Decision Making and Error Control for External Controls: A Statistical Analysis Plan for Estimation of Parameters for Studies Evaluating Treatments for Advanced Non-Small Cell Lung Cancer

Devin Incerti, Michael Bretscher, Ray Lin, Chris Harbon

2021-05-05

Contents

1 Research question and objectives

1.1 Study background and rationale

Randomized clinical trials (RCTs) are considered the gold standard for measuring the efficacy of treatments in medicine. Randomization of treatment assignment ensures that treatment effects have a causal interpretation. However, in some cases, it may not be feasible or ethical to perform a RCT. For example, a small target population may make enrollment of patients difficult. One manifestation of this is precision oncology, where single-arm trials have been used for accelerated or breakthrough regulatory approval.

Since there is no concurrent control group, interpretation of efficacy is difficult in single-arm trials. Although real-world data (RWD) can be used to construct "external controls" (ECs), naive comparisons to the trial data can be misleading due to differences in the underlying populations. Propensity score methods and regression adjustment aim to adjust for these differences so that treatment effects can be reliably estimated.

Still, estimating causal treatment effects with RWD is challenging because treatment assignment is not randomized. Bias in observational studies often results from factors including, but not limited to, unmeasured confounding, selection bias, and measurement error. Yet, while observational studies typically report uncertainty due to sampling variation, they rarely attempt to quantify the additional variability and bias resulting from the use of observational data.

We have developed a framework and meta-analytic model that can be used to incorporate these additional sources of bias and variability into EC analyses. The framework allows for prediction of true treatment effects in a new study given (i) EC treatment effects in the new study (i.e. a comparison of the trial experimental arm with an EC arm) and (ii) historical data comparing EC treatment effects to benchmark RCT treatment effects (i.e., a comparison of the trial experimental arm to the trial control arm). This document describes the analyses that will be used to estimate model parameters for a particular use case in advanced non-small cell lung cancer (NSCLC).

1.2 Objectives

1. Estimate the parameters needed to apply the statistical framework in advancd NSCLC using historical RCTs and RWD
2. Illustrate use of the framework for hypothetical new trials in advanced NSCLC

2 Estimation of hazard ratios

Parameterization requires estimates of the log hazard ratio of the internal control arm relative to the EC arm and the log hazard ratio of the experimental arm in the trial relative to the EC arm. Point estimates and standard errors of these quantities will be estimated from Cox models including a single binary covariate indicating whether a patient is in the trial or the EC data source.

A propensity score approach will be used, and will proceed in two parts. In part 1, propensity score methods will be used to adjust for observable differences in the trial and EC populations. In part 2, the average treatment effect on the treated (ATT) will be estimated so that estimates are based on the trial population. The remainder of this section outlines the propensity score methodology.

2.1 Data sources

Estimation will be performed using phase II and phase III RCTs for treatment of patients with advanced NSCLC. To ensure that patient-level data is available, trials will be restricted to those conducted by Roche. EC cohorts will consist of patients from Flatiron EHR data and datasets combining the EC and trial data will be built using Roche's internal `ecdata R` package. Table 1 provides a list of the complete analysis sample including the ClinicalTrials.gov ID numbers and the relevant experimental and comparator arms.

2.2 Outcome

The outcome is overall survival. Patients in the EC cohort will be right censored at the time of last follow-up from the trial.

2.3 Variable selection

The covariates included in the propensity score model will be based on those included in Carrigan et al. [2020]:

- Age
- Race (White, Black, Other)
- Sex
- Histology (Non-squamous, Squamous)
- Smoking status (Current/former, never)

Table 1: Analysis sample for parameter estimation

Number	Clinical trial	External control	Experimental	Comparator
1	NCT02008227	Flatiron	Atezolizumab	Docetaxel
2	NCT01903993	Flatiron	Atezolizumab	Docetaxel
3	NCT02366143	Flatiron	Atezolizumab + Bevacizumab + Carboplatin	Bevacizumab + Carboplatin
4	NCT01351415	Flatiron	Bevacizumab + SOC	SOC
5	NCT01519804	Flatiron	MetMAb + Platinum + Paclitaxel	Placebo + Platinum + Paclitaxel
6	NCT01496742	Flatiron	MetMAb + Bevacizumab + Platinum + Paclitaxel	Placebo + Bevacizumab + Platinum
7	NCT01496742	Flatiron	MetMAb + Platinum + Pemetrexed	Placebo + Platinum + Pemetrexed
8	NCT01366131	Flatiron	MEGF0444A + Bevacizumab + Carboplatin + Paclitaxel	Placebo + Bevacizumab + Carboplatin + Paclitaxel
9	NCT01493843	Flatiron	Pictilisib (340 mg) + Carboplatin + Paclitaxel	Placebo (340 mg) + Carboplatin + Paclitaxel
10	NCT01493843	Flatiron	Pictilisib (340 mg) + Carboplatin + Paclitaxel + Bevacizumab	Placebo (340 mg) + Carboplatin + Paclitaxel + Bevacizumab
11	NCT01493843	Flatiron	Pictilisib (260 mg) + Carboplatin + Paclitaxel + Bevacizumab	Placebo (260 mg) + Carboplatin + Paclitaxel + Bevacizumab
12	NCT02367781	Flatiron	Atezolizumab + Nab-Paclitaxel + Carboplatin	Nab-Paclitaxel + Carboplatin
13	NCT02367794	Flatiron	Atezolizumab + Nab-Paclitaxel + Carboplatin	Nab-Paclitaxel + Carboplatin
14	NCT02657434	Flatiron	Atezolizumab + Carboplatin or Cisplatin + Pemetrexed	Carboplatin or Cisplatin + Pemetrexed

- Cancer stage at initial diagnosis (Advanced - IIIB/IV, Early - IIIA or below)
- Time since initial diagnosis

Race categories may be collapsed into an "Other" category if sample sizes are too small (< 10 observations). Furthermore, race will be coded as Asian or Non-Asian for comparisons with NCT01351415 because race was coded in that manner in the trial. Histology will only be included if histology was not part of the inclusion and exclusion criteria for the trials. Variables will be excluded for a particular trial if they were not collected for that trial. If balance (see below) is deemed inadequate, interaction terms and nonlinear functions of continuous covariates will be considered: Specifically, age and time since initial diagnosis will be modeled with restricted cubic splines using 3 knots and time since initial diagnosis will be interacted with cancer stage at initial diagnosis.

2.4 Missing data

Missing data will be imputed using multivariate imputation by chained equations (MICE) [Buuren and Groothuis-Oudshoorn, 2010]. This multiple imputation approach has typically resulted in lower bias and variance than other methods when the data is missing at random (MAR) [White et al., 2011, Choi et al., 2019, Leyrat et al., 2019].

There are two options when performing a propensity score analysis on multiply imputed data: first, treatment effect estimates can be combined across datasets, and second, treatment effects can be estimated after combining the propensity score. We will use the former approach given that simulation evidence suggests it produces unbiased estimates and appropriate confidence intervals, while the latter does not (Leyrat et al. [2019]; Granger et al. [2019]). That is, for each imputed dataset, we (i) estimate the propensity score and (ii) estimate treatment effects given the estimated propensity score. Pooled point estimates and confidence intervals will be estimated by combining the treatment effects from each of the imputed datasets using Rubin's rule.

2.5 Propensity score estimation

Logistic regression will be used to estimate the propensity score using the covariates described in Section 2.3.

2.6 Inverse probability weighting

Inverse probability of treatment weighting (IPTW) permitting estimation of the ATT—commonly referred to as IPTW-ATT weighting—will be used. Trial patients receive a weight of 1 while EC patients receive a weight of $e/(1-e)$ where e is the propensity score. In other words, after weighting, the distribution of covariates in the EC is the same as in the trial.

2.7 Trimming

Propensity scores close to 0 or 1 can result in extreme weights. The impact of large weights can be reduced by "trimming", which can either refer to truncation of large weights downward (or small weights upward) or to exclusion of patients with extreme weights. We will use the latter approach and remove EC patients with propensity scores greater than the 99th percentile or less than the 1st percentile.

2.8 Assessment of balance

Baseline demographic and clinical characteristics will be compared across the trial and EC patients pre- and post-matching. Three types of diagnostics will be used. First, standardized mean differences (SMDs) will be computed for each covariate and displayed graphically. Thresholds of 0.1 [Nguyen et al., 2017] and 0.25 [Ho et al., 2007] will be used as visual aids for assessing balance, although we note that these are somewhat arbitrary. Second, density plots of the distributions of the propensity score will be displayed. Third, density plots will be provided for each continuous covariate since researchers have argued that balance diagnostics should extend beyond comparisons of means to comparisons of higher order moments [Imai et al., 2008, Austin, 2009].

2.9 Hazard ratios

Hazard ratios will be estimated using IPTW-ATT weighted Cox models. Models will be fit without covariate adjustment; that is, they will only include a single covariate for treatment assignment. This approach facilitates estimation of a marginal hazard ratio [Daniel et al., 2020], which is typically the estimand of interest from a RCT.

Bootstrapping will be used to estimate the variances of the point estimates. The entire propensity score methodology including estimation of the propensity score and estimation of hazard ratios using the Cox model will be implemented during each bootstrap sample.

3 Sensitivity analyses

3.1 Trimming

Estimates will be reported with and without trimming.

3.2 Variable selection

Two alternative specifications of the propensity score model will be used to assess sensitivity of the results to the "quality" of the propensity score methodology. First, hazard ratios will be estimated using unweighted Cox models so that no population adjustment is performed. Second, a less discriminating propensity score will be estimated using only a single variable, age.

3.3 Double adjustment

To facilitate simple estimation of marginal hazard ratios, the Cox models in the primary analyses do not adjust for baseline covariates. However a large literature dating back to Rubin [1973] has shown that regression combined with propensity score methods results in greater reduction in bias than when using either method alone, particularly when imbalance persists after population adjustment. For example, simulation evidence from Nguyen et al. [2017] suggests that adjusting for covariates with standardized differences greater than 0.1 after propensity score adjustment can remove residual confounding bias. Furthermore, even in RCTs, covariate adjustment increases power [Daniel et al., 2020]. A sensitivity analysis will consequently estimate hazard ratios with regression adjustment (i.e., by including all covariates used in the propensity score model in the Cox model).

3.4 Propensity score method

A number of propensity score methods have been suggested in the literature in addition to IPTW to adjust for differences in the treated and control populations. These include matching, stratification on the propensity score, and inclusion of the propensity score as a covariate. Simulation work has generally shown the weighting and matching are able to estimate conditional and marginal treatment effects with the least bias and smallest mean squared error [Austin, 2007, 2013]. The sensitivity analyses will consequently be limited to matching.

The first matching algorithm that will be evaluated is greedy 1:1 nearest neighbor matching using the linear propensity score. For a given treated subject, 1:1 nearest neighbor matching will select the control subject whose value of the linear propensity score is closest; that is, for a treated subject i, the control subject j with the minimum value of $D_{ij} = |logit(e_i) - logit(e_j)|$ will be chosen where e_k is the propensity score for subject k. If multiple control subjects are equally close to a treated subject, then the treated subject is chosen at random. Greedy matching implies that control subjects are chosen one at a time and the order in which they are chosen matters. Matching will be performed without replacement (except in cases where the number of control subjects is less than the number of treated subjects) to ensure that the matched controls are independent, although it is worth noting that there is given evidence that matching with replacement can reduce bias [Abadie and Imbens, 2006, Stuart, 2010].

A challenge of nearest neighbor matching—and propensity core methods more generally—is that it is difficult to specify a model that achieves satisfactory covariate balance. An iterative process of fitting the model, assessing balance, and respecifying the model has often been recommended [Rosenbaum and Rubin, 1984, Austin, 2008, Belitser et al., 2011]. Genetic matching can help overcome some of these challenges since the weight given to each covariate is based on an evolutionary search algorithm that iteratively checks and improves covariate balance [Sekhon, 2008, Diamond and Sekhon, 2013]. We will employ this algorithm using both the linear propensity score and all the covariates specified above, so that matching on the propensity score and the Mahalanobis distance are limiting cases. We will continue to use 1:1 matching given evidence from the simulation study by Austin [2010] showing that mean square error is typically minimized when matching either 1 or 2 controls to each treated subject.

Both algorithms will be performed without a caliper and without a caliper. Following the advice of Rosenbaum and Rubin [1985]—and the general consensus in the field—analyses will be performed with a caliper on the linear propensity score of 0.25 standard deviations. Note that calipers may remove trial patients and change the estimand.

4 Validation

Validation of the methodology will be performed using repeated k-fold cross validation. We will set $k = 5$ so that 80% of the data ($n = 12$) is used for parameter estimation and 20% ($n = 3$) is used for testing for each of the 5 splits of the data. The process will be repeated 10 times to reduce dependence on the chosen partitions.

The training data will be used to estimate all parameters of the meta-analytic model. The parameterized meta-analytic model will, in turn, be used to estimate true treatment effects for each of the studies in the test data; that is, EC treatment effects will be estimated for each of the studies in the test data and true treatment effects will be predicted conditional on the estimated EC treatment effects.

Performance will be evaluated by averaging across the 5*10=50 cross-validation iterations. The predicted true treatment effects will be compared to the estimated RCT treatment effects. Evaluation metrics will include bias and 95% coverage probabilities.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Peter C Austin. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine*, 26(16):3078–3094, 2007.

Peter C Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12):2037–2049, 2008.

Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

Peter C Austin. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American journal of epidemiology*, 172(9):1092–1097, 2010.

Peter C Austin. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849, 2013.

Svetlana V Belitser, Edwin P Martens, Wiebe R Pestman, Rolf HH Groenwold, Anthonius De Boer, and Olaf H Klungel. Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and drug safety*, 20(11):1115–1129, 2011.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

Gillis Carrigan, Samuel Whipple, William B Capra, Michael D Taylor, Jeffrey S Brown, Michael Lu, Brandon Arnieri, Ryan Copping, and Kenneth J Rothman. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. *Clinical Pharmacology & Therapeutics*, 107(2):369–377, 2020.

Jungyeon Choi, Olaf M Dekkers, and Saskia le Cessie. A comparison of different methods to handle missing data in the context of propensity score analysis. *European journal of epidemiology*, 34(1):23–36, 2019.

Rhian Daniel, Jingjing Zhang, and Daniel Farewell. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 2020.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3): 932–945, 2013.

Emily Granger, Jamie C Sergeant, and Mark Lunt. Avoiding pitfalls when combining multiple imputation and propensity scores. *Statistics in medicine*, 38(26):5120–5132, 2019.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

Kosuke Imai, Gary King, and Elizabeth A Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171 (2):481–502, 2008.

Clémence Leyrat, Shaun R Seaman, Ian R White, Ian Douglas, Liam Smeeth, Joseph Kim, Matthieu Resche-Rigon, James R Carpenter, and Elizabeth J Williamson. Propensity score analysis with partially observed covariates: how should multiple imputation be used? *Statistical methods in medical research*, 28 (1):3–19, 2019.

Tri-Long Nguyen, Gary S Collins, Jessica Spence, Jean-Pierre Daurès, PJ Devereaux, Paul Landais, and Yannick Le Manach. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC medical research methodology*, 17(1):1–8, 2017.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203, 1973.

Jasjeet S Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*, 2008.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.