FRIENDS of CANCER RESEARCH

A Friends of Cancer Research White Paper

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

CONTRIBUTORS

Ruthie Davi
Medidata Solutions

Andrea Ferris
LUNGevity Foundation

Andrew Howland
Medidata Solutions

Dominic Labriola
Bristol-Myers Squibb

Michael LeBlanc
Fred Hutchinson Cancer Research Center, University of Washington

David Lee
Medidata Solutions

Antara Majumdar
Medidata Solutions

Pallavi Mishra-Kalyani
U.S. Food and Drug Administration

Zhenming Shun
Daiichi Sankyo

Rajeshwari Sridhara
U.S. Food and Drug Administration

Larry Strianese
Medidata Solutions

Elizabeth Stuart
Johns Hopkins University

Joohee Sul
U.S. Food and Drug Administration

Xiang Yin
Medidata Solutions

Antoine Yver
Daiichi Sankyo

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 "crossover" 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 difficult or not feasible. These indications are often studied using single arm trials in which all enrolled patients receive the investigational agent.

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 niraparib in patients with breast cancer and germline BRCA mutation carriers.[4] The sponsor of the trial announced, "A large number of patients in the chemotherapy control arm did not continue in the trial long enough to receive their first radiological scan, which is required to assess disease progression, resulting in an unusually high rate of censoring in the control arm." While the early discontinuation of these patients could be related to a toxicity of the drug, the sponsor conjectures that "this is likely associated with the desire of patients who carry germline BRCA mutations to be treated with a PARP inhibitor rather than chemotherapy and the increased availability of PARP inhibitors." The trial sponsor concluded that the study is, "unlikely to produce data that is interpretable."

An example of the consequences of treatment cross-over are seen in a trial in patients with gastrointestinal stromal tumor (GIST) described in the labeling of sunitinib.[5] This trial was a double-blind, randomized study comparing sunitinib malate to placebo and appears to have been designed and conducted in accordance with very high standards typical of pharmaceutical drug development. After an interim analysis demonstrated a large effect on progression free survival in favor of the sunitinib arm (HR 0.33, 95% CI (0.24, 0.47)), patients on the placebo arm were offered open label sunitinib malate; 99 of the 118 patients (84%) assigned to placebo elected to receive sunitinib malate. At the protocol specified final analysis, there was no difference observed in OS (median OS 72.7 weeks for the sunitinib malate arm and 64.9 weeks for the placebo arm, HR 0.88, 95% CI (0.68, 1.1)) by the original randomized arms. The absence of an effect at the final analysis time point is likely a result of the treatment "cross-over" in the placebo arm.

One approach to circumvent these challenges introduced by loss of equipoise is to consider the use of historical data to facilitate the conduct of clinical trials. Historical patient level data generally has been gathered before the experimental product or similar products are available on the market and while effects of other rescue therapy after progression cannot be ruled out, effects due to the pure treatment "cross-over" to the experimental therapy or very similar therapy generally are not present.

The potential use of historical clinical data in the context of randomized clinical trials was first discussed in the literature by Pocock (1976).[6] More recently, Lim et al. (2018) provided a comprehensive review of well-known frequentist and Bayesian methodologies for leveraging historical 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.

DATA SOURCES

Data from two sources will be utilized in this project. Project Data Sphere[a] has provided patient level data from the control arms of three large randomized trials in non-small cell lung cancer (NSCLC). Medidata Solutions has provided patient level data from multiple clinical trials in NSCLC conducted by the pharmaceutical industry for purposes of drug development and are available in the Medidata Enterprise Data Store (MEDS). All patients in these trials presented at baseline with previously treated advanced NSCLC and were assigned to receive docetaxel in the control arm.

MEDS is a collection of thousands of previous clinical trials conducted by the pharmaceutical industry for drug or medical product development with patient level data recorded through the Medidata electronic data capture system. Per the legal agreements with the sponsors of these historical clinical trials and Medidata, these data are available for use in deidentified (i.e., patients and original sponsor of the trial cannot be identified) and aggregated (i.e., every analysis must include data from two or more sponsors) form.

ANALYSIS OBJECTIVES

The scope of this work is to explore the potential applications of historical clinical trials data in randomized clinical trials, with the aim of minimizing the number of patients required to be assigned to the control arm and providing a better understanding of the effects of the experimental therapy independent of the effect of treatment "cross-over" assuming the historical clinical trials data has been generated at a time when the current experimental therapy was not available.

This project will explore whether a *synthetic control arm* (SCA) can mimic the results of a traditional randomized control. This will be investigated with a case study in previously treated advanced NSCLC as follows.

- First, one of the three historical trials provided by Project Data Sphere will be selected and designated as the 'Target Trial A'. This selection is limited to Project Data Sphere studies since MEDS studies may not be displayed individually. Legal restrictions governing MEDS data require analyses to be aggregate, that is including data from two or more sponsors.

- Next, a SCA will be constructed using patient-level data from all other available historical data in NSCLC. Patients in the SCA will be selected to match the control patients in the Target Trial A based on important baseline characteristics and prognostic factors and with a propensity score matching approach.

[a] Project Data Sphere is a platform where the research community can share historical patient level data from academic and industry phase 3 cancer clinical trials. The analyses in this case study are at least partially based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the website have contributed to, approved, or are in any way responsible for the contents of this work.

- 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.

Table 1: Features of Historical Data

Table 1: Features of Historical Data							
	Design	**Region**	**Start/End of Trial(s)**	**Baseline Characteristics**	**Endpoints**	**Number of Patients**	**Control regimen**
Historical data (from multiple trials)	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.

Table 2: SCA Patient Eligibility Criteria

Table 2: SCA Patient Eligibility Criteria
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.

Table 3: Clinically Important Baseline Covariates Utilized in Propensity Score Matching

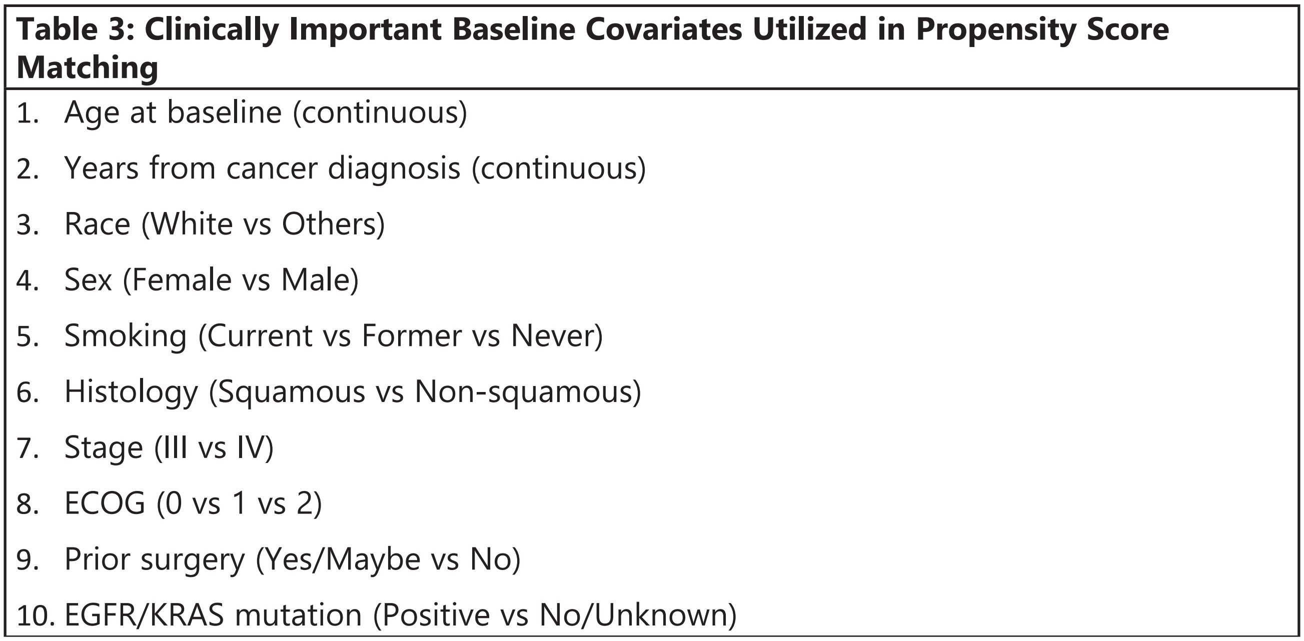


Table 3: Clinically Important Baseline Covariates Utilized in Propensity Score Matching
1. Age at baseline (continuous)
2. Years from cancer diagnosis (continuous)
3. Race (White vs Others)
4. Sex (Female vs Male)
5. Smoking (Current vs Former vs Never)
6. Histology (Squamous vs Non-squamous)
7. Stage (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 historical clinical trials data and matching procedures can stand-in for prospective patients and random assignment to treatment with standard of care in indications where there may be loss of equipoise.

Rubin and Thomas (1992) derive analytic expressions for the effect of matching using linear propensity scores with normally distributed covariates and find that substantial reductions in bias and variance are possible when these conditions are met.[12] Rubin and Thomas (1996) extend these results to covariates with a symmetrical ellipsoidal distribution, such as t-distributions.[13] Using Monte Carlo simulations, they confirm the accuracy of analytical approximations under normal and non-normal ellipsoidal distributions. Ho, et. al. (2007) further demonstrate that nonparametric matching using estimated propensity scores reduces the degree of model dependence, resulting in estimated treatment effects that exhibit greater robustness to researchers' parametric assumptions relative to analytic methods without data preprocessing by matching procedures.[14]

Using the guidelines proposed in Ho, et. al. (2007),[14] the following procedures will be used to carry out the propensity score matching:

Step 1: Estimate propensity scores. The propensity score is the probability of assignment of target trial control therapy conditional on the baseline characteristics (i.e., potential confounders) using logistic regression

$$p(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x})$$

where *T* denotes the control in the target trial (*T*=1) / historical control (*T*=0) and ***X*** is a vector representing the covariates to be included in the propensity score model (see Box 1 for an additional explanation of propensity score matching). The predictors included in the propensity score model are all available baseline characteristics described in Table 3. These baseline covariates will be utilized without further variable selection or trimming to obtain optimal balance between the matched subjects. Using a large set of covariates is recommended, even if some of the covariates are only related to self-selection and other covariates, and not necessarily to the outcome of interest.[15,16] Some researchers recommend using all available baseline covariates in the analysis if the sample size permits.[7]

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 control patient from the target trial and within a prespecified maximum distance (i.e., caliper). A caliper width equal to 0.25 of the pooled standard deviation of logit of the propensity score from the 2 groups, a widely utilized rule of thumb, will be used.[17] We will conduct matching without replacement, that is, the matched historical patients will be removed from consideration for further matching and next target trial control patients will be selected. This process will be repeated sequentially for all control patients in the target trial. The matched patients from the historical group are the components of SCA.

Step 3: Post-matching evaluation of covariate balance. The true propensity score should be a balancing score. We will examine whether the distribution of measured baseline covariates is similar between the matched target trial control arm and historical SCA subjects. Baseline demographic and disease characteristics will be summarized with descriptive statistics for the target trial control arm and SCA both before and after matching. Standardized difference in covariate means before matching and after matching will be computed and compared.

For a continuous covariate, the standardized difference is:

$$d = \frac{\overline{x_t} - \overline{x_c}}{\sqrt{(s_t^2 + s_c^2)/2}}$$

Where $\overline{x_t}$ and $\overline{x_c}$ denote the sample mean of the covariate for the target trial control and historical control groups, respectively; s_t^2 and s_c^2 denote the sample variance of the covariate for the target trial control and historical control groups, respectively.

For dichotomous (or categorical) variables, the standardized difference is defined as:

$$d = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\{\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_c)\}/2}}$$

Where $\hat{p}_t$ and $\hat{p}_c$ denote the prevalence of covariate (or a category of covariate) for the target trial control and historical control groups, respectively. For covariates with more than 2 categories, the standardized difference for each level of the categorical variable will be calculated.

The absolute standardized differences should generally be less than 0.25.[15] An absolute standardized difference of less than 0.10 has been taken to indicate a negligible difference in the mean or prevalence of a covariate between treatment groups.18 In addition, the matching process will be evaluated by examining the distribution of propensity scores, as well as individual baseline characteristics, including prognostic factors between the target trial control arm and SCA using graphical methods such as cloud plots, box plots, and quantile-quantile (Q-Q) plots. For continuous covariates, we will also summarize the mean and maximum deviation between the two empirical distributions in the Q-Q plots on the scale of the variables being measured.

Step 4: To explore whether OS observed in the control arm of the target trial is replicated by SCA, we will examine the similarity of OS between the SCA and target trial with the hazard ratio and associated 95% confidence interval for both before and after matching. Kaplan-Meier curves will be presented along with estimates of the median and other percentiles of survival times and 95% confidence intervals both before and after matching. Commonly used tests for differences in survival curves (i.e., log-rank test, Wilcoxon test, and likelihood ratio test) will also be presented both before and after matching.

Box 1. A non-technical description of propensity score matching and its possible effects

For illustration of the nature of propensity score matching, first consider a simplistic example where the number of important baseline characteristics is quite small, say age and ECOG score alone. Then for each patient in the target, we seek a patient from the historical pool with the same age and ECOG score. Assuming the amount of historical data is plentiful, this would lead to certain balance between the SCA and the target arm in terms of important baseline characteristics, age, and ECOG. However, the number of important baseline characteristics is rarely small and the scarcity of patients with exactly the same covariate pattern becomes problematic when the number of important covariates is larger. The propensity score can be thought of as a summarization of all the important baseline characteristics and their relationship to whether a patient is eligible to receive the therapy being studied. A key advantage of the propensity score approach is the reduction in dimension (i.e., many important baseline covariates) to a single value (i.e., propensity score). Achieving a match for most or all target patients on their propensity score is much more likely to be successful than requiring a direct match on many covariates at once. Matching on the propensity score likely will not provide exact balance between groups on all important baseline characteristics; rather, it will provide approximate balance for many baseline characteristics. Even with a propensity score approach there are some patients for whom an appropriate match will not be present in the available historical pool. In these cases, it is common practice to exclude these patients from the target matched set and proceed. 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 balance between treatment groups afforded by random treatment assignment. However, in this setting, randomization is not utilized and removing patients from the target improves balance between groups rather than threatens it. This practice of removing patients from the target could restrict the matched target patients to a set of patients with baseline characteristics that are not as wide ranging as is present in the overall disease population and so the appropriateness of extrapolating the analysis of this precise set and applying it to a more varied population should be considered.

RESULTS

PERFORMANCE OF SCA MATCHING PROCESS

Baseline Characteristics

The control arm in Target Trial A included 459 patients. As shown in Table 4, most patients were white (65%), male (63%), and current or former smokers (16% and 60%, respectively). Prior surgery was reported in 35% of patients and the rate of known EGFR or KRAS mutation was 7%. Patients commonly had non-squamous type NSCLC (78%), ECOG scores of 0 or 1 (24% and 67%, respectively), and disease stage 4 (87%).

The pool of historical clinical trial subjects available for possible inclusion in the SCA included 940 patients. As shown in Table 4, these patients were similar to the Target Trial A control arm in terms of age, years since cancer diagnosis, race, gender, ECOG score, and EGFR/KRAS mutation. Differences between the historical pool and Target Trial A control were evident though in the rate of current smokers (28% vs. 16%) and former smokers (46% vs. 60%), non-squamous disease (87% vs. 78%), disease stage 4 (77% vs. 87%), and prior surgery (9% vs. 35%).

Table 4. Baseline Characteristics by Arm Before and After Matching – Target Trial A

Baseline Characteristic	Before Matching		Matched		Unmatched
	Historical Pool (N=940)	Control in Target Trl A (N=459)	SCA (N=366)	Control in Target Trial A (N=366)	Control in Target Trial A (N=93)
Age at baseline, mean (std)	57.6 (10.5)	56.8 (11.0)	57.4 (11.0)	57.0 (10.7)	56.1 (12.1)
Years from cancer diagnosis, median (Q1, Q3)	0.7 (0.5, 1.0)	0.8 (0.5, 1.3)	0.7 (0.5, 1.0)	0.7 (0.5, 1.1)	1.3 (0.7, 1.9)
Race – White n (%)	645 (69%)	299 (65%)	239 (65%)	239 (65%)	60 (65%)
Sex – Female n (%)	316 (34%)	172 (37%)	128 (35%)	133 (36%)	39 (42%)
Smoking, n (%)					
Current	267 (28%)	74 (16%)	66 (18%)	71 (19%)	3 (3%)
Former	436 (46%)	276 (60%)	211 (58%)	208 (57%)	68 (73%)
Never	237 (25%)	109 (24%)	89 (24%)	87 (24%)	22 (24%)
Histology – Squamous, n (%)	120 (13%)	100 (22%)	65 (18%)	67 (18%)	33 (35%)
Stage – III, n (%)	213 (23%)	58 (13%)	54 (15%)	54 (15%)	4 (4%)
ECOG, n (%)					
0	334 (36%)	112 (24%)	85 (23%)	100 (27%)	12 (13%)
1	545 (58%)	306 (67%)	254 (69%)	233 (64%)	73 (78%)
2	61 (7%)	41 (9%)	27 (7%)	33 (9%)	8 (9%)
Prior surgery – Yes/Maybe, n (%)	83 (9%)	162 (35%)	66 (18%)	69 (19%)	93 (100%)
EGFR/KRAS mutation – Positive, n(%)	33 (4%)	33 (7%)	13 (4%)	16 (4%)	17 (18%)

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%).

Table 5. Baseline Characteristics by Arm Before and After Matching – Target Trial B

Baseline Characteristic	Before Matching	Matched		Unmatched	
	Historical Pool (N=857)	Control in Target Trial B (N=542)	SCA (N=417)	Control in Target Trial B (N=417)	Control in Target Trial B (N=125)
Age at baseline, mean (std)	58.0 (10.3)	56.2 (11.1)	57.1 (10.5)	57.0 (11.0)	53.6 (11.1)
Years from cancer diagnosis, median (Q1, Q3)	0.7 (0.5, 1.1)	0.7 (0.4, 1.0)	0.7 (0.5, 1.0)	0.7 (0.4, 1.0)	0.6 (0.4, 1.0)
Race – White n (%)	653 (76%)	291 (54%)	276 (66%)	270 (65%)	21 (17%)
Sex – Female n (%)	308 (36%)	180 (33%)	143 (34%)	140 (34%)	40 (32%)
Smoking, n (%) Current Former Never	 157 (18%) 503 (59%) 197 (23%)	 184 (34%) 209 (39%) 149 (28%)	 109 (26%) 199 (48%) 109 (26%)	 106 (25%) 196 (47%) 115 (28%)	 78 (62%) 13 (10%) 34 (27%)
Histology – Squamous, n (%)	104 (12%)	116 (21%)	65 (16%)	67 (16%)	49 (39%)
Stage – III, n (%)	186 (22%)	85 (16%)	78 (19%)	69 (17%)	16 (13%)
ECOG, n (%) 0 1 2	 266 (31%) 503 (59%) 88 (10%)	 180 (33%) 348 (64%) 14 (3%)	 142 (34%) 258 (62%) 17 (4%)	 134 (32%) 269 (65%) 14 (3%)	 46 (37%) 79 (63%) 0 (0%)
Prior surgery – Yes/ Maybe, n (%)	241 (28%)	4 (1%)	4 (1%)	4 (1%)	0 (0%)
EGFR/KRAS mutation – Positive, n (%)	33 (4%)	33 (6%)	14 (3%)	12 (3%)	21 17%)

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 balance between treatment groups afforded by random treatment assignment. However, in this setting, randomization is not utilized and removing patients from the target improves balance between groups rather than threatens it (in essence, prioritizing internal validity over external validity). This practice of removing patients from the target could restrict the matched patients to a set of patients with baseline characteristics that are not as wide ranging as is present in the target or overall disease setting and so the appropriateness of extrapolating the analysis of this precise set and applying it to a more varied population should be considered.

A similar display is shown for Target Trial B in Figure 1B.

Figure 1A. Cloud Plot for Target Trial A Illustrating Distribution of Propensity Scores

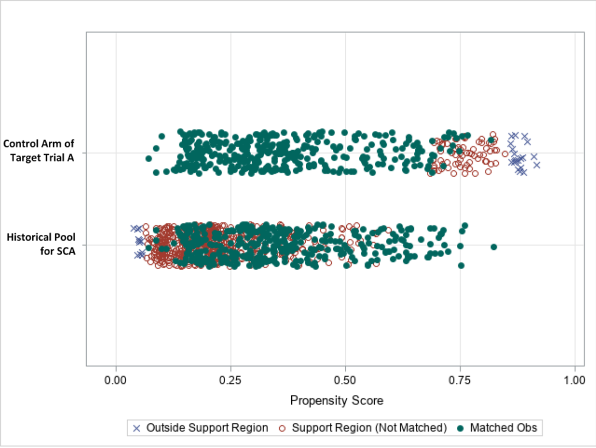
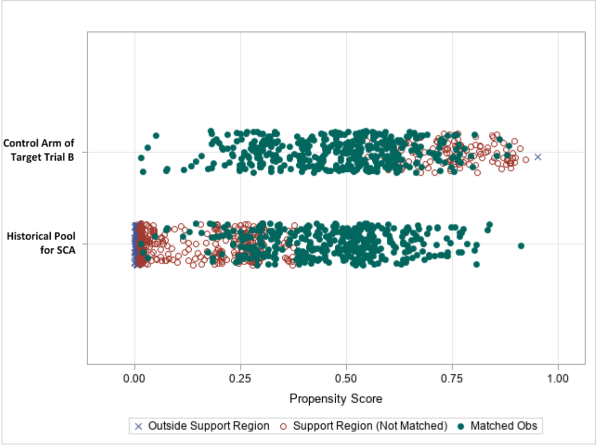


Figure 1B. Cloud Plot for Target Trial B Illustrating Distribution of Propensity Scores



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 45^0. 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 45^0 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-

Figure 2A. Distribution of Propensity Scores Before and After Matching – Target Trial A

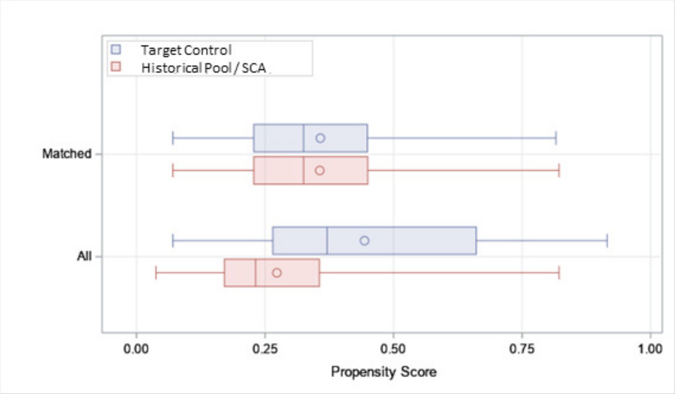


Figure 2B. Q-Q of Propensity Scores Before and After Matching – Target Trial A

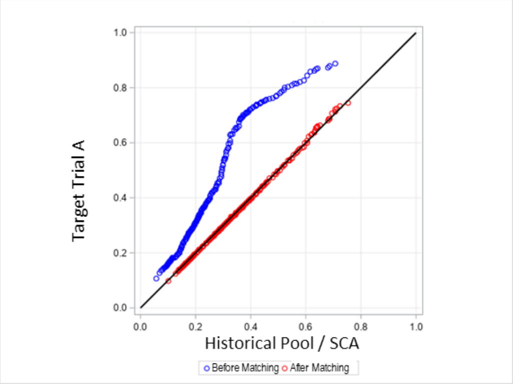


Figure 3A. Distribution of Propensity Scores Before and After Matching – Target Trial B

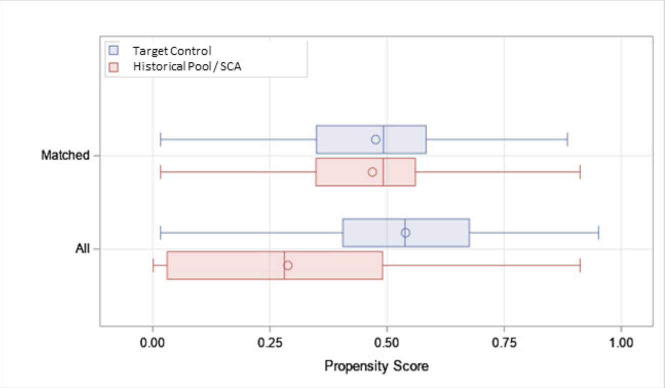
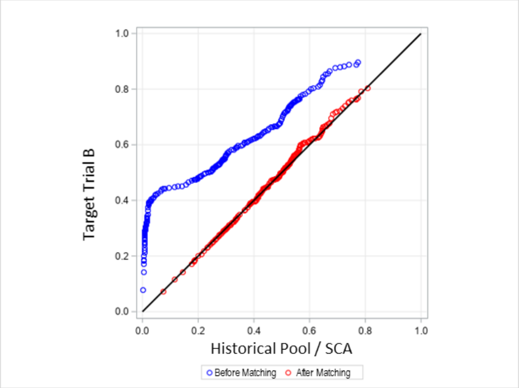


Figure 3B. Q-Q of Propensity Scores Before and After Matching – Target Trial B



tion) of deviation in propensity score between the two groups in the Q-Q plots changed from 0.121 (0.065) before matching to 0.001 (0.003) after matching. A similar result holds for Target Trial B.

Assessment of balance in terms of individual baseline covariates yields observations consistent with the conclusions afforded above by examination of the propensity scores. Figure 4A illustrates the standardized difference between the target trial and historical pool (before matching)/SCA (after matching) for each important baseline characteristic for Target Trial A. Figure 4B provides the same for Target Trial B. In all cases, reductions in the absolute standardized difference between groups for each variable are observed and the absolute standardized differences after matching are well below 0.10, the pre-specified threshold for designating a negligible difference in the mean or prevalence of a covariate between groups, for all but one instance.

Figure 4A. Plot of Standardized Difference of Important Baseline Covariates Before and After Matching – Target Trial A

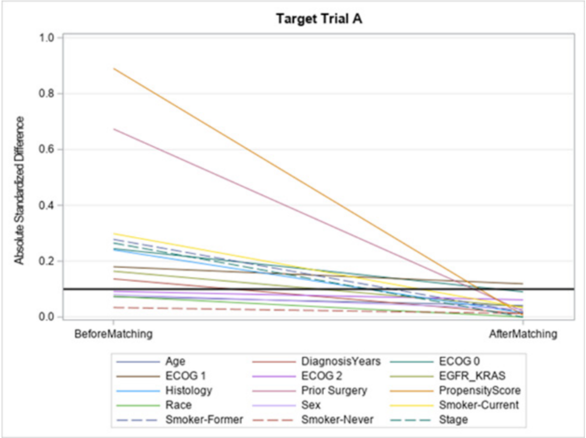
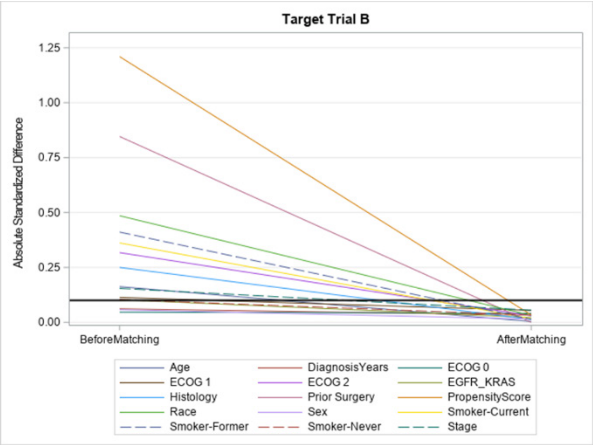


Figure 4B. Plot of Standardized Difference of Important Baseline Covariates Before and After Matching – Target Trial B

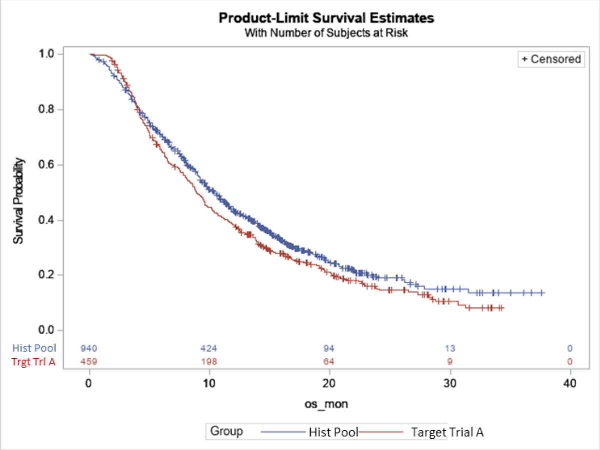


ASSESSMENT OF OVERALL SURVIVAL REPLICATION WITH SCA

In previous sections, we have demonstrated that the propensity score matching successfully balanced the distribution of baseline characteristics between the SCA and the control from the target trial. The main objective of this case study though is to explore whether the outcome of the randomized control arm from the target trial can be replicated using the SCA. This section details evidence that leads to the conclusion that indeed the OS for the SCA is very similar to that of the target trial. The same conclusion is reached for both Target Trial A and Target Trial B.

Figures 5A and 5B provide a comparison of the OS between the control arm of Target Trial A and the historical pool (before matching)/SCA (after matching), respectively. Before matching, there is a suggestion that the curves differ, as evidenced by little overlap of the Kaplan-Meier curves and space present between the curves suggesting that the OS for the Target Trial A is worse than that of the historical pool. The median survival was 8.9 months in the target versus 10.4 in the historical pool. The hazard ratio for the 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).

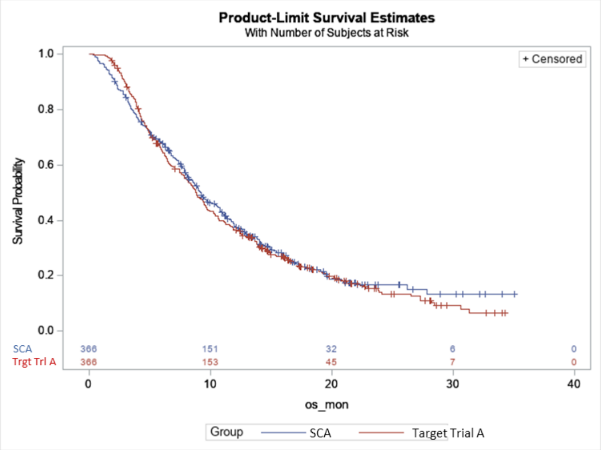
Figure 5A. Comparison of Overall Survival in Control Arm of Target Trial A versus Historical Pool (Before Matching)



Quartile	Hist Pool	Target
75	19.8 (18.4, 22.1)	17.4 (14.9, 20.1)
50	10.4 (9.6, 11.1)	8.9 (8.2, 9.6)
25	5.1 (4.4, 5.6)	4.6 (4.1, 5.0)

Log-Rank p-value: 0.03
Wilcoxon p-value: 0.07
-2Log(LR) p-value: 0.04
HR (Target vs Hist Pool), 95% CI: 1.16 (1.02, 1.32)

Figure 5B. Comparison of Overall Survival in Control Arm of Target Trial A versus SCA (After Matching)

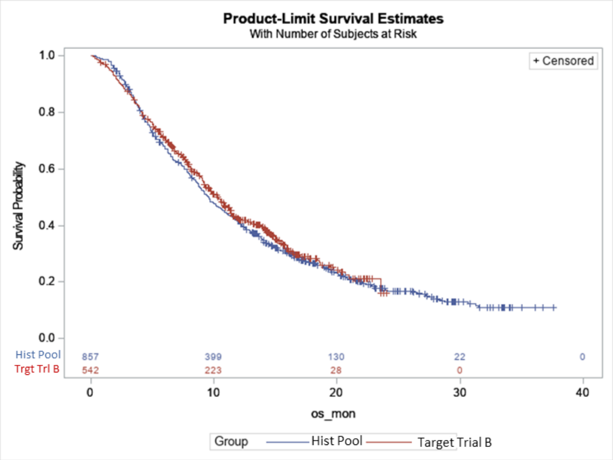


Quartile	SCA	Target
75	17.0 (14.9, 19.6)	16.6 (14.3, 19.6)
50	9.2 (8.2, 10.7)	8.8 (7.9, 9.6)
25	4.4 (3.6, 5.3)	4.6 (4.1, 5.0)

Log-Rank p-value: 0.65
Wilcoxon p-value: 0.97
-2Log(LR) p-value: 0.66
HR (Target vs SCA), 95% CI: 1.04 (0.88, 1.23)

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).

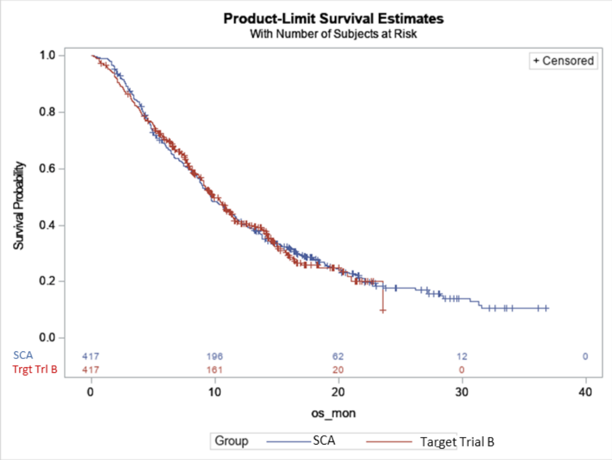
Figure 6A. Comparison of Overall Survival in Control Arm of Target Trial B versus Historical Pool (Before Matching)



Quartile	Hist Pool	Target
75	19.1 (16.9, 20.5)	19.7 (16.5, NE)
50	9.5 (8.9, 10.3)	10.4 (9.3, 11.3)
25	4.8 (4.2, 5.1)	5.1 (4.3, 5.9)

Log-Rank p-value: 0.35
Wilcoxon p-value: 0.37
-2Log(LR) p-value: 0.39
HR (Target vs Hist Pool), 95% CI: 0.94 (0.82, 1.07)

Figure 6B. Comparison of Overall Survival in Control Arm of Target Trial B versus SCA (After Matching)



Quartile	SCA	Target
75	19.6 (17.0, 22.1)	18.4 (15.8, NE)
50	9.6 (8.8, 11.0)	9.9 (9.0, 10.9)
25	4.8 (4.3, 5.3)	5.0 (4.1, 5.9)

Log-Rank p-value: 0.91
Wilcoxon p-value: 0.98
-2Log(LR) p-value: 0.94
HR (Target vs SCA), 95% CI: 1.01 (0.85, 1.19)

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] FDA Guidance for Industry Expedited Programs for Serious Conditions – Drugs and Biologics. 2014. Available online: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm358301.pdf

[2] Food and Drug Safety Innovation Act. 2012. Available online: https://www.congress.gov/bill/112th-congress/senate-bill/3187.

[3] Sridhara R., He K., Nie L., Shen Y, Tang S. 2015. Current Statistical Challenges in Oncology Clinical Trials in the Era of Targeted Therapy. *Statistics in Biopharmaceutical Research*. 7:348-356.

[4] Tesaro press release. 2017. Available online: http://ir.tesarobio.com/news-releases/news-release-details/tesaro-announces-expanded-development-program-niraparib-focused

[5] Sunitinib malate capsule prescribing information. 2006. Available online: http://labeling.pfizer.com/ShowLabeling.aspx?id=607

[6] Pocock, S. 1976. The combination of randomized and historical controls in clinical trials. J Chron Dis. 29:175-188.

[7] Lim J., Walley R., Yuan J., Liu J., Dabral A., Best N., Grieve A., Hampson L., Wolfram J., Woodward P., Yong F., Zhang X., Bowen E. 2018. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: review of methods and opportunities. DIA Therapeutic Innovation and Regulatory Science.

[8] Gökbuget, N, Kelsh, M, Chia, V, Advani, A, Bassan, R, Dombret, H, Doubek, M, Fielding, AK, Giebel, S, Haddad, V, Hoelzer, D, Holland, C, Ifrah, N, Katz, A, Maniar, T, Martinelli, G, Morgades, M, O'Brien, S, Ribera, JM, Rowe, JM, Stein, A, Topp, M, Wadleigh, M, Kantarjian, H. 2016. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood cancer journal, 6(9), e473.

[9] Neuenschwander, B., Capkun-Niggli, G., Branson, M., Spiegelhalter, D. 2010. Summarizing historical information on controls in clinical trials. *Clinical Trials*. 7:5-18.

[10] Rosmalen J., Dejardin D., van Norden Y., Lowenberg B., Lesaffre E. 2017. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*.

[11] Hobbs, B, Carlin, B., Sargent, D. 2013. Adaptive adjustment of the randomization ratio using historical control data. *Clin Trials*. 10:430-440.

[12] Rubin, D., Thomas, N. 1992. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*. 79:797–809.

[13] Rubin, D., Thomas, N. 1996. Matching using estimated propensity scores, relating theory to practice. *Biometrics*. 52:249–64.

[14] Ho, DE, Imai, K, King, G, Stuart, E. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*. 15(3):199–236.

[15] Stuart, E., Rubin, D. 2007. Best Practices in Quasi-Experimental Designs: Matching methods for causal inference. Chapter 11 in *Best Practices in Quantitative Social Science*. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications. 11:155-176.

[16] Harris, H., Horst, S. 2016. A Brief Guide to Decisions at Each Step of the Propensity Score Matching Process. *Practical Assessment, Research & Evaluation*. 21(4). Available online: http://pareonline.net/getvn.asp?v=21&n=4

[17] Rosenbaum, P., Rubin, D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 39:33–38.

[18] Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., McNeil, B. 2001. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*. 54:387–398.

ADDITIONAL REFERENCES

Borghaei H., Paz-Ares L., Horn L., Spigel D., Steins M., Ready N., Chow L., Vokes E., Felip E., Holgado E., Bariesi F., Kohlhaufl M., Arrieta O., Burgio M., Fayette J., Lena H., Poddubskaya E., Gerber D., Gettinger S., Rudin C., Rizvi N., Crino L., Blumenschein G., Antonia S., Dorange C., Harbison C., Finckenstein F., Brahmer J. 2015. Nivolumab versus Docetaxel in Advanced Non-Squamous Non-small Cell Lung Cancer. *N Engl J Med*. 373(17): 1627-1639.

Lin D, Psaty B, Kronmal R. 1998. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 54:948–963.

Liu, W., Kuramoto, S., Stuart, E.A. 2013. An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research. *Prevention Science* 14(6): 570-580.