



EXPLORING WHETHER A SYNTHETIC CONTROL ARM CAN BE DERIVED FROM HISTORICAL CLINICAL TRIALS THAT MATCH BASELINE CHARACTERISTICS AND OVERALL SURVIVAL OUTCOME OF A RANDOMIZED CONTROL ARM: CASE STUDY IN NON-SMALL CELL LUNG CANCER

INTRODUCTION

The U.S. Food and Drug Administration (FDA) aims to expedite the development and review of products intended to address an unmet medical need in the treatment of serious life-threatening conditions through the breakthrough therapy designation (BTD) as well as fast track, accelerated approval (AA), and priority review mechanisms.¹ In the case of AA, randomized trials meant to establish clinical benefit normally conducted before approval, may be conducted after AA, to confirm clinical benefit. For drugs and biologics intended to treat a serious or life-threatening condition, the FDA may grant BTD if preliminary clinical evidence indicates the product may provide substantial improvement over existing therapies, on ≥ 1 clinically significant endpoint.² Many products with BTD are approved through the AA pathway. Although AA may allow patients access to therapies that have demonstrated a substantial treatment effect, this introduces loss of clinical equipoise that may interfere with continued drug development. For example, patients may be reluctant to enroll in trials where they may be randomized to receive a perceived inferior therapy, or they may discontinue from ongoing clinical trials once the product is accessible through AA. FDA guidance states, “If it is clear during development that a product is intended to be approved under accelerated approval... confirmatory trial(s) should be underway at the time the marketing application is submitted.”¹ However, recruitment and conduct of the confirmatory trial must continue after AA. Data from the control arm may be compromised by early discontinuation or “cross-over” to the investigational therapy made available by AA, resulting in an inability to interpret the confirmatory clinical trial results. Finally, there are some clinical settings (e.g., rare diseases) where scarcity of patients or ethical concerns have demonstrated that a randomized control is not possible. These indications are often studied using single arm trials in which all enrolled patients receive the investigational agent.

CONTRIBUTORS

Medida a S...
LUNG...
Medida a S...
B...-Me e, S...
..

U.S. Food and D... Ad...
S...
Dauc...
U.S. Food and D... Ad...
S...
Medida a S...
J... H... U...
S...
U.S. Food and D... Ad...
Medida a S...
Dauc...
S...

ABOUT FRIENDS OF CANCER RESEARCH

Friends of Cancer Research drives collaboration among partners from every healthcare sector to power advances in science, policy, and regulation that speed life-saving treatments to patients.

The same impact on patient recruitment and retention may occur in circumstances where the drug is approved and available for off-label use, or when drugs with similar mechanisms of action, in the same drug class are approved. Interpretation of study results, such as overall survival (OS), are compromised when patients use alternate treatments (whether off-label use of the product under investigation or a newly marketed alternate treatment). This phenomenon has been coined “cross-over” or “treatment switch-over” and while some drugs have demonstrated benefits in OS even after cross over, “better methods to capture and summarize the OS benefit are needed” to address confounding bias introduced by this practice.³

Consider the example of the large, randomized trial (BRAVO study) assessing the PARP inhibitor

ical clinical trial data in a regulatory setting.⁷ Use of historical clinical trials data to enhance current research has some precedent. For instance, historical clinical trials data and propensity score methods were used to construct a reference response rate for a single-arm study of Blinatumomab for relapsed/refractory acute lymphoblastic leukemia, a rare disease.⁸ Lim et al. cite five drugs that incorporated historical control data in differing capacity, as part of a confirmatory clinical trial to obtain regulatory approvals between 2005 and 2015.⁵ None of those approvals; however, involved a direct comparison of the historical control arm to that of the treatment arm through a standard hypothesis testing procedure. The research proposed in this document aims to fill that gap. By choosing to retrospectively evaluate a carefully constructed synthetic control arm, not only against the actual control arm, but in future work, also against the treatment arm, we aim to understand the extent to which a synthetic control arm could be used for pragmatic purposes in cancer drug development.

An example of the use of historical control data for internal drug development decision making at a pharmaceutical company is presented in Neuenschwander et al.⁹ The discussion in that paper relates to non-confirmatory trials but can also be potentially used in a confirmatory trial setting. Rosmalen et al. present a comparative study of Bayesian methods to include historical data in the analysis of clinical trials data and stress the need to estimate the heterogeneity among trials and to satisfy criteria for comparability between the historical and current controls.¹⁰ Hobbs et al. investigate an adaptive randomization procedure that makes assignment to experimental therapy more likely when there is an absence of evidence for heterogeneity among the concurrent and historical controls.¹¹

Like any novel research initiative, the proposed use of historical control data to build a Synthetic Control Arm (SCA) has some associated risks. Selection bias and historical time effect are obvious risk factors. However, careful statistical planning and designing, along with a thorough understanding of the characteristics of the target population of interest, can help circumvent some of those risks. Pocock (1976) proposed a formal statistical plan for methodological inclusion of historical data in a randomized clinical trial.⁶ Appropriate statistical inference procedures for the context are also discussed. In addition, simulation studies can aid in understanding the bias-variance trade off and more generally, the influence of the historical control data.

This project is a unique collaboration of multiple stakeholders including contributions from

- Bristol-Myers Squibb
- Daiichi Sankyo
- Fred Hutchinson Cancer Research Center
- Friends of Cancer Research
- Johns Hopkins University
- LUNgevity Foundation
- Medidata Solutions
- Project Data Sphere
- U.S. Food and Drug Administration

We are grateful for the data, expertise, and resources each party has provided.

- Third, we will evaluate whether this matching has been successful by examining differences in baseline characteristics and prognostic scores in the target trial control arm and the SCA, as well as by exploring whether OS results observed for the target trial control arm are replicated in the SCA.
- Finally, additional evidence will be gained by repeating this process for a second Project Data Sphere trial designated as 'Target Trial B'. The process will not be repeated for the third Project Data Sphere trial since this trial is smaller than the others and fewer baseline variables are available for the matching processes.

Future research may be undertaken to explore whether a SCA can be used to mimic the treatment effect from a traditional randomized controlled trial. In that case, a SCA will be created to match the experimentally treated patients in the target trial and comparisons of the treatment effect using the randomized control and the same using the SCA will be made.

KEY FEATURES OF HISTORICAL DATA AND SCA ELIGIBILITY CRITERIA

Key features of the historical data and SCA eligibility criteria are described in this section. These studies were selected, and eligibility criteria were defined, based on clinical importance, balancing the need to identify a fairly homogenous set of historical clinical trial participants representative of a typical single indication in drug development and the desire to identify the largest volume of applicable historical data as possible.

As shown in Table 1, the historical data originated from open label or blinded phase 2 or 3 multinational trials, which began between 2004 and 2013. Enrollment in Target Trial A began in February of 2004 and the study reached its primary efficacy analysis time point in March 2007. Target Trial B began enrollment in May of 2006 and reached its primary efficacy analysis timepoint in August 2008. All patients were previously treated and presented at baseline with locally advanced or metastatic NSCLC. All patients were included in study arms that assigned treatment with docetaxel. Overall survival was measured as a key endpoint in all trials. One thousand three hundred ninety-nine (1,399) historical patients are available for this case study.



	Open label or blinded, phases 2 or 3	Multi-national	Began between 2004 and 2013. Ended btwn 2007 and 2016.	Previously treated locally advanced or metastatic non-small cell lung cancer	Overall survival measured	1399	Docetaxel

Eligibility criteria for the SCA are shown in Table 2. All patients in this set of 1,399 met these requirements at baseline. Historical patient level data, including assessments of eligibility criteria and other screening measurements from source historical trials were used to make these assessments.



a
1. Inclusion in a historical clinical trial accessible within this project
2. NSCLC stage III or IV at baseline
3. Received prior platinum-based chemotherapy
4. Men and women 18 years of age
5. Eastern Cooperative Oncology Group (ECOG) performance status of 2
6. Had measurable disease
7. Assigned to receive docetaxel as study treatment



ENDPOINTS AND COVARIATES

Because the historical data in this case study come from trials that have been conducted as part of clinical development programs and because methods for investigation of many indications in a regulatory setting are somewhat standardized by precedent, the populations, study design, data collection methods, and endpoints utilized in these trials are similar across trials. Overall survival is the endpoint of interest for this case study and was measured as a key outcome in all historical trials. Differences across studies in covariate definitions were present but have been reconciled as part of the data standardization process. Clinically important baseline covariates available across studies and to be used in the propensity score matching process are shown in Table 3.



<ol style="list-style-type: none">1. Age at baseline (continuous)2. Years from cancer diagnosis (continuous)3. Race (White vs Others)4. Sex (Female vs Male)5. Smoking (Current vs Former vs Never)6. Histology (Squamous vs Non-squamous)7. Stage (III vs IV)8. ECOG (0 vs 1 vs 2)9. Prior surgery (Yes/Maybe vs No)10. EGFR/KRAS mutation (Positive vs No/Unknown)

MATCHING METHODS AND EFFICACY ANALYSES

Propensity score matching is commonly used to analyze observational data to reduce bias due to confounding variables that are unbalanced between groups of interest (e.g., patients that received the treatment and those that did not). In the context of randomized clinical trials, the presumption is that the treatment groups will be generally balanced in terms of baseline covariates due to randomization and so differences between treatment and control can be reliably attributed to the treatment assignment. The intent of this project is to explore whether

Step 2: Create SCA by selecting historical patients to match control patients in the target trial using the estimated propensity scores. We will use greedy nearest-neighbor matching without replacement and a fixed 1-to-1 matching ratio, which aligns with the commonly used 1:1 randomization ratio in NSCLC historical trials. The control patients in the target trial will be randomly ordered. We will start from the first control patient in the target trial and will match the patient to a historical patient whose propensity score is closest to that of the

Where \hat{p}_c and \hat{p}_c denote the prevalence of covariate (or a category of covariate) for the

The control arm in Target Trial B included 542 patients. As shown in Table 5, most patients were white (54%), male (67%), and current or former smokers (34% and 39%, respectively). Prior surgery was reported in 1% of patients and the rate of known EGFR or KRAS mutation was 6%. Patients commonly had non-squamous type NSCLC (79%), ECOG scores of 0 or 1 (33% and 64%, respectively), and disease stage 4 (84%).

The pool of historical clinical trial subjects available for possible inclusion in the SCA included 857 patients. As shown in Table 5, these patients were similar to the Target Trial B control arm in terms of age, years since cancer diagnosis, gender, ECOG score, and EGFR/KRAS mutation. Differences between the historical pool and Target Trial B control were evident though in the rate of white patients (76% vs. 54%), the rate of current smokers (18% vs. 34%) and former smokers (59% vs. 39%), non-squamous type NSCLC (88% vs. 79%), disease stage 4 (78% vs. 84%), and prior surgery (28% vs. 1%).

Propensity Score Matching

As specified in the analysis plan, propensity score matching was utilized to attempt to select the appropriate patients from the historical pool for inclusion in the SCA so that the distribution of baseline characteristics would be well balanced between the SCA and the control from the target trial. This section details evidence that leads to the conclusion that indeed the matched groups are well balanced in terms of all observed baseline characteristics. The same conclusion is reached for both Target Trial A and Target Trial B.

The Cloud Plot in Figure 1A shows the distribution of propensity scores for the control arm of Target Trial A and the pool of historical patients available for inclusion in the SCA and the degree to which these distributions overlap. Green dots represent patients who are successfully matched with a patient in the opposite group with a similar propensity score. Red circles and blue x's represent patients for whom a match is not available. These are generally in the tails of the distribution of the target trial and visually we can see that there are no analogous patients available in this region of the historical pool. Patients in the target trial control arm who cannot be matched with a patient from the historical pool are excluded from further analysis.

Excluding unmatched target trial patients from further analysis is a common practice when utilizing matching methods. To many accustomed to analyzing clinical trials, this practice may seem alarming and in direct contradiction to the intent-to-treat principle normally relied upon in clinical trials to preserve the

The control arm in Target Trial A included 459 patients. Overlap in the distribution of propensity scores for the control arm of Target Trial A and the historical pool was significant but not complete. Three hundred sixty-six (80%) of the Target Trial A patients were successfully matched. The remaining 93 patients (20%) were not matched and were removed from further analysis. The baseline characteristics of the matched patients as well as the set of excluded unmatched patients from the target are described in Table 4. Baseline characteristics for the SCA and control arm in Target Trial A after matching now appear to be well balanced between groups, even for characteristics where differences were observed between the historical pool and target trial before matching. The most notable characteristic of the set of target patients who are not matched and are excluded from further analysis is the rate of patients with prior surgery. Attention should be given to the question of whether an analysis of patients with low rates of prior surgery can be extrapolated to the overall population, including patients with prior surgery.

The control arm in Target Trial B included 542 patients. Overlap in the distribution of propensity scores for the control arm of Target Trial B and the historical pool was significant but not complete. Four hundred seventeen (77%) of the target trial patients were successfully matched. The remaining 175 patients (23%) were not matched and were removed from further analysis. The baseline characteristics of the matched patients as well as the set of excluded unmatched patients from the target are described in Table 5. Baseline characteristics for the SCA and control arm in Target Trial B after matching now appear to be well balanced between groups, even for characteristics where differences were observed between the historical pool and target trial before matching. The most notable characteristics of the set of target patients who are not matched and are excluded from further analysis is the rate of white patients and rate of current smokers. Attention should be given to the question of whether an analysis of patients with differences in these characteristics be extrapolated to the overall population.


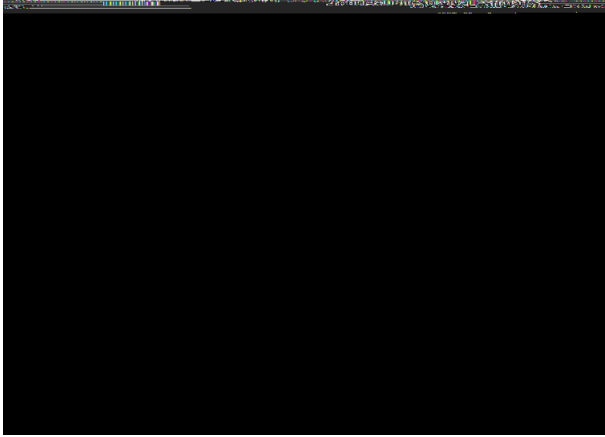
The balance between groups noted by numerical examination of the baseline characteristics can be explored further through graphical displays commonly used for the evaluation of the degree of success of the propensity score matching approach. Figures 2A and 2B provide a box plot and Q-Q plot respectively of the distribution of the propensity score before and after matching for Target Trial A. Figures 3A and 3B provide the same for Target Trial B. In all cases, significant gains in the comparability of the groups after matching are evident.

The distributions of the propensity score for the target trial and historical pool including all patients before matching are shown in the lower set of boxplots in Figures 2A and 3A. The analogous distributions after matching are shown in the upper region of these figures. There is considerable discordance between the target and historical pool before matching. In the case of Target Trial A, the median for the control is higher than that of the historical pool and the variability in scores is larger in the control than the historical pool. However, after matching, both the median and variability of the groups are very similar as evidenced by the similar placement of the median line and width of the 'box' in the boxplots for the groups. In the case of Target Trial B, the median for the control is higher than that of the historical pool and the variability in scores is smaller in the control than the historical pool. However, after matching, both

the median and variability of the groups are very similar.

Q-Q plots are scatterplots created by plotting the quantiles for one group of data against another. Quantiles are cut points that divide the range of a probability distribution into continuous intervals with equal probabilities. For example, a commonly used set of quantiles are 'quartiles', and they divide the distribution into quarters. The first quartile is defined as the middle number between the smallest number and the median of the data set. The second quartile is the median of the data. The third quartile is the middle value between the median and the highest value of the data set. Although this may seem a complex derivation, the Q-Q plot provides a straightforward interpretation for assessing similarity between groups. If both sets of quantiles come from the equal distributions, we will see the points forming a line that's roughly straight from the origin at 45° . The blue dots in the Q-Q plots in Figures 2B and 3B are a comparison of the quantiles in the historical pool to that of the Target Trial A control before matching. The red dots are the analogous comparison after matching. As evidenced by the red dots falling right along the 45° reference line and the blue dots not forming a straight line and being some distance from the reference line, we conclude that the degree of similarity in the distributions after matching is better than before matching. The mean (standard devia-

target relative to the historical pool was 1.16 with confidence interval that excludes 1 (95% CI 1.02, 1.32). This difference between groups is further supported by the log rank, Wilcoxon, and likelihood ratio tests comparing the difference in these curves ($p=0.03$, 0.07, and 0.04, respectively). After matching; however, there is significant overlap in the Kaplan-Meier curves for the target and SCA. The median survival was 8.8 months in the target versus 9.2 months in the SCA. The hazard ratio for the target relative to the SCA was 1.04 with confidence interval that includes 1 and indicates the plausible range for the HR is between 0.88 and 1.23, suggesting similarity of the SCA and target trial control arm in terms of OS. This similarity between groups is further supported by the log rank, Wilcoxon, and likelihood ratio tests comparing the difference in these curves ($p=0.65$, 0.97, and 0.66, respectively).

					
e					
75	19.8 (18.4, 22.1)	17.4 (14.9, 20.1)	75	17.0 (14.9, 19.6)	16.6 (14.3, 19.6)
50	10.4 (9.6, 11.1)	8.9 (8.2, 9.6)	50	9.2 (8.2, 10.7)	8.8 (7.9, 9.6)
25	5.1 (4.4, 5.6)	4.6 (4.1, 5.0)	25	4.4 (3.6, 5.3)	4.6 (4.1, 5.0)

Similar results are observed for Target Trial B (Figures 6A and 6B). Although the difference in OS between the control in Target Trial B and historical pool before matching is not clear, as it was with Target Trial A, there is still evidence that the similarity in OS is enhanced by the propensity score matching. After matching, the median survival was 9.9 years in the target versus 9.6 years in the SCA. The hazard ratio for the target relative to SCA was 1.01 with confidence interval that includes 1 and indicates the plausible range for the HR is between 0.85 and 1.19, suggesting similarity of the SCA and target control. This similarity between groups is further supported by the log rank, Wilcoxon, and likelihood ratio tests comparing the difference in these curves (p=0.91, 0.98, and 0.94, respectively).



CONCLUSIONS

With this case study in NSCLC, we have demonstrated that it is possible to produce “matched” cohorts of patients from historical clinical trials using propensity scores derived from observed baseline characteristics. In these examples, the OS for the SCA was observed to be very similar to that of the randomized control. Further research is needed to build a broader body of experience and to identify the circumstances under which this approach is feasible and appropriate. An assessment of whether a synthetic control can be used to replicate the treatment effect (difference between arms) of a randomized controlled trial, as well as an assessment of sensitivity to unknown or unobserved confounders is planned by this working group. Exploration of alternative matching methods, in addition to the 1-1 nearest neighbor caliper matching without replacement used in this case study, may make it possible to reduce the proportion of unmatched patients and resolve extrapolation concerns.

Overall, the results of this case study are promising and represent an important step toward understanding whether the use of SCA can inform the design of a randomized trial, potentially minimizing the number of patients required to be assigned to a control arm. This approach may mitigate many of the challenges faced when enrolling or maintaining a concurrent control arm is difficult due to rarity of the disease, or availability of the investigational agent outside the study.

REFERENCES

1. S. ... OL. ...
2. S. ... OL. ...
3. S. ... OL. ...
4. S. ... OL. ...
5. S. ... OL. ...
6. S. ... OL. ...
7. S. ... OL. ...
8. S. ... OL. ...
9. S. ... OL. ...
10. S. ... OL. ...
11. S. ... OL. ...
12. S. ... OL. ...