


ORIGINAL ARTICLE

Matching methods in precision oncology: An introduction and illustrative example

Deirdre Weymann¹  | Janessa Laskin^{2,3} | Steven J.M. Jones^{4,5} | Howard Lim^{2,3} | Daniel J. Renouf^{2,3} | Robyn Roscoe⁴ | Kasmintan A. Schrader^{5,6} | Sophie Sun^{2,3} | Stephen Yip^{7,8} | Marco A. Marra^{4,5} | Dean A. Regier^{1,9}

¹Canadian Centre for Applied Research in Cancer Control, Cancer Control Research, BC Cancer, Vancouver, BC, Canada

²Division of Medical Oncology, BC Cancer, Vancouver, BC, Canada

³Department of Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

⁴Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada

⁵Department of Medical Genetics, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

⁶Department of Molecular Oncology, BC Cancer, Vancouver, BC, Canada

⁷Department of Pathology & Laboratory Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

⁸Department of Pathology, BC Cancer, Vancouver, BC, Canada

⁹School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

Correspondence

Dean A. Regier, BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC, Canada, V5Z 1L3. Email: dregier@bccrc.ca

FUNDING INFORMATION

This research was supported by Genome British Columbia (B20POG) and Genome British Columbia/Genome Canada (G05CHS).

ABSTRACT

Background: Randomized controlled trials (RCTs) are uncommon in precision oncology. We provide an introduction and illustrative example of matching methods for evaluating precision oncology in the absence of RCTs. We focus on British Columbia's Personalized OncoGenomics (POG) program, which applies whole-genome and transcriptome analysis (WGTA) to inform advanced cancer care.

Methods: Our cohort comprises 230 POG patients enrolled between 2014 and 2015 and matched POG-naïve controls. We generated our matched cohort using 1:1 propensity score matching (PSM) and genetic matching prior to exploring survival differences.

Results: We find that genetic matching outperformed PSM when balancing covariates. In all cohorts, overall survival did not significantly differ across POG and POG-naïve patients ($p > 0.05$). Stratification by WGTA-informed treatment indicated unmatched survival differences. Patients whose WGTA information led to treatment change were at a reduced hazard of death compared to POG-naïve controls in all cohorts, with estimated hazard ratios ranging from 0.33 (95% CI: 0.13, 0.81) to 0.41 (95% CI: 0.17, 0.98).

Conclusion: These results signal that clinical effectiveness of precision oncology approaches will depend on rates of genomics-informed treatment change. Our study will guide future evaluations of precision oncology and support reliable effect estimation when RCT data are unavailable.

KEYWORDS

administrative data, genomic sequencing, matching, precision medicine, quasi-experimental methods

1 | INTRODUCTION

Precision oncology promises to improve health outcomes by tailoring treatments to an individual's own genomic profile. To date, the available evidence for genomic testing focuses on clinical validity and feasibility rather than effectiveness (Burke et al., 2006; Khoury et al., 2007). The development of a robust evidence base supporting the idea of precision oncology has been limited by a lack of randomized controlled trials (RCTs; Schork, 2015).

RCTs are the “gold standard” in efficacy research (Silverman, 2009), but are uncommon in precision oncology primarily because the genomic-level heterogeneity shaping patients' treatment responses is difficult to account for in trial designs (Schork, 2015). N-of-1 trials are a proposed solution but their validity strongly depends on the assumption of clinical stability, which is often violated in oncology (Collette & Tombal, 2015). Combined, these challenges limit the availability of reliable counterfactuals. In the absence of RCT data, researchers are beginning to use administrative data to evaluate precision oncology (Weymann et al., 2018).

Owing to non-randomization, observational studies of precision oncology may be confounded by subjects' characteristics, introducing selection bias into effect estimates. Quasi-experimental matching methods combined with population-based administrative data can begin to meet the challenge of producing reliable estimates. Past research has found that well-designed observational studies produce similar results to RCTs and that matching is a valid tool for generating causal effectiveness estimates (Anglemyer et al., 2014; Benson & Hartz, 2000; Golder et al., 2011; Lonjon et al., 2014; Stuart, 2010).

Our study provides an introduction to matching methods in the context of precision oncology. We outline two methodological approaches and illustrate how matching can be used to mitigate selection bias when estimating the health impacts of omics-guided cancer care. Our study setting is the BC Cancer Personalized OncoGenomics (POG) program, a

single-group research study using whole-genome and transcriptome analysis (WGTA) to guide treatment planning for patients with advanced stage cancers (Laskin et al., 2015).

1.1 | Matching overview

Matching methods are used to minimize or eliminate selection bias resulting from confounding. While several methods exist, propensity score matching (PSM) is among the most common (Stuart, 2010). PSM estimates a patient's underlying probability (or propensity score) of receiving treatment. Individuals in the treatment group are then matched with controls who have similar propensity scores (Rosenbaum & Rubin, 1983). Other methods to adjust for propensity scores are available, including stratification, weighting, and regression adjustment. Compared to these methods, matching performs relatively well when balancing baseline covariates and generating effect estimates (Austin, 2009, 2013). Matching also allows for data reduction when treatment is rare, which is a common characteristic of precision oncology.

The application of PSM involves a series of steps and each step requires a series of considerations summarized in Table 1. Steps include: (a) specifying the propensity score model; (b) determining the matching method and algorithm; (c) assessing covariate balance; and (d) repeating steps 1 to 3 until balance on key covariates is achieved.

1.1.1 | Step 1 – Specifying the propensity score model

PSM begins with estimating a regression model for a binary dependent variable indicating treatment. When treatments are provided at more than one time point, researchers must assess whether a patient's probability of treatment is likely to change over time. In oncology, patients often begin first-line treatment soon after diagnosis and advance to new treatment

Steps	Relevant considerations
1. Specify propensity score model	<ul style="list-style-type: none"> - Time constant vs. varying probability - Model type (e.g., logistic) - Covariate selection
2. Determine matching method and algorithm	<ul style="list-style-type: none"> - Nearest neighbor vs. optimal matching - Ratio matching - Matching with ties and/or replacement - Caliper widths
3. Assess covariate balance	<ul style="list-style-type: none"> - Standardized differences <0.10 - $0.50 < \text{variance ratios} < 2.00$ - Empirical quantile–quantile plots - Nonparametric hypothesis tests
4. Repeat steps 1 to 3 until balance on key covariates achieved	

TABLE 1 Steps and considerations for propensity score matching.

lines following disease progression, relapse, or toxicity, resulting in different treatment initiation dates across patients. If the probability that patients begin a treatment is constant over time and unlikely to be affected by time-varying covariates, a regression model that produces a time-constant propensity score should be estimated, the most common being logistic regression (Austin, 2011a). If this probability is likely to vary, a model that produces a time-varying score is more appropriate, such as a Cox proportional hazards model (Lu, 2005). Time-varying models may become increasingly relevant if precision oncology becomes widely applied and the probability of accessing omics-guided treatment changes over time.

Following the decision on type of regression model, researchers determine which covariates to include in the model. PSM critically assumes there are no unobserved differences across treated and control patients conditional on the propensity score, termed “ignorability.” To meet this assumption, all covariates that correlate with both the probability of treatment and the final outcome probability must be modeled (Rubin & Thomas, 1996). After fitting the propensity score model and assessing model fit, researchers estimate propensity scores for each individual and assess overlap of scores across treatment and control groups. Overlap refers to the area of common support across propensity score distributions.

1.1.2 | Step 2 – Determining the matching method and algorithm

Once satisfied with the level of overlap across groups, researchers must decide on a matching method and algorithm. Common methods include nearest neighbor matching and optimal matching. Nearest neighbor matching selects the control individual with the smallest distance in propensity score from the treated individual. Optimal matching minimizes a global rather than individual distance. Compared to nearest neighbor matching, optimal matching produces more closely matched pairs but does not improve overall balance across matched cohorts (Austin, 2014; Gu & Rosenbaum, 1993).

When deciding between matching algorithms, researchers can consider ratio matching, matching with or without replacement or ties, and caliper widths. Ratio matching selects the number of matched controls for each treated individual. Ratios range from one-to-one matching, where each treated patient is matched to a single control, to many-to-one matching, to variable ratio matching, where the number of matched controls varies across treated patients (Ming & Rosenbaum, 2001). The decision around number of allowable matches involves a trade-off between bias and statistical efficiency. As a result, the literature typically recommends matching either one or, at most, two controls to each treated patient (Austin, 2010).

Matching with replacement allows the same control to be matched to multiple treated patients. This approach can improve the average quality of matches, reduce bias, and is useful when few controls are similar to treated patients (Caliendo & Kopeinig, 2008; Stuart, 2010). Matching with ties allows multiple controls with equal propensity scores to be matched to a treated patient. The alternative, to randomly select a single control patient from those with tied values, is generally not recommended because it will underestimate the variance of the final outcome variable (Sekhon, 2011). If matching with replacement or ties, the data should be weighted to avoid false imprecision.

Caliper widths specify a maximum distance between propensity scores for a treated individual and their matched controls. Optimal caliper widths vary across applications. When at least some of the covariates are continuous, Austin (2011b) recommends using a caliper width equal to 0.2 of the standard deviation of the logit of the propensity score. When all covariates are binary, estimation performance is less sensitive to the choice of caliper width.

1.1.3 | Step 3 – Assessing covariate balance

The propensity score is a balancing score (Rosenbaum & Rubin, 1983). After matching on the true score, baseline characteristics will asymptotically have the same distribution across treated and control patients. If the propensity score model is misspecified, covariates in the matched sample will be imbalanced leading to biased estimates. It is critical to assess balance of entire covariate distributions, through standardized differences in means and medians, variance ratios, empirical quantile–quantile (QQ) plots, higher order, and interaction terms (Austin, 2011a). While hypothesis tests, such as paired *t*-tests or Kolmogorov–Smirnov (KS) tests are often used to assess differences, conclusions can be misleading because matching reduces sample sizes and increases *p*-values (Imai et al., 2008; Kolmogorov, 1933; Smirnov, 1939).

To assess whether balance on key covariates is achieved, “rules of thumb” are often used. Standardized differences less than 10 and variance ratios close to 1 are taken to suggest no evidence of imbalance (Cohen, 1977; Rubin, 2001). Standardized differences exceeding 10 and variance ratios outside the range of 0.5 to 2 indicate imbalance. The literature generally recommends minimizing imbalance without limit (Imai et al., 2008). If there remains any evidence of imbalance on key variables after matching, the propensity score model and/or matching algorithm can be respecified. Given that the propensity score is a function of covariates rather than outcomes, repeated analyses attempting to balance covariate distributions across treated and control patients will not bias final effect estimates (Rubin, 2001). The iterative process of PSM can be time consuming

and covariate distributions may never balance. Even when balance is achieved, it may not be maximized.

1.1.4 | Optimizing balance with genetic matching

Genetic matching automates the process of maximizing balance on observed covariates through supervised learning (Diamond & Sekhon, 2013). As a generalized form of Mahalanobis distance matching, this method matches individuals based on their weighted Mahalanobis distance rather than the difference between their propensity scores (Mahalanobis, 1936). Through the use of an evolutionary genetic search algorithm (Sekhon & Mebane, 1998), genetic matching iteratively checks different weights for each covariate when calculating the generalized Mahalanobis distance and considers balance statistics at each iteration. This algorithm selects the final weights to minimize baseline differences across groups according to pre-specified optimization criteria. Most steps for genetic matching are similar to PSM, with the exception of selecting criteria for optimization. For any criteria specified, genetic matching will asymptotically converge to the optimal matched cohort.

1.2 | Analysis after matching

Following matching, researchers can estimate an intervention's impacts using nonparametric and parametric analyses, including regression (Stuart, 2010). This two-step approach involving matching as a data preprocessing step can result in effect estimates that are less sensitive to modeling choices and final specifications (Ho et al., 2007). There is some debate around whether final outcomes analysis must account for the matched nature of the data and whether variance estimation must account for propensity score estimation and the matching procedure (Stuart, 2008, 2010). If outcomes are analyzed across treatment groups as a whole rather than across individual matched pairs, researchers typically do not adjust their analytic methods. If variance estimation does not account for the initial matching step, standard errors will be overestimated and confidence intervals will be relatively conservative. If desired, standard bootstrapping can be used to account for the uncertainty associated with matching, but these methods are invalid when matching with replacement (Abadie & Imbens, 2008; Austin & Small, 2014). To enable consistent standard error estimation in these scenarios, additional methods development is required.

1.3 | Application to precision oncology

In July 2012, POG began to explore the feasibility of integrating WGTA data into clinical decision-making (Laskin

et al., 2015). After July 2014, the program expanded to offer WGTA to a larger number of BC patients diagnosed with advanced cancers. POG generates WGTA data from a patient's fresh tumor biopsy sample (frequently a metastatic site) and genome sequence data from matched normal DNA. Clinical and pathology information are combined with genomic data to produce a report of genomic aberrations as well as candidate pathways dysregulated in patient's tumors. The analysis reveals detailed information on aberrant genes and related biological processes that may underpin malignant progression. One objective of POG-initiated WGTA is to provide clinicians with insight into a patient's tumor genome and transcriptome such that potential therapeutic targets and resistance mechanisms can be identified and optimal treatment options can be considered. Whether POG's approach results in measurable survival benefit remains unexamined, largely owing to the lack of an available control cohort and the heterogeneity of diseases considered. In the following case study, we illustrate how matching methods and population-based administrative data can be used to generate a control cohort for POG's evaluation. Our analytic approach is depicted in Figure 1.

2 | MATERIALS AND METHODS

2.1 | Ethical compliance

Our study was approved by the University of British Columbia-BC Cancer Research Ethics Board.

2.2 | Data sources

We based our retrospective analysis on de-identified administrative data sets provided by the BC Cancer Registry. We obtained data for adult patients diagnosed with cancers of varying histologies in BC. Our study cohort comprised adults who participated in POG between July 2014 and December 2015 and matched POG-naive controls. POG inclusion criteria are: (a) metastatic disease considered incurable by the oncologist; (b) good performance status; and (c) life expectancy greater than 6 months. POG-naive controls were eligible for matching after receiving systemic therapy treatment for advanced stage disease, indicated by BC Cancer systemic therapy protocol codes.

We identified eligible control patients from the BC Cancer Registry, a population-based provincial cancer registry recording disease, demographic, and mortality information for all cancer diagnoses in BC. We identified POG patients from the BC Cancer Outcomes and Surveillance Integration System POG Module Database. To assess treatment history, we obtained prescription records containing information

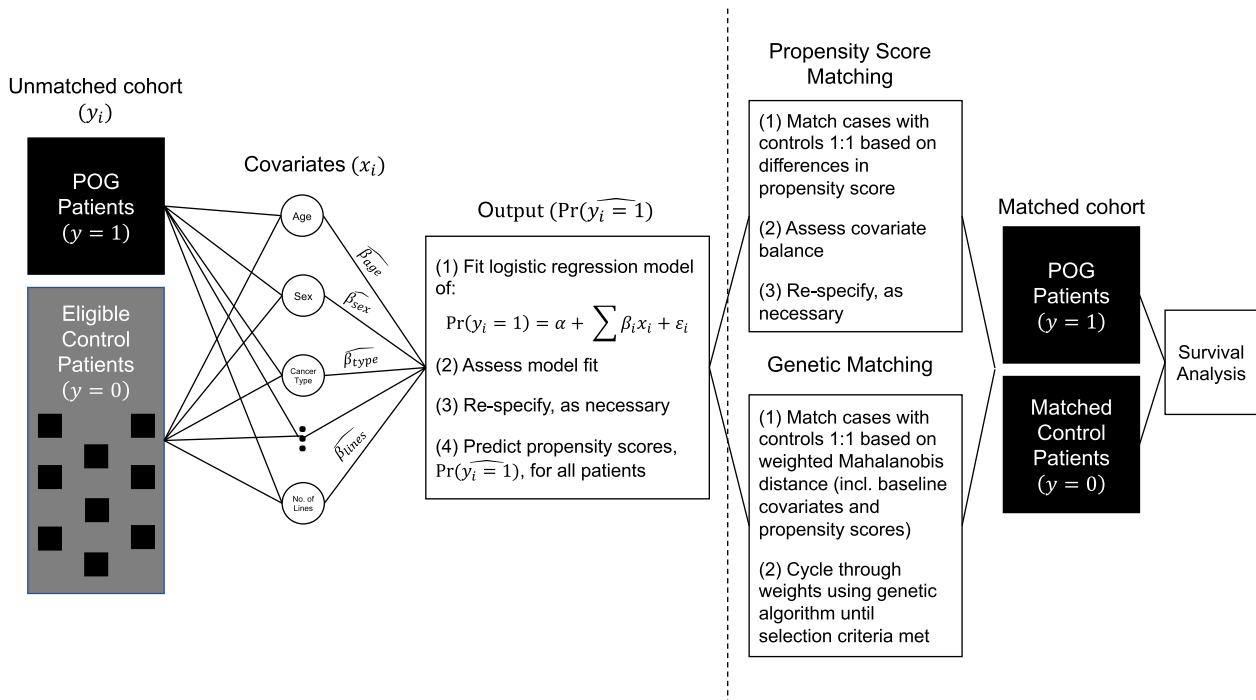


FIGURE 1 Overview of analytic approach.

about drug type, dispensation date, and protocol code for systemic therapy drugs dispensed by BC Cancer pharmacies. In BC, all approved systemic therapy treatments administered in regional cancer centers, community hospitals, or taken at home are dispensed by BC Cancer pharmacies.

2.3 | Control selection

Cases (POG patients) and controls (POG-naive patients) were matched based on the index date, which was the date of POG biopsy. Biopsy date indicates when patients who have consented to POG begin undergoing WGTA. We excluded 60 patients who consented to POG but either withdrew or were ineligible for the study (21% of patients). For the index date, each control was randomly assigned a pseudo-biopsy date, as described in Figure S1. To ensure sufficient overlap across groups, we excluded patients diagnosed with stage 0, benign, noninvasive, or in situ cancers, and patients whose performance status indicated they were completely disabled at initial diagnosis (0.01% of patients).

To match each case to a control, we used logistic regression models to calculate propensity scores estimating each patient's probability of participating in POG. Models were adjusted for a selection of baseline covariates hypothesized to correlate with both probability of participating in POG and mortality, including patient demographics, clinical characteristics, and treatment histories at the index date. Genetic testing history and mutational status do not influence POG eligibility and were not included in our propensity score

models. Specifications for our covariates, including category levels, were chosen to maximize model fit according to Akaike information criterion (Akaike, 1974). Squared terms were included in regression models to allow for nonlinear relationships between continuous variables and the probability of participating in POG.

Owing to the high degree of variation across patients diagnosed with breast cancer and those diagnosed with other cancers, we stratified propensity score models and matching analyses. Among patients with breast cancer, the regression model accounted for index date, age, rurality, year of diagnosis, grade at initial diagnosis, cancer stage at initial diagnosis, and number of lines of systemic therapy treatment received prior to index date. Among patients with other cancers, the regression model also accounted for sex, primary cancer site, and performance status at initial diagnosis. In the absence of manually coded number of lines of therapy data for the cohort, we applied an automated algorithm validated for use in administrative prescription drug data (Weymann et al., 2019).

We applied two matching techniques: 1:1 nearest neighbor matching on propensity scores and 1:1 genetic matching on propensity scores and baseline covariates. We selected final regression models and matching algorithms that maximized balance of baseline covariates, quadratic terms, and relevant interaction terms across cases and controls. We compared balance across matched and unmatched cohorts using standardized mean differences, variance ratios, QQ plots, bootstrapped KS tests in continuous variables, and paired *t*-tests in binary variables. Further details are available in Supplemental Material.

TABLE 2 Demographic and clinical characteristics.

Characteristics	No. (%) of cases (n = 230)	No. (%) of eligible unmatched controls (n = 5,224)	No. (%) of propensity score matched controls (n _{weighted} = 230)	No. (%) of genetic matched controls (n _{weighted} = 230)
Sex, female	141 (61.3)	2,899 (55.5) [~]	51.5 (65.9)	143 (62.2)
Age at index date, mean (SD)	56.2 (SD=12.8)	66.4 (SD=12.2) ^{*α}	55.8 (SD=14.1)	56.5 (SD=11.4)
Rurality				
Urban	182 (79.1)	3,141 (60.1) ^{*α}	179 (77.8)	184 (80.0)
Rural	36 (15.7)	1,581 (30.3) ^{*α}	39.5 (17.2)	35 (15.2)
Mixed	10 (4.3)	414 (7.9) ^{*~}	10.5 (4.6)	9 (3.9)
LHA missing	2 (0.9)	88 (1.7)	1 (0.4)	2 (0.9)
Primary cancer site				
Gastrointestinal	69 (30.0)	1,185 (22.7) ^{*~}	64.5 (28.0)	71 (30.9)
Breast	49 (21.3)	931 (17.8)	49 (21.3)	49 (21.3)
Lung	28 (12.2)	721 (13.8)	25.5 (11.0)	26 (11.3)
Pancreas	20 (8.7)	173 (3.3) ^{*~p}	16 (7.0)	20 (8.7)
Other	64 (27.8)	2,214 (42.4) ^{*~}	75 (32.6) [~]	64 (27.8)
Year of diagnosis, mean (SD)	2012.0 (SD=4.4)	2011.1 (SD=0.8) ^{*~}	2011.9 (SD=3.3) [*]	2012.2 (SD=3.6)
Stage at initial diagnosis				
Stage I	21 (9.1)	310 (5.9) [~]	34 (7.4)	21 (9.1)
Stage II	15 (6.5)	421 (8.1)	22 (9.6) [~]	17 (7.4)
Stage III	13 (5.7)	272 (5.2)	6 (2.6) ^{~p}	13 (5.7)
Stage IV	45 (19.6)	517 (9.9) ^{*α}	41 (17.8)	42 (18.3)
REC, UNK, NCR	136 (59.1)	3,704 (70.9) ^{*α}	127 (62.6)	137 (59.6)
Number of lines prior to index date, mean (SD)	1.6 (SD=1.2)	1.8 (SD=1.3) ^{*~}	1.5 (SD=1.0) [*]	1.7 (SD=1.1) [*]

Counts and frequencies reported for categorical variables. Means and standard deviations reported for continuous variables. Differences from cases are statistically significantly different at p -value < 0.05^{*}, < 0.10[~] (bootstrapped Kolmogorov–Smirnov tests, paired t -tests). Standardized differences are > 0.10[~], > 0.20^α. Variance ratio < 0.50 or > 2.00. No., number; REC, UNK, NCR, recurrent, stage unknown, or no classification recommended; SD, standard deviation.

2.4 | Survival analysis

Survival was estimated from the index date in matched and unmatched cohorts. All analyses accounted for weights related to our matching structure and censoring resulting from varying index dates and lengths of follow-up until the end point of interest, death. We estimated Kaplan–Meier (KM) survival functions and used log rank tests to assess differences (Kaplan & Meier, 1958). We identified statistical significance using a threshold of $p < 0.05$. One-year survival rates were inferred based on estimated KM survival functions. We also estimated unadjusted, Weibull regression models of probability of mortality to determine the hazard ratio (HR) associated with POG versus POG-naïve care. Subgroup analysis according to whether POG patients received a treatment change based on their WGTA results was applied to explore differences in overall survival patterns. For simplicity, we present pooled survival analyses. Analyses stratified according to cancer type (breast cancer vs. other cancers) and covariate-adjusted Weibull regression results are available in Figures S2 and S3 and Table S7, respectively. Adjusting for covariates for which some evidence of imbalance remained resulted in no material differences in study findings or conclusions.

To estimate average 1 year survival time in the presence of incomplete follow-up data, we applied inverse probability of censoring weighting (Bang & Tsiatis, 2000; Lin, 2003; Willan et al., 2005). Inverse probability weighting reduces estimation bias by recreating the sample population we would expect to see in the absence of censoring. We estimated inverse probability weights for each 1-month time interval using KM product limit estimates of probability of censoring. We used weighted regression methods to generate mean estimates and applied nonparametric bootstrapping to simulate corresponding sampling distributions. We conducted all statistical analyses in R and Stata 15 (Team, 2017; StataCorp, 2017).

3 | RESULTS

3.1 | Matching

From July 2014 to December 2015, 230 patients participated in the POG program. We identified 5,224 POG-naïve patients as eligible for matching. Table 2 summarizes the demographic and clinical characteristics of the pooled unmatched and matched cohorts after weighting. We report detailed balance statistics in Supplemental Material. Prior to matching, 32 of the 54 covariates examined showed evidence of imbalance (Table S2). Compared to unmatched control patients, a higher proportion of POG patients were women, lived in urban areas, were diagnosed with either gastrointestinal, pancreas, or other cancers, and were stage IV or had unknown

stage, recurrent, or unstageable cancers at initial diagnosis. On average, POG patients were younger than unmatched controls and were diagnosed with cancer more recently.

All POG patients were matched following PSM and genetic matching resulting in the identification of two counterfactuals for this single-arm application of precision oncology. Propensity score models are reported in Table S1. The propensity score matched cohort included 210 controls ($n_{\text{weighted}}=230$) and the genetic matched cohort included 204 controls ($n_{\text{weighted}}=230$). Covariate balance improved after matching, with genetic matching outperforming PSM for higher order terms, interaction terms, and in stratified cohorts. In the PSM cohort, eight of 54 covariates examined showed some evidence of residual imbalance across POG patients and matched controls compared to four of 54 covariates in the genetic-matched cohort (Table S2). In stratified cohorts, these differences increased to 13 of 54 versus 3 of 54 (Table S3) and 10 of 54 versus 3 of 54 (Table S4). Unbalanced covariates frequently related to number of lines of prior therapy, year of diagnosis, and/or performance status.

3.2 | Survival

During follow-up, 59 (25.7%) deaths occurred in POG patients, 1,265 (24.2%) deaths occurred in unmatched POG-naïve patients, 57 (24.8%) occurred in propensity score matched POG-naïve patients, and 61.5 (26.7%) occurred in genetic matched POG-naïve patients. Given that POG enrollment was ongoing throughout the period, observation times ranged from 1 month to a maximum of 1.5 years. On average, patients were observed for 0.6 years prior to death or censoring. Among POG patients, 15% ($n = 35$) experienced a treatment change during the study period as a result of WGTA. WGTA-informed treatments included approved therapies reimbursed by BC's public health-care system, off-label therapies either provided through Compassionate Access Programs or paid for out-of-pocket, and therapies provided in clinical trials (Laskin et al., 2015). The remaining 85% ($n = 195$) of POG patients did not receive WGTA-informed treatment owing to a number of factors, including no clinically actionable findings generated, no targeted treatments available or accessible to patients, currently responding to a non-targeted treatment option, or declining health status. Additional details on which WGTA-informed treatments were dispensed during the study period and why WGTA-informed treatments were not given are provided in Tables S5 and S6.

Overall, few significant survival differences were identified across POG patients and POG-naïve controls before matching and none were identified after matching. Figure 2 shows that KM survival functions were overlapping in all matched cohorts. While POG patients appeared to have

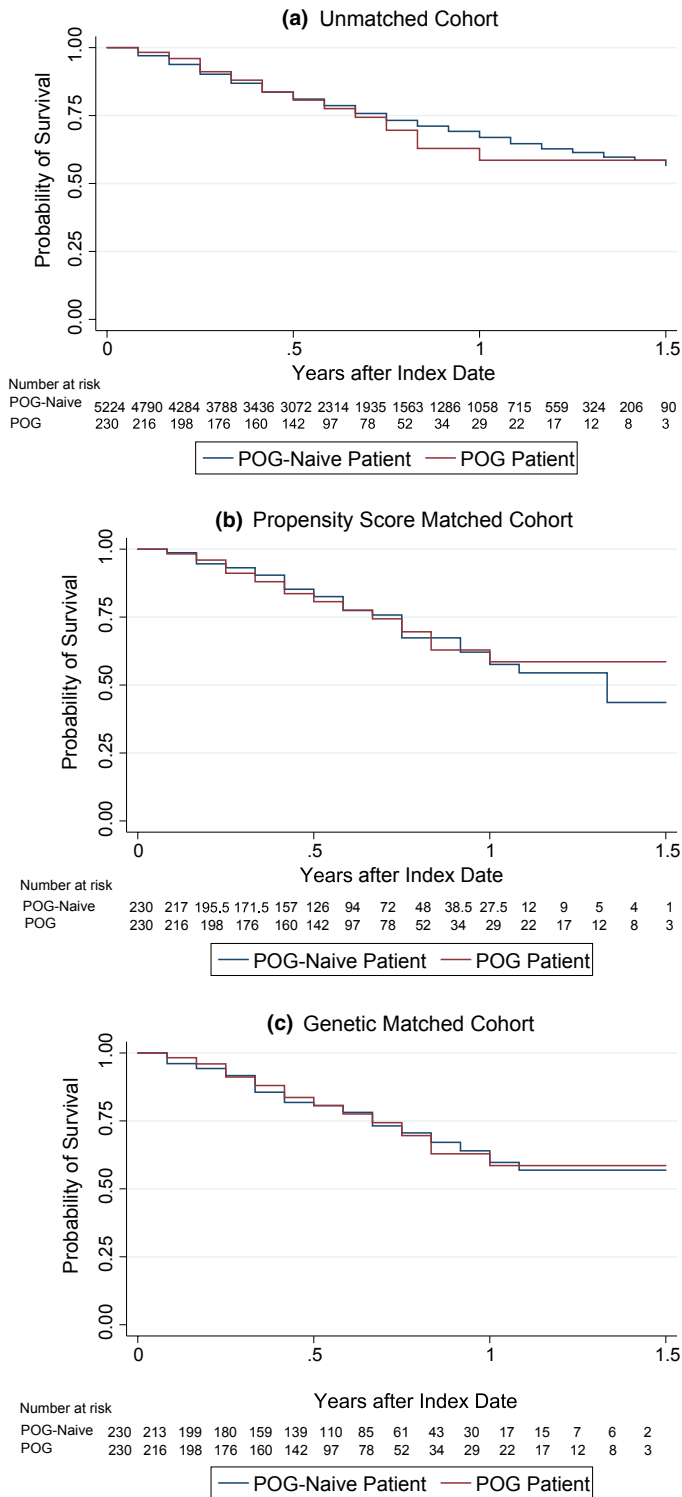


FIGURE 2 Kaplan–Meier survival estimates for POG patients and POG-naïve patients in matched and unmatched cohorts. Each subgraph depicts survival functions across POG patients and POG-naïve patients in the different cohorts. Risk tables present the number of uncensored patients at risk of death at the beginning of each interval across groups.

reduced survival compared to unmatched POG-naïve patients, survival functions were not statistically significantly different across cases and controls in any cohort ($p = 0.511$

in unmatched cohort; $p = 0.931$ in propensity score matched cohort; $p = 0.966$ in genetic matched cohort). Estimated 1-year survival rates were somewhat higher in unmatched POG-naïve patients than in POG patients (66.96% (95% CI: 65.21%, 68.65%) versus 58.56% (95% CI: 47.98%, 67.72%), respectively). These differences were eliminated after matching, with rates falling to 57.60% (95% CI: 46.33%, 67.32%) in propensity score matched controls and 59.73% (95% CI: 49.35%, 68.65%) in genetic matched controls.

According to the Weibull regression results summarized in Table 3, POG patients did not have statistically significantly different hazards of death compared with POG-naïve patients in any cohort. HRs associated with POG patients compared to POG-naïve patients ranged from 1.12 (95% CI: 0.86, 1.45) in the unmatched cohort to 0.95 (95% CI: 0.66, 1.37) after PSM, to 0.97 (95% CI: 0.68, 1.39) after genetic matching. Similarly, average 1-year survival was not statistically significantly different across POG and POG-naïve patients in any cohort ($p = 0.147$ in unmatched cohort; $p = 0.801$ in propensity score matched cohort; $p = 0.997$ in genetic matched cohort). Average 1-year survival times ranged from 8.13 months (95% CI: 7.43, 8.84) in POG patients to 8.65 months (95% CI: 8.50, 8.80) in unmatched POG-naïve controls, to 8.24 months (95% CI: 7.56, 8.91) in propensity score matched controls, to 8.16 months (95% CI: 7.49, 8.83) in genetic matched controls.

When stratifying analyses based on whether POG patients received WGTA-informed treatment, we found evidence of survival differences. POG patients who received WGTA-informed treatment had improved KM estimated survival compared to POG-naïve patients in all cohorts, shown in Figure 3. POG patients who did not receive informed treatment experienced either reduced survival when compared to unmatched POG-naïve patients or overlapping survival when compared to matched controls. Differences in KM survival functions were statistically significant in all cohorts ($p = 0.023$ in unmatched cohort; $p = 0.021$ in propensity score matched cohort; and $p = 0.028$ in genetic matched cohort). Estimated 1-year survival rates ranged from 72.79% (95% CI: 44.69%, 88.23%) in POG patients who received WGTA-informed treatment to 55.86% (95% CI: 44.19%, 66.03%) in those who did not.

In all cohorts, Weibull regression indicated that POG patients who received WGTA-informed treatment experienced a statistically significant reduction in their hazard of death compared to POG-naïve patients (HR: 0.41, 95% CI: 0.17, 0.98 in unmatched cohort; HR: 0.33, 95% CI: 0.13, 0.81 in propensity score matched cohort; and HR: 0.34, 95% CI: 0.14, 0.86 in genetic matched cohort; Table 3). Non-informed treatment correlated with a significant increase in the hazard (HR: 1.33, 95% CI: 1.01, 1.75) in the unmatched cohort but had no significant effect in the matched cohorts (HR: 1.16, 95% CI: 0.80, 1.68 in the propensity score matched cohort and HR: 1.17, 95% CI: 0.81, 1.69 in the genetic matched cohort).

TABLE 3 Weibull regression estimates for hazard ratios.

	Unmatched cohort (n = 5,454)	Propensity score matched cohort (n _{weighted} =460)	Genetic matched cohort (n _{weighted} =460)
Model 1			
POG naïve	(Ref.)	(Ref.)	(Ref.)
POG enrolled	1.12 (SE: 0.15)	0.95 (SE: 0.18)	0.97 (SE: 0.18)
Constant	0.42* (SE: 0.01)	0.53* (SE: 0.07)	0.50* (SE: 0.08)
Log-likelihood	-3,890	-306	-333
Likelihood-ratio χ^2 statistic	0.67	0.07	0.02
p-value for likelihood-ratio test	0.412	0.793	0.889
AIC	7,786	618	672
Model 2			
POG naïve	(Ref.)	(Ref.)	(Ref.)
POG enrolled, WGTA informed	0.41* (SE: 0.18)	0.33* (SE: 0.15)	0.34* (SE: 0.16)
POG enrolled, WGTA non-informed	1.33* (SE: 0.19)	1.16 (SE: 0.22)	1.17 (SE: 0.22)
Constant	0.42* (SE: 0.01)	0.54* (SE: 0.07)	0.50* (SE: 0.07)
Log-likelihood	-3,885	-301	-328
Likelihood-ratio χ^2 statistic	9.59	10.42	9.66
p-value for likelihood-ratio test	0.008	0.005	0.008
AIC	7,779	610	664

AIC, Akaike Information Criterion; SE, standard error.

Hazard ratio estimates are statistically significant at p -value < 0.05*, < 0.10°.

In all cohorts, 1-year survival was significantly greater in POG patients who received WGTA-informed treatment than in POG-naïve patients. While average 1-year survival was statistically significantly less in POG patients who received non-informed treatment compared to POG-naïve controls in the unmatched cohort, this difference was not significant in either matched cohort. Average 1-year survival times ranged from 7.51 months (95% CI: 6.76, 8.25) in POG patients who received non-informed treatment to 10.99 months (95% CI: 10.17, 11.80) in those who received WGTA-informed treatment.

4 | DISCUSSION

We provide an introduction to matching methods and illustrate how matching can be used to analyze precision oncology interventions using real-world data. We focus on the use of PSM and genetic matching to reduce selection bias in a case study of BC's POG program, which applies WGTA to inform treatment planning for patients with advanced cancers. We find both methods are valid tools for generating well-balanced counterfactuals for precision oncology, even in highly heterogeneous patient cohorts. Genetic matching outperformed PSM when achieving balance on covariates of interest, coinciding with past literature comparing these approaches (Radice et al., 2012).

To our knowledge at the time of writing, our study is the first to examine the survival benefits of applying WGTA to inform treatment planning across multiple tumor sites. Analyses on unmatched and matched cohorts indicate that overall survival did not significantly differ across POG and POG-naïve patients. These results broadly align with past observational studies of less comprehensive forms of genomic testing to guide cancer care and may reflect low WGTA-informed treatment rates observed within our study period (Presley et al., 2018). Of the patients enrolled in POG, 15% (n = 35) received a treatment change informed by their genomic results.

Stratification by WGTA-informed treatment revealed differences in estimated survival. Estimated 1-year survival ranged from 8.16 months (95% CI: 7.49, 8.83) to 8.65 months (95% CI: 8.50, 8.80) in POG-naïve patients compared to 10.99 months (95% CI: 10.17, 11.80) in POG patients who received WGTA-informed treatment. In all cohorts, POG patients whose WGTA led to treatment change were at a statistically significantly reduced hazard of death compared to POG-naïve patients. Estimated HRs ranged from 0.33 (95% CI: 0.13, 0.81) to 0.41 (95% CI: 0.17, 0.98). While this subgroup analysis may influence overall covariate balance achieved through matching, we found little evidence of residual imbalance on observable characteristics. Sensitivity analysis involving estimation of covariate-adjusted Weibull regression models led to no substantive differences in estimates or conclusions (Table S7). Although these results do

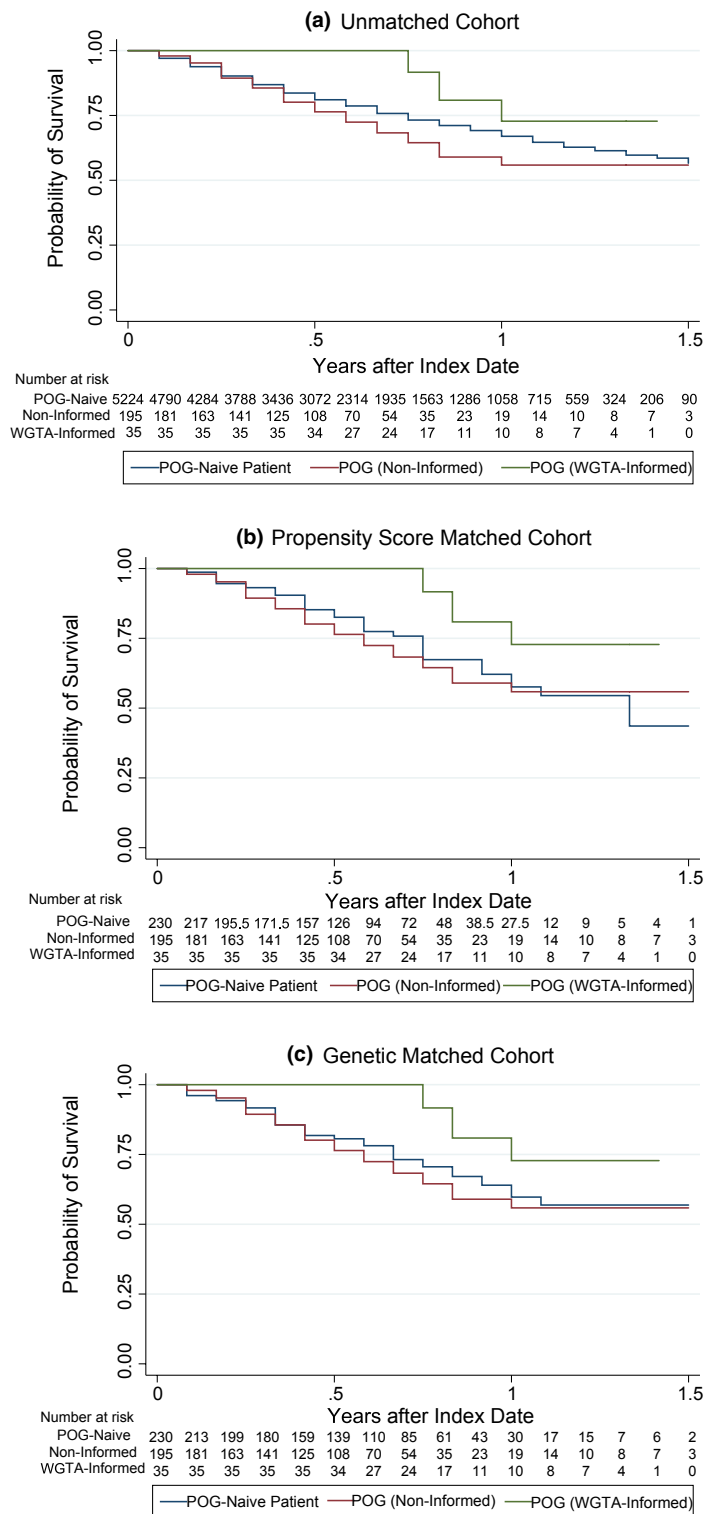


FIGURE 3 Kaplan–Meier survival estimates for POG patients stratified by WGTA-informed treatment and POG-naive patients in matched and unmatched cohorts. Each subgraph depicts survival functions across POG patients who received WGTA-informed treatment, POG patients who did not receive WGTA-informed treatment and POG-naive patients in the different cohorts. Risk tables present the number of uncensored patients at risk of death at the beginning of each interval across groups.

not guarantee the absence of unobserved confounding, they do signal that the proportion of patients receiving WGTA-guided care will impact precision oncology's ability to deliver on promises of improving patients' health.

POG patients whose WGTA information did not lead to treatment change had an average 1-year survival time of 7.51 months (95% CI: 6.76, 8.25) and were at a significantly increased hazard of death compared to unmatched

POG-naïve patients (HR: 1.33, 95% CI: 1.01, 1.75). In the matched cohorts, there were no statistically significant differences in hazards across these two groups. Our study's small sample size and short follow-up period for our clinical endpoints, overall and 1-year survival, may limit the precision of our final estimates. Yet, the observed variation demonstrated through matching provides evidence of selection bias in the unmatched control cohort and compared to covariate adjustment, matching produces less biased effect estimates. (Austin, 2013; Ho et al., 2007) Failing to account for apparent baseline differences in patients' characteristics will lead to inaccurate conclusions about the health impacts of precision oncology.

Compared to the clinical end points examined in this study, progression-free survival analyses may yield additional estimates of the short-term benefits of WGTA-informed treatment. Disease progression on molecularly guided treatment is less likely to be impacted by crossover and requires a smaller sample size and shorter follow-up period for data collection than overall survival (Villaruz & Socinski, 2013). However, the accuracy of progression-free survival estimates relies on objective, consistent measures of disease progression, which are unavailable in many jurisdictions including BC. Future evaluations of precision oncology would benefit greatly from routine collection and standardization of disease progression information in clinical outcomes databases.

Studies are beginning to apply matching when evaluating precision oncology but often do not report detailed balance statistics (Barcenas et al., 2017; Presley et al., 2018). Matching relies on the strong underlying assumption of ignorability. Careful balance assessment for entire covariate distributions is critical to obtain reliable results. Even if final cohorts appear well-balanced on observable characteristics, unobserved confounding can remain, which may introduce selection bias. For example, disease prognosis may guide the selection of POG patients and influence eligibility for WGTA-informed treatments, but this information is not routinely recorded in administrative databases. Further research exploring the use of matching in combination with quasi-experimental methods that adjust for certain forms of unobserved confounding are necessary to support health-care decision-making in precision oncology.

In conclusion, matching methods combined with population-based administrative data offer a solution to the challenge of non-randomized enrollment observed in many precision oncology studies. In the absence of RCTs, PSM and genetic matching can be used to mitigate the selection bias present in observational studies. We find that while both methods are able to identify a counterfactual for a single-arm application of precision oncology, genetic matching outperformed PSM

when balancing observable characteristics. Genetic matching will thus result in more reliable effect estimates than PSM alone. After matching, we detected no significant overall survival differences POG patients and matched controls. Instead, our analyses signaled that the clinical effectiveness of precision oncology approaches will depend on the influence of genomic information on subsequent treatment change. Our study will guide future applications of matching in precision oncology and help to ensure reliability of final effect estimates in the absence of RCT data.

ACKNOWLEDGMENTS

We gratefully acknowledge the participation of our patients and families, the POG team, the GSC platform, and the generous support of the BC Cancer Foundation and Genome British Columbia (project B20POG). MAM acknowledges infrastructure investments from the Canada Foundation for Innovation and the support of the Canada Research Chairs and the CIHR Foundation (FDN-143288) programs. SJMJ acknowledges the support of the Canada Research Chairs program. This research was supported by Genome British Columbia (B20POG) and Genome British Columbia/Genome Canada (G05CHS).

CONFLICTS OF INTERESTS

Deirdre Weymann, Steven J.M. Jones, Robyn Roscoe, Sophie Sun, Kasmintan A. Schrader, and Marco A. Marra report no conflicts of interest. Janessa Laskin has received honoraria, ad boards, and institutional grant funding from Roche, BI, AstraZeneca, and Takeda. Howard Lim has received honoraria from Eisai, Taiho, Roche, Lilly, Amgen, and Leo and is an investigator on trials with Bayer, BMS, Lilly, Roche, AstraZeneca, and Amgen. Daniel J. Renouf has received research funding and honoraria from Bayer and Roche, as well as travel funding and honoraria from Servier, Celgene, Taiho, and Ipsen. Sophie Sun has received research grant and honoraria funding from AstraZeneca. Stephen Yip is an advisory board member of Roche, Bayer and Pfizer and has received honoraria from Amgen. Dean A. Regier has received travel support from Illumina.

DATA AVAILABILITY STATEMENT

Patient-level administrative data used in this retrospective study are confidential and are not available in a public repository, in accordance with institutional policies.

ORCID

Deirdre Weymann  <https://orcid.org/0000-0002-3072-5657>

REFERENCES

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76, 1537–1557.

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.
- Anglemyer, A., Horvath, H. T., & Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews*, *4*, 1465–1858.
- Austin, P. C. (2009). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, *29*, 661–677.
- Austin, P. C. (2010). 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*, 1092–1097.
- Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.
- Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*, 150–161.
- Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, *32*, 2837–2849.
- Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*, 1057–1069.
- Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, *33*, 4306–4319.
- Bang, H., & Tsiatis, A. A. (2000). Estimating medical costs with censored data. *Biometrika*, *87*, 329–343.
- Barcenas, C. H., Raghavendra, A., Sinha, A. K., Syed, M. P., Hsu, L., Patangan Jr, M. G., Chavez-Macgregor, M., Shen, Y., Hortobagyi, G. H., & Valero, V. (2017). Outcomes in patients with early-stage breast cancer who underwent a 21-gene expression assay. *Cancer*, *123*, 2422–2431.
- Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, *342*, 1878–1886.
- Burke, W., Houry, M. J., Stewart, A., & Zimmern, R. L. (2006). The path from genome-based research to population health: Development of an international public health genomics network. *Genetics in Medicine*, *8*, 451–458.
- Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31–72.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. [chapter 2]. : Academic Press, Inc.
- Collette, L., & Tombal, B. (2015). N-of-1 trials in oncology. *The Lancet Oncology*, *16*, 885–886.
- Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*, 932–945.
- Golder, S., Loke, Y. K., & Bland, M. (2011). Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLoS Med*, *8*, e1001026.
- Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–420.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as non-parametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.
- Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, *171*, 481–502.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.
- Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genetics in Medicine*, *9*, 665–674.
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale Dell'istituto Italiano Degli Attuari*, *4*, 461.
- Laskin, J., Jones, S., Aparicio, S., Chia, S., Ch'Ng, C., Deyell, R., Eirew, P., Fok, A., Gelmon, K., & Ho, C. (2015). Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. *Molecular Case Studies*, *1*, a000570.
- Lin, D. (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine*, *22*, 1181–1200.
- Lonjon, G., Boutron, I., Trinquart, L., Ahmad, N., Aim, F., Nizard, R., & Ravaud, P. (2014). Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Annals of Surgery*, *259*, 18–25.
- Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics*, *61*, 721–728.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Academy of Sciences, India*, *2*, 49–55.
- Ming, K., & Rosenbaum, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, *10*, 455–463.
- Presley, C. J., Tang, D., Soulos, P. R., Chiang, A. C., Longtine, J. A., Adelson, K. B., Herbst, R. S., Zhu, W., Nussbaum, N. C., Sorg, R. A., Agarwala, V., Abernethy, A. P., & Gross, C. P. (2018). Association of broad-based genomic sequencing with survival among patients with advanced non-small cell lung cancer in the community oncology setting. *JAMA*, *320*, 469–477.
- R Core Team, (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014. <https://www.R-project.org/>
- Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., & Sekhon, J. S. (2012). Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *The International Journal of Biostatistics*, *8*, 1–43.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249–264.
- Schork, N. J. (2015). Personalized medicine: time for one-person trials. *Nature*, *520*, 609–611.

- Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42, 1–52.
- Sekhon, J. S., & Mebane, W. R. (1998). Genetic optimization using derivatives. *Political Analysis*, 7, 187–210.
- Silverman, S. L. (2009). From randomized controlled trials to observational studies. *The American Journal of Medicine*, 122, 114–120.
- Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2, 3–16.
- Statacorp (2017). *Stata statistical software: Release 15*, College Station, TX: StataCorp LLC.
- Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine*, 27, 2062–2065.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1.
- Villaruz, L. C., & Socinski, M. A. (2013). The clinical viewpoint: Definitions, limitations of RECIST, practical considerations of measurement. *Clinical Cancer Research*, 19(10), 2629–2636. <https://doi.org/10.1158/1078-0432.CCR-12-2935>
- Weymann, D., Costa, S., & Regier, D. A. (2019). Validation of a cyclic algorithm to proxy number of lines of systemic cancer therapy using administrative data. *JCO Clinical Cancer Informatics*, 3, 1–10.
- Weymann, D., Pataky, R., & Regier, D. A. (2018). Economic evaluations of next-generation precision oncology: A critical review. *JCO Precision Oncology*, 2, 1–23. <https://doi.org/10.1200/PO.17.00311>
- Willan, A. R., Lin, D., & Manca, A. (2005). Regression methods for cost-effectiveness analysis with censored data. *Statistics in Medicine*, 24, 131–145.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Weymann D, Laskin J, Jones SJ, et al. Matching methods in precision oncology: An introduction and illustrative example. *Mol Genet Genomic Med*. 2021;9:e1554. <https://doi.org/10.1002/mgg3.1554>