A Friends of Cancer Research Whitepaper

CHARACTERIZING THE USE OF EXTERNAL CONTROLS FOR AUGMENTING RANDOMIZED CONTROL ARMS AND CONFIRMING BENEFIT

OBJECTIVE

Friends of Cancer Research (*Friends*) convened a working group to characterize methodological processes and to discuss the implementation and opportunities for formal regulatory use of external controls. This whitepaper describes several approaches to constructing an external control and also considers the use of hybrid designs that supplement or augment the control group in the randomized control trials (RCT) with data from an external population. This whitepaper further discusses statistical methodology to help address potential biases and improve the usefulness of the data as well as other adjustment methods that rely on patient summary data. In addition, we describe several scenarios where the use of external controls may be advantageous and practices that can help guide the implementation within a clinical study. A use case was prepared that characterizes the construction of an external control using clinical trial data in multiple myeloma to compare the treatment effect with a randomized control versus an external control and assesses the potential impact of unmeasured confounders.

INTRODUCTION

In drug development, RCT are the gold standard for evaluating the safety and efficacy of medical treatments. However, oncology drug development increasingly relies on the use of single-arm clinical trials especially in certain settings where there are ethical or feasibility challenges with deploying a concurrent control arm. While single-arm trials alone may yield important safety and efficacy signals and can be relied on for regulatory decision making in certain clinical and regulatory contexts, external controls (sometimes referred to as synthetic controls) may provide additional context and supplementary evidence. Expanding the use of external controls to other difficult-to-study indications may reduce patient burden where research may be slowed or unin-

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.

WORKING GROUP

Francois Beckers
EMD Serono

Adrian Cassidy
Novartis

Ruthie Davi
Acorn AI, A Medidata Company

Alaknanda Joshi
Novartis

Laura Koontz
Flatiron Health

Michael LeBlanc
University of Washington

Michael Menefee
FDA

Reena Nadpara
Novartis

Erik Pulkstenis
AbbVie

Dirk Reitsma
PPD

Meghna Samant
Flatiron Health

Xiang Yin
Acorn AI, A Medidata Company

William Capra
Genentech

Frank Cihon
Bayer U.S.

Ritesh Jain
EMD Serono

Bindu Kanapuru
FDA

Dominic Labriola
Bristol-Myers Squibb

Nicole Mahoney
Flatiron Health

Pallavi Mishra-Kalyani
FDA

Jim Omel
Cancer Research Advocate

Jeremy Rassen
Aetion

Gary Rosner
Johns Hopkins University

Chenguang Wang
Johns Hopkins University

We are grateful for the data, expertise, and/or review each working group member has provided.

terpretable due to the use of a concurrent randomized control. The latter may be the case with some confirmatory trials of medical products made available through the accelerated approval pathway where the control arm may be compromised by early discontinuation or treatment crossover to the investigational therapy made available by an accelerated approval.[1]

Study designs that deviate from the traditional RCT, such as single arm or externally controlled trials, are considered in guidance for regulatory approval when their use is justified.[2,3] These types of trial designs can be warranted for scenarios where randomization may be difficult or infeasible due to the rarity of the disease, scarcity of patients, scientific concerns about treatment switching/crossover, or ethical considerations. For example, challenges introduced by treatment crossover can be observed in the double-blind, randomized study comparing sunitinib to placebo. An interim analysis demonstrated a large effect on progression free survival (PFS) and patients on the placebo arm were offered sunitinib. During the final analysis, the treatment effect size for overall survival (OS) was diminished, which was likely due to treatment crossover.[4]

As described in regulation, external controls have generally been allowed only in, "special circumstances;" for example, "diseases with high and predictable mortality" and when, "effect of the drug is self-evident." This restricted use is due, in part, to the perceived inability for external controls to be "well assessed with respect to pertinent variables as can concurrent control populations," as stated in FDA guidance and regulation.[2] However, our ability to electronically store and manage continually aggregating real-world data (RWD) from electronic medical records, claims data, prior clinical trials data, and other sources is opening opportunities that were not possible before. Moreover, higher quality external controls are more available today than in the past due to the availability of patient level data and statistical methods for achieving balance in baseline characteristics between the clinical trial and external controls.

There are several examples of the use of external controls for regulatory applications evaluating effectiveness, but most have been used for informal, rather than direct, statistical comparison. The use of external controls is most common in orphan disease settings where it can be difficult to accrue patients, especially for a randomized clinical trial. There are some notable examples of the use of external controls in oncology drug development:

1. Blinatumomab (Blincyto)[5,6]: Historical clinical trial site patient data and propensity score methods were used to construct complete remission and OS reference rates for comparison to the single-arm study of blinatumomab for Ph-negative B-precursor cell relapsed/refractory acute lymphoblastic leukemia. As per the sponsor, the historical clinical trial data of 1,139 patients from the EU and the US was used to support the FDA's breakthrough therapy designation and accelerated approval in December 2014.

2. Bavencio (Avelumab)[7]: In 2017, Bavencio received accelerated approval for Merkel Cell Carcinoma on the basis of an 88-patient single arm Phase II trial. Real-world evidence (RWE),

contributed by external data from a registry, was used as supportive evidence, but the regulatory approval was based primarily on data from the Phase II trial.

Additional efforts and case studies have helped inform methodology for constructing external controls and describe limitations and opportunities with these types of analyses. For instance, a case study in non-small cell lung cancer demonstrated that it is possible to produce a "matched" cohort to a randomized control arm.[1] More experience and understanding of the circumstances where external data may serve as an external control are needed to characterize the full utility and potential of external controls. This paper explores the design and analyses of studies leveraging an external control built from external historical or contemporaneous patient-level data selected to be similar in important prognostic (or clinical) characteristics to patients treated with the experimental regimen.

METHODOLOGICAL APPROACHES AND CONSIDERATIONS FOR CONSTRUCTING AN EXTERNAL CONTROL

Multiple sources of data exist to populate an external control cohort. These data sources include clinical trial data, published clinical data, and real-world data derived from electronic health records (EHRs) and other sources. As external data sources are considered, the advantages and limitations associated with the various sources and whether patient-level data is available will need to be evaluated when designing a clinical study. Methods discussed in this whitepaper focus on the use of individual patient-level data rather than aggregate-level data.

COHORT SELECTION AND ADJUSTMENT METHODS

Careful cohort selection is critical to developing a robust external control to control for potential biases that can be encountered in clinical research (Table 1). Lack of randomization can result in several potential biases. In particular, selection bias and confounding bias need to be considered when selecting patients in the external control cohort. Selection bias occurs when the observed patients are not representative of the broader population of interest and thereby can challenge the external validity of the results.

Some examples include selecting patients from a specific geographic region or with certain clinical characteristics such as age, comorbidities, prognostic indices or prior/concurrent therapies that are not representative of the clinical trial population. It is also important to select a cohort in which we can account for confounding that may arise due to lack of randomization. Confounding bias occurs when there is an imbalance in the distribution of key baseline characteristics that are associated with both the outcome and exposure to treatment. Such characteristics are called confounders and are typically characterized as "measured" and "unmeasured." The presence of confounders is particularly important to consider when using controls from RWD sources, like electronic medical record data, since certain patient characteristics that are likely to impact outcomes

Table 1: Select biases encountered in clinical research

Bias	Explanation	Methods to Reduce Bias
Confounding Bias	Selection of experimental and control patients completed in such a way that the patient characteristics are systematically different across treatment groups, perhaps with those with better prognoses preferentially receiving one therapy over another.	Randomization
Selection Bias	Occurs when the observed patients are not representative of the broader population of interest and thereby can challenge the external validity of the results.	Randomization; Improved sampling
Performance Bias	Follow-up differs by treatment. Differential care according to treatment beyond the treatment itself. Systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest.	Standardization of treatment and follow-up plans for all patients
Detection Bias	Outcome assessment differs by treatment leading to systematic differences in outcome determination.	Masking
Attrition Bias	Systematic differences between groups in withdrawals from a study or treatment exist.	Analysis by intention to treat
Time-trend Bias	Prognostic characteristics of available patients change during the course of the trial especially for trials with long recruitment periods.	Maintain randomization

(e.g., age, access to clinical care, socioeconomic status) are also likely to influence treatment exposure. For example, we may see a distribution of patients in the clinical trial skewed more toward younger and fitter patients, while the population of real-world patients may comprise a much broader patient population including the elderly and patients with more comorbidities. Even within the real-world population, confounding by indication may occur as certain types of patients may be more likely to be prescribed certain treatments because of their characteristics. Finally, differences in the characteristics of sites participating in clinical trials (e.g., site effect, site volume, clinical care protocols, access to multimodality care, academic vs. community centers, etc.) can also confound outcomes and bias results. The myriad of considerations discussed above make it challenging to isolate the treatment effect in externally controlled studies, and analytical approaches need to be considered to mitigate such biases, where possible.

A fundamental step when considering external controls is thoughtful and rigorous planning in the design phase. This involves careful identification of key baseline prognostics and confounding factors through tools such as directed acyclic graphs (DAGs), and accordingly pre-specifying the key inclusion/exclusion criteria for external cohort selection.[8,9] The identification and prioritization of key criteria for selection is critical because not all criteria typically applied in clinical trials may be available or possible to collect using completed historical trials or retrospective real-world datasets. It is thus important to align, as much as possible, the criteria between the clinical trial and the external control. Once a prioritized list of criteria is identified, all efforts must be made to collect the relevant data to a high degree of completeness and accuracy, noting that in some cases prospective approaches may be needed to intentionally collect the required data element. Sponsors should clearly and transparently document these efforts (e.g., through patient attrition diagrams, data deficiencies), hypothesize the impact of missing data elements on overall outcomes, and have plans to address this impact.

Despite careful selection of the external cohort in alignment with the trial eligibility criteria, imbalances in key confounding factors may still exist that need to be further mitigated through thoughtful consideration and pre-specification of appropriate statistical methodologies. There is no one adjustment method universally preferred over others; Table 2 below outlines methods that are commonly used to drive greater balance in patient distributions among measured covariates. This is not intended to be a comprehensive list and variations of these methods are common. The choice of the statistical method in a particular context ultimately depends on a variety of factors including available external cohort size, number of key variables to consider, tolerance for complexity, etc. Propensity score methods are especially important in the creation of external controls and are further discussed in the next section.[10,11] A propensity score (PS) is the probability of being treated with one drug versus another, based on the measured factors known about the patient. In a single number, the PS captures much of the nuance about treatment choice and allows us to control for a substantial amount of confounding using a single variable (a detailed description of propensity scores, propensity score matching, and propensity score weighting is in Appendix 1).

For hybrid designs (randomized controls augmented with external controls), some statistical methods determine the degree to which external information enters the analysis of a clinical trial in a data-dependent way. If the external data, particularly outcomes or covariate-adjusted outcomes, seem consistent with the outcomes of the current trial's control group, the algorithms will give relatively more weight to the external data than when there appear to be heterogeneities. Some example methods are commensurate priors, power priors, and meta-analytic predictive priors and are commonly used in trial designs with hybrid controls.[12–16]

Table 2. Commonly used statistical methods to balance baseline factors[8]

Method	Description	Key Benefits	Key Limitations
Exact matching	Trial patients are matched 1:1 or 1:many to external controls on a set of important baseline characteristics	- Simple and intuitive	- Need large external cohort sample size to find matched controls for all patients and some trial patients may remain unmatched - Often very limited number of baseline factors can be used for matching - Inefficient use of data from unmatched trial and control patients

Table 2 continues on Page 8

Propensity score matching[17]	Trial patients are matched with fixed or various ratios to external controls on propensity scores (probability a patient is in the trial cohort vs external control conditional on baseline covariates). Since scores are continuous, calipers/ intervals are commonly used	- Can be simple and intuitive - Large number of baseline factors can be captured and balanced through propensity score - Matching is based on one single score rather than on the full multivariate set of baseline factors - Calipers provide flexibility to relax matching requirements and enable more efficient use of external controls	- Some matching algorithms need a large external cohort sample size to find matched controls for all patients and some trial patients may remain unmatched - Inefficient use of data from unmatched trials and control patients when insufficient number of matches are found - Requires correct specification of the propensity score model - Pre-specifying width of the caliper may be challenging depending on the context and sample size
Propensity score weighting – Inverse probability of treatment weights (IPTW)[18]	Propensity scores are typically used to weight patients in the trial and external cohorts in a way that achieves balance in the baseline characteristics	- Efficient use of all trial and external control patients	- Distorts the original distribution of the trial patients since they are also weighted along with the external controls, thereby changing the target population for which treatment efficacy is being assessed - Requires correct specification of the propensity score model - May require more complex analytic decisions, e.g. trimming, in case of extreme propensity scores
Propensity score weighting – Weighting by odds[19]	Patients in the trial arm are given a weight of 1 (i.e. all information is included) while odds of propensity scores are used to weight patients in the external cohorts	- Distribution of trial patients remains intact and full information from all trial patients is utilized - Efficient use of all trial and external control patients	- Requires correct specification of the propensity score model

Outcome regression models	Association between treatment and outcome is modeled adjusting for baseline covariates Doubly robust regression models are sometimes considered whereby a function of propensity scores is used as weights in the model, making it more robust to model misspecification[20]	- Generally easy to understand as familiarity with regression models is high among research community - Efficient use of all trial and external control patients - Doubly robust models provide insurance against model misspecification (i.e the results are unbiased so long as either the outcome model or propensity model are correctly specified)	- The outcome model must be correctly specified if no propensity score weighting is used - Either the outcome or propensity score model must be correctly specified if a weighted model is used - No separation of design for balancing baseline factors from the outcome analysis

NOTES ON EFFECT ESTIMATES

Matching and weighting are on their surface very similar, but there is a subtle difference in the values one estimates from each approach. The matching approach will estimate the average treatment effect in the treated, which can be more tangibly thought of as the treatment effect among those patients who were reasonable candidates for either treatment choice: this is a notion of clinical equipoise. On the other hand, IPTW weighting estimates the average treatment effect in the entire population and considers what would happen if all patients were moved from control to treatment.[21]

For the questions considered here, we would expect there to be little difference between the average treatment effect in the treated and the average treatment effect in the entire population, as all patients would have met stringent inclusion/exclusion criteria, and thus would in all likelihood be eligible for either treatment pathway. As such, considerations of the feasibility of matching should outweigh considerations of the estimated treatment effect.

CONTROL OF CONFOUNDING BIAS

Confounding bias results from not accounting for factors that are associated with both the treatment choice and the outcome, independent of any effect via treatment.[22] In studies of medications, some of the strongest confounding comes from confounding by indication, in which patients’ level of illness drives treatment choice (sicker patients may get “stronger” treatments) as well as outcome (sicker patients may experience worse outcomes).[23] This can be particularly difficult to address, though design approaches such as fit-for-purpose data, RCT-like study design,[24] new user cohorts,[25] and principled process,[26] as well as analytic approaches, such as multivariable regression, propensity scores, and high-dimensional propensity scores,[27] can eliminate the effect of measured (or measureable) confounding.

However, unmeasured confounding may yet remain, and control of a factor that is ultimately unmeasurable is a substantial challenge. Approaches that can control for this unmeasurable confounding, such as instrumental variable analysis, are not frequently seen in the medical literature but can be effective.[28] Separately, high dimensional propensity scores can "uncover" previously-unmeasured confounders and reduce bias. There are powerful techniques that allow us to assess unmeasured confounding.

Causal diagrams can help elucidate potential sources of bias.[8] More quantitatively, sensitivity analyses allow us to ask ourselves questions like, "If we had an unmeasured confounder (or group of confounders) of strength x, how much would our results be affected?" and "How powerful would an unmeasured confounder (or group of confounders) have to be to meaningfully alter our interpretation of the situation we've observed?"[29] E-values and tipping point analyses may be potential solutions for assessing the impact of unmeasured confounders on the overall treatment effect. The use case included in Appendix 2 illustrates a tipping point analysis, which shows the strength a confounder would need in order to change the statistical significance or numerical direction of the original estimates of the treatment effect. By addressing these questions, we can better reason about the robustness of our results to issues like unmeasured confounding; presenting such results can strengthen readers' and reviewers' confidence in the evidence.

While the potential for unmeasured confounding is a key issue in any non-randomized study, in the single-arm study with external controls scenario, a more important issue is whether the experience of the controls truly represents the counterfactual experience of the treated patients. That is, would standard of care patients have been treated with the single-arm treatment had the single-arm treatment been available to them, and vice-versa? To ensure this, we implement strong inclusion/exclusion criteria, draw controls from populations similar to that of treated patients, and apply other key design approaches.

OUTCOMES AND ENDPOINT CONSIDERATIONS

Even when a set of patients comparable to the experimentally treated patients can be identified for the external control, to create valid inference regarding the treatment effect, one must also ensure comparable ascertainment and measurement of the outcomes of interest for the external control and experimentally treated patients. Differences between arms in endpoint collection methods and endpoint definitions can bias the treatment effect estimates. But the way the endpoints are captured for the external control patients generally is not within the researcher's control and may not be completely consistent with the experimentally treated patients. Additionally, assessment of response or progression free survival endpoints may be performed locally or centrally, and those assessment differences should be a consideration when external control data are utilized. In addition, some response criteria, especially in hematologic malignancies, are complex, which may result in differences in implementation from study to study.

These inconsistencies may be more or less challenging based on the source of data. For example, external controls built from historical clinical trial data enjoy the benefit of similar collection and definition of efficacy and safety endpoints while endpoints representing similar clinical concepts may be captured differently in external controls built from real-world data. Endpoints that are objective may be less affected by different measurement techniques, timing, or settings and may be preferred when using an external control. For instance, progression free survival may have more complex considerations than an endpoint related to tumor shrinkage when considering options for external controls.[30,31]

OPERATIONAL CONSIDERATIONS

Situations that may support the use of an external control include those where randomization may not be feasible due to ethical, scientific, or operational considerations (Table 3).[32] For example, for certain orphan diseases, rare diseases, or rare biomarker-defined cohorts, it may not be possible to enroll a sufficient number of patients to have a concurrent control, meriting consideration of external data sources. Hybrid designs could also be considered to reduce the number of patients assigned to the control by augmenting with external data. In some cases, it may be unethical to randomize patients to the control arm. All patients could receive a promising experimental drug in an externally controlled trial, making this type of study more attractive to patients, and lessening the risk of trials closing due to poor accrual. Externally controlled studies may also be valuable when treatment crossover from a deployed control arm to the experimental arm of an RCT, or to off-study treatments including new treatments approved during the course of the study, compromises the interpretability of treatment effects. In some respects, externally controlled data may be preferable to single-arm studies that are often employed to address the limitations noted in the situations above and given that some time to event endpoints may be difficult to interpret in a single-arm study.

Table 3. Select scenarios that may benefit from the use of an external control

Scenario	Challenge	Role of External Controls
Uncontrolled studies (e.g., single-arm trial, expanded access)	Outcomes of the experimentally treated patients are difficult to interpret without an understanding of expected outcomes for patients who did not receive the experimental treatment	-To provide context needed to interpret outcomes of experimentally treated patients by comparing to a group of patients who did not receive experimental treatment
Studies of orphan diseases, rare diseases or rare biomarker-defined cohorts	Recruitment of patients is very difficult due to rarity of defined disease so that a concurrent control may not be possible and resulting single arm data is difficult to interpret	-To improve patient recruitment and allow a design where all patients can be treated with the experimental product -To provide context needed to interpret outcomes of experimentally treated patients by comparing to a group of patients who did not receive experimental treatment -To function as a natural history cohort to describe patient characteristics and outcomes in these settings
Post-marketing confirmatory study following accelerated approval	Recruitment and/or retention to a randomized controlled trial when the experimental product is available on the market is very difficult and sometimes impossible	-To augment or replace the randomized control of the confirmatory trial so that an external control may be constructed and confirmatory studies could be completed. An additional benefit would be that patients enrolling in the trial have a higher probability or even assurance of receiving the experimental therapy
High rate of treatment cross-over	Patients assigned to the control arm of a randomized controlled trial may use the experimental product or a similar product in the same class when the experimental product or a similar product in the same class is available on the market thereby diluting the ability of the study to demonstrate a difference between arms	-To augment or replace a randomized control with patients who did not receive the experimental product (since perhaps they were studied at a time when the experimental product was not available) so that the difference between arms is a more accurate estimate of the actual treatment effect

To date, from a regulatory perspective, external controls have been used to provide a benchmark or context for interpreting single arm effectiveness studies. With careful planning and scientifically rigorous approaches, external controls may be compared through formal statistical methods and support regulatory decisions. The clinical questions and regulatory decisions sought should drive the selection of data source, study design, and analytic approaches.

Bias inherent in externally controlled studies may be difficult to account for, but certain approaches may increase the credibility of such studies and reduce concerns (see above). Once a decision has been made to use an external comparator, there are specific considerations that may strengthen or limit the credibility of resulting data. These considerations, described in current FDA guidance, include ensuring similarity between the external populations and those receiving the experimental drug with respect to critical baseline characteristics such as disease severity, duration of illness, prior treatments, and other critical prognostic factors.

Another important consideration is the comparability of endpoint assessments regarding both definitions and ascertainment (timing, measurement). Historical clinical trial data may have more applicable data than data derived from EHRs or registries, which may not collect the sorts of endpoints used in clinical trials or collect them at consistent time points. For example, an endpoint like overall survival is less likely to suffer from ascertainment bias than is expected from more complex endpoints like response rate or progression free survival, which may differ in definition and ascertainment as well as analytical approaches across different datasets or physician assessments.

Patient management also matters, especially for cancer types in which the standard of care is not agreed upon or has rapidly changed over time. For example, as toxicity management improves over time, this may in turn impact patient outcomes. It would be beneficial if management of patients from historical data sources was similar enough to the current clinical trial to limit any resulting bias. This may be assessed by looking at the constancy of treatment outcomes historically for the control regimen. Consistency would lead to a higher level of confidence that if a randomized control had been deployed in the current trial, then it would have behaved similarly. To the extent that patient management in clinical trials differs from patient management in clinical practice, this also may result in differences between using historical clinical trial data vs. RWD. It should also be noted that the patient population(s) are rapidly changing in many areas. The immunotherapy revolution has dramatically changed many patient populations available for clinical trials relative to data that may be available historically. A complete and transparent assessment of these issues will help researchers and reviewers understand the scientific strength of the evidence of safety and effectiveness resulting from the study.

REGULATORY CONSIDERATIONS

FDA regulations explicitly recognize the use of external controls, including a hybrid approach where a clinical trial control group is augmented with external data, to support regulatory decision-making in limited circumstances.[2,3] While the use of external control data matures to the point where it may support regulatory approval more broadly, careful consideration should be given to near-term uses in appropriate regulatory and clinical contexts. Rather than replacing RCTs in situations where randomization is feasible, new methodological approaches and data sources may allow the use of external comparators, in situations where randomization would be unethical or infeasible. For example, external patient level data may be used to augment randomized control arms as part of a hybrid approach that could reduce the number of patients that are randomized to the control arm within a study. Such data may come from completed RCTs or from real-world sources such as electronic medical records.

Given a solid rationale for an external control, and a careful assessment of whether an external control would be scientifically feasible based on the considerations just outlined, the actual implementation of the external control requires care and planning. Several procedural best practices are advised as part of the regulatory process to increase the credibility of externally controlled studies. Pre-specification of protocols and statistical analysis plans provide confidence that the external control group selection process follows a prospective methodology and plan that could be independently performed or duplicated. This should include a detailed protocol with clear objectives and description of the study population, as well as details regarding data sources and critical features of the study design and analysis plan. The approach should be specified in the statistical analysis plan or other companion document and should not be biased by actual analysis of candidate external control group data that may be perceived to introduce selection bias. This may happen for example if historical data/trials with superior results are preferentially omitted. As a result, it is important that the entire selection process of a dataset and patient-level data be prespecified independent of outcome data.[33,34] The final statistical analysis and any sensitivity analyses should also be clearly pre-specified consistent with good statistical practice.

Early discussions with regulators and review of key planning documents is likely to result in valuable feedback for sponsors using external controls. Sponsors should consider soliciting FDA feedback by means of protocol submissions or formal product meetings. Sponsors may also explore opportunities for participation in the Agency’s Complex Innovative Trial Designs pilot program, which exists to further the use of new trial designs.[35] When external data comes from real-world data sources, sponsors may request input on study designs from the FDA’s RWE Subcommittee and should note any submission of RWD to the agency for tracking purposes.

CONCLUSIONS AND RECOMMENDATIONS

In oncology, there are clinical settings and scenarios where randomization may be difficult or not feasible (e.g., rare disease, small patient population, loss of equipoise, availability of the investigational agent outside of the clinical trial). Additionally, patients with serious, life threatening diseases may often seek trials where the likelihood of receiving the investigational agent is high (e.g., single arm studies, designs that allow treatment crossover). However, these scenarios (described in Table 3) may make interpreting the clinical trial results difficult or could introduce uncertainty in the results. The use of external controls in clinical studies represents an opportunity to potentially reduce the number of patients in the control arm, enhance data obtained from clinical trials, and improve the interpretability of results.

This whitepaper describes methodological approaches for constructing an external control cohort and reducing or managing potential biases that can be introduced in these types of analyses as well as operational and regulatory considerations to help guide their successful use. The case study developed for this whitepaper (Appendix 2) also helps demonstrate how to operationalize several of the concepts described in this whitepaper and inform the design of future clinical studies.

Additional considerations may also need to be explored to further facilitate the use of external control cohorts more formally in oncology drug development and regulatory discussions:

- Identify methods and mechanisms to share patient-level data to facilitate robust analyses
- Clarify how sponsors and investigators can incorporate external controls for formal analyses to support regulatory decisions
- Establish best practices for the use of specific data sources and appropriate methodologies to help develop and promote standards
- Characterize appropriate uses of specific endpoints in external controls and the ability to compare across studies

Glossary

Control Arm – In a clinical trial, the group of participants that is not given the experimental intervention being studied is the control arm. A control arm is used to establish the expected outcome without the effect of the new experimental therapy, and the result in the experimentally treated patients is judged relative to this. The control arm may receive an intervention that is considered effective (the standard), a placebo, or no intervention.

Randomized Control Arm – In a randomized controlled clinical trial, the group of participants who are randomly selected to not receive the experimental intervention is a randomized control arm. Random selection of patients and concurrent study of the randomized control arm with the study of the experimental intervention group provides high levels of assurance that differences between the randomized control arm and the experimental intervention arm are attributable to the intervention, not imbalances in baseline characteristics or differences in time, place, or circumstances of treatment.

External Control Arm – An umbrella term referring to any control that is not a randomized control. Can be used as a reference for interpretation of a set of experimental data especially when randomization is unethical or unfeasible.

Concurrent Control Arm – A type of external control. A group chosen from the same or similar population as the experimental intervention group and treated over the same period of time as the experimentally treated patients. Ideally, the experimental intervention and control groups should be similar with regard to all baseline and on-treatment variables that could influence the outcome, except for the study treatment. May be patient-level data or summary information gained from medical literature or other sources.

Historical Control – A type of external control. A non-concurrent comparator group of patients who received treatment (placebo or active treatments) in the past or for whom data are available through records. May be patient-level data or summary information gained from medical literature or other sources.

Synthetic Control Arm – A type of external control consisting of patient level data from patients external to the trial and selected with statistical methods such as propensity scores to provide confidence that the baseline characteristics of the selected external patients are balanced and comparable with the baseline characteristics of the experimentally treated patients. Can be formed from external clinical trials data, real-world data, or other data sources.

References

1. Friends of Cancer Research. Exploring Whether a Synthetic Control Arm can be Deriving from Historical Clinical Trials that Match Baseline Characteristics and Overall Survival Outcome of a Randomized Control Arm. (2018). Available at: https://www.focr.org/sites/default/files/SCA White Paper.pdf.
2. 21 CFR 314.126.
3. FDA. Guidance for Industry: E10 Choice of Control Group and Related Issues in Clinical Trials. (2001). Available at: https://www.fda.gov/media/71349/download.
4. Faivre, S. et al. Sunitinib in pancreatic neuroendocrine tumors: updated progression-free survival and final overall survival from a phase III randomized study. Ann. Oncol. 28, 339–343 (2016).
5. Barlev, A. et al. Estimating Long-Term Survival of Adults with Philadelphia Chromosome-Negative Relapsed/Refractory B-Precursor Acute Lymphoblastic Leukemia Treated with Blinatumomab Using Historical Data. Adv. Ther. 34, 148–155 (2017).
6. Gökbuget, N. et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood Cancer J. 6, e473–e473 (2016).
7. Cowey, C. L. et al. Real-world treatment outcomes in patients with metastatic Merkel cell carcinoma treated with chemotherapy in the USA. Futur. Oncol. 13, 1699–1710 (2017).
8. Greenland, S., Pearl, J. & Robins, J. M. Causal Diagrams for Epidemiologic Research. Epidemiology 10, (1999).
9. Miguel A. Hernán, J. M. R. Causal Inference: What If. (Chapman & Hall/CRC, 2019).
10. Ho, D. E., Imai, K., King, G. & Stuart, E. A. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Polit. Anal. 15, 199–236 (2007).
11. Stuart, E. A. Matching methods for causal inference: A review and a look forward. Stat. Sci. 25, 1–21 (2010).
12. Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. & Sargent, D. J. Hierarchical Commensurate and Power Prior Models for Adaptive Incorporation of Historical Information in Clinical Trials. Biometrics 67, 1047–1056 (2011).
13. Murray, T. A., Hobbs, B. P. & Carlin, B. P. Combining nonexchangeable functional or survival data sources in oncology using generalized mixture commensurate priors. Ann. Appl. Stat. 9, 1549–1570 (2015).
14. Ibrahim, J. G., Chen, M.-H., Gwon, Y. & Chen, F. The power prior: theory and applications. Stat. Med. 34, 3724–3749 (2015).
15. Schmidli, H. et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 70, 1023–1032 (2014).
16. Viele, K. et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm. Stat. 13, 41–54 (2014).
17. Rubin, D. B. & Thomas, N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. Biometrics 52, 249–264 (1996).
18. Lunceford, J. K. & Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. 23, 2937–2960 (2004).

19. Li, F., Morgan, K. L. & Zaslavsky, A. M. Balancing Covariates via Propensity Score Weighting. J. Am. Stat. Assoc. 113, 390–400 (2018).
20. Bang, H. & Robins, J. M. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics 61, 962–973 (2005).
21. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav. Res. 46, 399–424 (2011).
22. K Rothman, S Greenland, T. L. Modern Epidemiology, 3rd Edition. (RTI International, 2007).
23. Walker, A. M. Confounding by Indication. Epidemiology 7, (1996).
24. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am. J. Epidemiol. 183, 758–764 (2016).
25. Ray, W. A. Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. Am. J. Epidemiol. 158, 915–920 (2003).
26. Schneeweiss, S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol. Drug Saf. 19, 858–868 (2010).
27. Schneeweiss, S. et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 20, 512–522 (2009).
28. Rassen, J. A., Brookhart, M. A., Glynn, R. J., Mittleman, M. A. & Schneeweiss, S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. J. Clin. Epidemiol. 62, 1226–1232 (2009).
29. Schneeweiss, S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol. Drug Saf. 15, 291–303 (2006).
30. Seymour, L. et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clin. Cancer Res. 16, 1764–1769 (2010).
31. FDA. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. 2018 Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics. (Accessed: 1st November 2019)
32. FDA. FRAMEWORK FOR FDA'S REAL-WORLD EVIDENCE PROGRAM. (2018). Available at: https://www.fda.gov/media/120060/download.
33. FDA. Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products. (2018). Available at: https://www.fda.gov/media/117976/download.
34. FDA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry. (2019). Available at: https://www.fda.gov/media/124795/download.
35. FDA. Complex Innovative Trial Designs Pilot Program. (2019). Available at: https://www.fda.gov/drugs/development-resources/complex-innovative-trial-designs-pilot-program.
36. ROSENBAUM, P. R. & RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55 (1983).
37. Stürmer, T., Wyss, R., Glynn, R. J. & Brookhart, M. A. Propensity scores for confounder

adjustment when assessing the effects of medical interventions using nonexperimental study designs. J. Intern. Med. 275, 570–580 (2014).

38. Cepeda, M. S., Boston, R., Farrar, J. T. & Strom, B. L. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. Am. J. Epidemiol. 158, 280–287 (2003).
39. Seeger, J. D., Bykov, K., Bartels, D. B., Huybrechts, K. & Schneeweiss, S. Propensity Score Weighting Compared to Matching in a Study of Dabigatran and Warfarin. Drug Saf. 40, 169–181 (2017).
40. Stürmer, T., Rothman, K. J., Avorn, J. & Glynn, R. J. Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study. Am. J. Epidemiol. 172, 843–854 (2010).
41. Hernán, M. Á., Brumback, B. & Robins, J. M. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. Epidemiology 11, (2000).

Appendix 1: Detailed Description of Propensity Scores, Propensity Score Matching, and Propensity Score Weighting

Propensity Scores

The propensity score is a method developed in the early 1980's, and further developed substantially over the past decades, to reduce bias due to confounding in observational (non-randomized) studies.[21,36,37] A more novel application of propensity scores is to create balance between a clinical trial treatment arm and an external control group (see Table 2).[6] While it does not control for unmeasured confounding, there are several advantages:

- A propensity score makes it possible to create balance across many factors simultaneously, avoiding issues of cutting data "too thin" when exact matching on many factors.
- In scenarios where patient n is limited but confounding is strong, the single propensity score value allows us to capture a large amount of confounding using substantially fewer degrees of freedom in an outcome model.[38]
- The propensity score can be effectively used in a variety of ways, including matching, weighting, or regression.

Propensity Score Matching

The most common use of the propensity score is in matching. The idea is straightforward: if we are able to estimate the probability of a patient being treated with the investigational treatment, as compared to the standard of care (SOC) measured in external controls, then if we match patients who had similar probabilities of being treated with the investigational treatment and treated with the SOC, then the choice between investigational treatment and SOC for that patient can be thought of as essentially random. That is, if we can (1) estimate a propensity score using all relevant confounders, and then (2) take each patient treated with the investigational treatment and find a similar patient treated with the SOC, we will (3) create a cohort of patients in which each confounder tends to be balanced between the investigational and SOC treatment groups, with no need to further adjust for confounding. As a result, the data can be analyzed like that of an RCT, even though we created the balance by construction rather than design.

The advantages of matching are substantial, and include:

- Clear methodology that is easily understood by readers, reviewers, and others.
- A table of baseline characteristics that can be verified for balance, building confidence in results.
- Simple analytics that do not require, for example, bootstrapped variances or other statistical nuances.

However, matching can also introduce a challenge: if we seek to match all patients treated with the investigational treatment but fail to find a match among those treated with the SOC, then we will lose one or more patients that received the investigational treatment in the analysis. In cases where there is a substantial n, this is often manageable, but in small trials where each patient's clinical experience is of extraordinary value, losing patients is highly unfavorable.

Propensity Score Weighting

An alternative to matching in which no data are lost is propensity score weighting. Weighting is a propensity score-based approach to standardization; while there are a wide variety of weighting techniques that can be used, the one most commonly seen (and the one used in the Blincyto example) is inverse probability of treatment (IPTW) weighting. Another technique, weighting by propensity odds, is discussed in Table 2.

As a propensity score estimates a patient's probability of receiving a given treatment, the inverse probability of treatment weight is the inverse of the propensity score (that is, 1/PS) for patients receiving the investigational treatment and 1/(1-PS) for SOC patients. (We use 1-PS because that is the probability of being treated with the SOC.) When using IPTW weights, we model a population in which both treated patients and control patients are "standardized" to resemble the entire study population, such that treated patients may be standardized to more resemble controls and vice-versa.

The clear advantage of this technique is that no patient data are lost; we are able to use all data and achieve confounding control. However, there are several disadvantages:

- The method appears somewhat opaque and may not create confidence by readers, reviewers, and other stakeholders.
- Because this method counts certain patients more than others (those with high weights versus those with low weights), it is possible that it may overweight the experience of one or more patients.[39] Control of the maximum assigned weight is often necessary.[40]
- This method will change the weight of both patients receiving the investigational agent and SOC patients, and as such, the data as weighted will not represent patients' actual experience in the single-arm trial.
- Technical adjustments are often needed to stabilize weights and to accurately report variance.[41]

Appendix 2: Developing a Synthetic Control Arm Derived from Historical Multiple Myeloma Clinical Trials and Assessing Unobserved Confounders

1. CASE STUDY OBJECTIVES

This case study builds on previous work (Friends of Cancer Research whitepaper, 2018, case study in non-small cell lung cancer) and continues exploration of whether a synthetic control arm (SCA) can be useful for assessment of medical product efficacy and safety in indications where a randomized control presents ethical or practical challenges. This case study has two primary objectives.

- Objective 1: To explore whether the treatment effect based on a SCA (i.e., investigational arm vs. SCA) can mimic the treatment effect based on the randomized control (i.e., investigational arm vs. randomized control).
- Objective 2: To develop and illustrate statistical methods (e.g., tipping point analyses) useful for assessing the impact of unobserved confounders on the demonstration of efficacy in the setting of a SCA.

This case study will also address some of the concerns regarding incomplete matching in the previous work by utilizing matching methods that do not require exclusion of a large proportion of investigational product (IP) treated patients and by extending SCA exploration to additional indications.

2. DATA SOURCES

This case study is based on patient-level data from multiple historical clinical trials in relapsed/refractory multiple myeloma. These trials have been conducted by the pharmaceutical industry for the purposes of drug development and are available through the Medidata Enterprise Data Store (MEDS). MEDS is a collection of thousands of previous clinical trials with patient-level data recorded through the Medidata electronic data capture system, Rave. Per the legal agreements with the sponsors of these historical clinical trials and Medidata, these data are available for use in deidentified (e.g., patients and original sponsor of the trial cannot be identified) and aggregated (e.g., every analysis must include data from two or more sponsors) form.

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 phase 3 multinational trials that were conducted between 2010 and 2017. At baseline, all patients had:

- Relapsed or refractory multiple myeloma
- Received at least 2 prior lines of treatment
- Received prior treatment with lenalidomide and bortezomib
- Age ≥ 18 years

Including both investigational and control arms from the historical trials, there were 946 historical patients available for this case study.

Table 1: Features of Historical Data

	Design	Region	Start/End of Trial(s)	Baseline Characteristics	Endpoints	Number of Patients in All Arms	Control Regimen
Historical Data (from multiple trials)	Open label, phase 3	Multi-national	Trial conducted between 2010 and 2017	- Relapsed or refractory multiple myeloma - Received at least 2 prior lines of treatment - Received prior treatment with lenalidomide and bortezomib - Age ≥ 18 years	Overall survival	946	Dexamethasone

Because the historical data in this case study came from trials that had been conducted as part of clinical development programs, the populations, study design, data collection methods, and endpoints utilized in these trials are fairly consistent across trials. Nevertheless, differences across studies in some variable 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 creation of the SCA are shown in Table 2. Overall survival is the endpoint of interest for this case study and was measured as a key outcome in all historical trials that had similar study designs, such as the disease population and follow-up time.

Table 2: Clinically Important Baseline Covariates Available Across Historical Trials

1. Race (White vs. Others/unknown)
2. Region (Europe vs. Others/unknown)
3. ECOG=0 vs 1 vs 2 or 3
4. Number of drug classes refractory (≥4 vs. <4)
5. Cytogenetic risk (High vs. Standard/unknown)
6. Prior stem cell transplant (Yes vs. No/unknown)
7. Age (continuous)
8. Days since last PD/relapse to first study dose (continuous)
9. Sex (F vs. M)
10. Bone lesion (Yes vs. No/unknown)
11. Best response to last therapy (≥PR vs. <PR/unknown)
12. Number of prior lines of therapy (continuous)
13. Years since diagnosis (continuous)
14. Weight (continuous)

3. RATIONALE AND METHODS

3.1 For objective 1, we explored whether the treatment effect based on a SCA can mimic the treatment effect based on a randomized control using a historical randomized controlled trial in multiple myeloma. This trial, the 'Target Randomized Trial', had a 2:1 treatment assignment ratio and included 294 patients assigned to investigational treatment and 149 patients assigned to dexamethasone as a control. An SCA was selected from the remaining 201 patients assigned to dexamethasone control in all other studies available within this project. Patients assigned to investigational therapies in all trials except the target trial made up the remainder of the total 946 patients referenced above (table 1) and were not utilized in this case study. Historical patients were selected for inclusion in the SCA to balance the baseline characteristics of the IP treated patients in the Target Randomized Trial and the SCA using propensity score methods. Selection of the historical patients for the SCA was completed using only baseline characteristics without knowledge of any post-randomization data.

While appealing in its simplicity and similarity to a randomized design, the commonly used approach to propensity score matching, Greedy 1-1 matching, was not possible for this case study due to the limited number of historical control patients available. Rather, we used a matching method called optimal full matching (often referred to as full matching), which was introduced by Rosenbaum (Rosenbaum 1991) and recommended recently (Hansen 2004, Austin and Stuart 2015a). Full matching subdivides the subjects into strata of different sizes, consisting of either one IP treated subject and at least one control subject or one control subject and at least one IP treated subject. The algorithm of full matching is to minimize the average differences within a matched set in the propensity score between IP treated and control subjects. An attractive feature of this approach is that it can use most or even the entire set of all IP treated subjects in the analysis. This contrasts with conventional matching approaches such as Greedy matching where a portion of treated subjects cannot be matched and therefore are excluded from the final analysis. As a result, full matching might avoid potential bias due to incomplete matching, which can occur when some treated subjects are excluded from the matched sample.

Step 1: Estimate propensity scores. The propensity score is the probability of assignment of target trial investigational product 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 investigational product 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. The predictors included in the propensity score model are all available baseline characteristics described in Table 2. 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 (Stuart & Rubin 2008, Harris 2016). Some researchers recommend using all available baseline covariates in the analysis (Lim 2018) if the sample size permits.

Step 2: Create SCA by selecting historical patients to match investigational patients in the Target Randomized Trial using full matching. SAS PROC PSMATCH (SAS/STAT® 15.1) will be used for matching, and the maximum number of historical controls to be matched with each IP treated patients and the maximum number of IP treated patients to each historical control will be determined based on the ratio of the number of subjects between IP treated patients and historical controls (Hansen 2004) as well as the performance of balancing baseline characteristics listed in table 2.

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 Target Randomized Trial investigational arm and SCA subjects. Baseline demographic and disease characteristics will be summarized with descriptive statistics for the Target Randomized Trial investigational arm and SCA. Standardized difference in covariate means before matching and after matching will be computed and compared.

For a continuous covariate, the standardized difference is:

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

Where $\overline{x_t}$ and $\overline{x_c}$ denote the sample mean of the covariate for the Target Randomized Trial investigational arm and historical control groups, respectively; s_t^2 and s_C^2 denote the sample variance of the covariate for the Target Randomized Trial investigational arm and his torical 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 Randomized Trial investigational arm 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.

To account for the difference in the number of treated and control subjects within each matched set in full matching, a weighted standardized difference will be used and weights will be derived from the strata imposed by the full matching and constructed as follows: IP treated patients are assigned a weight of one, while each historical control patient has a weight calculated as the number of IP treated patients in its matched set divided by the number of controls in the matched set. (Ho 2007) The weights of controls are scaled such that the sum of the weights from matched controls across all the matched sets is equal to the number of uniquely matched treated subjects.

Each sample estimate (sample means, variances, and prevalences) in the above formulas will be replaced by its weighted equivalent. The weighted mean $\bar{x}_{WT} = \frac{\sum w_i x_i}{\sum w_i}$ and weighted sample variance $s_{WT}^2 = \frac{\sum w_i}{(\sum w_i)^2 - \sum w_i^2} \sum w_i (x_i - \bar{x}_{WT})^2$ will be used, where w_i is the weight assigned to the ith subject (Austin and Stuart 2015b).

The absolute standardized differences should generally be less than 0.25 (Stuart et al., 2008). 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 (Normand et al., 2001). 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 Randomized Trial investigational arm and SCA using graphical methods such as cloud plots.

The treatment effect on overall survival based on the SCA will be described alongside the treatment effect from the Target Randomized Trial using a Kaplan Meier curve, log rank test, hazard ratio, and 95% confidence interval for the hazard ratio. Weighted estimates incorporating the weights induced by the full matching will be examined.

3.2 Objective 2 is undertaken to illustrate an approach for testing the robustness of the treatment effect to an unobserved or unknown covariate, a potential confounder. While methods such as propensity score matching can adjust for observed confounding, unobserved confounding or unavailable measurement is often a concern compared to the gold standard randomized clinical trial where both observed and unobserved confounders can be balanced. When a key variable is not available for historical patients used to build the SCA, balance between groups in this factor cannot be assured or even described. For example, there may be situations where a key biomarker discovered to have prognostic value only in recent years is available in today's investigational patients, but was not measured or is otherwise unavailable in historical trials. Imbalance in this known or unknown factor could bias the comparison between groups. Under this objective, we illustrate a special type of sensitivity analyses (i.e., tipping point analyses) designed to assess how strong the association of an unobserved confounder with the treatment assignment and the outcome would have to be to change the study inference. If the effects of the investigational product (efficacy or safety) is insensitive over a wide range of plausible assumptions regarding the confounding, then the qualitative effects can be concluded to be secure despite the possibility of unobserved confounders.

Utilizing methods proposed by Lin (Lin 1998), we will adjust the observed treatment effect (HR and 95% confidence intervals) for overall survival to reflect the impact of a theoretical unobserved confounder. Let β and β* denote the true and apparent regression parameters for the treatment effects, respectively. The β is the parameter of interest adjusting for the potential unobserved confounder; while β*, obtained from the observed analysis and necessarily produced by a reduced model due to the unavailability of the unobserved confounder will be adjusted by specifying the distributions of the unobserved confounder among the treatment arms as well as the effects of the unobserved confounder on outcome as

$$\beta \approx \beta^{*} - log\frac{e^{\gamma_1}P_1 + (1 - P_1)}{e^{\gamma_0}P_0 + (1 - P_0)}$$

where P_0 and P_1 are the assumed prevalence of the unmeasured confounder among the investigational group and SCA respectively, and the assumed hazard ratio of the unmeasured confounder on the event of interest among the investigational group and SCA is $\Gamma_0 = e^{\gamma_0}$ and $\Gamma_1 = e^{\gamma_1}$, respectively. Without loss of generalizability, we can assume $\Gamma = e^{\gamma_0} = e^{\gamma_1}$. The strength of these assumed relationships between the potential confounder and treatment arm imbalance (ie, prevalences P_0 and P_1) and the potential confounder and overall survival (ie, hazard ratio Γ) will be varied over a range of relevant values so that the point where the conclusion regarding the effect of the drug is changed can be identified.

The assumptions that result in a loss of statistical significance of the treatment effect with the SCA will be highlighted as the ‘statistical tipping point’. Assumptions at which the numerical direction of the treatment effect is changed will be highlighted as the ‘clinical tipping point’. These tipping points allow an understanding of how imbalanced and influential an unobserved confounder would have to be in order to change the qualitative conclusion. The reader may then make a judgement regarding whether a confounder with this degree of imbalance and impact is likely to exist in the clinical setting and therefore whether the efficacy conclusion is robust against unobserved confounders.

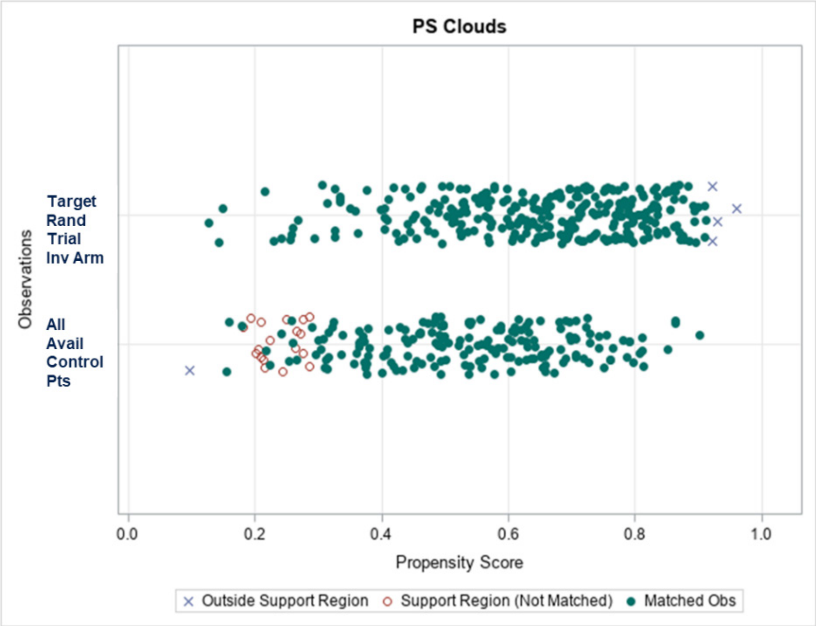
4 RESULTS

4.1 SCA CREATION AND BASELINE BALANCE ACHIEVED

As described in the methods section, full matching was used to select, match, and weight the appropriate patients from the historical pool for inclusion in the SCA to balance the distribution of baseline characteristics between the SCA and the investigational arm from the Target Randomized Trial. Propensity scores were calculated as described in the methods section and utilizing the covariates listed in Table 2. The Cloud Plot in Figure 1 shows the distribution of propensity scores for the investigational arm from the Target Randomized Trial (top) and all available control patients from other trials (bottom). The figure illustrates the degree to which these distributions overlap. The investigational arm from the Target Randomized Trial included 294 patients. Overlap in the distribution of propensity scores for the investigational arm in the Target Randomized Trial and the historical controls was nearly complete. 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 distributions and visually we can see that there are no analogous patients available in this region in the opposite group. Two hundred ninety (99%) in the investigational arm in the Target Randomized Trial were successfully matched. The remaining 4 patients (1%) were not matched and were removed from further analysis. A larger number of control patients are not matched and are excluded from further analysis, but this is of no consequence since our interest is inference regarding the investigational treatment, not the controls themselves.

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 concerning 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 characteristi 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. But with only 4 patients excluded in this case, there is likely to be very little impact on extrapolation and may illustrate a possible advantage of full matching over greedy 1-1 matching, which may resul in more patient exclusions in certain cases.

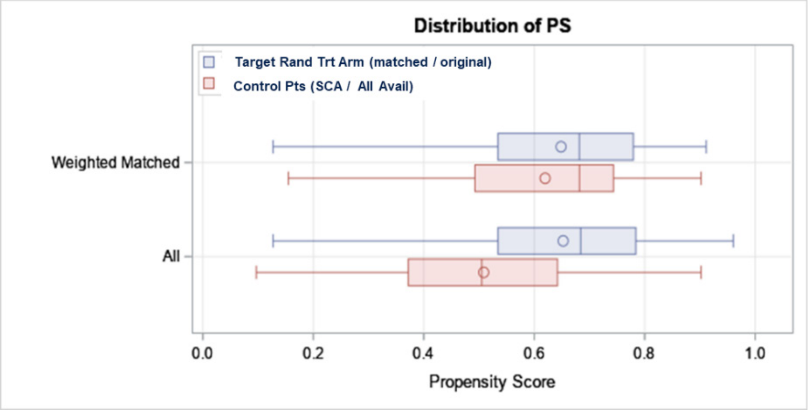
Figure 1. Cloud Plot: Distribution of Propensity Scores for Investigational Arm of Target Randomized Trial Versus All Available Historical Control Patients



We now consider the degree of balance that has been achieved by the propensity score full matching. The propensity score can be considered a summarization of all baseline characteristics and so we begin by examining the balance achieved in the propensity score.

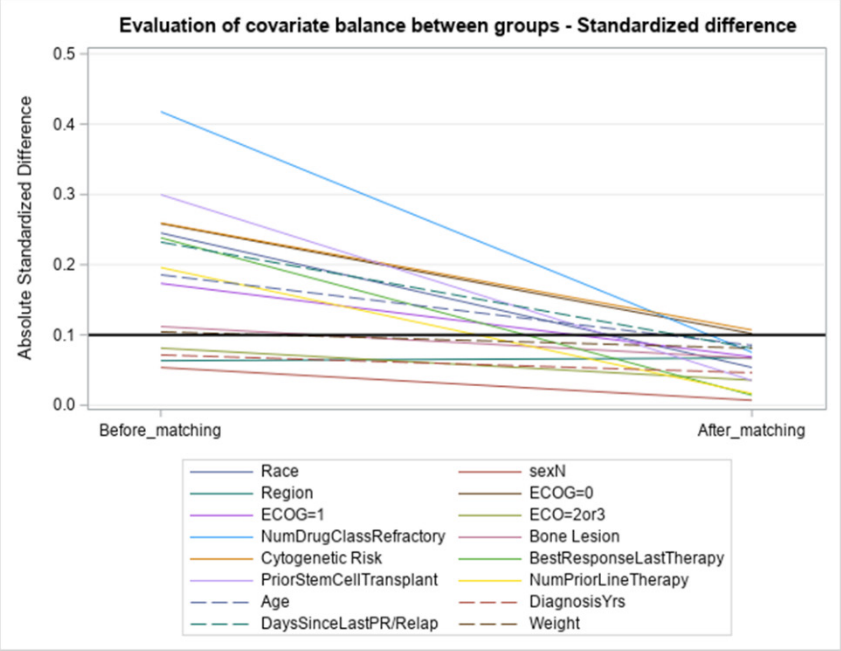
The distributions of the propensity score for the investigational arm of the Target Randomized Trial and all available historical control patients before matching are shown in the lower set of boxplots in Figure 2. The analogous distributions after matching are shown in the upper region of these figures. There is considerable discordance between the investigational arm of the Target Randomized Trial and all available historical controls before matching. For example, the median for the investigational arm is higher than that of the historical pool. However, after matching, the medians of the groups are very similar.

Figure 2. Distribution of Propensity Scores Before and After Matching



Assessment of balance in terms of individual baseline covariates yields observations consistent with the conclusions afforded above by examination of the propensity scores and indicates very good balance between groups after matching. Figure 3 illustrates the standardized difference between the investigational arm of the Target Randomized Trial and historical controls (before matching) on the left and the same between the matched investigational arm of the Target Randomized Trial and SCA (after matching) for each baseline characteristic examined in this case study. In all cases, reductions in the absolute standardized difference between groups for each variable are observed and the absolute standardized differences after matching are equal to or below 0.10, a commonly used threshold for designating a negligible difference in the mean or prevalence of a covariate between groups, for all but two instances.

Figure 3. Plot of Standardized Difference of Important Baseline Covariates Before and After Matching



Similarly, examination of the baseline characteristics (on their original scales) for the matched investigational arm of the Target Randomized Trial and the SCA reveals good balance between groups.

The matched investigational arm of the Target Randomized Trial includes 290 patients. As shown in Table 3, most patients were white males from the Europe region with an average age of 63.5 years. Many patients were refractory to 4 or more drug classes (72.1%) and/or had prior stem cell transplant (61%) at baseline. The SCA is quite well balanced with the investigational arm and is weighted to represent 290 patients. Similar to the investigational arm, white males from the Europe region were common in the baseline estimates for the SCA and the average age for the SCA was 64.3 years. Also like the investigational arm, the SCA includes many patients who were refractory to 4 or more drug classes (68.6%) and/or had prior stem cell transplant (59.3%) at baseline. Overall, very good balance in baseline characteristics is achieved between the investigational arm and SCA.

Table 3. Baseline Characteristics - SCA vs. Matched Investigational Arm in Target Randomized Trial

Baseline Characteristic	**Matched Investigational Arm in Target Randomized Trial (N=290)**	**SCA Weighted Summary (N=290)**
Race (White)	235 (81.0)	241 (83.1)
Region (Europe)	232 (80.0)	224 (77.2)

ECOG=0
ECOG=1
ECOG=2 or 3
	105 (36.2)
	134 (46.2)
	51 (17.6)
		91 (31.4)
		144 (49.7)
		55 (19.0)

Number Drug Classes Refractory (>=4)	209 (72.1)	199 (68.6)
Cytogenetic Risk (high)	29 (10.0)	39 (13.4)
Prior Stem Cell Transplant	177 (61.0)	172 (59.3)
Age (continuous)	63.5 (9.4)	64.3 (9.6)
Days since last PD/relapse to first study dose (continuous)	64.6 (80.1)	71.2 (104.5)
Sex (Male)	174 (60.0)	175 (60.3)
Bone lesion	204 (70.3)	195 (67.2)
Best response to last therapy (≥PR vs. <PR/unknown)	106 (36.6)	104 (35.9)
Number of prior lines of therapy (<4)	64 (22.1)	62 (21.4)
Years since diagnosis (continuous)	6.3 (4.1)	6.1 (4.4)
Weight (continuous)	74.5 (15.3)	73.3 (18.4)

4.2 REPLICATION OF TREATMENT EFFECT ON OVERALL SURVIVAL WITH SCA (OBJECTIVE 1)

In previous sections, we have demonstrated that the propensity score full matching successfully balanced the distribution of baseline characteristics between the SCA and the investigational arm of the Target Randomized Trial. We now move to the first primary objective of this case study, to explore whether the treatment effect based on a SCA (i.e., matched investigational arm from Target Randomized Trial vs. SCA) can mimic the treatment effect based on the randomized control (i.e., investigational arm vs. randomized control in Target Randomized Trial).

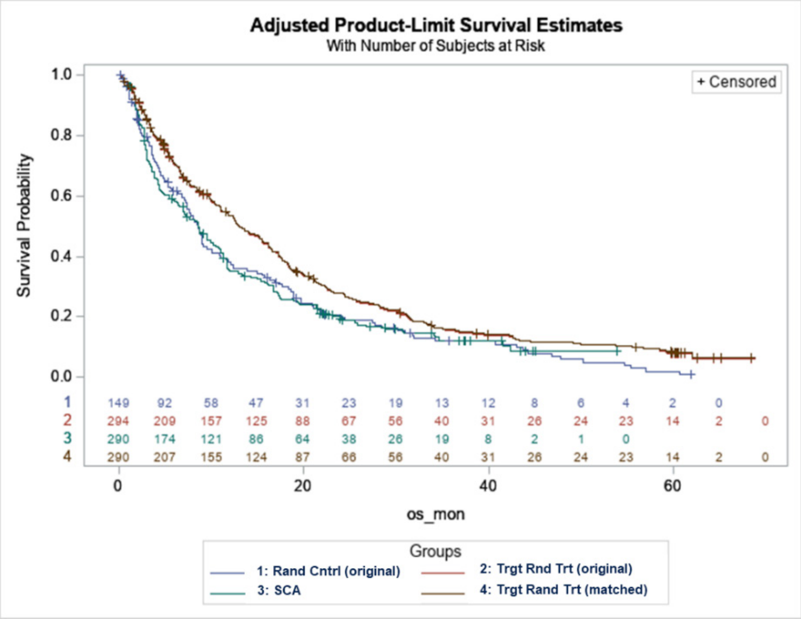
Figure 4 provides a description of OS for four groups:

- Investigational arm of the Target Randomized Trial (red)
- Randomized control arm of the Target Randomized Trial (blue)
- Matched investigational arm of the Target Randomized Trial (brown)
- SCA (teal)

The Target Randomized Trial demonstrated a positive treatment effect on overall survival, as evidenced by a separation of the Kaplan Meier curves representing the investigational and randomized control arms of a Target Randomized Trial. The hazard ratio for the investigational arm versus the randomized control is 0.743 with a confidence interval that excludes 1 (95% CI: (0.60, 0.92)). This difference between groups is also supported by the log rank test (p=0.0061).

The treatment effect utilizing SCA is very similar. The Kaplan Meier curve for the SCA visually overlaps and crosses with that of the randomized control and the quantified differences between SCA and the matched investigational arm of the Target Randomized Trial are very similar to the original trial. The hazard ratio for the matched investigational arm versus the SCA is 0.758 with a confidence interval that excludes 1 (95% CI: (0.63, 0.91)). This difference between groups is also supported by the log rank test (p=0.0158).

Figure 4 Overall Survival Treatment Effect Using the Randomized Control Versus Using SCA



Investigational vs. Control from Target Randomized Trial		Matched Investigational from Target Rand Trial vs. SCA	
Log-Rank 2-sided P-value	0.0063	Log-Rank 2-sided P-value	0.0158
Hazard Ratio (95% CI)	0.74 (0.60, 0.92)	Hazard Ratio (95% CI)	0.76 (0.63, 0.91)

5.0 TIPPING POINT ANALYSES FOR UNOBSERVED CONFOUNDERS – OBJECTIVE 2

This section illustrates an approach for testing the robustness of the treatment effect to an unobserved or unknown covariate. While propensity score matching can be used to balance observed covariates, it cannot guarantee to balance or describe balance for unobserved covariates. The HR and 95% confidence interval for the effect of treatment in the investigational arm of the Target Randomized Trial relative to SCA was estimated to be 0.76 (0.63, 0.91) but one may question whether this is due to the investigational product or due to an imbalance in an unknown or unmeasured confounder.

Using the methods of Lin (Lin, 1998), as described in section 3.2, the observed treatment effect can be adjusted to reflect the possibility of an unknown confounder when the prevalence of the confounder in each treatment arm is known (or assumed) and the influence the confounder has on outcomes is known (or assumed). For example, suppose an unknown confounder is present for only 10% of the investigational arm in this case study while it is present for 30% of the SCA and that the confounder is moderately predictive of overall survival with a hazard ratio for overall survival for those with versus without the confounder of 1.5. Then the adjusted treatment effect separate from the effect of this confounder is estimated to be HR=0.83 with 95% CI (0.69, 0.99). This leads to a conclusion that is qualitatively consistent with that of the original unadjusted treatment effect, that the investigational product is providing a statistically significant benefit. If, however, we had assumed a little stronger imbalance between groups and set the prevalence of the confounder in the SCA slightly higher, say 35%, while all other assumptions remained the same, the adjusted treatment effect separate from the effect of this confounder is estimated to be HR=0.85

with 95% CI (0.71,1.02). These results indicate no statistically significant difference between the investigational arm and the SCA and is qualitatively inconsistent with the original unadjusted analysis. That is the assumption of a 35% prevalence in the SCA rather than 30% is the 'statistical tipping point' where statistical significance of the treatment effect is changed from the original unadjusted analysis. A similar threshold, a 'clinical tipping point', exists where the numerical estimate of the HR exceeds 1 and the numerical trend for the treatment effect is no longer consistent with the original unadjusted analysis.

The example provided above represents just a few possible sets of assumptions regarding the unobserved confounder. To fully understand the possible impact of an unobserved confounder, many sets of assumptions, a grid across all possible or plausible assumptions should be considered. Tables 4 and 5 provide estimates of the treatment effect (HR and 95% confidence interval) adjusted for a theoretical unobserved confounder. The prevalence of this unobserved confounder in the investigational group and SCA are assigned all possibilities, between 0 and 0.8 in increments of 0.05 and are included in the rows and columns of Tables 4 and 5. The relationship between the theoretical unobserved confounder and overall survival is assumed moderate (hazard ratio for those with and without the confounder set to 1.5) in Table 4 and strong (hazard ratio for those with and without the confounder set to 2.0) in Table 5. Entries in each of the cells are the adjusted treatment effects (HR and 95% CI) under these sets of conditions.

The diagonal entries indicated in red text are under the assumption that the unobserved confounder is balanced between the investigational arm and the SCA and therefore the adjusted treatment effect is identical to the original analysis. Moving to the right of the diagonal, as the prevalence of the confounder is assumed to be higher in the SCA than the investigational arm, the HR and 95% confidence intervals initially provide the same conclusion as the original analysis, that there is a statistically significant benefit of the investigational product. Eventually though the imbalance in the theoretical confounder becomes enough to lead to the conclusion that the treatment effect is not statistically significant. This is the 'statistical tipping point' and is represented in Tables 4 and 5 by yellow shading. Moving even further to the right and increasing the discrepancy in prevalence of the confounder between arms even further eventually leads to a numerical estimate of the HR that is bigger than 1 and is no longer directionally consistent with the original analysis. This is the 'clinical tipping point' and is represented in Tables 4 and 5 by green shading.

These tipping points allow an understanding of how imbalanced and influential an unobserved confounder would have to be in order to change the qualitative conclusion regarding the statistical significance or numerical direction of the original unadjusted treatment effect. With this information, the reader may make a judgement regarding whether a confounder with this degree of imbalance and impact is likely to exist in the clinical setting and therefore whether the efficacy conclusion is robust against unobserved confounders.

6.0 CONCLUSION

In this case study in relapsed/refractory multiple myeloma, we have demonstrated that it is possible to produce an SCA from historical clinical trial data using propensity score methods that is well balanced with the investigational arm at baseline. This case study further illustrated that this is possible even when the historical data size is limited and without excessive exclusion of nonmatched patients from the investigational arm, both benefits possibly attributable to the full matching approach.

Importantly, this case study also demonstrated the treatment effect on OS estimated in comparison to the randomized control was very closely matched by that of the SCA, suggesting that SCA could be used to augment or replace a randomized control in future trials in indications where a randomized control is ethically or practically challenging.

Tipping point analyses illustrated in this case study are an effective way of understanding the possible impact of unobserved confounders on the treatment effect estimates and whether the statistical and numerical direction of those effects are reliable despite a reasonable degree of confounding expected in the particular clinical setting.

Table 4: Statistical and Clinical Tipping Points for Overall Survival Analysis (when HR for overall survival of those with and without confounder set to 1.5)

Trt Effect HR (95% CI)		Assumed prevalence of unobserved confounder in SCA																
		0.0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
Assumed prevalence of unobserved confounder in investigational arm	0.0	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.80 (0.66,0.95)	0.81 (0.68,0.98)	0.83 (0.70,1.00)	0.85 (0.71,1.02)	0.87 (0.73,1.04)	0.89 (0.74,1.07)	0.91 (0.76,1.09)	0.93 (0.77,1.11)	0.95 (0.79,1.14)	0.97 (0.81,1.16)	0.99 (0.82,1.18)	1.00 (0.84,1.20)	1.02 (0.85,1.23)	1.04 (0.87,1.25)	1.06 (0.88,1.27)
	0.05	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.68,0.97)	0.83 (0.69,1.00)	0.85 (0.71,1.02)	0.87 (0.72,1.04)	0.89 (0.74,1.06)	0.91 (0.76,1.09)	0.92 (0.77,1.11)	0.94 (0.79,1.13)	0.96 (0.80,1.15)	0.98 (0.82,1.17)	1.00 (0.83,1.20)	1.02 (0.85,1.22)	1.04 (0.86,1.24)
	0.1	0.72 (0.60,0.86)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.68,0.97)	0.83 (0.69,0.99)	0.85 (0.71,1.02))	0.87 (0.72,1.04)	0.88 (0.74,1.06)	0.90 (0.75,1.08)	0.92 (0.77,1.10)	0.94 (0.78,1.12)	0.96 (0.80,1.15)	0.97 (0.81,1.17)	0.99 (0.83,1.19)	1.01 (0.84,1.21)
	0.15	0.71 (0.59,0.84)	0.72 (0.60,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.68,0.97)	0.83 (0.69,0.99)	0.85 (0.71,1.01)	0.86 (0.72,1.03)	0.88 (0.73,1.06)	0.90 (0.75,1.08)	0.92 (0.76,1.10)	0.93 (0.78,1.12)	0.95 (0.79,1.14)	0.97 (0.81,1.16)	0.99 (0.82,1.18)
	0.2	0.69 (0.57,0.83)	0.71 (0.59,0.85)	0.72 (0.60,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.68,0.97)	0.83 (0.69,0.99)	0.84 (0.70,1.01)	0.86 (0.72,1.03)	0.88 (0.73,1.05)	0.90 (0.75,1.07)	0.91 (0.76,1.09)	0.93 (0.78,1.11)	0.95 (0.79,1.14)	0.96 (0.80,1.16)
	0.25	0.67 (0.56,0.81)	0.69 (0.58,0.83)	0.71 (0.59,0.85)	0.72 (0.60,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.67,0.97)	0.83 (0.69,0.99)	0.84 (0.70,1.01)	0.86 (0.72,1.03)	0.88 (0.73,1.05)	0.89 (0.74,1.07)	0.91 (0.76,1.09)	0.93 (0.77,1.11)	0.94 (0.79,1.13)
	0.3	0.66 (0.55,0.79)	0.68 (0.56,0.81)	0.69 (0.58,0.83)	0.71 (0.59,0.85)	0.73 (0.60,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.67,0.97)	0.82 (0.69,0.99)	0.84 (0.70,1.01)	0.86 (0.71,1.03)	0.87 (0.73,1.05)	0.89 (0.74,1.07)	0.91 (0.76,1.09)	0.92 (0.77,1.11)
	0.35	0.65 (0.54,0.77)	0.66 (0.55,0.79)	0.68 (0.56,0.81)	0.69 (0.58,0.83)	0.71 (0.59,0.85)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.67,0.97)	0.82 (0.69,0.99)	0.84 (0.70,1.00)	0.85 (0.71,1.02)	0.87 (0.73,1.04)	0.89 (0.74,1.06)	0.90 (0.75,1.08)
	0.4	0.63 (0.53,0.76)	0.65 (0.54,0.78)	0.66 (0.55,0.79)	0.68 (0.57,0.81)	0.69 (0.58,0.83)	0.71 (0.59,0.85)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.65,0.93)	0.79 (0.66,0.95)	0.81 (0.67,0.96)	0.82 (0.68,0.98)	0.84 (0.70,1.00)	0.85 (0.71,1.02)	0.87 (0.72,1.04)	0.88 (0.74,1.06)
	0.45	0.62 (0.52,0.74)	0.63 (0.53,0.76)	0.65 (0.54,0.78)	0.67 (0.55,0.80)	0.68 (0.57,0.82)	0.70 (0.58,0.83)	0.71 (0.59,0.85)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.93)	0.79 (0.66,0.95)	0.80 (0.67,0.96)	0.82 (0.68,0.98)	0.84 (0.70,1.00)	0.85 (0.71,1.02)	0.87 (0.72,1.04)
	0.5	0.61 (0.51,0.73)	0.62 (0.52,0.74)	0.64 (0.53,0.76)	0.65 (0.54,0.78)	0.67 (0.56,0.80)	0.68 (0.57,0.82)	0.70 (0.58,0.84)	0.71 (0.59,0.85)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.93)	0.79 (0.66,0.94)	0.80 (0.67,0.96)	0.82 (0.68,0.98)	0.83 (0.70,1.00)	0.85 (0.71,1.02)
	0.55	0.59 (0.50,0.71)	0.61 (0.51,0.73)	0.62 (0.52,0.75)	0.64 (0.53,0.77)	0.65 (0.55,0.78)	0.67 (0.56,0.80)	0.68 (0.57,0.82)	0.70 (0.58,0.84)	0.71 (0.59,0.85)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.93)	0.79 (0.66,0.94)	0.80 (0.67,0.96)	0.82 (0.68,0.98)	0.83 (0.69,1.00)
	0.6	0.58 (0.49,0.70)	0.60 (0.50,0.72)	0.61 (0.51,0.73)	0.63 (0.52,0.75)	0.64 (0.53,0.77)	0.66 (0.55,0.79)	0.67 (0.56,0.80)	0.69 (0.57,0.82)	0.70 (0.58,0.84)	0.71 (0.60,0.86)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.93)	0.79 (0.66,0.94)	0.80 (0.67,0.96)	0.82 (0.68,0.98)
	0.65	0.57 (0.48,0.69)	0.59 (0.49,0.70)	0.60 (0.50,0.72)	0.61 (0.51,0.74)	0.63 (0.52,0.75)	0.64 (0.54,0.77)	0.66 (0.55,0.79)	0.67 (0.56,0.81)	0.69 (0.57,0.82)	0.70 (0.58,0.84)	0.72 (0.60,0.86)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.93)	0.79 (0.66,0.94)	0.80 (0.67,0.96)
	0.7	0.56 (0.47,0.67)	0.58 (0.48,0.69)	0.59 (0.49,0.71)	0.60 (0.50,0.72)	0.62 (0.51,0.74)	0.63 (0.53,0.76)	0.65 (0.54,0.77)	0.66 (0.55,0.79)	0.67 (0.56,0.81)	0.69 (0.57,0.82)	0.70 (0.59,0.84)	0.72 (0.60,0.86)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.92)	0.79 (0.66,0.94)
	0.75	0.55 (0.46,0.66)	0.57 (0.47,0.68)	0.58 (0.48,0.69)	0.59 (0.49,0.71)	0.61 (0.51,0.73)	0.62 (0.52,0.74)	0.63 (0.53,0.76)	0.65 (0.54,0.78)	0.66 (0.55,0.79)	0.68 (0.56,0.81)	0.69 (0.57,0.83)	0.70 (0.59,0.84)	0.72 (0.60,0.86)	0.73 (0.61,0.87)	0.74 (0.62,0.89)	0.76 (0.63,0.91)	0.77 (0.64,0.92)
	0.8	0.54 (0.45,0.65)	0.55 (0.46,0.66)	0.57 (0.47,0.68)	0.58 (0.49,0.70)	0.60 (0.50,0.71)	0.61 (0.51,0.73)	0.62 (0.52,0.75)	0.64 (0.53,0.76)	0.65 (0.54,0.78)	0.66 (0.55,0.79)	0.68 (0.56,0.81)	0.69 (0.58,0.83)	0.70 (0.59,0.84)	0.72 (0.60,0.86)	0.73 (0.61,0.88)	0.74 (0.62,0.89)	0.76 (0.63,0.91)

Table 5: Statistical and Clinical Tipping Points for Overall Survival Analysis (when HR for overall survival of those with and without confounder set to 2)

Trt Effect HR (95% CI)	Assumed prevalence of unobserved confounder in SCA																
Assumed prevalence of unobserved confounder in investigational arm	0.0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
0.0	0.76 (0.63,0.91)	0.80 (0.66,0.95)	0.83 (0.70,1.00)	0.87 (0.73,1.04)	0.91 (0.76,1.09)	0.95 (0.79,1.14)	0.99 (0.82,1.18)	1.02 (0.85,1.23)	1.06 (0.88,1.27)	1.10 (0.92,1.32)	1.14 (0.95,1.36)	1.17 (0.98,1.41)	1.21 (1.01,1.45)	1.25 (1.04,1.50)	1.29 (1.07,1.54)	1.33 (1.11,1.59)	1.36 (1.14,1.63)
0.05	0.72 (0.60,0.86)	0.76 (0.63,0.91)	0.79 (0.66,0.95)	0.83 (0.69,0.99)	0.87 (0.72,1.04)	0.90 (0.75,1.08)	0.94 (0.78,1.12)	0.97 (0.81,1.17)	1.01 (0.84,1.21)	1.05 (0.87,1.25)	1.08 (0.90,1.30)	1.12 (0.93,1.34)	1.16 (0.96,1.38)	1.19 (0.99,1.43)	1.23 (1.02,1.47)	1.26 (1.05,1.51)	1.30 (1.08,1.56)
0.1	0.69 (0.57,0.83)	0.72 (0.60,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.95)	0.83 (0.69,0.99)	0.86 (0.72,1.03)	0.90 (0.75,1.07)	0.93 (0.78,1.11)	0.96 (0.80,1.16)	1.00 (0.83,1.20)	1.03 (0.86,1.24)	1.07 (0.89,1.28)	1.10 (0.92,1.32)	1.14 (0.95,1.36)	1.17 (0.98,1.40)	1.21 (1.01,1.44)	1.24 (1.03,1.49)
0.15	0.66 (0.55,0.79)	0.69 (0.58,0.83)	0.73 (0.60,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.95)	0.82 (0.69,0.99)	0.86 (0.71,1.03)	0.89 (0.74,1.07)	0.92 (0.77,1.11)	0.96 (0.80,1.14)	0.99 (0.82,1.18)	1.02 (0.85,1.22)	1.05 (0.88,1.26)	1.09 (0.91,1.30)	1.12 (0.93,1.34)	1.15 (0.96,1.38)	1.19 (0.99,1.42)
0.2	0.63 (0.53,0.76)	0.66 (0.55,0.79)	0.69 (0.58,0.83)	0.73 (0.61,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.95)	0.82 (0.68,0.98)	0.85 (0.71,1.02)	0.88 (0.74,1.06)	0.92 (0.76,1.10)	0.95 (0.79,1.13)	0.98 (0.82,1.17)	1.01 (0.84,1.21)	1.04 (0.87,1.25)	1.07 (0.90,1.29)	1.11 (0.92,1.32)	1.14 (0.95,1.36)
0.25	0.61 (0.51,0.73)	0.64 (0.53,0.76)	0.67 (0.56,0.80)	0.70 (0.58,0.84)	0.73 (0.61,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.94)	0.82 (0.68,0.98)	0.85 (0.71,1.02)	0.88 (0.73,1.05)	0.91 (0.76,1.09)	0.94 (0.78,1.13)	0.97 (0.81,1.16)	1.00 (0.83,1.20)	1.03 (0.86,1.23)	1.06 (0.88,1.27)	1.09 (0.91,1.31)
0.3	0.58 (0.49,0.70)	0.61 (0.51,0.73)	0.64 (0.53,0.77)	0.67 (0.56,0.80)	0.70 (0.58,0.84)	0.73 (0.61,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.94)	0.82 (0.68,0.98)	0.85 (0.70,1.01)	0.87 (0.73,1.05)	0.90 (0.75,1.08)	0.93 (0.78,1.12)	0.96 (0.80,1.15)	0.99 (0.83,1.19)	1.02 (0.85,1.22)	1.05 (0.88,1.26)
0.35	0.56 (0.47,0.67)	0.59 (0.49,0.71)	0.62 (0.51,0.74)	0.65 (0.54,0.77)	0.67 (0.56,0.81)	0.70 (0.59,0.84)	0.73 (0.61,0.87)	0.76 (0.63,0.91)	0.79 (0.66,0.94)	0.81 (0.68,0.98)	0.84 (0.70,1.01)	0.87 (0.73,1.04)	0.90 (0.75,1.08)	0.93 (0.77,1.11)	0.95 (0.80,1.14)	0.98 (0.82,1.18)	1.01 (0.84,1.21)
0.4	0.54 (0.45,0.65)	0.57 (0.47,0.68)	0.60 (0.50,0.71)	0.62 (0.52,0.75)	0.65 (0.54,0.78)	0.68 (0.56,0.81)	0.70 (0.59,0.84)	0.73 (0.61,0.88)	0.76 (0.63,0.91)	0.79 (0.65,0.94)	0.81 (0.68,0.97)	0.84 (0.70,1.01)	0.87 (0.72,1.04)	0.89 (0.74,1.07)	0.92 (0.77,1.10)	0.95 (0.79,1.14)	0.97 (0.81,1.17)
0.45	0.52 (0.44,0.63)	0.55 (0.46,0.66)	0.58 (0.48,0.69)	0.60 (0.50,0.72)	0.63 (0.52,0.75)	0.65 (0.54,0.78)	0.68 (0.57,0.81)	0.71 (0.59,0.85)	0.73 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.94)	0.81 (0.68,0.97)	0.84 (0.70,1.00)	0.86 (0.72,1.03)	0.89 (0.74,1.06)	0.91 (0.76,1.10)	0.94 (0.78,1.13)
0.5	0.51 (0.42,0.61)	0.53 (0.44,0.64)	0.56 (0.46,0.67)	0.58 (0.48,0.70)	0.61 (0.51,0.73)	0.63 (0.53,0.76)	0.66 (0.55,0.79)	0.68 (0.57,0.82)	0.71 (0.59,0.85)	0.73 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.94)	0.81 (0.67,0.97)	0.83 (0.70,1.00)	0.86 (0.72,1.03)	0.88 (0.74,1.06)	0.91 (0.76,1.09)
0.55	0.49 (0.41,0.59)	0.51 (0.43,0.62)	0.54 (0.45,0.64)	0.56 (0.47,0.67)	0.59 (0.49,0.70)	0.61 (0.51,0.73)	0.64 (0.53,0.76)	0.66 (0.55,0.79)	0.68 (0.57,0.82)	0.71 (0.59,0.85)	0.73 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.94)	0.81 (0.67,0.97)	0.83 (0.69,1.00)	0.86 (0.71,1.03)	0.88 (0.73,1.05)
0.6	0.47 (0.39,0.57)	0.50 (0.41,0.60)	0.52 (0.43,0.62)	0.54 (0.45,0.65)	0.57 (0.47,0.68)	0.59 (0.49,0.71)	0.62 (0.51,0.74)	0.64 (0.53,0.77)	0.66 (0.55,0.79)	0.69 (0.57,0.82)	0.71 (0.59,0.85)	0.73 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.94)	0.81 (0.67,0.96)	0.83 (0.69,0.99)	0.85 (0.71,1.02)
0.65	0.46 (0.38,0.55)	0.48 (0.40,0.58)	0.51 (0.42,0.61)	0.53 (0.44,0.63)	0.55 (0.46,0.66)	0.57 (0.48,0.69)	0.60 (0.50,0.72)	0.62 (0.52,0.74)	0.64 (0.54,0.77)	0.67 (0.56,0.80)	0.69 (0.57,0.83)	0.71 (0.59,0.85)	0.74 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.94)	0.80 (0.67,0.96)	0.83 (0.69,0.99)
0.7	0.45 (0.37,0.53)	0.47 (0.39,0.56)	0.49 (0.41,0.59)	0.51 (0.43,0.61)	0.54 (0.45,0.64)	0.56 (0.46,0.67)	0.58 (0.48,0.69)	0.60 (0.50,0.72)	0.62 (0.52,0.75)	0.65 (0.54,0.77)	0.67 (0.56,0.80)	0.69 (0.58,0.83)	0.71 (0.59,0.85)	0.74 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.93)	0.80 (0.67,0.96)
0.75	0.43 (0.36,0.52)	0.45 (0.38,0.54)	0.48 (0.40,0.57)	0.50 (0.42,0.60)	0.52 (0.43,0.62)	0.54 (0.45,0.65)	0.56 (0.47,0.67)	0.58 (0.49,0.70)	0.61 (0.51,0.73)	0.63 (0.52,0.75)	0.65 (0.54,0.78)	0.67 (0.56,0.80)	0.69 (0.58,0.83)	0.71 (0.60,0.86)	0.74 (0.61,0.88)	0.76 (0.63,0.91)	0.78 (0.65,0.93)
0.8	0.42 (0.35,0.50)	0.44 (0.37,0.53)	0.46 (0.39,0.55)	0.48 (0.40,0.58)	0.51 (0.42,0.61)	0.53 (0.44,0.63)	0.55 (0.46,0.66)	0.57 (0.47,0.68)	0.59 (0.49,0.71)	0.61 (0.51,0.73)	0.63 (0.53,0.76)	0.65 (0.54,0.78)	0.67 (0.56,0.81)	0.69 (0.58,0.83)	0.72 (0.60,0.86)	0.74 (0.61,0.88)	0.76 (0.63,0.91)

CASE STUDY REFERENCES

Austin PC and Stuart EA. 2015a. Optimal full matching for survival outcomes: A method that merits more widespread use. Stat Med; 34: 3949–3967.

Austin PC and Stuart EA. 2015b. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine; 34(28):3661–3679.

Friends of Cancer Research whitepaper 2018: available online at https://www.focr.org/events/friends-cancer-research-annual-meeting-2018

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

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.

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.

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.

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.

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

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

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

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.