



Propensity score matching versus coarsened exact matching in observational comparative effectiveness research

Journal of **Comparative Effectiveness Research**

David Guy^{*,1,2} , Igor Karp¹ , Piotr Wilk¹, Joseph Chin³ & George Rodrigues^{1,2}

¹Department of Epidemiology and Biostatistics, Schulich School of Medicine & Dentistry, Western University, London, ON N6A 3K7, Canada

²Department of Radiation Oncology, London Health Sciences Centre, London, ON N6A 5W9, Canada

³Department of Urology, Schulich School of Medicine & Dentistry, Western University, London, ON N6A 3K7, Canada

*Author for correspondence: dguy4@uwo.ca

Aim & methods: We compared propensity score matching (PSM) and coarsened exact matching (CEM) in balancing baseline characteristics between treatment groups using observational data obtained from a pan-Canadian prostate cancer radiotherapy database. Changes in effect estimates were evaluated as a function of improvements in balance, using results from randomized clinical trials to guide interpretation. **Results:** CEM and PSM improved balance between groups in both comparisons, while retaining the majority of original data. Improvements in balance were associated with effect estimates closer to those obtained in randomized clinical trials. **Conclusion:** CEM and PSM led to substantial improvements in balance between comparison groups, while retaining a considerable proportion of original data. This could lead to improved accuracy in effect estimates obtained using observational data in a variety of clinical situations.

First draft submitted: 16 March 2021; Accepted for publication: 10 May 2021; Published online: 1 June 2021

Keywords: balance • coarsened exact matching • comparative effectiveness research • observational data • propensity score matching

Comparative effectiveness research aims to generate evidence on the relative effectiveness and safety of different treatment approaches. Since the implementation of electronic patient health records, the proliferation of medical record and administrative claims databases has led to observational research occupying a large proportion of comparative effectiveness research [1]. However, since treatment in this context is not random but influenced by factors that also influence outcomes of interest, confounding must be accounted for when estimating treatment effects. This is usually done through regression modeling, wherein variation in the outcome is modeled as a function of variation in treatment and confounders [2]. Statistical software then estimates treatment effects, while accounting for confounders through algorithms such as maximum likelihood estimation [3]. However, the outcome could vary as a function of the confounders in a manner that does not tightly adhere to or is difficult to identify with functional forms or simple two-way interactions commonly used in health research [3]. Inability to accurately quantify outcome variation attributable to confounding through appropriate modeling leads to residual confounding and biased effect-estimates [4].

Issues of residual confounding are further exacerbated with decreasing balance in the distribution of baseline covariates between treatment groups, as accurate estimation of regression coefficients becomes more reliant on model specification [5]. Data preprocessing techniques can increase the overlap and balance in the distribution of baseline covariates between treatment groups [6,7]. Propensity score matching (PSM) is an example of such a technique that has become increasingly popular in recent years [8]. The propensity score is defined as the probability of receiving treatment according to baseline covariates [9]. Matching on the propensity score can balance baseline covariates between treatment groups [10]. Coarsened exact matching (CEM) is another data preprocessing technique [11], wherein continuous and ordinal characteristics are categorized and values of categorical variables are collapsed

(i.e., coarsened). Afterward, observations between groups with the same values of coarsened variables are retained for comparison.

Previous studies have compared the utility of PSM and CEM [12,13]. However, previous studies examined high-dimensional datasets and often relied upon quantile-based rules for coarsening continuous variables [13,14]. In contrast, the number of variables that consistently influence treatment variation for prostate cancer (PCa and many other clinical decisions) is relatively small; furthermore, there is often *a priori* information enabling creation of strata that are more prognostically meaningful and thus more efficient in matching (i.e., greater data retention for the same or better improvements in balance) than strata formed from quantile-based rules.

The objective of this study was to compare the performance of CEM and PSM in the context of observational datasets, using two examples from PCa research.

Methodology

Database

Data were abstracted from the Prostate Cancer Risk Stratification database (ProCaRS) [15]. This database contains information on 7974 patients diagnosed with PCa and treated with different forms of primary radiotherapy between 1994 and 2010 from four Canadian institutions in Toronto, Quebec City, Montreal and Vancouver [16]. Details regarding ethics approval, database construction and quality assurance have been previously described [15].

Comparison one

The first comparison was informed by a randomized clinical trial (RCT) that compared the rate of biochemical failure among men diagnosed with intermediate-risk PCa and treated with either internal seed radiotherapy and hormone therapy (ISRT + HT) or external beam radiotherapy in addition to HT (EBRT + HT) [17]. Patients were included from the ProCaRS dataset if they met the PCa-specific eligibility criteria specified by Morris *et al.*, with two modifications. This involved a range for HT duration (4–16 vs 12 months specifically) and EBRT dose of ≥ 74 Gy or ISRT dose of ≥ 144 Gy instead of the specific dose-escalation protocols as investigated by Morris *et al.* The final sizes of the ISRT and EBRT + HT groups were 433 and 132, respectively (Supplementary Figure 1A).

Comparison two

The second comparison was informed by a RCT that compared rate of biochemical failure among men diagnosed with localized PCa and treated with EBRT alone (EBRT) or in combination with HT (EBRT + HT) [18]. The analysis was restricted to intermediate-risk PCa due to observed effect modification by risk category. Patients were included from the ProCaRS dataset if they met the PCa-specific eligibility criteria specified by Jones *et al.*, with two modifications: a total EBRT dose of ≥ 66 Gy instead of the 66.6 Gy, and HT duration of 3–6 months of HT compared with the 4-month duration implemented by Jones *et al.* The final sample size included 126 and 579 men in the EBRT + HT and EBRT groups, respectively (Supplementary Figure 1B).

Covariate selection

We included baseline covariates with prognostic value in relation to the rate of biochemical failure [19]. This included baseline PSA, clinical stage, Gleason sum, EBRT dose (if applicable) and treatment year. Gleason sum was divided into three groups: 6 (3 + 3), 7 (3 + 4) and 7 (4 + 3), while clinical tumor stage (T-stage) was divided into clinically inapparent and unilateral disease versus bilateral disease not extending outside the prostate [20]. Age was not included as a covariate, as it did not demonstrate an association with biochemical failure in either comparison (adjusted for treatment received: hazard ratio (HR), [95% CI]: 0.98 [0.93, 1.03]; $p = 0.50$ and HR [95% CI]: 1.00 [0.98, 1.02]; $p = 0.99$ for datasets used in comparison one and two, respectively). Moreover, age was strongly associated with treatment choice, which would bias effect estimates if it was adjusted for [21].

Propensity score matching

A logistic model was used to estimate the propensity score using the above covariates and treatment received as a binary-dependent variable [21]. Locally weighted scatterplot smoothers were used to assess for departures from linearity. Improvements in the model fit were assessed using the likelihood ratio test and pseudo- R^2 . We checked for interactions and nonlinearity in baseline covariates [9,22]. Baseline PSA was modeled as a restricted cubic spline with four knots, treatment year was treated as a discrete variable with 2-year categories and an interaction term between baseline PSA and Gleason sum was added. DFBETA statistics revealed one outlier, wherein a subject

received EBRT + HT instead of ISRT + HT despite a very low PSA, clinical T-stage and Gleason sum. This patient was retained, as they did not have any contraindications to EBRT. Ratios of 1:3 and 1:4 were used for comparison one and two, respectively, given the ratio of index to reference observations available. Caliper widths included a range of 0.5–0.005 standard deviations of the logit of the propensity score (Supplementary Table 1A & B). Nearest-neighbor matching was used without replacement [23].

Coarsened exact matching

Baseline covariates were coarsened according to prognostic value [20]. Gleason sum and clinical T-stage were matched on directly. PSA was divided starting with ranges >0, 0–4, 4–10 and 10–20 ng/ml, and progressively divided finer within those ranges [20]. The range for treatment year was divided into halves, then thirds. EBRT dose was split into low (≥ 66 and < 73 Gy) and high (≥ 73 to < 79.8 Gy) dosage. Coarsening ranges are presented in Supplementary Table 2A & B.

Balance diagnostics

We chose four balance diagnostics that measure different characteristics of the multivariable distribution of baseline covariates to measure improvements in balance when further restricting matching strategies. The standardized mean difference (SMD) in proportion of patients having high-intermediate versus low-intermediate risk PCa as defined by the ProCaRS system was used as a prognostic score-based balance measure [19,24]. The ProCaRS risk groups capture imbalance in combinations of specific values for baseline covariates to the extent that each is associated with variation in the rate of biochemical failure. Since this balance measure is limited in capturing subtle differences in individual variables, we also examined the absolute SMD for individual variables, the average absolute SMD for baseline covariates and the variance ratio of individual variables [24–26].

Statistical analysis

Statistical analyses were performed using RStudio version 3.6.0 [27]. Descriptive statistics were calculated for baseline covariates involved in matching. The MatchIt package in R was used to match participants [28]. Cox proportional-hazards regression analyses for estimating the effect of treatment group on the hazard of biochemical failure were performed using the survival package [29]. Log-minus-log survival plots and scaled Schoenfeld residuals were examined for violations of proportional hazards, which, when present, were handled by modeling variables as a function of time. Model log-likelihoods were examined for improvements in fit after incorporating higher-order terms and transformations for continuous covariates. No influential observations were identified through DBETA statistics. HRs and 95% CIs were estimated from unmatched data before and after adjustment for treatment year, EBRT dose (if applicable), the natural logarithm of PSA and categorized values of clinical T-stage and Gleason sum. For matched data, we employed Cox models clustered by the matched sets with associated weights, using robust variance estimators to generate confidence intervals [29,30]. After CEM, continuous covariates were included in the model to control for possible residual confounding. After PSM strategies, all covariates were included in the Cox model.

Results

Comparison one

Descriptive statistics for the unmatched treatment groups are reported in Table 1A. Briefly, men treated with ISRT + HT compared with EBRT + HT were, on average, younger, treated at earlier dates, had less advanced tumor characteristics and received HT for a similar duration.

Figures 1A and 2A depict the selection processes for PSM and CEM, respectively, with the red data points representing the matching strategies that led to optimal balance to sample size trade-off. Sixty-three and 66% of the source population were retained through PSM and CEM, respectively. The associated average absolute SMDs for individual variables were 0.025 and 0.0073, while the risk-group SMDs were 0.046 and 0, respectively (Figure 3A). Both matching strategies balanced all individual covariates according to the commonly accepted threshold of < 0.1 , while only CEM led to balance in the variance ratios in the acceptable range of 0.92–1.08 [22].

Changes in baseline characteristics for treatment groups due to stricter matching are presented in Supplementary Figures 2A & 3A. Briefly, as matching became stricter, the ISRT + HT group characteristics tended toward those of the EBRT + HT until a certain point, wherein characteristics in both groups tended toward those of the

Table 1. Descriptive statistics for patient and treatment characteristics.

(A) Descriptive statistics for comparison one						
Patient and treatment characteristics	ISRT + HT (n = 433)		EBRT + HT (n = 132)		SMD	Variance ratio
Age						
– Median	68		72		0.7644	0.55
– Q1, Q3	63, 72		69, 75			
Clinical T-stage						
– T1a–2a	373	86.14%	122	92.42%	0.2041	
– T2b–c	60	13.86%	10	7.58%		
PSA (ng/ml)						
– Median	7.60		9.02		0.3468	0.92
– Q1, Q3	5.70, 10.50		5.88, 12.60			
Gleason sum						
– 6 (3 + 3)	126	29.10%	23	17.42%	0.2790	
– 7 (3 + 4)	249	57.51%	64	48.48%	0.1815	
– 7 (4 + 3)	58	13.39%	45	34.09%	0.5014	
Treatment year						
– Median	2002		2003		0.3029	0.67
– Q1, Q3	2001, 2004		2002, 2004			
HT duration (months)						
– Median	5.98		5.45		0.1285	2.84
– Q1, Q3	5.55, 6.81		4.82, 8.46			
ProCaRS risk groups						
– Low intermediate	404	93.30%	99	75%	0.5177	
– High intermediate	29	6.70%	33	25%		
(B) Descriptive statistics for comparison two						
	EBRT + HT (n = 126)		EBRT (n = 579)		SMD	Variance ratio
Age						
– Median	72		72		0.046	1.16
– Q1, Q3	68.25, 75		69, 75			
Clinical T-stage						
– T1a–2a	109	86.51%	453	78.24%	0.22	
– T2b–c	17	13.49%	126	21.76%	0.22	
PSA (ng/ml)						
– Median	8.75		9.20		0.0061	1.26
– Q1, Q3	5.71, 12.46		6.00, 12.70			
Gleason sum						
– 6 (3 + 3)	23	18.25%	224	38.69%	0.4649	
– 7 (3 + 4)	67	53.18%	260	44.91%	0.1660	
– 7 (4 + 3)	36	28.57%	95	16.41%	0.2945	
EBRT dose (Gy)						
– Median	7560		7000		0.5899	1.46
– Q1, Q3	7400, 7980		6600, 7980			
Treatment year						
– Median	2001		2000		0.3237	1.49
– Q1, Q3	2000, 2004		1997, 2003			
ProCaRS risk groups						
– Low intermediate	88	69.84%	424	73.23%	0.0751	
– High intermediate	38	30.16%	155	26.77%	0.0751	
Clinical T-stage: Clinical tumor-stage; EBRT: External beam radiotherapy; HT: Hormone therapy; ISRT: Internal seed radiotherapy; ProCaRS: Prostate cancer risk stratification; SMD: Standardized mean difference.						

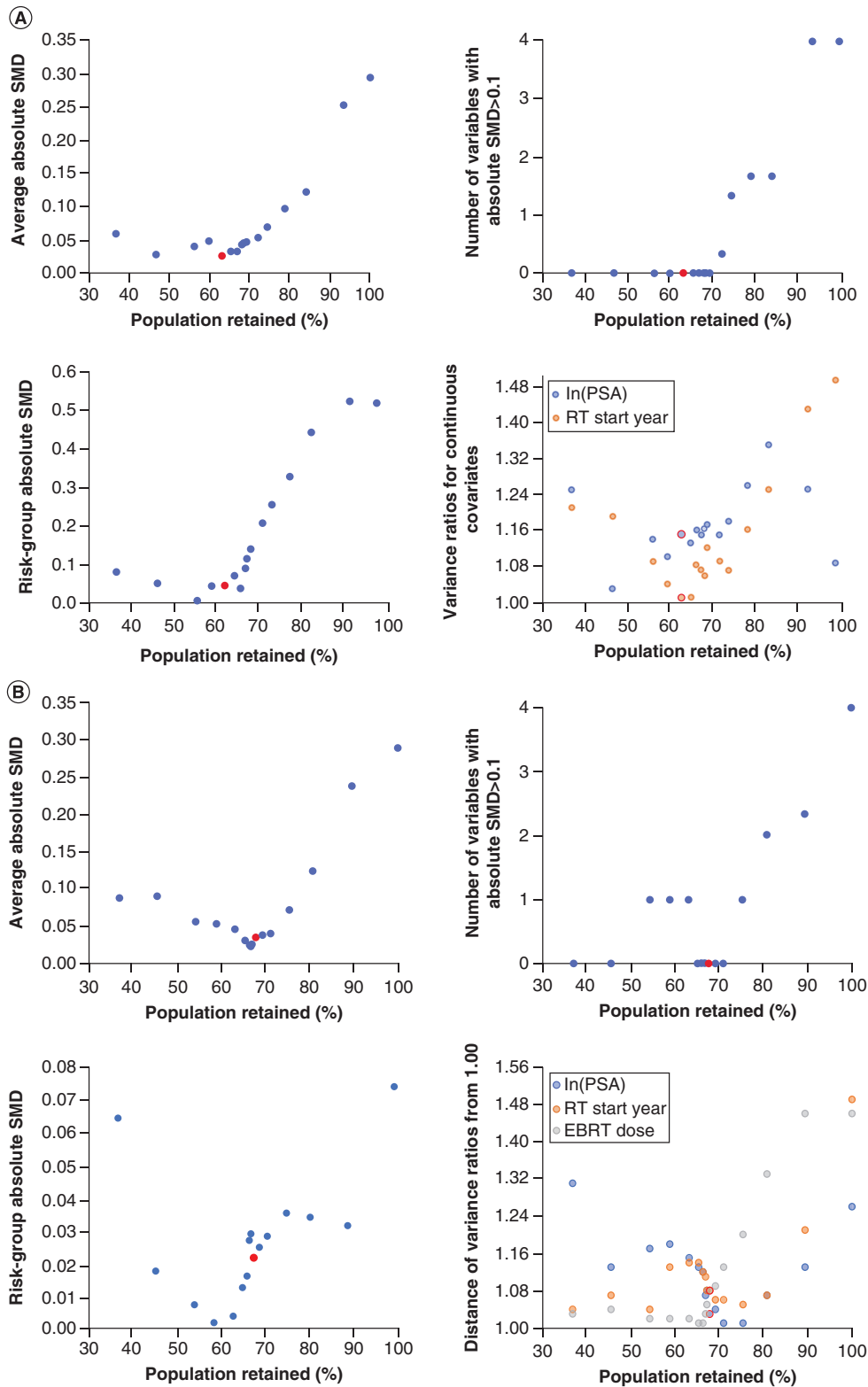


Figure 1. Balance achieved with each propensity score matching strategy by percent of the original population retained for comparison (A) one and (B) two. The red dot indicates the chosen matching strategy.

*Variance ratio values for EBRT Dose and In(PSA) in the selected matching strategy in comparison two overlap (both have a value of 1.01) so are not discernible.

SMD: Standardized mean difference.

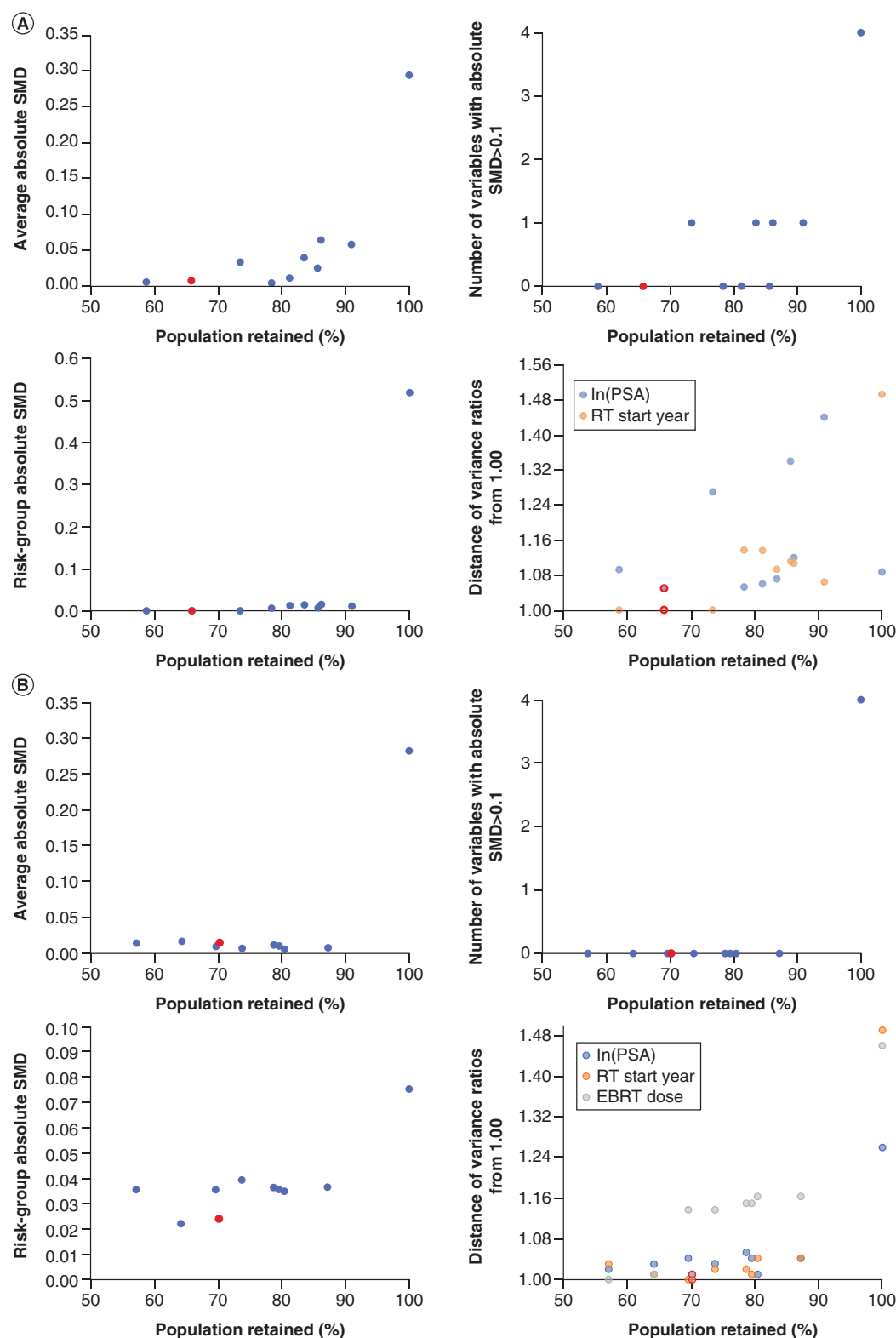


Figure 2. Balance achieved with each coarsened exact matched strategy by percent of the original population retained for comparison (A) one and (B) two. The red dot indicates the chosen matching strategy.

*Variance ratio values for EBRT Dose and RT Start Year in the selected matching strategy in comparison two overlap (both have a value of 1.08) so are not discernible

EBRT: External beam radiotherapy; PSM: Propensity score matched; SMD: Standardized mean difference.

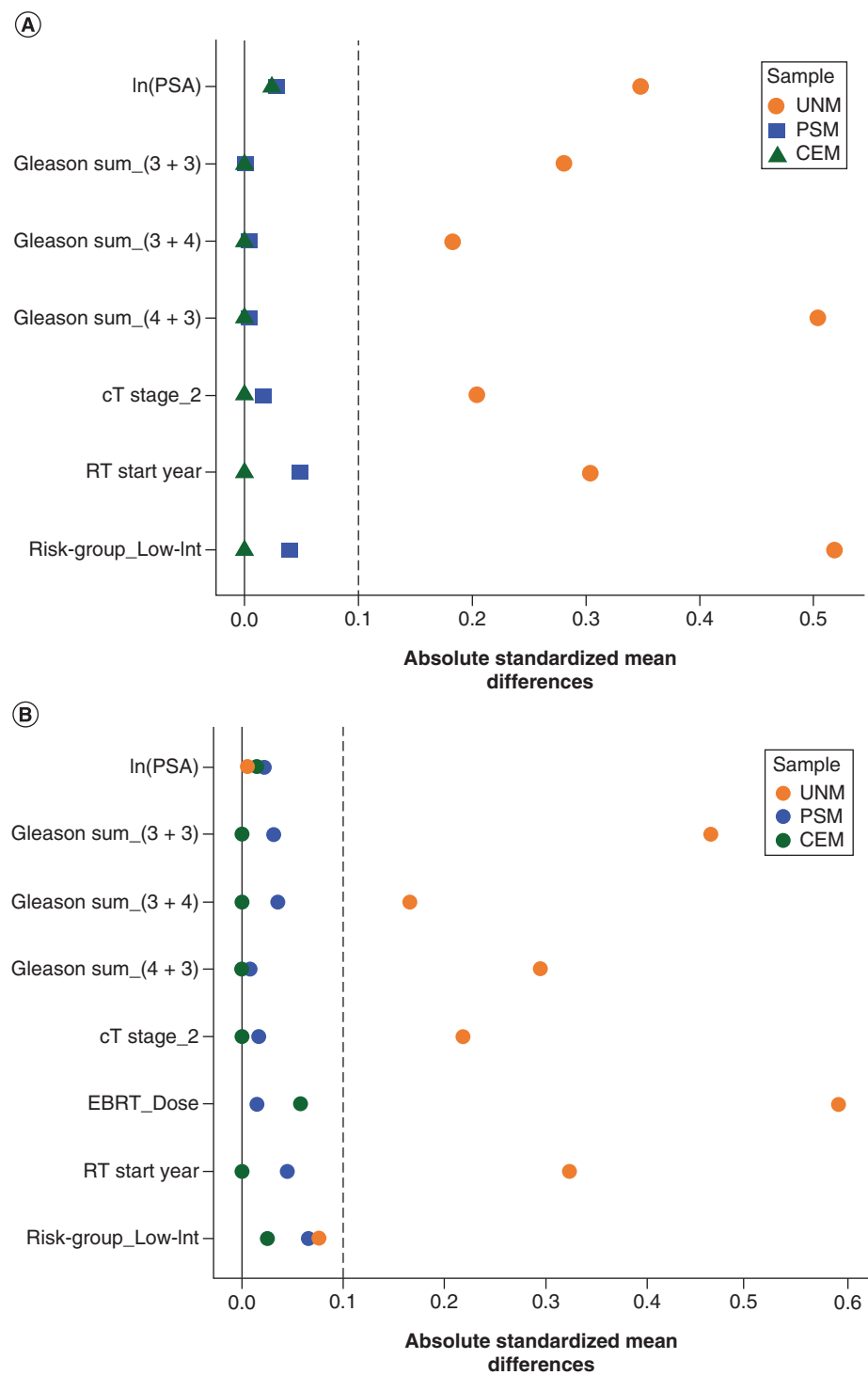


Figure 3. Love plot of the absolute SMD for individual baseline covariates before matching and after PSM and CEM in comparison (A) one and (B) two. CEM: Coarsened exact matched; PSM: Propensity score matched; SMD: Standardized mean difference; UNM: Unmatched.

ISRT + HT group. In the matching strategy chosen, values for characteristics in both groups fell between values in both groups before matching (Tables 2A & 2B).

The effect estimates are presented in Table 3A. Compared with the benchmark RCT HR (95% CI) of 2.04 (1.25, 3.33), the unadjusted HR (95% CI) from our data was 6.55 (3.82, 11.26), which attenuated to 4.48 (2.44, 8.22) with adjustment. The unadjusted and multivariable adjusted estimates after PSM were 4.06 (1.98, 8.11) and 3.84 (1.91, 8.71), respectively, while those after CEM were 4.04 (1.88, 8.66) and 3.84 (1.77, 8.34), respectively. Other candidate matching strategies for both PSM and CEM that resulted in similar balance led to similar results (Table 3A).

Comparison two

Descriptive statistics for unmatched treatment groups in analysis two are reported in Table 1B. Treatment groups were similar (SMD <0.1) with respect to age, PSA and proportion of low- versus high-intermediate risk-group status. The EBRT + HT compared with the EBRT group had a slightly greater proportion of clinically inapparent and unilateral disease, Gleason sum 7 (3 + 4) and 7 (4 + 3), received higher radiotherapy doses and were treated later in calendar time.

Figures 1B & 2B show the selection processes for PSM and CEM, respectively, with the red data points representing matching strategies that led to optimal balance to sample size trade-off. Sixty-eight and 70% of the source population were retained through PSM and CEM, respectively. The associated mean SMDs were 0.034 and 0.015, while the risk-group SMDs were 0.022 and 0.024, respectively (Figure 3A). Both strategies maintained SMD for all individual covariates under <0.1, and variance ratios for continuous covariates within the acceptable range of 0.92–1.08 [22].

Changes in baseline characteristics for treatment groups due to stricter matching are presented in Supplementary Figures 2B & 3B. As matching became stricter, the EBRT group characteristics tended toward those of the EBRT + HT group until a certain point, wherein characteristics in both groups tended toward those of the EBRT group. In the matching strategy chosen, values for characteristics in both groups fell between values in both groups before matching (Table 2B).

The effect estimates are presented in Table 3B. Compared with the benchmark RCT HR (95% CI) of 1.79 (1.45, 2.21) the unadjusted effect estimate was 1.40 (0.99, 1.98), which increased to 1.52 (1.06, 2.16) after adjustment. The unadjusted and adjusted effect estimates after PSM were 1.39 (0.97, 1.99) and 1.44 (1.00, 2.05), respectively, and after CEM were 1.53 (0.95, 2.46) and 1.55 (0.98, 2.45), respectively. Other candidate matching strategies for both PSM and CEM that demonstrated similar balance led to similar results (Table 3B).

Discussion

We sought to compare the performance of two popular data preprocessing techniques in the context of observational data, using examples from PCa research. Balance in the distributions of individual variables, as measured by SMD, was improved with both PSM and CEM. CEM generally led to smaller SMDs for individual covariates and overall average SMD when compared with PSM with similar levels of data retention. Furthermore, the risk-group SMD was improved through both PSM and CEM, but to a greater extent after CEM. Likewise, the variance ratio for continuous covariates was closer to one after both matching strategies but more so after CEM for baseline PSA and treatment start date; however, PSM led to a variance ratio closer to one for EBRT dose. These findings are consistent with other studies, wherein large improvements in balance were observed after CEM (compared with PSM) using balance diagnostics based on the comparison of multivariable distributions between treatment groups [12,13].

We found that both CEM and PSM led to matched samples with average values of characteristics falling in a range of observed values of characteristics in each treatment group. This is expected since it represents areas of common support. It is also favorable since results are more ‘generalizable’ to patient groups with characteristics that are amenable for either treatment under comparison. Furthermore, the precision of effect estimates did not differ notably between PSM and CEM. In contrast, Fullerton *et al.* found that CEM led to matched samples that differed greatly from the original population and either treatment group in their baseline characteristics as well as greater precision for PSM [12]. This seeming discrepancy is likely explained by the difference in the number of baseline covariates between datasets used for matching. This explanation is supported by findings from Ripollone *et al.*, who reported that smaller covariate sets of eight covariates used in CEM retained a substantially greater proportion of the original population and improved precision compared with larger covariate sets with up to 119 covariates [13].

Table 2. Descriptive statistics for patient and treatment characteristics after matching

(A) Descriptive statistics for PSM strategy 10 and CEM strategy 8 in comparison one								
Patient and treatment characteristics	PSM strategy 10				CEM strategy 8			
	ISRT + HT (n = 248)		EBRT + HT (n = 109)		ISRT + HT (n = 276)		EBRT + HT (n = 96)	
Age (years)								
– Median	69		72		69.5		72	
– Q1, Q3	64, 72		69, 75		64, 72.25		69, 75	
Clinical T-stage								
– T1a–c	231	93.12%	102	93.58%	264	95.83%	92	95.83%
– T2b–c	17	6.88%	7	6.42%	12	4.17%	4	4.17%
PSA (ng/ml)								
– Median	7.60		8.06		7.95		8.47	
– Q1, Q3	5.65, 10.60		5.63, 10.70		6.08, 11.00		5.72, 11.78	
Gleason sum								
– 6 (3 + 3)	52	21.10%	23	21.10%	63	22.92%	22	22.92%
– 7 (3 + 4)	119	47.86%	52	47.71%	152	55.21%	53	55.21%
– 7 (4 + 3)	77	31.04%	34	31.19%	60	21.88%	21	21.88%
Treatment year								
– Median	2003		2003		2003.5		2003.5	
– Q1, Q3	2002, 2004		2002, 2004		2002, 2005		2002, 2005	
HT duration (months)								
– Median	5.98		5.49		5.95		5.67	
– Q1, Q3	5.55, 6.78		4.80, 8.08		5.48, 6.83		4.93, 8.51	
ProCaRS risk groups								
– Low intermediate	26	10.55%	13	11.93%	236	85.42%	92.89	85.42%
– High intermediate	222	89.45%	96	88.07%	40	14.58%	3.11	14.58%
(B) Descriptive statistics for PSM strategy 4 and CEM strategy 7 in comparison two								
	PSM strategy 6				CEM strategy 7			
	EBRT + HT (n = 126)		EBRT (n = 347)		EBRT + HT (n = 118)		EBRT (n = 377)	
Age (years)								
– Median	72		72		72		72	
– Q1, Q3	68.25, 75		68, 75		69, 75		66, 74	
Clinical T-stage								
– T1a–c	109	86.51%	298	85.91%	105	88.98%	335	88.98%
– T2b–c	17	13.49%	49	14.09%	13	11.02%	42	11.02%
PSA (ng/ml)								
– Median	8.75		8.47		8.81		8.50	
– Q1, Q3	5.71, 12.46		5.87, 12.05		5.75, 12.15		6.10, 12.30	
Gleason sum								
– 6 (3 + 3)	23	18.25%	68	19.64%	22	18.64%	70	18.64%
– 7 (3 + 4)	67	53.18%	179	51.46%	66	55.93%	211	55.93%
– 7 (4 + 3)	36	28.57%	100	28.90%	30	25.42%	96	25.42%
Treatment year								
– Median	2001		2001		2001		2001	
– Q1, Q3	2000, 2004		2000, 2004		2000, 2004		2000, 2004	
EBRT dose (Gy)								
– Median	7560		7560		7560		7560	
– Q1, Q3	7400, 7980		7400, 7980		7400, 7980		7400, 7980	
ProCaRS risk groups								
– Low intermediate	88	69.84%	252	72.75%	82	69.49%	258	68.40%
– High intermediate	38	30.16%	95	27.25%	36	30.51%	119	31.60%
CEM: Coarsened exact matching; Clinical T-stage: Clinical tumor-stage; EBRT: External beam radiotherapy; HT: Hormone therapy; ISRT: Internal seed radiotherapy; ProCaRS: Prostate cancer risk-stratification; PSM: Propensity score matching.								

CEM: Coarsened exact matching; Clinical T-stage: Clinical tumor-stage; EBRT: External beam radiotherapy; HT: Hormone therapy; ISRT: Internal seed radiotherapy; ProCaRS: Prostate cancer risk-stratification; PSM: Propensity score matching.

Table 3. Estimated treatment effects before and after matching in relation to the benchmark trial

(A) Effect estimates obtained from unmatched and matched samples from comparison one, and the benchmark trial						
Matching strategy	Unadjusted			Adjusted		
	Hazard ratio	Lower bound	Upper bound	Hazard ratio	Lower bound	Upper bound
RCT	2.17	1.33	3.45	2.04	1.25	3.33
UNM	6.55	3.82	11.26	4.48	2.44	8.22
CEM 6	3.79	1.78	8.08	3.67	1.68	8.02
CEM 8	4.04	1.88	8.66	3.84	1.77	8.34
CEM 9	2.81	1.17	6.81	2.74	1.12	6.73
PSM 9	4.25	2.23	8.08	3.76	1.94	7.27
PSM 10	4.06	1.98	8.11	3.84	1.91	7.71
PSM 11	3.86	1.85	8.05	3.87	1.84	8.15
(B) Effect estimates obtained from unmatched and matched samples from comparison two, and the benchmark trial						
Matching strategy	Unadjusted			Adjusted		
	Hazard ratio	Lower bound	Upper bound	Hazard ratio	Lower bound	Upper bound
RCT	–	–	–	1.79	1.45	2.21
UNM	1.40	0.99	1.98	1.52	1.06	2.16
CEM 7	1.53	0.95	2.46	1.55	0.98	2.45
CEM 8	1.49	0.98	2.26	1.52	1.00	2.29
CEM 9	1.48	0.93	2.36	1.52	0.92	2.43
PSM 4	1.43	1.01	2.01	1.47	1.04	2.07
PSM 5	1.43	1.00	2.04	1.46	1.02	2.08
PSM 6	1.39	0.97	1.99	1.44	1.00	2.05

CEM: Coarsened exact matching; PSM: Propensity score matching; RCT: Randomized clinical trial; UNM: Unmatched.

In comparison one, the rate of biochemical progression was greater in the EBRT + HT than the ISRT + HT group. This can, in part, be attributed to differences in the risk of biochemical progression following treatment as reflected by baseline PSA, clinical T-stage and Gleason sum. After adjusting for these variables, the estimate was attenuated and more consistent with that of the benchmark RCT. Similar results were obtained from a similar comparison using observational data and PSM, wherein the rate of biochemical failure was elevated among men diagnosed with intermediate-risk PCa and treated with EBRT alone relative to those treated with combination therapy with ISRT (HR [95% CI]: 2.27 [1.43, 3.57]) [31]. After both PSM and CEM, however, the adjusted effect was closer to that of the benchmark RCT. This might be due to limitations in confounding control afforded through regression modeling. To clarify, appropriate model specification would require adequate representation of the functional forms of the relations between the outcome and the treatment and confounders at issue, which may necessitate inclusion of polynomial terms for continuous characteristics, as well as adequate inclusion of the requisite – and possibly multiway – interaction terms between the independent variables. However, these relations do not necessarily operate according to such specifications. Furthermore, accurate modeling of effect estimates rests on the assumption of positivity, with even small violations of which potentially resulting in biased effect estimates [32]. Even without further adjustment for confounding, matching led to effect estimates closer to those obtained from the benchmark RCT than regression modeling. This might show the bias reduction potential offered through PSM and CEM even without further adjustment. However, further attenuation in the effect estimate after regression modeling after matching shows the remaining confounding not entirely managed through matching.

In the second comparison, the rate of biochemical progression was elevated in the EBRT compared with the EBRT + HT group. Similar results were obtained from a similar comparison using observational data, wherein the rate of biochemical failure was elevated among men diagnosed with intermediate-risk PCa and treated with EBRT alone relative to those who also received HT (HR [95% CI]: 1.67 [1.02, 2.75]) [33]. This difference was likely underestimated since EBRT + HT group had poorer prognostic characteristics at baseline. Regression adjustment for potential confounders led to an increased effect estimate. The relative difference in unadjusted and adjusted effect estimates compared with the first comparison was much smaller. This is likely attributable to the greater balance between treatment groups in comparison two relative to comparison one. This notion is further supported in that matching did not substantially change effect estimates.

An alternative explanation for the observed differences in the effect estimates might be that each approach estimates a different parameter. Specifically, regression modeling estimates – albeit approximately – the average treatment effect in the study population. In contrast, PSM and CEM provide for estimates of the average treatment effect among the index-treatment group (i.e., EBRT + HT). In our case, however, since some observations from the index-treatment group were dropped after PSM and CEM, we estimated the average treatment effect among the treated who remained after matching, which has been the feasible sample average treatment effect among the treated [34]. Random variation of the HR might also explain the findings, at least partly. However, effect estimates drawn from several candidate PSM and CEM strategies consistently estimated effects closer to the benchmark RCT in comparison one where imbalance was substantial; whereas effect estimates provided through several candidate PSM and CEM strategies consistently estimated effects similar to that provided through regression modeling where imbalance was not as substantial.

Our study had several strengths. First, we used a systematic approach to identifying an optimal matching strategy through identifying the ‘plateau’ in the association between balance and percentage of data retention with progressively stricter matching criteria. Second, we used matching ratios that retained a greater number of reference-treatment observations to enhance precision of the effect estimates after PSM. Third, we took advantage of *a priori* knowledge to inform our decisions on CEM coarsenings for baseline variables rather than rely on quantile-based rules, as in previous studies [12,35]. This has the potential to optimize the efficiency of matching strategies by reducing imbalance while retaining a greater part of the original data. Finally, the use of effect estimates from real-world evidence provided from RCTs performed in a similar era among patients with similar characteristics provided further guidance in the interpretation of the results.

Conclusion

Both matching strategies appear to be effective at enhancing the management of confounding in observational data with few covariates. The use of regression adjustment which should be used in conjunction with matching strategies, as shown here, has potential to control for residual confounding after matching. In contrast with recent reports, CEM appears to be a feasible strategy for preprocessing of observational data with fewer baseline covariates and *a priori* knowledge to inform coarsening of such variables that can result in retention of a large proportion of the original data from which to generate effect estimates with reasonable precision and utility to inform clinical practice in the absence of RCTs.

Summary points

- We compared the performance of propensity score matching (PSM) and coarsened exact matching (CEM) in balancing baseline covariates and data retention.
- Two treatment comparisons informed from randomized clinical trials (RCTs) were drawn from an observational pan-Canadian radiotherapy database.
- CEM and PSM led to increased balance in baseline covariates, while retaining a majority of the original data.
- Improvements in balance after matching were associated with a shift in the effect estimate closer to benchmark RCTs, compared with traditional regression alone.
- Adjustment of effect estimates after matching through multivariable regression modeling led to a further shift in the effect estimate closer to the benchmark RCT.
- PSM and CEM are effective in reducing imbalance in observational data.
- Improvements in balance through matching could hold potential to improve accuracy in effect estimation compared with traditional regression alone.
- Further adjustment of effect estimates through multivariable regression modeling has potential to control for residual confounding and should be implemented after matching.

Author contributions

Conceptualization was done by D Guy, I Karp, G Rodrigues, P Wilk and J Chin. Data curation was accomplished by D Guy and G Rodrigues. D Guy and I Karp contributed to the formal analysis. Funding acquisition was supported by D Guy and G Rodrigues. D Guy, I Karp, G Rodrigues, J Chin and P Wilk were responsible for the investigation. Methodology was completed by D Guy, I Karp, G Rodrigues and J Chin. Drafting and review of article was done by D Guy, I Karp, G Rodrigues, P Wilk and J Chin. Final approval of submitted article was provided by D Guy, I Karp, G Rodrigues, P Wilk and J Chin.

Acknowledgments

D Guy thanks the Physicians' Service Inc. Foundation or the Ontario Graduate Scholarship program for supporting research training. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Physicians' Service Inc. Foundation or the Ontario Graduate Scholarship program.

Financial & competing interests disclosure

The Physicians' Service Inc. Foundation and the Ontario Graduate Scholarship program supported research training for D Guy. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

- Schaumburg DA, Shah S, Nordstrom BL, McDonald L, Ramagopalan SV, Stokes M. Evaluation of comparative effectiveness research: a practical tool. *J. Comp. Eff. Res.* 7(5), 503–515 (2018).
- Etz A. Introduction to the concept of likelihood and its applications. *Adv. Methods Pract. Psychol. Sci.* 1(1), 60–69 (2018).
- Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. *Regression Methods in Biostatistics, 2nd Edition*. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W (Eds). Springer, NY, USA, 1–527 (2012).
- Greenland S, Schwartzbaum J, Finkle W. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am. J. Epidemiol.* 151(5), 531–539 (2000).
- King G, Lucas C, Nielsen R. Optimizing balance and sample size in matching methods for causal inference (2013). <https://gking.harvard.edu/files/gking/files/frontier.0.pdf>
- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15(3), 199–236 (2007).
- **Reviews the issue of model dependence and researcher bias and demonstrates how preprocessing through matching can reduce model dependence and thus researcher bias.**
- Grijalva CG, Roumie CL, Murff HJ *et al.* The role of matching when adjusting for baseline differences in the outcome variable of comparative effectiveness studies. *J. Comp. Eff. Res.* 4(4), 341–349 (2015).
- Yao XI, Wang X, Speicher PJ *et al.* Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *J. Natl Cancer Inst.* 109(8), 1–9 (2017).
- Rubin DB, Rosenbaum PR. Reducing bias in observational studies using score on the propensity subclassification. *J. Am. Stat. Assoc.* 79(387), 516–524 (1984).
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).
- **Defines what the propensity score estimates (i.e., probability of treatment decision given a patient's baseline characteristics) and how balancing on the propensity score is an effective way to reduce confounding by indication.**
- King G, Nielsen R. Why propensity scores should not be used for matching. *Polit. Anal.* 27(4), 435–454 (2019).
- **Reviews the limitations of matching on the propensity score to improve balance in the distribution of baseline characteristics and demonstrates the propensity matching paradox, which motivates the need for a systematic development and evaluation of matching strategies to optimize balance and sample size trade-off.**
- Fullerton B, Boris P, Krohn R, Adams JL, Gerlach FM, Erler A. The comparison of matching methods using different measures of balance: benefits and risks exemplified within a study to evaluate the effects of german disease management programs on long-term outcomes of patients with Type 2 diabetes. *Health Serv. Res.* 51(5), 1960–1980 (2016).
- Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Jessica M. Evaluating the utility of coarsened exact matching for pharmacoepidemiology using real and simulated claims data. *Am. J. Epidemiol.* 189(6), 613–622 (2020).
- **Compares the performance of coarsened exact matching relative to propensity score preprocessing techniques, using simulated and observational data.**
- Sturges H. The choice of a class interval. *J. Am. Stat. Assoc.* 21(153), 65–66 (1926).
- Rodrigues G, Gonzalez-Maldonado S, Lukka H *et al.* The Prostate Cancer Risk Stratification (ProCaRS) Project: database construction and outcome analysis. *Int. J. Radiat. Oncol.* 84(3), S57 (2012).
- Smith GD, Pickles T, Crook J *et al.* Brachytherapy improves biochemical failure-free survival in low- and intermediate-risk prostate cancer compared with conventionally fractionated external beam radiation therapy: a propensity score matched analysis. *Int. J. Radiat. Oncol. Biol. Phys.* 91(3), 505–516 (2015).

17. Morris WJ, Tyldesley S, Rodda S *et al.* Androgen Suppression Combined with Elective Nodal and Dose Escalated Radiation Therapy (the ASCENDE-RT Trial): an analysis of survival endpoints for a randomized trial comparing a low-dose-rate brachytherapy boost to a dose-escalated external beam boost for high- and intermediate-risk prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 98(2), 275–285 (2017).
18. Jones C, Hunt D, McGowan D *et al.* Radiotherapy and short-term androgen deprivation for localized prostate cancer. *N. Engl. J. Med.* 365(2), 107–118 (2011).
19. Rodrigues G, Lukka H, Warde P *et al.* The prostate cancer risk stratification (ProCaRS) project: recursive partitioning risk stratification analysis. *Radiother. Oncol.* 109(2), 204–210 (2013).
20. Stephenson AJ, Kattan MW, Eastham JA *et al.* Prostate cancer-specific mortality after radical prostatectomy for patients treated in the prostate-specific antigen era. *J. Clin. Oncol.* 27(26), 4300–4305 (2009).
21. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am. J. Epidemiol.* 163(12), 1149–1156 (2006).
22. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28, 3083–3107 (2009).
23. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33(6), 1057–1069 (2014).
24. Stuart E, Lee B, Leacy F. Prognostic score-based balance measures for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* 66(Suppl. 8), S84–S90 (2013).
25. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat. Med.* 33, 1685–1699 (2014).
26. Belitser SV, Martens EP, Pestman WR, Groenwold RHH, De Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. *Pharmacoepidemiol. Drug Saf.* 20, 1115–1129 (2011).
27. Team RS. RStudio: Integrated Development Environment for R (2021). www.rstudio.com/
28. Ho D, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* 42(8), 1–43 (2011).
29. Therneau TM, Lumley T, Atkinson E, Crowson C. Package 'survival'. 1–176 (2020). <https://github.com/therneau/survival>
30. Gayat E, Resche-Rigon M, Mary J. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm. Stat.* 11(3), 222–229 (2012).
31. Khor R, Duchesne G, Tai K *et al.* Direct 2-arm comparison shows benefit of high-dose-rate brachytherapy boost vs external beam radiation therapy alone for prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 85(3), 679–685 (2013).
32. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am. J. Epidemiol.* 171(6), 674–677 (2010).
33. Ludwig M, Kuban D, Du X, Lopez D, Yamal J, Strom S. The role of androgen deprivation therapy on biochemical failure and distant metastasis in intermediate-risk prostate cancer: effects of radiation dose escalation. *BMC Cancer* 15(190), 1–8 (2015).
34. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. *J. Am. Stat. Assoc.* 106(493), 345–361 (2011).
35. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin M. Implications of the propensity score matching paradox in pharmacoepidemiology. *Am. J. Epidemiol.* 187(9), 1951–1961 (2018).