X_{2i} + \beta_3 {*} X_{3i} + \beta_4 {*} X_{4i} + \beta_5 {*} X_{5i} + \beta_6 {*} X_{6i} + \beta_7 {*} X_{7i} \\ & + \beta_8 {*} X_{8i} + \beta_9 {*} X_{9i} + \beta_{10} {*} X_{10i} + \beta_{11} {*} X_{11it} + \beta_{12} {*} X_{12i} + \beta_{13} {*} X_{11i} {*} X_{12t} \\ & + \beta_{14} {*} X_{3i} {*} X_{11i} {*} X_{12it} \end{aligned} \quad \text{(Eqn. 1)}$$

where X_{11} and X_{12} refer to time of FEV1 measurement and treatment arm, respectively.

We used different matching schemes to evaluate four alternative designs as follows:

Scheme 1: limits the trial size to 89 patients, but replaces placebo arm patients with matched historical controls reducing the required number of prospectively enrolled patients.

Scheme 2: includes the original 89 patients, but increases the trial sample size by shifting concurrent controls to the sirolimus arm and replacing the shifted concurrent controls with matched historical controls.

We investigated both 1:1 and 1:2 concurrent control/ historical control matching ratios for the two schemes. Thus, we have four alternative designs: (a) Scheme 1 – 1:1 Matching, (b) Scheme 1 – 1:2 Matching, (c) Scheme 2 – 1:1 Matching, (d) Scheme 2 – 1:2 Matching. All designs maximize the number of prospectively enrolled patients who receive treatment. The actual time intervals occurring between prospectively enrolled patients were used to compute a realistic estimate of the total length of enrollment for each alternative design. Historical controls were immediately available and did not contribute any delays.

Data analysis.—The primary endpoint in the MILES trial was the rate of FEV1 change at 12 months. The rate of FEV1 change per month (FEV1 slope) was re-estimated using the baseline, 3, 6, 9, 12 month measures (5 time points) for the MILES trial and the baseline and 12-month measures (2 time points) for the registry. We re-analyzed the archived MILES trial data first to reproduce the results and then obtained the results under the four hypothetical alternative designs imputing the counterfactual FEV1 values. We used a linear mixed effects model to fit the FEV1 values measured over time as a function of time in months, treatment arm, and treatment by time interaction with random intercept and time effects, to be consistent with the final analysis of the actual trial. The regression model was specified as follows:

$$FEV1_{it} = \mu_{it} + b_{0it} + b_{1it} * X_{11it} + \varepsilon_{it} \qquad \text{(Eqn. 2)}$$

The standard error of the FEV1 slope were 2 for both arms of the trial and 3 for the historical controls indicating that the treatment effect was evaluated with sufficient precision.

In the MILES trial, a planned interim analysis was conducted when 40 patients had completed the 12-month visit. We added hypothetical interim analyses after 20 and 60 patients had completed the 12-month visit to further investigate the effect size. The Lan-Demets alpha spending function[33] was used to obtain a nominal type I error rate of 0.05 at the final analysis. The total number of patients in the new trial and historical controls borrowed were used to calculate the adjusted type I error rates at the interim analyses and all confidence limits. Details on the computation of the confidence limits are in the online appendix (A2).

Simulations

To investigate the operating characteristics and design properties of the alternative designs, we simulated 2000 datasets. The data were simulated under:

1. the null hypothesis (H_0) setting the FEV1 slopes for both arms equal to that observed in the MILES placebo arm, that is, no treatment difference, and
2. the alternative hypothesis (H_1) setting the FEV1 slopes of each arm as estimated in the final report of the MILES trial.[26]

The type I error was calculated as the proportion of 95% confidence intervals of the estimated between-arm FEV1 slope difference not including zero under H_0. Power was calculated as the proportion of confidence intervals not including zero under H_1.

The FEV1 values for the *i*-th patient at the *t*-th follow up time were calculated as:

$$FEV1_{it} = \mu_{it} + b_{0it} + b_{1it} {}^{*} X_{11it} + \varepsilon_{it} \quad \text{(Eqn. 3)}$$

where, b_{0it}, b_{1it}, and ε_{it} are randomly generated assuming a $N(0, \sigma^2)$ distribution with σ^2 = 2000 under H_0 and σ^2 = 400 under H_1. These values of σ^2 were chosen to yield a 5% type I error and 85% power with a total sample size of 89. This was done by computing the error rates for different values of σ^2. We calculated the grand mean (μ_{it}) for the data generative model using the same regression model as in Eqn. 1 using the regression parameters (βs) estimated from the MILES trial data. The FEV1 values for the registry controls were not simulated and the actual data were used for analysis. The actual baseline covariates for all patients as observed in the MILES trial and the registry were used for matching and imputing the counterfactual FEV1 values for the trial. Details are in the online appendix (A3).

Results

We conducted a preliminary analysis to investigate if the replacement of concurrent controls with propensity score-matched historical controls would change the average FEV1 slope among the controls while maintaining the covariate balance, using the nearest neighbor algorithm with (a) no caliper width (η) specified, (b) η = 0.2, and (c) η = 0.1 (Table 1). Thirty-two concurrent controls from the USA/Canada were replaced with 32 (1:1) or 64 (1:2) matched historical controls without replacement when no caliper width was specified. Similarly, 43 (1:1) or 86 (1:2) historical controls were matched when we attempted to replace Japanese controls with historical controls. The overall covariate balance achieved was similar for all η values while the number of available matches decreased with increasingly smaller caliper widths. The point estimates of the FEV1 slope differences between control groups are close to the null value of zero and quite different from the treatment effect found in the MILES trial. The standard errors of the slope difference estimates are larger with smaller caliper widths as expected from the smaller sample sizes after matching. This analysis confirmed not only that covariate balance was achieved through propensity score matching, but also that replacing concurrent controls with historical controls did not appreciably change the FEV1 slope in the control arm. We chose not to specify caliper width in subsequent analyses for simplicity.

We present the results of the re-analysis of the MILES trial adding hypothetical interim analysis results in Table 2. The number of patients on the MILES trial was 89 by the end of the enrollment period of 964 days; the treatment allocation ratio and the proportion treated approached the designed value of 0.5 by the second interim analysis. The FEV1 slope estimates and standard errors at the third interim analysis were very similar to the final analysis estimates for both sirolimus and placebo arms. The confidence intervals for the FEV1 slope difference included zero for the first and second interim analyses but were significant at the third interim and final analyses.

Table 2 also displays results from the re-analyses using the four alternative designs not substituting Japanese controls. In the “Scheme 1 – 1:1 Matching” design, the trial would require 7, 10, 15, and 22 fewer concurrent controls, respectively, at the three interim and final analyses. This design would shorten the enrollment period by about two months (Figure 2). At least 80% of the prospectively enrolled patients would have received active treatment. The FEV1 slope difference would have been significantly different from zero by the second interim analysis. Using the “Scheme 1 – 1:2 Matching” design would have resulted in 10, 18, 22, and 34 fewer concurrent controls at the three interim and final analyses, respectively, but these analyses would have occurred considerably sooner. The target number of patients would have been enrolled about five months earlier (Figure 2). The total sample size would have increased to 121 under “Scheme 2 – 1:1 Matching” and 153 under “Scheme 2 – 1:2 Matching” designs within the same enrollment period of 964 days. The FEV1 slope difference would have been significantly different from zero by the third interim analysis for all alternative designs, similar to the observed design.

The results from the re-analyses of the MILES trial using the alternative designs substituting Japanese controls are presented in Table 3. The Scheme 1 designs would lead to shorter enrollment periods while the Scheme 2 designs would lead to larger study sizes, 132 and 175 patients for the two matching ratios, respectively. Substituting Japanese controls resulted in larger standard errors and wider confidence intervals for the estimates of FEV1 slopes and differences. The FEV1 slope difference would be significantly different from zero by the third interim analysis only for the Scheme 2 designs, requiring a much larger sample size to mitigate between-study heterogeneity.

Table 4 presents the results of the final analysis for the simulated trials for the observed and the four alternative designs under H_0 and H_1. The type I error was reduced using the alternative designs compared to the observed design whether or not Japanese controls were replaced. The Scheme 2 designs produced estimates of FEV1 slope differences closer to zero and with narrower confidence intervals when Japanese controls were not replaced. Under H_1, the FEV1 slope differences, standard errors, and confidence intervals were consistent with the observed design when Japanese controls were not replaced. The power increased considerably with the Scheme 2 designs. When Japanese controls were replaced, a desirable power was achieved only with the largest sample size under the “Scheme 2 – 1:2 Matching” design.

The performance of the alternative designs compared to the observed design under both H_0 and H_1 from the simulations in Table 4 are presented in Figure S1. The mean biases, standard deviations, root mean squared errors of the FEV1 slope differences were calculated from the 2000 simulations. The mean biases were farther from zero for the Scheme 1 designs compared to the Scheme 2 designs. The standard deviations and root mean squared errors were lower for all alternative designs compared to the observed design under H_0 for all scenarios. The Scheme 1 designs yielded slightly higher standard deviations and root mean squared errors compared to the observed design under H_1.

Discussion

Under some conditions, it may be advantageous to design randomized controlled trials incorporating high quality historical control data from previous clinical trials or well-designed longitudinal natural history studies. Dynamically replacing concurrent controls with comparable historical controls can reduce the number of patients assigned to the control arm or needed for overall enrollment, while maintaining covariate balance and power. More patients can receive a new or potentially superior treatment, possibly allowing early trial completion.

Adaptive designs using Bayesian dynamic borrowing methods[1–3] to achieve desirable treatment ratios assess similarity between the study outcomes while ignoring the covariate distribution. Covariate adjusted methods exist but are similarly contingent upon the exchangeability assumption.[34–37] Recently, propensity score-augmented designs have been proposed to balance covariate distributions discounting dissimilarities on treatment effect.[38–44] Previous work on incorporating historical controls using propensity score methods has focused on comparing trial controls with historical controls. One article focuses on demonstrating how historical controls can be used to adjust the sample size in single-arm studies at the end of the trial. A few recent papers have also used propensity score-matched historical controls to make decisions at interim analysis.[45,46]

In our analysis, we used propensity score matching to identify historical controls from a registry with similar baseline characteristics in order to assign a larger number of prospectively enrolled patients to the new treatment arm using alternative designs. We implemented interim monitoring rules using frequentist group sequential designs for trial monitoring. The results suggest comparable and often superior performance of the alternative designs over the traditional RCT design.

Because the MILES trial was completed in the past we had access to the characteristics of all participants at once, which made propensity score modeling and matching easier to accomplish. In a prospective trial, it may be necessary to enroll a certain number of patients to reliably determine the patient characteristics and effect modifiers to be included in the propensity model. Additional care would be required if there are unmeasured confounders or if all covariates are not measured in both datasets. There may be subgroups requiring integration of different sources of external controls. In the MILES trial, having suitable control patients for the Japanese patients could have required borrowing controls from two disparate studies. Additional analysis and matching methods would be required to make the historical cohorts from different studies similar.

It is well known that propensity scores are sensitive to model specification.[47] Correctly specifying the propensity model including important baseline characteristics and interaction terms is crucial for ensuring the validity of the comparisons. The model has to be constructed ad hoc from historical data and may not produce optimal results in an ongoing trial. Model averaging[48,49] and ensemble learning methods[50] can be used to estimate propensity scores in a data-driven manner and may be useful in the context of dynamic borrowing.

Although intuitively appealing, replacing concurrent controls in an interventional clinical trial with historical controls warrants careful scrutiny of the rigor and quality of data collection in the studies that are the source of historical controls. For example, in natural history studies efficacy assessments may provide estimates as accurate as in a trial, but the information collected on adverse effects may not be as thorough or as frequent. Incorporation of meticulous safety data collection methods into natural history studies would facilitate incorporating historical controls in the designs evaluated in this paper. Care needs to be taken to maintain blinding of treatment assignment when a large proportion of subjects in the study get assigned to the interventional arm.

The above-mentioned problems could be partly alleviated by limiting the number of concurrent controls to be replaced or by adaptively allocating more patients to the interventional arm. Alternatively, one could give different weights to the concurrent and historical controls when analyzing the data. A two-stage design[40] has been recommended in which a predetermined number of current controls are enrolled to provide a more accurate measure of similarity with the historical controls; next, the effective sample size estimated as the number of comparable historical controls from propensity score matching can be used to adaptively update the allocation ratio, assigning more patients to the new treatment arm.[1–3] The effective sample size quantifies the number historical controls comparable to concurrent controls; thus, the total number of controls needed to obtain a pre-determined allocation ratio can be adjusted using effective sample size at interim time points.

It took the MILES trial seven years to be completed at a cost of over $5M. Historical control borrowing methods appear to be a potentially useful strategy for making clinical trials more efficient and affordable, and highlight the importance to establish and maintain high quality natural history registries. Finally, we note that our methods are highly relevant to the design of clinical trials in rare diseases. The Rare Disease Clinical Research Network (www.rdcrn.org), funded by the National Institutes of Health, comprises over 20 current or past research consortia that conduct long-term natural history studies of multiple rare diseases, and are encouraged to bring new treatments to trial. It provides a logical context for conducting small-size clinical trials augmented with historical controls borrowed from the natural history studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The MILES trial was funded grants from the NIH Office of Rare Disease Research, administered by the National Center for Research Resources (RR019498 andRR019259), the Food and Drug Administration (FD003362), Canadian Institutes of Health Research, Pfizer Pharmaceuticals, the Japanese Ministry of Health, Labor, and Welfare, the LAM Foundation, the Tuberous Sclerosis Alliance, Cincinnati Children's Hospital Medical Center, Vi and John Adler, and the Adler Foundation. The National Heart, Lung and Blood Institute (NHLBI)

Lymphangioleiomyomatosis (LAM) Registry was supported by a National Heart, Lung and Blood Institute (U01HL58440) and the National Heart, Lung, and Blood Institute Intramural Research Program. We thank all investigators involved in the MILES trial and the NHLBI LAM Registry team for their contributions in the successful conduct of both endeavors. Special acknowledgements to Dr. Joel Moss and the team at the NHLBI Intramural Program for leading and collecting the majority of the data in the NHLBI LAM Registry. Dr. Jeffrey Krischer of the University of South Florida was the lead statistician of the MILES trial and his approach to the analysis was instrumental in defining the key covariates and effect estimates. Dr. John L. Thompson of Columbia University provided thoughtful advice and comments on the manuscript. Lastly, the authors express their sincere gratitude to all the patients with LAM who participated in the MILES trial and the NHLBI LAM Registry for their efforts and their enduring dedication and commitment to research.

Funding

This research was partly funded by the National Center for Advancing Translational Sciences (NCATS) for the Data Management and Coordinating Center (DMCC) of the Rare Diseases Clinical Research Network and by the National Institute of Neurological Disorders and Stroke (NINDS) under a cooperative agreement (U02TR002818).

Grant Support:

Maurizio Macaluso was supported in part by the cooperative agreement for the Data Management and Coordinating Center of the Rare Disease Clinical Research Network sponsored by the National Center for Advancing Translational Sciences (NCATS) and by the National Institute of Neurological Disorders and Stroke (NINDS) (U02TR002818).

Appendix

A1. Propensity Score Model for Treatment Assignment

We used logistic regression to calculate the propensity score for a patient in the trial and a registry control given a pre-specified function of the baseline covariates as $e(\mathbf{X}) = P(S = s / \mathbf{X})$. The propensity score model included age at enrollment (X_1), race (X_2), menopausal status (X_3), disease subtype (X_4), history of angiomyolipomas (X_5), history of pneumothorax (X_6), need for supplemental oxygen (X_7), time since diagnosis to enrollment (X_8), baseline FEV1 values (X_9), enrollment site (X_{10}), and few two-way interaction terms with baseline FEV1. Specifically, the following model was used to match the trial and registry controls:

$$logit(P(\text{Trial Patient })) = \beta_0\beta_1{*} + X_1 + \beta_2{*}X_2 + \beta_3{*}X_3 + \beta_4{*}X_4 + \beta_5{*}X_5 + \beta_6{*}X_6 + \beta_7{*}X_7 + \beta_8{*}X_8 + \beta_9{*}X_9 + \beta_{10}{*}X_{10} + \beta_{11}{*}X_1{*}X_9 + \beta_{11}{*}X_3{*}X_9 + \beta_{13}{*}X_4{*}X_9 + \beta_{14}{*}X_7{*}X_9$$

The main effects were included based on the published report of the MILES trial [26] that were examined to assess balance achieved by randomization. Effect modification by menopausal status and change in FEV1 have been reported in prior publications [28, 30]. Therefore, we included an interaction with menopausal status and baseline FEV1 in the propensity score model. We checked for effect modification by some other covariates by examining descriptive statistics (Table S3). We included an interaction term with baseline FEV1 if there was a significant association either in the trial or registry with baseline FEV1 (age, need for supplemental oxygen) or the difference in means appeared large (disease subtype). The enrollment site was dropped from the model when only the control patients from USA/Canada were matched on. The “nearest neighbor” greedy matching method was employed using the “matchit” function in the R package “MatchIt”. The default method of “no caliper width” matches a trial control with a registry control that has the closest distance measure. The distance measure used was the propensity score calculated using the above logistic regression model.

A2. Data Analysis and Confidence Limit Calculation Methods

In order to obtain the slope estimates, the FEV1 values measured over time were modeled as a function of time in months, treatment arm, and treatment by time interaction using a linear mixed effects model assuming random intercept and time. The regression model was specified as follows:

$$FEV1_{it} = (b_{0it} + \beta_0) + (b_{1it} + \beta_1)^{*} time_{it} + \beta_{2*} treatment_i + \beta_3{}^{*} treatment_i{}^{*} time_{it} + \varepsilon_{it} \quad \text{(Eqn. 2)}$$

The analysis was limited to 12 month data for both the trial and the registry using the available data at all time points. The model was fitted using the restricted maximum likelihood method in R using the "lmer" function in the "nlme" package. The confidence intervals were calculated using the "contest1D" function in the "lmerTest" package.

In the MILES trial, a planned interim analysis was conducted with the use of the O' Brien–Fleming stopping boundary when 40 patients had completed the 12-month visit [26]. We added hypothetical interim analyses after 20 and 60 patients had completed the 12-month visit to further investigate the effect size to obtain a nominal type I error rate of 0.05 at the final analysis. This was done to investigate the effect of historical control borrowing on interim analysis results. We implemented this using "ldbounds" package in R that uses the Lan-Demets alpha spending function to calculate the O' Brien–Fleming stopping boundaries. The adjusted type I error was determined using the proportion of patients at each interim analysis time. This proportion was calculated as the (number of patients enrolled + number of matched patients available from the registry) at the pre-specified interim analysis time divided by the (total number of patients that would be available from the trial + total number of matched registry controls).

A3. Simulation Methods

In the simulation studies, we keep the baseline covariates fixed for the MILES and registry patients as observed. We construct a predictive model from the actual trial data including the following baseline covariates: age at enrollment (X_1), race (X_2), menopausal status (X_3), disease subtype (X_4), history of angiomyolipomas (X_5), history of pneumothorax (X_6), need for supplemental oxygen (X_7), time since diagnosis to enrollment (X_8), and baseline FEV1 values (X_9):

$$\begin{aligned} FEV1_{it} = {} & \beta_0 + \beta_1{}^{*}X_{1i} + \beta_2{}^{*}X_{2i} + \beta_3{}^{*}X_{3i} + \beta_4{}^{*}X_{4i} + \beta_5{}^{*}X_{5i} + \beta_6{}^{*}X_{6i} + \beta_7{}^{*}X_{7i} \\ & + \beta_8{}^{*}X_{8i} + \beta_9{}^{*}X_{9i} + \beta_{10}{}^{*}X_{10i} + \beta_{11}{}^{*}X_{11it} + \beta_{12}{}^{*}X_{12i} + \beta_{13}{}^{*}X_{11i}{}^{*}X_{12t} \\ & + \beta_{14}{}^{*}X_{3i}{}^{*}X_{11i}{}^{*}X_{12it} \end{aligned} \quad \text{(Eqn. 1)}$$

where X_{11} and X_{12} refer to time of FEV1 measurement and treatment arm, respectively. We calculate the mean FEV1 values for the i-th patient at the t-th follow up visit (μ_{it}) using Eqn. 1 incorporating the covariate values from the MILES patients in the order they enrolled in the trial. The regression coefficients (βs) were estimated using the actual MILES trial data by fitting a linear mixed effect model with random intercept and time. The FEV1 values for the i-th patient at the t-th follow up visit (FEV1$_{it}$) for each simulation is drawn from a normal distribution with mean $\mu_{it} + b_{0it} + b_{1it}{}^{*}\ X_{11it}$ and variance σ^2 of the residuals, ε_{it}. Specifically, we use the following equation:

$$FEV1_{it} = \mu_{it} + b_{0it} + b_{1it}{}^{*}X_{11it} + \varepsilon_{it} \quad \text{(Eqn. 3)}$$

The random effect terms, b_{0it} and b_{1it}, and residuals, ε_{it}, were all assumed to have a $N(0, \sigma^2)$ distribution with common σ^2. The σ^2 values were tuned to yield a 5% type I error (under

H_0) and 85% power (under H_1) with a total sample size of 89 at the final analysis. The FEV1 values under the alternative hypothesis was generated using Eqn. 2 incorporating the same β values yielded by fitting the model in Eqn. 1 on the MILES data. This centered the FEV1 slope at approximately the slope estimates reported in the MILES trial in both arms. In the null scenario, all terms in Eqn. 1 related to treatment effect or interactions were set to 0 yielding slope estimates approximating to the placebo arm slope in both arms.

A4. SAS Code and Output for Rate of FEV1 Change

The SAS code and output (A.4.1) below demonstrate the model specification for the linear mixed model as described in the MILES original report [26]. The same model specification was used to estimate the FEV1 slope in the registry (A.4.2). The results are reported in Table S2.

A.4.1 Model for the MILES Patients

SAS Code:

```
proc mixed data=miles order=data;
class treatment_id subject_id time_point_cd;
model fev1_v1=treatment_id time1 treatment_id*time1/ solution chisq ddfm=kr;
random intercept time1/ subject=subject_id type=un g v vcorr;
estimate ‘slope Placebo’ time1 1 treatment_id*time1 0 1/cl;
estimate ‘slope Sirolimus’ time1 1 treatment_id*time1 1 0/cl;
where time1 < 15;
run;
```

SAS Output:

The SAS System

Model Information	
Data Set	WORK.MILES
Dependent Variable	fev1_v1
Covariance Structure	Unstructured
Subject Effect	subject_id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	**Levels**	**Values**
treatment_id	2	Sirolimus Placebo
subject_id	89	101014 101058 101584 101806 101810 102011 102655 102726 102735 102824 103177 103260 103503 103646 103702 103727 103824 103888 103994 104206 104268 104578 104650 104657 104693 104737 104762 104931 104961 105045 105051 105158 105169 105181 105220 105227 105230 105288 105295 105299 105315 105324 105336 105389 105403 105445 101022 101384 101583 101686 101807 102121 102516 102827 103343 103424 103542 103572 103660 103671 103723 103735 103775 103905 104048 104088 104185 104311 104330 104729 104804 104809 104882 104974 105105 105157 105170 105185 105199 105286 105293 105325 105327 105333 105346 105347 105377 105390 105419
time_point_cd	5	Baseline 3 months 6 months 9 months 12 months

Dimensions	
Covariance Parameters	4
Columns in X	6
Columns in Z per Subject	2
Subjects	89
Max Obs per Subject	5

Number of Observations	
Number of Observations Read	410
Number of Observations Used	407
Number of Observations Not Used	3

Iteration History			
Iteration	**Evaluations**	**−2 Res Log Like**	**Criterion**
0	1	6029.86869919	
1	2	5216.79324706	0.00311406
2	1	5208.09835040	0.00088670
3	1	5205.75458458	0.00011151
4	1	5205.48388572	0.00000240
5	1	5205.47844311	0.00000000

Convergence criteria met.

Estimated G Matrix				
Row	**Effect**	**subject_id**	**Col1**	**Col2**
1	Interce pt	101014	175535	324.39
2	time1	101014	324.39	69.2824

Estimated V Matrix for subject_id 101014					
Row	**Col1**	**Col2**	**Col3**	**Col4**	**Col5**
1	182771	176183	177157	178454	179427
2	176183	184346	178498	180350	181739
3	177157	178498	187747	183194	185206
4	178454	180350	183194	194222	189829
5	179427	181739	185206	189829	200533

Estimated V Correlation Matrix for subject_id 101014					
Row	**Col1**	**Col2**	**Col3**	**Col4**	**Col5**
1	1.0000	0.9598	0.9563	0.9472	0.9372
2	0.9598	1.0000	0.9595	0.9531	0.9452
3	0.9563	0.9595	1.0000	0.9593	0.9545
4	0.9472	0.9531	0.9593	1.0000	0.9619
5	0.9372	0.9452	0.9545	0.9619	1.0000

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
UN(1,1)	subject_id	175535
UN(2,1)	subject_id	324.39
UN(2,2)	subject_id	69.2824
Residual		7236.60

Fit Statistics	
−2 Res Log Likelihood	5205.5

Fit Statistics	
AIC (Smaller is Better)	5213.5
AICC (Smaller is Better)	5213.6
BIC (Smaller is Better)	5223.4

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
3	824.39	<.0001

Solution for Fixed Effects						
Effect	**treatment_id**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept		1359.26	64.7075	86.8	21.01	<.0001
treatment_id	Sirolimus	17.8737	89.9947	86.7	0.20	0.8430
treatment_id	Placebo	0	.	.	.	.
time1		−11.6450	2.0176	82.1	−5.77	<.0001
time1*treatment_id	Sirolimus	12.7216	2.7956	77.9	4.55	<.0001
time1*treatment_id	Placebo	0	.	.	.	.

Type 3 Tests of Fixed Effects						
Effect	**Num DF**	**Den DF**	**Chi-Square**	**F Value**	**Pr > ChiSq**	**Pr > F**
treatment_id	1	86.7	0.04	0.04	0.8426	0.8430
time1	1	77.9	14.29	14.29	0.0002	0.0003
time1*treatment_id	1	77.9	20.71	20.71	<.0001	<.0001

Estimates								
Label	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
slope Placebo	−11.6450	2.0176	82.1	−5.77	<.0001	0.05	−15.6585	−7.6315
slope Sirolimus	1.0766	1.9352	73.6	0.56	0.5797	0.05	−2.7797	4.9330

A.4.2 Registry Patients

SAS Code:

```
proc sort data=Registry; by descending pid time1; run;
proc mixed data=Registry order=data;
class pid time;
```

```
model fev1_v1= time1/ solution chisq ddfm=kr cl;
random intercept time1/ subject=pid type=un g v vcorr;
where time1 <= 15;
run;
```

SAS Output:

The SAS System

Model Information	
Data Set	WORK.REGISTRY
Dependent Variable	fev1_v1
Covariance Structure	Unstructured
Subject Effect	pid
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	**Levels**	**Values**
pid	109	6007 6006 6005 6003 6002 5004 5002 4007 4006 4004 4003 4002 4001 3220 3219 3217 3212 3211 3208 3207 3205 3199 3198 3196 3192 3191 3188 3185 3178 3177 3176 3173 3172 3171 3161 3159 3154 3147 3146 3140 3139 3137 3136 3132 3130 3128 3127 3125 3124 3121 3120 3118 3114 3111 3110 3108 3097 3096 3088 3083 3077 3074 3073 3072 3070 3065 3060 3058 3056 3054 3053 3052 3051 3050 3046 3045 3042 3041 3039 3037 3034 3032 3031 3030 3027 3026 3023 3022 3021 3019 3018 3016 3015 3013 3011 3008 3007 3005 3004 2015 2012 2009 2001 1011 1010 1008 1007 1004 1001
time	2	0 12

Dimensions	
Covariance Parameters	4
Columns in X	2
Columns in Z per Subject	2
Subjects	109
Max Obs per Subject	2

Number of Observations	
Number of Observations Read	192

Number of Observations	
Number of Observations Used	192
Number of Observations Not Used	0

Iteration History			
Iteration	**Evaluations**	**−2 Res Log Like**	**Criterion**
0	1	2889.68243965	
1	4	2820.12864270	0.00460641
2	1	2813.05616633	0.00120137
3	1	2811.33458249	0.00011182

Iteration History			
Iteration	**Evaluations**	**−2 Res Log Like**	**Criterion**
4	1	2811.18806885	0.00000123
5	1	2811.18654903	0.00000000

Convergence criteria met.

Estimated G Matrix				
Row	**Effect**	**pid**	**Col1**	**Col2**
1	Interce pt	6007	147441	2974.50
2	time1	6007	2974.50	

Estimated V Matrix for pid 6007		
Row	**Col1**	**Col2**
1	196309	186109
2	186109	273646

Estimated V Correlation Matrix for pid 6007		
Row	**Col1**	**Col2**
1	1.0000	0.8030
2	0.8030	1.0000

Covariance Parameter Estimates		
Cov Parm	**Subject**	**Estimate**
UN(1,1)	pid	147441
UN(2,1)	pid	2974.50
UN(2,2)	pid	3.97E-14
Residual		48868

Fit Statistics	
−2 Res Log Likelihood	2811.2
AIC (Smaller is Better)	2817.2
AICC (Smaller is Better)	2817.3
BIC (Smaller is Better)	2825.3

Null Model Likelihood Ratio Test		
DF	**Chi-Square**	**Pr > ChiSq**
2	78.50	<.0001

Solution for Fixed Effects								
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**
Intercept	1437.27	42.4653	108	33.85	<.0001	0.05	1353.09	1521.44
time1	−6.4278	2.8156	82.6	−2.28	0.0250	0.05	−12.0284	−0.8271

Type 3 Tests of Fixed Effects						
Effect	**Num DF**	**Den DF**	**Chi-Square**	**F Value**	**Pr > ChiSq**	**Pr > F**
time1	1	82.6	5.21	5.21	0.0224	0.0250

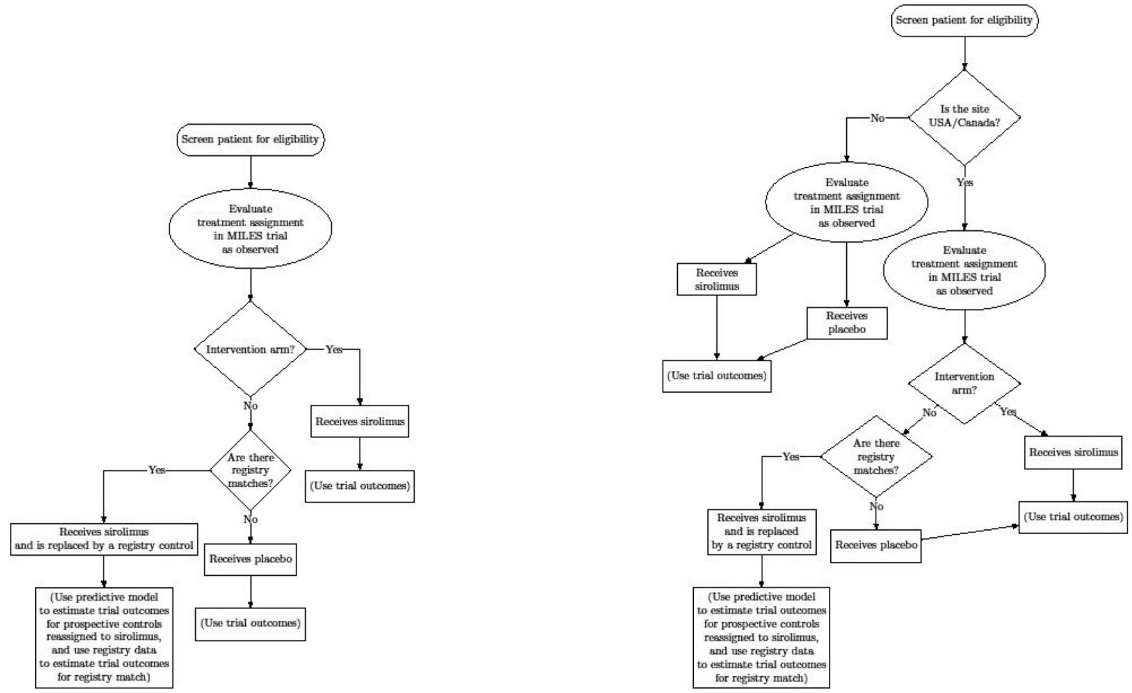
References

1. Hobbs BP, Carlin BP and Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. Clin Trials 2013; 10, 430–440. [PubMed: 23690095]
2. Kaizer AM, Hobbs BP and Koopmeiners JS. A multi-source adaptive platform design for testing sequential combinatorial therapeutic strategies. Biometrics 2018; 74(3):1082–1094. [PubMed: 29359450]
3. Normington J, Zhu J, Mattiello F, et al. An efficient Bayesian platform trial design for borrowing adaptively from historical control data in lymphoma. Contemp Clin Trials 2020 Feb;89:105890. [PubMed: 31740427]

4. US Food and Drug Administration Center for Devices and Radiological Health. Guidance for Industry and Food and Drug Administration Staff. Use of real-world evidence to support regulatory decision-making for medical devices. Centre for Devices and Radiological Health, Rockville; 2017.
5. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program. MD: Rockville; December 2018.
6. Ibrahim JG and Chen M-H. Power Prior Distributions for Regression Models. Statistical Science 2000; 15:46–60.
7. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics 2011; 67(3):1047–1056. [PubMed: 21361892]
8. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 2014; 70(4):1023–1032. [PubMed: 25355546]
9. Kaizer AM, Koopmeiners JS and Hobbs BP. Bayesian hierarchical modeling based on multisource exchangeability. Biostatistics 2018; 19(2):169–184. [PubMed: 29036300]
10. Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chronic Dis 1976; 29(3):175–188. [PubMed: 770493]
11. Viele K, Berry S, Neuenschwander S, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat 2014; 13(1):41–45. [PubMed: 23913901]
12. van Rosmalen J, Dejardin D, van Norden Y, et al. Including historical data in the analysis of clinical trials: Is it worth the effort? Stat Methods Med Res 2018;27(10):3167–3182. [PubMed: 28322129]
13. Lim J, Walley R, Yuan J, et al. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. Ther Innov Regul Sci 2018; 52(5):546–559. [PubMed: 29909645]
14. Ghadessi M, Tang R, Zhou J, et al. A roadmap to using historical controls in clinical trials - by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). Orphanet Journal of Rare Diseases 2020; 15(1):69. [PubMed: 32164754]
15. Schmidli H, Häring DA, Thomas M, et al. Beyond randomized clinical trials: Use of external controls. Clin Pharmacol Ther 2020; 107(4):806–816. [PubMed: 31725899]
16. Seifu Y, Gamalo-Siebers M, Barthel FM, et al. Real-world evidence utilization in clinical development reflected by US product labeling: Statistical review. Ther Innov Regul Sci 2020; 54(6):1436–1443. [PubMed: 32514737]
17. Hatswell A, Freemantle N, Baio G, et al. Summarizing salient information on historical controls: A structured assessment of validity and comparability across studies. Clin Trials 2020; 17(6):607–616. [PubMed: 32957804]
18. Cooner F, Gamalo-Siebers M, Xia A, et al. Use of alternative designs and data sources for pediatric trials. Stat Biopharm Res 2020; 12(2):210–223.
19. Jahanshahi M, Gregg K, Davis G, et al. The Use of External Controls in FDA Regulatory Decision Making. Ther Innov Regul Sci 2021; 55(5):1019–1035. [PubMed: 34014439]
20. Hall KT, Vase L, Tobias DK, et al. Historical controls in randomized clinical trials: opportunities and challenges. Clin Pharmacol Ther 2021; 109(2): 343–351. [PubMed: 32602555]
21. Rosenbaum P and Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41–55.
22. Rosenbaum P and Rubin D. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984; 79:516–524.
23. Rosenbaum P and Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985; 39:33–38.
24. Stuart EA and Rubin DB. Matching with multiple control groups and adjusting for group differences. J Educ Behav Stat 2008; 33:279–306.
25. McCarthy C, Gupta N, Johnson SR, et al. Lymphangioleiomyomatosis: pathogenesis, clinical features, diagnosis, and management. Lancet Respir Med 2021; 9(11):1313–1327. [PubMed: 34461049]

26. McCormack FX, Inoue Y, Moss J, et al. Efficacy and safety of sirolimus in lymphangioleiomyomatosis. N Engl J Med 2011; 364(17):1595–1606. [PubMed: 21410393]
27. McCormack FX, Gupta N, Finlay GR, et al. Official American Thoracic Society/Japanese Respiratory Society clinical practice guidelines: lymphangioleiomyomatosis diagnosis and management. Am J Respir Crit Care Med 2016; 194(6):748–761. [PubMed: 27628078]
28. Gupta N, Finlay GA, Kotloff RM, et al. ATS Assembly on Clinical Problems. Lymphangioleiomyomatosis Diagnosis and Management: High-Resolution Chest Computed Tomography, Transbronchial Lung Biopropensity scorey, and Pleural Disease Management. An Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guideline. Am J Respir Crit Care Med 2017;196(10):1337–1348. [PubMed: 29140122]
29. Ryu JH, Moss J, Beck GJ, et al. ; NHLBI LAM Registry Group. The NHLBI lymphangioleiomyomatosis registry: characteristics of 230 patients at enrollment. Am J Respir Crit Care Med 2006;173(1):105–11. [PubMed: 16210669]
30. Gupta N, Lee H-S, Ryu JH, et al. The NHLBI LAM Registry: Prognostic physiologic and radiologic biomarkers emerge from a 15-year prospective longitudinal analysis. Chest 2019; 155(2):288–296. [PubMed: 29940164]
31. Hankinson JL, Odencrantz JR and Fedan KB. Spirometric reference values form a sample of the general U.S. population. Am J Respir Crit Care Med 1999; 159:179–187. [PubMed: 9872837]
32. Gupta N, Lee HS, Young LR, et al. Analysis of the MILES cohort reveals determinants of disease progression and treatment response in lymphangioleiomyomatosis. Eur Respir J 2019;53(4):1802066. [PubMed: 30846465]
33. DeMets DL and Lan KK. Interim analysis: the alpha spending function approach. Stat Med 1994;13(13–14):1341–52; discussion 1353–1356. [PubMed: 7973215]
34. Hobbs BP, Sargent DJ and Carlin BP. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. Bayesian Anal 2012; 7(3): 639–674. [PubMed: 24795786]
35. Han B, Zhan J, John Zhong Z, et al. Covariate-adjusted borrowing of historical control data in randomized clinical trials. Pharm Stat 2017;16(4):296–308. [PubMed: 28560815]
36. Psioda MA, Soukup M and Ibrahim JG. A practical Bayesian adaptive design incorporating data from historical controls. Stat Med 2018;37(27): 4054–4070. [PubMed: 30033617]
37. Kotalik A, Vock DM, Donny EC, et al. Dynamic borrowing in the presence of treatment effect heterogeneity. Biostatistics 2021; 22(4): 789–804. [PubMed: 31977040]
38. Lin J, Gamalo-Siebers M and Tiwari R. Propensity score matched augmented controls in randomized clinical trials: A case study. Pharm Stat 2018; 17(5):629–647. [PubMed: 30066459]
39. Lin J, Gamalo-Siebers M and Tiwari R. Propensity-score-based priors for Bayesian augmented control design. Pharm Stat 2019; 18(2):223–238. [PubMed: 30537087]
40. Yuan J, Liu J, Zhu R, et al. Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. J Biopharm Stat 2019; 29(3):558–573. [PubMed: 30612514]
41. Chen WC, Wang C, Li H, et al. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. J Biopharm Stat 2020; 30(3):508–520. [PubMed: 32370640]
42. Wang C, Li H, Chen W-C, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. J Biopharm Stat 2019; 29:731–748. [PubMed: 31530111]
43. Wang C, Lu N, Chen W-C, et al. Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. J Biopharm Stat 2020; 30:495–507. [PubMed: 31707908]
44. Liu M, Bunn V, Hupf B, et al. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. Stat Med 2021; 40(22):4794–4808. [PubMed: 34126656]
45. Ventz S, Comment L, Louv B, et al. The use of external control data for predictions and futility interim analyses in clinical trials. Neuro Oncol 2022; 24(2): 247–256. [PubMed: 34106270]

46. Sawamoto R, Oba K and Matsuyama Y. Bayesian adaptive randomization design incorporating propensity score-matched historical controls. Pharm Stat 2022;21(5):1074–1089. [PubMed: 35278032]
47. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Political Analysis 2019; 27(4): 435–454.
48. Xie Y, Zhu Y, Cotton CA, et al. A model averaging approach for estimating propensity scores by optimizing balance. Stat Methods Med Res 2019; 28(1): 84–101. [PubMed: 28712346]
49. Kaplan D and Chen J. Bayesian Model Averaging for Propensity Score Analysis. Multivariate Behav Res 2014; 49(6):505–17. [PubMed: 26735355]
50. Lee BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010; 29(3):337–346. [PubMed: 19960510]



(a) All Patients without regards for recruitment site

(b) Stratified by recruitment site (USA/Canada vs. Japan)

Figure 1: Treatment Assignment Algorithm in the Re-analysis of the MILES Trial Data

(a) All Patients without regards for recruitment site

(b) Stratified by recruitment site (USA/Canada vs. Japan)

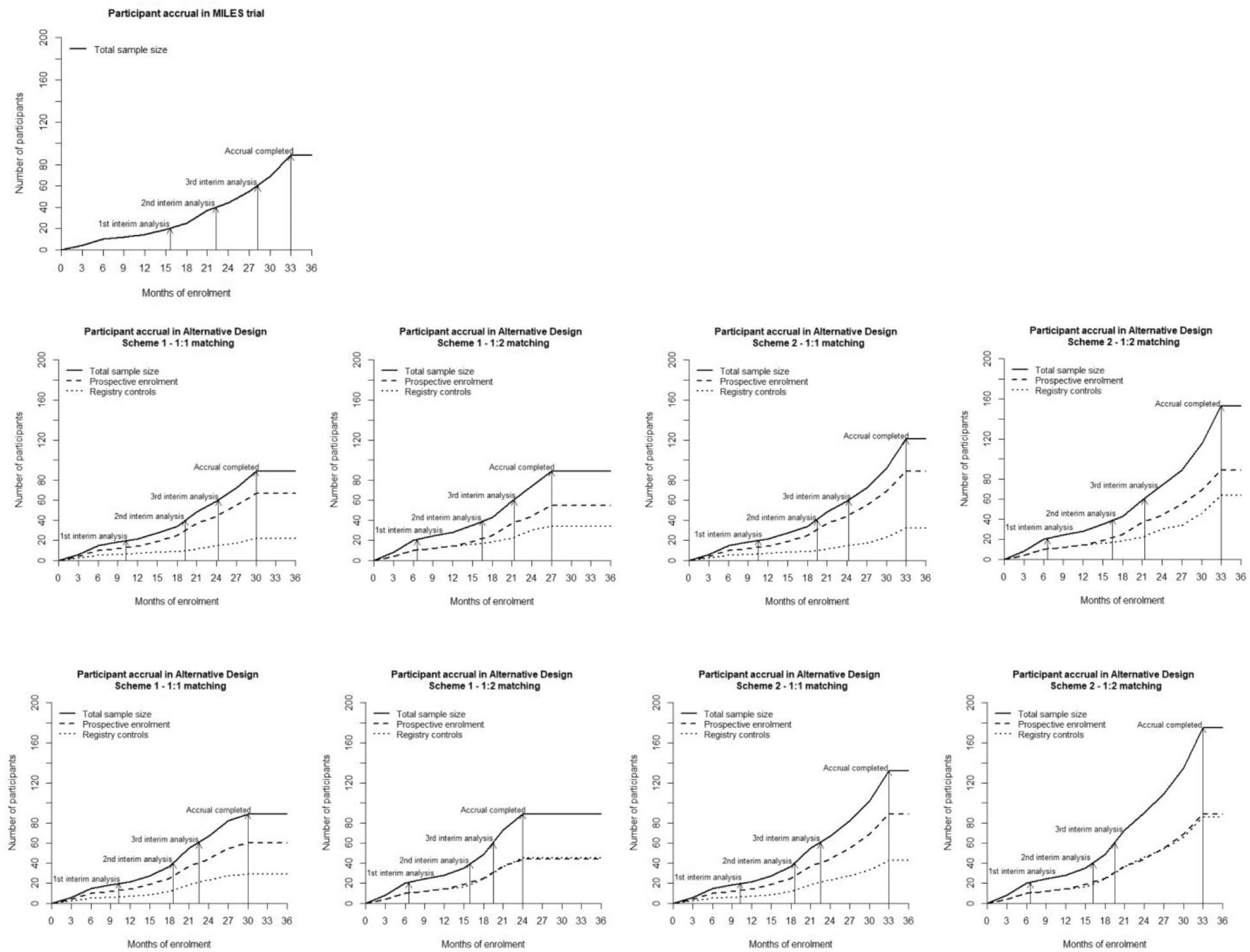


Figure 2: Participant accrual, timing of interim analyses and total sample size in the MILES trial and in four hypothetical alternative designs

Top row: Miles Trial

Middle row: Alternative Designs not replacing Japanese controls

Bottom row: Alternative Designs replacing Japanese controls

Table 1:

Covariate Distribution and FEV1 Slopes of the Patients in the Placebo Arm of the MILES trial and Registry after Matching

		MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)	MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)	MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)
		No Caliper Width Specified			**$\eta = 0.2$**			**$\eta = 0.1$**		
Sample Size	NRJC	32	32 (1:1)	64 (1:2)	22	22 (1:1)	42 (1:2)	22	22 (1:1)	36 (1:2)
	RJC	43	43 (1:1)	86 (1:2)	22	22 (1:1)	42 (1:2)	22	22 (1:1)	39 (1:2)
Propensity Score	NRJC	0.48±0.30	0.34±0.17	0.24±0.16	0.32±0.20	0.31±0.20	0.29±0.17	0.32±0.19	0.31±0.19	0.29±0.17
Mean±SD	RJC	0.61±0.34	0.31±0.16	0.19±0.16	0.32±0.21	0.31±0.20	0.28±0.17	0.32±0.19	0.31±0.19	0.34±0.18
Age (years)	NRJC	47.5±10.3	44.0±10.1	45.5±10.5	46.7±11.0	45.4±9.8	45.5±9.7	46.1±10.6	45.7±9.7	45.6±10.0
Mean±SD	RJC	45.9±10.3	44.0±9.8	45.3±11.2	45.9±10.5	45.0±9.7	44.0±8.4	46.1±10.6	45.6±9.7	45.6±10.0
Race, N (%) White	NRJC	30 (94) 1 (3)	27 (84) 3 (9)	57 (89) 3 (5)	20 (91) 1 (5)	20 (91) 1 (5)	38 (91) 2 (5)	20 (91) 1 (5)	20 (93) 1 (4)	33 (92) 1 (4)
Asian	RJC	30 (70) 12 (28)	37 (86) 3 (7)	78 (91) 4 (5)	20 (91) 1 (5)	20 (90) 1 (6)	38 (90) 2 (5)	20 (91) 1 (5)	20 (91) 1 (5)	36 (90) 2 (4)
TSC, N (%)	NRJC	2 (6)	4 (12)	8 (12)	2 (9)	2 (11)	4 (11)	2 (9)	3 (13)	4 (11)
	RJC	4 (9)	4 (9)	9 (10)	2 (9)	2 (10)	5 (11)	2 (9)	3 (12)	3 (9)
Postmenopausal, N (%)	NRJC	14 (44)	11 (34)	33 (52)	11 (50)	10 (46)	20 (48)	10 (45)	10 (47)	17 (49)
	RJC	16 (37)	17 (40)	54 (63)	10 (45)	10 (46)	21 (45)	10 (45)	10 (46)	20 (50)
Angiomyolipoma, N (%)	NRJC	17 (53)	13 (41)	25 (39)	9 (41)	9 (42)	16 (39)	9 (41)	10 (44)	14 (39)
	RJC	22 (51)	17 (40)	29 (34)	9 (41)	9 (42)	17 (39)	9 (41)	9 (42)	16 (40)
Pneumothorax, N (%)	NRJC	24 (75)	21 (66)	41 (64)	15 (68)	15 (66)	27 (64)	14 (64)	13 (60)	22 (60)
	RJC	29 (67)	29 (67)	51 (59)	15 (68)	17 (57)	27 (65)	14 (64)	14 (54)	24 (62)
SOU, N (%)	NRJC	16 (50)	16 (50)	29 (45)	13 (59)	10 (47)	20 (47)	13 (59)	9 (42)	15 (41)
	RJC	23 (53)	20 (47)	45 (52)	14 (64)	10 (47)	20 (48)	13 (59)	13 (50)	18 (45)
FEV1 Volume (ml)	NRJC	1475±421	1485±401	1519±449	1499±453	1496±434	1491±426	1489±446	1498±418	1477±424
Mean±SD	RJC	1378±446	1469±443	1457±439	1456±433	1479±450	1495±427	1489±446	1452±416	1489±425
Diagnosis to enrollment time (years), Median	NRJC	5.54	4.29	2.95	4.58	3.86	3.76	4.94	4.50	3.99
	RJC	5.50	4.21	3.07	4.94	3.72	3.64	4.94	4.35	3.28
FEV1 Slope ± SE (CI)	NRJC	−11±3 (−16, −5)	−12±3 (−18, −6)	−9±2 (−14, −5)	−11±4 (−19, −3)	−9±4 (−18, −1)	−10±3 (−15, −4)	−12±4 (−20, −5)	−10±4 (−18, −2)	−10±3 (−16, −4)

		MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)	MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)	MILES (Placebo)	Registry (1:1 matching)	Registry (1:2 matching)
		No Caliper Width Specified			**$\eta = 0.2$**			**$\eta = 0.1$**		
	RJC	−12±2 (−16, −7)	−10±3 (−15, −5)	−8±2 (−12, −5)	−12±4 (−19, −4)	−9±4 (−18, −1)	−9±3 (−15, −3)	−12±4 (−20, −5)	−10±4 (−18, −2)	−9±3 (−15, −3)
FEV1 Slope Difference ± SE (CI)	NRJC	NA	1±4 (−7, 9)	−1±5 (−10, 9)	NA	2±6 (−10, 13)	−1±5 (−10, 9)	NA	0±6 (−11, 11)	−2±5 (−11, 8)
	RJC	NA	−2±3 (−8, 5)	−3±3 (−9, 2)	NA	1±6 (−11, 12)	−2±5 (−11, 8)	NA	0±6 (−11, 11)	−3±5 (−12, 7)

TSC: Tuberous sclerosis complex

SOU: Supplemental oxygen use

NRJC: Not replacing Japanese controls

RJC: Replacing Japanese controls

SD: Standard Deviation

SE: Standard Error

CI: Confidence Interval

Table 2:

FEV1 Slopes by Arm and Between-arm Slope Differences for the Observed and Alternative Designs for the MILES trial

Analysis	Interim (First)	Interim (Second)	Interim (Third)	Final
Observed Design[a]				
Total Trial N (ESS[†])	20 (0)	40 (0)	60 (0)	89 (0)
Treatment Ratio	0.60	0.50	0.52	0.52
Proportion Treated in Current Trial	0.60	0.50	0.52	0.52
Days to Enroll All Patients	474	664	883	964
Sirolimus Slope: Mean±SE (CI)	4±4 (−27, 35)	1±2 (−7, 9)	2±2 (−4, 7)	1±2 (−3, 5)
Placebo Slope: Mean±SE (CI)	−12±5 (−50, 26)	−10±2 (−18, −1)	−12±2(−17, −7)	−12±2 (−16, −8)
Slope Difference: Mean±SE (CI)	16±6 (−33, 65)	11±3 (−1, 22)	13±3 (6, 21)[*]	13±3 (7, 18)[*]
Alternative Design (Scheme 1 – 1:1 Matching)[a]				
Total Trial N (ESS[†])	13 (7)	30 (10)	45 (15)	67 (22)
Treatment Ratio	0.65	0.63	0.61	0.64
Proportion Treated in Current Trial	1	0.83	0.82	0.85
Days to Enroll All Patients	320	566	748	909
Sirolimus Slope: Mean±SE (CI)	3±4 (−20, 26)	3±3 (−5, 11)	2±2 (−3, 7)	3±2 (0.02, 8)
Placebo Slope: Mean±SE (CI)	−13±7 (−48, 22)	−11±4 (−23, 0.05)	−11±3 (−18, −4)	−10±2 (−15, −6)
Slope Difference: Mean±SE (CI)	16±8 (−24, 57)	14±4 (0.22, 28)[*]	13±3 (0.84, 22)[*]	14±3 (8, 19)[*]
Alternative Design (Scheme 1 – 1:2 Matching)[b]				
Total Trial N (ESS[†])	10 (10)	22 (18)	38 (22)	55 (34)
Treatment Ratio	0.50	0.53	0.50	0.51
Proportion Treated in Current Trial	1	0.95	0.79	0.81
Days to Enroll All Patients	180	504	642	819
Sirolimus Slope: Mean±SE (CI)	5±8 (−37, 48)	4±5 (−13, 21)	3±4 (−6, 11)	3±3 (−3, 9)
Placebo Slope: Mean±SE (CI)	−5±9 (−52, 42)	−10±6 (−30, 11)	−11±4 (−20, −0.8)	−12±3 (−18, −5)
Slope Difference: Mean±SE (CI)	10±12 (−52, 72)	14±8 (−13, 40)	13±5 (0.08, 26)[*]	15±5 (6, 23)[*]
Alternative Design (Scheme 2 – 1:1 Matching)[c]				
Total Trial N (ESS[†])	20 (8)	40 (12)	60 (19)	89 (32)
Treatment Ratio	0.71	0.62	0.63	0.64
Proportion Treated in Current Trial	1	0.80	0.83	0.87
Days to Enroll All Patients	474	664	883	964
Sirolimus Slope: Mean±SE (CI)	4±3 (−11, 20)	2±2 (−5, 9)	3±2 (−2, 8)	2±2 (−1, 5)
Placebo Slope: Mean±SE (CI)	−13±6 (−40, 14)	−10±3 (−19, −0.4)	−12±3 (−18, −5)	−12±2 (−16, −8)
Slope Difference: Mean±SE (CI)	17±6 (−15, 48)	12±4 (0.26, 24)[*]	15±3 (7, 23)[*]	14±3 (9, 19)[*]
Alternative Design (Scheme 2 – 1:2 Matching)[d]				

Analysis	Interim (First)	Interim (Second)	Interim (Third)	Final
Total Trial N (ESS[†])	20 (16)	40 (24)	60 (38)	89 (64)
Treatment Ratio	0.56	0.50	0.51	0.51
Proportion Treated in Current Trial	1	0.80	0.83	0.88
Days to Enroll All Patients	474	664	883	964
Sirolimus Slope: Mean±SE (CI)	4±5 (−21, 30)	2±3 (−9, 13)	3±3 (−5, 11)	2±2 (−3, 6)
Placebo Slope: Mean±SE (CI)	−11±7 (−44, 22)	−10±4 (−23, 2)	−11±3 (−20, −3)	−11±2 (−16, −6)
Slope Difference: Mean±SE (CI)	15±9 (−26, 56)	12±5 (−4, 29)	14±4 (3, 26)[*]	13±3 (6, 19)[*]

[†]Effective sample size from matched historical controls

[*]Significant slope difference between treatment arms

SE: Standard Error

CI: Confidence Interval

[a]Adjusted α at interim analyses were 0.0000045, 0.0017, and 0.013

[b]Adjusted α at interim analyses were 0.0000045, 0.0017, and 0.013

[c]Adjusted α at interim analyses were 0.000006, 0.0013, and 0.011

[d]Adjusted α at interim analyses were 0.000008, 0.0011, and 0.010

Table 3:

FEV1 Slopes by Arm and Between-arm Slope Difference for the Alternative Designs of the MILES trial when replacing Japanese Controls

Analysis	Interim (First)	Interim (Second)	Interim (Third)	Final
Alternative Design (Scheme 1 – 1:1 Matching)[a]				
Total Trial N (ESS[†])	13 (7)	27 (13)	40 (20)	60 (29)
Treatment Ratio	0.65	0.68	0.67	0.67
Proportion Treated in Current Trial	1	1	1	1
Days to Enroll All Patients	320	552	664	883
Sirolimus Slope: Mean±SE (CI)	−2±5 (−29, 25)	−2±3 (−29, 25)	−4±3 (−11, 4)	−4±2 (−9, 0.3)
Placebo Slope: Mean±SE (CI)	−13±8 (−53, 26)	−13±5 (−29, 3)	−15±5 (−27, −3)	−12±4 (−20, −5)
Slope Difference: Mean±SE (CI)	11±9 (−35, 57)	11±6 (−7, 29)	11±6 (−3, 26)	8±4 (−0.8, 16)
Alternative Design (Scheme 1 – 1:2 Matching)[b]				
Total Trial N (ESS[†])	10 (10)	22 (20)	30 (30)	44 (45)
Treatment Ratio	0.50	0.52	0.50	0.49
Proportion Treated in Current Trial	1	1	1	1
Days to Enroll All Patients	180	504	566	692
Sirolimus Slope: Mean±SE (CI)	5±8 (−37, 48)	3±5 (−13, 20)	2±6 (−14, 17)	2±5 (−7, 11)
Placebo Slope: Mean±SE (CI)	−5±9 (−52, 42)	−8±6 (−29, 12)	−3±6 (−20, 13)	−7±5 (−17, 3)
Slope Difference: Mean±SE (CI)	10±12 (−52, 72)	12±8 (−14, 38)	5±9 (−18, 27)	9±7 (−5, 26)
Alternative Design (Scheme 2 – 1:1 Matching)[c]				
Total Trial N (ESS[†])	20 (8)	40 (20)	60 (29)	89 (43)
Treatment Ratio	0.71	0.67	0.67	0.67
Proportion Treated in Current Trial	1	1	1	1
Days to Enroll All Patients	474	664	883	964
Sirolimus Slope: Mean±SE (CI)	4±3 (−12, 20)	2±4 (−9, 15)	4±3 (−4, 11)	3±2 (−2, 7)
Placebo Slope: Mean±SE (CI)	−13±6 (−42, 16)	−9±6 (−27, 10)	−9±4 (−20, 1)	−11±3 (−17, −4)
Slope Difference: Mean±SE (CI)	17±6 (−16, 50)	11±7 (−11, 34)	13±5 (−0.2, 26)	13±4 (5, 21)[*]
Alternative Design (Scheme 2 – 1:2 Matching)[d]				
Total Trial N (ESS[†])	20 (16)	40 (40)	60 (58)	89 (86)
Treatment Ratio	0.56	0.5	0.51	0.51
Proportion Treated in Current Trial	1	1	1	1
Days to Enroll All Patients	474	664	883	964
Sirolimus Slope: Mean±SE (CI)	4±5 (−23, 32)	1±4 (−12, 15)	3±4 (−6, 13)	3±3 (−3, 8)
Placebo Slope: Mean±SE (CI)	−11±7 (−47, 24)	−5±5 (−20, 10)	−5±4 (−16, 5)	−9±3 (−15, −3)
Slope Difference: Mean±SE (CI)	15±9 (−29, 60)	6±6 (−13, 26)	9±6 (−5, 23)	12±4 (3, 20)[*]

[†]Effective sample size from matched historical controls

[*]Significant slope difference between treatment arms

SE: Standard Error

CI: Confidence Interval

[a] Adjusted α at interim analyses were 0.0000045, 0.0017, and 0.013

[b] Adjusted α at interim analyses were 0.0000045, 0.0022, and 0.013

[c] Adjusted α at interim analyses were 0.0000023, 0.0018, and 0.013

[d] Adjusted α at interim analyses were 0.0000015, 0.0018, and 0.013

Table 4:

Simulated FEV1 Slopes by Arm and Between-Arm Slope Difference for the Observed and Alternative Designs for MILES at the Final Analysis

Analysis	**Not Replacing Japanese Controls**		**Replacing Japanese Controls**	
	H_0	**H_1**	**H_0**	**H_1**
Observed Design				
Sirolimus Slope: Mean±SE (CI)	−12±6 (−25, 0)	0.7±3 (−5, 7)		
Placebo Slope: Mean±SE (CI)	−11±7 (−24, 2)	−12±3 (−18, −6)		
Slope Difference: Mean±SE (CI)	−1±9 (−19, 17)	12±4 (4, 21) *		
Type I Error/Power	0.05	0.84		
Alternative Design (Scheme 1 – 1:1 Matching)				
Sirolimus Slope: Mean±SE (CI)	−12±5 (−22, −1)	2±3 (−3, 7)	−11±5 (−21, −0.6)	2±3 (−4, 8)
Placebo Slope: Mean±SE (CI)	−10±7 (−24, 4)	−10±4 (−17, −3)	−8±8 (−24, 8)	−8±5 (−18, 2)
Slope Difference: Mean±SE (CI)	−2±8 (−18, 15)	12±4 (3, 20) *	−3±10 (−21, 16)	16±6 (−2, 22)
Type I Error/Power	0.02	0.82	0.02	0.37
Alternative Design (Scheme 1 – 1:2 Matching)				
Sirolimus Slope: Mean±SE (CI)	−11±6 (−23, 0.7)	2±4 (−5, 10)	−12±6 (−24, 1)	2±5 (−8, 11)
Placebo Slope: Mean±SE (CI)	−13±7 (−26, 0.4)	−12±4 (−20, −3)	−8±7 (−22, 6)	−7±6 (−18, 3)
Slope Difference: Mean±SE (CI)	1±9 (−16, 19)	14±6 (3, 25) *	−3±9 (−22, 15)	9±7 (−5, 23)
Type I Error/Power	0.02	0.80	0.02	0.17
Alternative Design (Scheme 2 – 1:1 Matching)				
Sirolimus Slope: Mean±SE (CI)	−12±4 (−20, −3)	2±2 (−3, 6)	−11±4 (−14, −8)	2±3 (−3, 7)
Placebo Slope: Mean±SE (CI)	−12±6 (−24, 0.1)	−12±3 (−18, −5)	−9±7 (−22, 3)	−9±4 (−17, −1)
Slope Difference: Mean±SE (CI)	0.4±7 (−14, 15)	13±4 (6, 21) *	−2±8 (−17, 13)	11±5 (2, 20) *
Type I Error/Power	0.02	0.98	0.01	0.68
Alternative Design (Scheme 2 – 1:2 Matching)				
Sirolimus Slope: Mean±SE (CI)	−12±4 (−21, −4)	1±3 (−4, 7)	−11±5 (−20, −2)	1±3 (−4, 7)
Placebo Slope: Mean±SE (CI)	−12±5 (−21, −3)	−11±3 (−17, −5)	−8±5 (−18, 2)	−11±3 (−17, −5)
Slope Difference: Mean±SE (CI)	−0.1±6 (−13, 12)	12±4 (4, 20) *	−3±7 (−17, 10)	12±4 (5, 20) *
Type I Error/Power	0.02	0.97	0.03	0.97

*Significant slope difference between treatment arms

SE: Standard Error

CI: Confidence Interval