6-29-2021 2:00 PM

Addressing Bias in Non-Experimental Studies Assessing Treatment Outcomes in Prostate Cancer

David E. Guy, *The University of Western Ontario*

Supervisor: Rodrigues, George B., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Epidemiology and Biostatistics

Abstract

Confounding is a critical concern in non-experimental comparative effectiveness research. Although regression can reduce confounding, issues of non-positivity and model dependence remain when baseline characteristics between treatment groups vary considerably. As such, we evaluated the ability of matching techniques to balance baseline characteristics between treatment groups using non-experimental data. We identified a set of balance diagnostics that assessed key differences in baseline covariates with potential for confounding. These diagnostics were used in a novel systematic approach to developing and evaluating models for use in propensity score matching that optimized balance and data retention. We then compared the performance of propensity score and coarsened exact matching strategies in optimizing balance and data retention, using non-experimental data from a pan-Canadian prostate cancer database. Both matching techniques balanced baseline covariates adequately and retained approximately 70% of the data. Improvements in balance after matching were associated with closer agreement in the effect estimate with the associated RCT compared to regression modeling alone. Furthermore, regression modelling after matching led to even closer agreement compared to matching alone. To study the role of treatment selection and prostate cancer outcomes, we performed a systematic review and meta-analysis that examined the rate of prostate cancer-specific mortality among those with high-risk non-metastatic prostate cancer who were initially treated with radiation or surgery. No statistically significant difference was found between groups in this analysis; however, this might be explained by the moderator variables of radiation type. In follow-up to this analysis, we compared the rate of metastatic progression following treatment between those with unfavorable-risk non-metastatic prostate cancer and treated with radiation or surgery, using data acquired from two Ontario cancer centers. The novel approaches to matching developed in this thesis were used to balance baseline characteristics between groups. Results from this comparison showed no statistically significant difference between treatment groups. In summary, a systematic approach to matching can be effective in balancing baseline covariates and producing more accurate effect estimates from non-experimental data. Moreover, initial treatment selection between radiation and surgery in the realm of higher risk prostate cancer does not appear to significantly influence important oncological outcomes.

Keywords: Comparative effectiveness research; non-experimental data; confounding; balance; propensity score matching; coarsened exact matching; regression analysis; prostate cancer; unfavorable-risk non-metastatic; high-risk non-metastatic; radiation therapy; external beam radiation therapy; brachytherapy; radical prostatectomy; androgen deprivation therapy

Summary for Lay Audience

Accounting for bias in research performed using nonrandomized data is necessary to validly quantify differences in treatment effectiveness. Although statistical techniques can reduce bias, they are of limited value when treatment groups vary substantially. Matching individuals between treatment groups can overcome this issue; however, depending on how matching is accomplished, different issues may persist. For example, matching directly on all patient characteristics can lead to too few matches from which to draw valid conclusions. Alternatively, using a simple score derived from patient characteristics (e.g., a health score as defined by the presence of multiple illnesses, health behaviors such as smoking, exercise and diet, and age) might be limited in its ability to differentiate between those with similar scores who might still vary considerably in important ways. As such, we compared the ability of different matching strategies to balance important patient characteristics between treatment groups, while generating enough matches. To accomplish this, a set of tests were identified from previous research that adequately quantify important differences between treatment groups when attempting to estimate treatment effects. We developed a systematic approach to matching that optimized similarity in characteristics between groups, while maximizing the number of matches made. Finally, the ability of two different matching strategies were compared using nonrandomized data obtained from a pan-Canadian radiotherapy database of men diagnosed with prostate cancer. Both strategies performed well, leading to minimal differences between treatment groups, while generating enough matches to validly estimate treatment effects. In follow-up to the matching project using prostate cancer data, we

aggregated effect estimates from studies comparing the effectiveness of radiation and surgery in treating high-risk prostate cancer, using a research technique called a systematic review and meta-analysis. No difference was found in the effectiveness of these two treatment modalities in this patient population. The last project in this thesis used the matching strategies developed in earlier chapters to compare the effectiveness of radiation and surgery in the treatment of higher-risk prostate cancer with newly acquired patient data. Like other studies, no difference in effectiveness was identified between these treatment modalities. However, due to data limitations, these estimates could not account for several potential biases which are explored in this thesis.

Co-Authorship Statement

David Guy wrote all chapters of this doctoral thesis dissertation as part of the fulfillment of requirements for their Doctor of Philosophy from the Department of Epidemiology and Biostatistics. Chapters four and five made use of secondary data obtained from the pan-Canadian radiotherapy database entitled, “Prostate Cancer Risk Stratification database” (for whom Dr. Rodrigues is the “keeper” of the data), to develop and evaluate methodological techniques for the purposes of guiding analysis performed in chapter seven. Data analyzed in chapter seven were collected by David Guy and updated by Dr. Glicksman. David Guy was responsible for the development and conceptualization of all research questions, writing of literature reviews and introductions for each research project, development and execution of methodology, statistical analyses, results, and interpretation through discussion. Their supervisory committee (Dr. George Rodrigues, Dr. Igor Karp, Dr. Joseph Chin, and Dr. Piotr Wilk) provided guidance and feedback in the conceptualization of research questions and interpretation of results. In addition to the members of the supervisory committee, colleagues from Sunnybrook Health Sciences Centre and the North York General Hospital (Dr. Chen, Dr. Glicksman, Dr. Loblaw, Dr. Buckley, Dr. Cheung, Dr. Chung, Dr. Flax, Dr. Hajek, Dr. Morton, Dr. Noakes, Dr. Spevack) aided in clarifying concepts, interpreting findings, and revising the manuscript presented in chapter seven so were awarded co-authorship for this project. Dr. Chen also contributed to the systematic review and meta-analysis performed in chapter six.

Acknowledgements

I would like to thank those who provided guidance and support throughout the process of this PhD thesis:

- To begin, my supervisor, Dr. Rodrigues, has been instrumental in the initiation and completion of my thesis. From the first request to supervise my PhD to the submission of this thesis, he has always been available and enthusiastic in providing comprehensive feedback toward the initiation and completion of numerous grant applications, and research ethics applications. He has also facilitated connections between numerous institutions across Canada, which were pivotal to the success of these projects. Finally, he has provided thorough guidance toward the completion of all of my research proposals, protocols, and manuscripts.
- Dr. Karp has provided meticulous mentorship in the methodological content of this thesis, challenging each step along the way to strengthen project quality. It has been a privilege to converse with him on a number of Epidemiological topics and develop a better understanding of the theory of Epidemiology and causal inference.
- Dr. Wilk has provided invaluable professional guidance. His wisdom has improved my understanding of the relations in academic work, which has been very helpful in navigating conflict throughout my experience at Western University. His numerous revisions have also improved the quality and clarity in my work.
- Dr. Chin has been invaluable in providing the opportunity to collect data from his experience as a Uro-oncological surgeon in treating men diagnosed with prostate

cancer. His thorough feedback on the topic content for each of my projects has improved my understanding on the pathological and clinical aspects of prostate cancer.

- The administrative team at the Department of Epidemiology and Biostatistics has provided structure and support in scheduling courses and meetings and provided information and updates to enable a smooth completion of my program.
- My classmates and colleagues at the Department of Epidemiology and Biostatistics have provided feedback on my research proposal and interesting discussions that have facilitated my academic and professional growth.
- The faculty members: Dr. Anderson, Dr. Klar, Dr. Stranges, Dr. Wilk, Dr. Karp, Dr. Mark Speechley, Dr. GY Zou, Dr. Ali, Dr. Heisel, Dr. Martin, Dr. Thind and others who have always been supportive in both my thesis work and academic development. I have always been surprised at the consistent high-quality of the teaching in this department, which has made my experience here quite valuable.
- My partner, Delia Dragomir, for being my strongest support and best example of hard work, dedication, and compassion. You have made the challenges I have faced throughout this program more manageable and meaningful. I would also like to thank Doina, Stefan and Ioana Dragomir for their kindness and support.
- My friends here in London (ON) for all their encouragement and accompaniment on all of our experiences together.
- My siblings, nieces, and nephews for the many opportunities we had to smile and laugh.
- My parents for showing me the value of hard work and dedication.

List of Tables

List of Figures

List of Appendices

Table of Contents

Thesis Map and Orientation

The goal of this thesis was to investigate methods used to mitigate bias when comparing the effectiveness of radiation therapy relative to radical prostatectomy as initial management options in treating unfavorable-risk non-metastatic prostate cancer. The first chapter provided a general review of the carcinogenesis, epidemiology, and clinical management of prostate cancer to provide understanding of the nature of this disease, information on the relative health burden and the standard of care in screening, diagnosis, prognosis, and management. This chapter also provides a review of the current state of research on the effectiveness of available treatments used in the management of unfavorable-risk non-metastatic prostate cancer and associated knowledge gaps. The challenges to performing RCTs when comparing radiation therapy and radical prostatectomy, which are the upfront standard of care options, are reviewed and illustrate the importance of evidence generated from non-experimental data when estimating their relative treatment effectiveness. A discussion on how confounding manifests when comparing the effectiveness of radiation therapy and radical prostatectomy using non-experimental data obtained from routine clinical practice is provided.

In chapter two, commonly employed methods for preventing and controlling confounding when performing comparative effectiveness research using non-experimental data (e.g., regression modeling and matching strategies) are explored. To measure the effectiveness of matching strategies in preventing confounding, balance in the distribution of baseline covariates with potential for confounding needed to be assessed. As such, in chapter three, available balance diagnostics are reviewed and a set that comprehensively measured imbalances in the multivariable distribution of baseline covariates with prognostic value when comparing treatments using non-experimental data was subsequently identified. The identified set of balance diagnostics is then used in the research work described in chapter four to inform a systematic approach to developing and evaluating propensity score models for matching. An illustration is provided using a treatment comparison from the prostate cancer literature. In chapter five, the performance of propensity score matching versus coarsened exact matching in balancing baseline covariates and thus preventing confounding is compared. This involved the use of a pan-Canadian radiation therapy database to provide two treatment comparisons. Treatment comparisons were informed by two RCTs to enable guidance in the interpretation of effect estimates before and after matching so that one could infer whether matching led to effect estimates closer to or further from those obtained from RCTs.

In chapter six, a systematic review and meta-analysis of available non-experimental studies comparing the rate of prostate cancer-specific mortality between men diagnosed with high-risk non-metastatic prostate cancer who were initially treated with radiation therapy or radical prostatectomy is performed. This analysis was performed to assess the quality and to survey the findings available in this patient population which is lacking randomized data to guide treatment selection. In chapter seven, the relative rate of metastatic progression between men diagnosed with unfavorable-risk non-metastatic prostate cancer who initially underwent radiation therapy or radical prostatectomy is estimated. For this project, non-experimental data were obtained from a Canadian academic multidisciplinary clinic where men eligible for both radiation therapy and radical prostatectomy were consulted by both a surgeon and radiation oncologist so were less likely to differ prognostically than those seen in traditional clinics. The matching methods developed in chapter four and five were used to mitigate differences in the distribution of baseline covariates and the balance diagnostics identified in chapter three were used to measure their effectiveness.

This thesis concluded with an integrated discussion of the findings from the projects performed and provided suggestions for future research on this basis.

Chapter 1: An Epidemiological and Clinical Review of Prostate Cancer

1.1 Prostate Carcinogenesis and Epidemiology

The prostate is an exocrine gland that makes part of the male reproductive tract.(1) It serves to secrete an alkaline fluid as part of the ejaculate, which protects sperm from the acidic vaginal environment to promote successful fertilization.(1) Prostate cancer (PCa) typically develops in the prostatic epithelial tissue, and growth is typically dependent on androgen signaling.(2) PCa was the second leading cancer diagnosis and fifth leading cause of cancer death globally in 2018.(3) In 2020, the Canadian Cancer Society estimated that about 23,300 men would be diagnosed with PCa in Canada and over 4,200 of those diagnosed would die from their disease.(4) Age is a well-established risk factor for PCa, with incidence increasing sharply after 50 years of age.(5) The prevalence in Canada is approximately 100 per 100,000 men between 50 and 54 years of age, increasing to 700 per 100,000 among men aged 60-64 and over 700 per 100,000 for men older than 80 years.(6) Family history is also a risk factor for PCa, increasing the incidence at an earlier age in men with compared to without a family history of PCa.(7,8) Differences in incidence rates and severity of disease have also been noted for different races.(9,10) In 2015, 157.6 per 100 000 black men were diagnosed with PCa, compared with 93.9 per 100 000 white men in the United States.(10) Black men also tend to have more aggressive disease on diagnosis and are more likely to die from their PCa than white men.(9)

1.2 Screening and Diagnosis

In earlier stages of PCa, cases are generally asymptomatic.(11) In later stages of the disease, urinary symptoms include hesitancy, nocturia and retention.(11) In the case of advanced metastatic PCa, systemic signs and symptoms may arise, including fatigue, weight loss, and bone pain.(11) Most cases of PCa are identified in their earlier stages through prostate-specific antigen (PSA) testing.(12)

The PSA is an androgen-regulated serine protease produced by the epithelial cells of the prostate.(13) The PSA is often elevated in the context of PCa, benign prostatic hyperplasia and prostatitis, and transiently after prostate biopsy, acute urinary retention, physical activity, and other activities.(1) PCa screening using PSA has been a controversial issue over the past few decades.(14) Although it increases diagnosis of indolent disease that may lead to 'unnecessary' treatment, it also increases diagnosis of aggressive PCa in its earlier and more treatable stages.(15) The most recent guidelines for PCa screening in Canada come from the Canadian Urological Association, which were published in 2017.(15) Since evidence remains equivocal surrounding the relative benefits and harms of PSA screening for patients, it is suggested that men over the age of 50 with a greater than 10-year life expectancy be engaged in a shared decision-making process of whether or not to undergo PSA screening.(15) Men younger than 45 years with a first or second-degree family history of PCa should also be considered for screening as they are at an increased risk for clinically significant PCa. Intervals between screening tests should be individualized according to previous PSA levels (e.g., PSA < 1 ng/ml and 1-3 ng/ml should repeat in four and two years, respectively).(15) More frequent intervals or adjunctive testing strategies should be considered for a significantly elevated PSA (i.e. > 3 ng/ml).(15) Deciding when to discontinue PSA screening should involve consideration of life expectancy and

PSA levels. For instance, men aged 60 with a PSA < 1 ng/ml, and men who are aged over 70 years or have a less than 10-year life expectancy should have PSA screening discontinued.(15)

Physical examination of the prostate through digital-rectal examination allows the size of the prostate gland to be assessed, nodules or lumps to be detected, and a clinical tumor stage to be assigned.(11) The digital-rectal examination and PSA testing as screening measures are often carried out by a general practitioner such as a family physician.(16) Results from these tests together with consideration of the patient's age, family history of PCa and race are used to inform whether the patient should be referred to the urologist for consideration of more definitive forms of diagnosis through prostate biopsy and histological examination.(16) If appropriate, the urologist performs a prostate biopsy and sends tissue samples to the pathologist. The pathologist examines the microscopic appearance of the prostate's glandular architecture. It is the responsibility of the pathologist to diagnose the PCa and assign a grade or score based on the Gleason system.(17) A Gleason grade of 1 to 2 is assigned when the glandular appearance is very organized and non-dysplastic upon microscopic examination.(18) A Gleason grade of 3 indicates that the glandular appearance is sufficiently disorganized to warrant a diagnosis of PCa.(18) Further information on Gleason grading will be covered in Section 1.3.3.

1.3 Risk-Stratification

Risk-stratification of PCa is defined by the association of pre-treatment variables with the trajectory of disease following treatment. To objectively define and validate the prognostic value of risk-strata, characteristics of the disease are measured before treatment and are correlated with measures of disease progression following treatment.(19) Risk-stratification serves important purposes in research and clinical contexts.(20) First, clinical trials can condition/adjust estimates of treatment effect on risk strata to reduce heterogeneity, therefore improving statistical power to identify treatment effects. This also has the added benefit of improving precision and accuracy in effect estimates for specific subpopulations and allows advances in treatment protocols to target different levels of disease appropriately (e.g., multimodal treatment for more aggressive diseases vs unimodal treatment for more indolent disease). Clinicians can use this information to guide treatment decisions regarding selection of modality and the use of multimodal therapies. This ultimately reduces the variability in treatment decisions because of clinician bias, experience, and knowledge. Risk-stratification also provides a common nomenclature to define PCa characteristics, improving communication between clinicians and institutions. This facilitates collaboration in research and patient management to ultimately improve progress and patient care.

Recommended baseline characteristics used to define risk-strata in the clinical context include measurements of PSA, clinical stage derived from digital-rectal exam, Grade Group, amount of cancer on biopsy and imaging.(21) Their use in defining risk-strata is based on their demonstrated predictability of important disease endpoints following treatment. The following paragraphs describe each factor and how they relate to clinical endpoints and thus the rationale for their contribution in risk-stratification. A summary table is provided outlining current definitions of risk-strata.

1.3.1 The Prognostic Role of PSA

Studies investigating the association of PSA and pathological outcomes following surgical resection of the prostate through radical prostatectomy (RP) have demonstrated significant correlation between rising PSA from 4 ng/ml and important clinical and surgical endpoints (i.e., pathological stage, organ-confined disease, extraprostatic extension, seminal vesicle invasion and disease-free survival).(22,23) Similar results were found following PCa radiation therapy (RT) in two clinical trials where high-risk PCa patients with PSA levels greater than 50 ng/ml had inferior clinical and biochemical endpoints (i.e., overall survival, distant metastasis, and biochemical failure) compared to other high-risk PCa patients.(20) Common ranges of PSA that are used in risk-stratification include <10, 10.1-20, and >20 ng/ml;(21,24,25) however, new cut-off values are also being evaluated.(20)

1.3.2 The Prognostic Role of Clinical Staging

Clinical staging is accomplished through digital-rectal examination, imaging, and histological results from tissue biopsy. Upon digital-rectal examination, presence of prostatic nodules, asymmetry and/or induration elevate suspicion of PCa and can help determine the extent of tumour involvement.(26) Imaging modalities such as transrectal ultrasound and magnetic resonance imaging can also detect prostatic lesions which might represent malignant disease.(21) The most widely used system for clinical staging is the American Joint Committee on Cancer/Union Internationale Contre le Cancer (AJCC/UICC) tumor, node and metastasis staging system.(27) Since the work in this thesis focuses on non-metastatic PCa, only the tumor (T) aspect of the AJCC/UICC staging system will be reviewed. Clinical stage T1 (cT1) is a clinically inapparent tumor that is neither detectable on digital-rectal exam nor imaging.

Subcategorization of cT1a and cT1b correspond to an incidental tumor finding on transurethral resection of the prostate involving ≤5% and >5%, respectively, of the number of chips or amount of tissue resected out for clinically benign disease. Patients undergo transurethral resection of the prostate to relieve their obstructive urinary symptoms and function, usually due to benign prostatic hypertrophy.(18) cT1c corresponds to identification from needle biopsy secondary to an elevated PSA test without palpable disease. Clinical stage T2 (cT2) are tumors detectable by digital-rectal exam or imaging but that are perceived to be confined within the prostate. cT2a tumor involves up to one-half of one lobe, cT2b involves more than one-half of one lobe and cT2c involves both lobes. Clinical stage T3 indicates that the tumor has grown outside of the prostate and either has not spread to the seminal vesicles (T3a) or has spread to the seminal vesicles (T3b). Clinical stage T4 indicates that the tumor has spread to tissues next to the prostate other than the seminal vesicles.

The AJCC/UICC staging system describes the anatomic extent of disease, which has value in treatment planning. For instance, a greater cT stage indicates a greater anatomic spread of the disease that might warrant delivering RT to the tissues surrounding the prostatic capsule or pelvic lymph node dissection through RP to evaluate whether the cancer has metastasized beyond adjacent tissues. Increased classification, however, is not necessarily associated with poorer prognosis.(28) For instance, many studies have demonstrated similar PSA recurrence rates post-RP between cT1c and cT2a tumors.(29–34) The lack of differentiation between cT1c and cT2a likely results from the low sensitivity of digital-rectal examination in assessing the presence and extent of PCa since it relies on the clinician’s ability to detect aspects of the tumor and the patient’s anatomy to facilitate examination (e.g., prostates of overweight compared to normal weight patients are generally more difficult to palpate). Obek and colleagues found that in PCa tumors characterized by digital-rectal examination as unilateral, 69% were pathologically bilateral and 4% were cancer free in the lobe with a palpable abnormality.(35) Moreover, increasing tumor involvement laterally has not significantly correlated with biochemical failure independent of other established prognostic factors.(36) Finally, subcategorization of T1a and T1b is derived from a single study, involving 117 patients that found subdividing disease into transurethral resection of the prostate specimens with more or less than 5% tumor was associated with noticeable differences in clinical outcomes.(37) Since very few instances of PCa are detected through transurethral resection of the prostate, these cT1 subcategorizations have limited applicability.

1.3.3 The Prognostic Role of Grade Group

Grade Group is determined by the most and second most predominant pathological Gleason grade on biopsy.(21) Gleason grade is based on glandular differentiation, with well-differentiated tumors representing lower-risk PCa that tends to grow and spread slower, thus indicating a better prognosis.(24,25) Although Gleason grading has been reported in a variety of ways, the Grade Group system has been recommended for risk-stratification to inform decisions regarding treatment of localized prostate cancer.(17,21) However, since the data analyzed in the this thesis arose from before the latest Grade Group system was adopted, the traditional Gleason sum will be used instead. A Grade Group of 1 corresponds to the most and second most predominant Gleason score on biopsy of 3 and 3, respectively. This corresponds to a Gleason sum of 6 (3+3). A Grade Group of 2 corresponds to a Gleason sum of 7 (3+4), while a

Grade Group of 3 is defined by a Gleason sum of 7 (4+3). Finally, Grade Groups of 4 and 5 correspond to Gleason sums of 8 (3+5, 4+4, or 5+3) and 9 (4+5 or 5+4) to 10 (5+5), respectively.(17) The prognostic ability of the Grade Group system has been validated by multiple institutions on the basis of 4-year biochemical progression-free survival (BPFS) rates with increased clinical Grade Group values corresponding to significantly decreased rates of BPFS.(38)

Other measures used in the subcategorization of low-risk PCa into very low-risk PCa include PSA density, percentage of biopsy cores positive for malignancy, and highest involvement of malignant tissue in a biopsy core.(21) PSA density is a quotient of serum PSA and prostate volume. Specifically, a value exceeding 0.15 ng/ml of PSA per cm^3 of prostatic tissue increases the suspicion for clinically significant PCa.(15,39) Percentage of positive biopsy cores has been strongly associated with PCa death when assessed as a continuous and ordinal variable.(40) Moreover, the extent of core involvement on tissue biopsy has also been associated with important endpoints (e.g., ≤10, >10-25, >25-75 and >75% core involvement rates corresponded to PCa death rates of 8, 21, 38, and 56%, respectively).(40) Multiple studies have found that men presenting with low-risk PCa who also had a PSA density <0.15ng/ml/cm^3, ≤2 positive cores on biopsy and no core with >50% involvement had a very low probability of adverse pathology at surgery and rate of metastatic disease when managed with active surveillance.(41–43)

1.3.4 Summary

Risk-categories are intended to simplify decision making and have both research and clinical value.(19,21) Very low-risk PCa has a metastatic progression rate of <1% while on active surveillance at 15 years whereas PCa progression of men with low-risk PCa while on active surveillance is not as clear.(42,44,45) As a result, patients with very low-risk disease should be strongly recommended active surveillance whereas select patients with low-risk may be offered definitive therapy.(21) Further subcategorization of the traditional intermediate-risk category into favorable and unfavorable intermediate-risk was precipitated by significant differences in recommended management options in imaging, pelvic node dissection during RP, and advisability of androgen deprivation therapy (ADT) in conjunction with RT.(21) Finally, substratification of high-risk PCa into high and very high-risk does not provide much clinical utility, as management plans are very similar even though outcomes differ significantly.(19,21) As a result, these risk strata have been collapsed for clinical purposes, but retain value for research purposes.

Table 1.1 Risk-stratification of non-metastatic prostate cancer

	*AUA/ASTRO/SUO Guidelines	ProCaRS Risk-Stratification	NCCN
Very low risk	PSA <10 AND Grade Group 1 AND clinical stage T1-T2a AND <34% of biopsy cores positive AND no core with >50% involved, AND PSA density <0.15 ng/ml/cc	Clinical stage T1–T2a AND PSA ≤6 AND Gleason score ≤6	Clinical stage T1c, Gleason sum ≤6, PSA <10, <3 biopsy cores positive, ≤50% cancer in each core, and PSA density <0.15 ng/ml/g
Low risk	Clinical stage T1-2a AND PSA <10 AND Grade Group 1 AND not very-low-risk	Clinical stage T1–T2a AND PSA >6 to ≤10 AND Gleason sum ≤6	Clinical stage T1-2a, Gleason sum ≤6, and PSA <10 AND not very-low-risk
Favorable intermediate risk	Clinical stage T2b-c AND Grade Group 1 (with PSA 10 to <20) OR Grade Group 2 (with PSA <10)	PSA ≤10 (with T2b-c OR Gleason sum =7) OR PSA >10 to ≤20 (with T1-2a AND Gleason sum ≤6)	Clinical stage T1–T2, PSA ≤20, and Gleason score ≤7 not otherwise low-risk
Unfavorable intermediate risk	Grade Group 2 (with either PSA 10 to <20 or clinical stage T2b-c) OR Grade Group 3 (with PSA <20)	Gleason sum=7 and at least one of PSA >10 to ≤20 OR T2b-c	
High risk	PSA ≥20 or Grade Group 4-5 OR clinical stage ≥T3	Core positivity <87.5% AND Clinical stage T3–T4 OR (PSA >20 to PSA <30) OR Gleason sum 8 to 10	Clinical stage T3–T4 or PSA >20, or Gleason score 8–10
Very high risk	-	Clinical stage T3–T4 OR (PSA >20 to PSA <30) OR Gleason sum 8 to 10	Clinical stage T3b–4

		AND (PSA >30 OR core positivity >87.5%)	

*AUA = American Urological Association; ASTRO = American Society for Radiation Oncology; SUO = Society of Urologic Oncology; ProCaRS = Prostate Cancer Risk-Stratification; NCCN = National Comprehensive Cancer Network

1.4 Oncological Outcomes in Prostate Cancer

Assessing treatment outcomes in PCa is necessary to inform patients and clinicians of the relative harm to benefit ratio for different therapeutic options. Outcome measures should be meaningful or reliably determine meaningful outcomes in PCa such as disease-free status and survival. As well, they should occur frequently enough to allow comparison between therapies and standardized to enable valid comparisons between studies. In the context of localized PCa, the optimal outcome is PCa-specific survival (CSS) rather than overall survival (OS), as PCa is slow growing and many patients with PCa die from other causes.(46) However, OS is also important as treatments can directly and indirectly lead to death. Unfortunately, meaningful comparisons of CSS might only be feasible many years after follow-up due to the slow growing nature of PCa.(47,48) Surrogate markers are often used to assess short-term outcomes associated with CSS such as changes in PSA following therapy, and development of radiographic or bone scan evidence of metastasis.(46) These measures have been found to antedate CSS by approximately 13 and 5 years, respectively, while occurring approximately 2- and 10-years post-intervention, respectively.(47–49) Moreover, metastatic progression has been found to be the primary determinant of CSS.(47,49) As a result, studies comparing new therapies or prognostic markers in the realm of localized PCa use such definitions to accomplish their work in a more reasonable time frame.

1.4.1 Biochemical Failure

In the context of localized PCa, values indicating biochemical failure depend on the type of intervention. In the setting of RP, surgeons attempt to remove all benign and malignant prostatic tissue, so PSA should fall to an undetectable level if treatment is successful.(50) However, this does not occur for about 4 weeks since serum PSA half-life is about 2.6 days.(50) As such, it is recommended that the first PSA-test be performed 3-months after surgery.(51) Approximately, 20-40% of cases will demonstrate rises in PSA level post-intervention.(52,53) Minimal rises in PSA may indicate incomplete resection of benign prostate tissue whereas more significant and rising PSA levels raise concern of persistent local or distant metastatic disease.(54) Although a consensus has not been reached on what denotes a significant PSA level post-RP, different thresholds have been proposed for indicating biochemical recurrence following RP, ranging from ≥0.2 to ≥0.5 ng/ml.(46) Patients with a PSA of ≥0.2 ng/ml post-RP are said to have biochemical failure and are 53% likely to develop clinical recurrence over time.(47) Others have suggested ≥0.4 ng/ml as a more clinically relevant cut-off since it has been shown to optimally discriminate men who later present with metastatic progression as well as strongly correlate with continued PSA progression, secondary therapy, and a rapid PSA doubling time compared to competing definitions.(49,55,56) As such, they suggest that this should be the standard definition of biochemical recurrence for reporting outcomes following RP for biochemical endpoints in clinical trials using combined modality treatment strategies and to identify patients suitable for systemic therapy in clinical trials post-RP.(49) However, other factors should be considered when deciding to initiate salvage treatment such as life-expectancy as determined by age and general health status as well as Gleason score, pathological stage, surgical margin, and lymph-node status, since these also predict biochemical and clinical progression, and disease-specific and overall mortality.(57)

Changes in PSA following RT differ compared to RP and even among different RT approaches, making classification of biochemical failure complicated.(58,59) As prostatic tissue is irradiated, cells are damaged and become inflamed, releasing PSA into the circulation.(58) A phenomenon known as PSA bounce also commonly occurs in the setting of radioactive seed implantation followed by external beam radiation (EBRT).(59) In 1996, the American Society for Therapeutic Radiology and Oncology initially defined biochemical failure after EBRT as three consecutive PSA rises after nadir.(60) The date of failure was defined as the halfway point between the nadir date and the first rise. This definition posed several short-comings, including not being linked to clinical progression or survival, performing poorly among those receiving ADT, biasing estimates of event-free survival and violated the proportional hazards assumption leading to statistical limitations in its reporting.(61) As such, the definition has been revised to address these shortcomings in a second consensus panel by the American Society for Therapeutic Radiation Oncology - Radiation Therapy Oncology Group.(61) They recommended that a rise of PSA by ≥2 ng/ml above the nadir should be the standard definition of biochemical failure after EBRT with or without ADT, which is now convention.(61) Thompson et al further evaluated this definition in the context of brachytherapy (BT) and found that 44% of patients whose PSA rose ≥2 ng/ml above nadir, subsequently fell to ≤0.5 ng/ml without intervention.(62) They defined this as the benign phenomenon known as PSA bounce. Other definitions of PSA bounce include specified increases in PSA above nadir by 0.1,(63,64) 0.2,(64–66) 15%,(67) and 35%.(64) Interestingly, patients who experience PSA bounce after BT tend to have better outcomes in terms of biochemical failure, metastatic progression, and OS than those not demonstrating PSA bounce.(68)

1.5 Management

The most established management methods include active surveillance, RT, and RP, while more investigative novel ablative techniques include high-intensity focused ultrasound and cryotherapy.(21) Many factors are involved when deciding upon treatment for PCa such as risk-strata of PCa, age, life expectancy, pre-treatment general function and genitourinary symptoms, expected post-treatment function, and potential for salvage therapy.(21) Many treatments are clinically acceptable in the realm of PCa with the ratio of harm to benefit often being equivalent.(69,70) As such, guidelines developed by the American Urological Association, American Society for Therapeutic Radiation Oncology and Society of Urologic Oncology stress the importance of a shared decision-making process that incorporates best evidence with patient values to determine the appropriate course of therapy.(21) This involves consideration of the aforementioned factors that influence treatment decision and consultation with different PCa care specialists.(12,21)

1.5.1 Favourable-Risk Prostate Cancer

The standard of care for a diagnosis of very low-risk or low-risk PCa is active surveillance.(71) Active surveillance entails delayed treatment in the presence of continuous monitoring of PCa growth and progression wherein reclassification to a higher risk of disease progression prompts consideration of definitive intervention.(18) This differs from watchful waiting wherein the consideration of treatment does not occur until symptoms and/or signs emerge.(18) Watchful waiting might be the appropriate choice for an individual who has competing illnesses such as cardiovascular and respiratory diseases that decrease their life-expectancy, reducing the utility of active treatment to improve life-expectancy or quality of life. Several longitudinal studies have indicated that delayed intervention in the context of active surveillance compared to immediate definitive therapy does not lead to significant differences in biochemical recurrence rates, positive surgical margins, extraprostatic extension,(72–74) or risk of incurable disease.(75,76) Moreover, multiple active surveillance programs have demonstrated OS rates to be between 97% to 100% at 15-years.(42,77–82)

Based on the international success of many active surveillance programs, a best practice guideline was created outlining appropriate steps in active surveillance monitoring.(71) Guidelines suggest that following a diagnosis of very low- and low-risk PCa, patients should receive a PSA test every 3 to 6 months, an annual digital-rectal examination, and a 12 to 14 core confirmatory transrectal ultrasound biopsy within 6 to 12 months of diagnosis with biopsies repeated every 3 to 5 years thereafter.(71) Clinicians may also consider multiparametric MRI if pathologic and clinical findings are discordant,(71) as research has shown high negative predictive values between 83% and 100%,(83) and has shown multiparametric MRI to be a good predictor of disease reclassification in the context of active surveillance.(84,85)

Upon re-biopsy, an increase in Gleason sum or volume should prompt consideration of definitive therapy.(86,87) Specifically, a Gleason sum of at least 7 (3 + 4) with greater than 10% involvement of Gleason 4 warrants definitive treatment.(71) Moreover, significant increases in Gleason sum 6 volume should also prompt consideration of definitive therapy; however, clear criteria for total volume are currently not available. These recommendations are supported by results showing reduced risk of distant metastasis and PCa-specific mortality for men undergoing RP versus watchful waiting in the Scandinavian Prostate Cancer Group-4 Randomized Trial.(87) The following modes of definitive treatment can be considered in the context of reclassification while on active surveillance.

1.5.2 Intermediate-Risk Prostate cancer

As mentioned previously, intermediate-risk PCa can be categorized as favourable or unfavourable. Patients diagnosed with favourable intermediate-risk PCa should consider active surveillance.(71) The standard therapy for unfavourable intermediate-risk PCa is either RP or RT with adjuvant ADT.(21) Other options include RT alone, whole gland cryosurgery, high intensity focused ultrasound and focal therapy; however, evidence surrounding these options is less robust.(21) Unfavourable intermediate-risk PCa patients may also be offered active surveillance if life expectancy is ≤ 5 years.(21)

RP involves surgical removal of the prostate gland with curative intent of PCa. Derivations of the procedure exist, including open, laparoscopic, and robotic approaches. In the Scandinavian Prostate Cancer Group Study Number 4 (SPCG-4), men randomized to RP with a diagnosis of Gleason sum of ≥ 7 and PSA ≥10 ng/ml had a reduced relative risk of death from any cause and PCa-specific mortality compared to men randomized to watchful waiting.(87) However, watchful waiting in this study did not emulate active surveillance in that men were not continuously monitored and were offered trans-urethral resection of the prostate (TURP) in the presence of obstructive voiding symptoms or ADT upon image detected metastases. Observations from the Prostate Cancer Intervention versus Observation Trial (PIVOT) also indicated significant reduction in PCa-specific mortality among intermediate-risk PCa or with a baseline PSA ≥10 ng/ml who underwent RP.(86) Similarly, however, the observation arm did not include continuous monitoring but rather palliative and chemotherapy in the presence of symptomatic or metastatic progression. Ten-year results from the ProtecT trial demonstrated no difference in PCa-specific mortality among patients with localized PCa who were randomized to active surveillance, RP or RT+ADT.(88,89) These results did not vary according to PSA level (< or ≥6 ng/ml), Gleason sum (6 or ≥7), or clinical stage (T1c or T2) at diagnosis. However, incidence of metastatic disease and clinical progression were significantly higher in the active surveillance compared to the RP and RT groups. Since PCa is more often a slow growing tumor, the authors note that longer follow-up is necessary to ascertain results regarding CSS.

The practice of RT has evolved over the past few decades and encompass a number of techniques including EBRT, and BT.(90) EBRT approaches include intensity modulated and stereotactic body RT, while BT includes both low and high dose rate.(90) Low dose rate BT is reserved for less aggressive PCa as a monotherapy wherein radioactive seeds are placed in the prostate and left permanently.(18) High dose rate BT, on the other hand, involves the insertion of thin metal rods to direct high doses of radioactive rays to the prostate.(18) It can be used in conjunction with EBRT to deliver the total required RT dose in fewer treatments.(18)

The RTOG 9408 randomized trial found benefit in RT with compared to without short-term ADT in 10-year OS and CSS among 1979 men diagnosed with early stage localized PCa.(91) Post-hoc analysis revealed a stronger benefit in intermediate- compared to low-risk men. These findings were consistent with a smaller trial which randomized men to RT with versus without six-months of ADT and found a 15-year OS benefit.(90) A series of comparative outcome studies based on retrospective data suggest benefit in survival for RP compared to EBRT and BT.(92,93,102,94–101)

Since RT and RP are not appropriate for all patients, certain circumstances warrant cryosurgery, HIFU and focal therapy. Recent guidelines for management of clinically localized PCa suggest cryotherapy may be appropriate in select patients depending on preferences, comorbidities and life expectancy.(90) Further, HIFU and focal therapy offer quality of life advantages to traditional approaches; however, studies comparing effectiveness with traditional approaches are lacking so are recommended to be offered only in the setting of a clinical trial.(90)

1.5.3 High-Risk Prostate Cancer

Recommended therapy for high-risk PCa is RP or RT with ADT.(90) Results from the SPCG-4 trial have demonstrated benefit in OS and CSS for RP over watchful waiting.(87) Results from PIVOT indicated reduced metastases at 10 and 12-year follow up among those who underwent RP over watchful waiting, while men with high-risk disease specifically had a lower PCa-specific mortality rate (9.1% v 17.5%).(86) A randomized trial has shown benefit in clinical disease-free survival among locally advanced PCa patients receiving RT and ADT over RT alone.(103) Moreover, the rate of PCa-specific mortality among men with locally advanced PCa was elevated for those receiving short-term (6 months) compared to long-term (3 years) ADT (HR: 1.71 [1.14 – 2.57]).(104) OS and CSS were also improved among men with Gleason 8-10 who were randomized to long-term compared to short-term ADT and RT (31.9% vs 45.1% and 83.9% vs 88.7%, respectively).(105)

Data comparing cryosurgery and traditional therapies (i.e., RP and RT) in high-risk patients is insufficient for the triage of patients between these two modalities.(90,106) Similarly, data comparing effectiveness between HIFU and traditional therapies is non-existent, while clinical studies vary in outcome reporting.(107) As such, guidelines recommend that these therapies be offered to high-risk patients only in the context of a clinical trial.(90)

1.6 Gaps in Evidence Informing Care of Men Diagnosed with Unfavorable-Risk Prostate Cancer

Despite the advances surrounding treatment selection and sequencing for unfavorable intermediate- and high-risk PCa noted above (referred to collectively as unfavorable-risk PCa), optimal initial therapy remains an area of intense academic and clinical debate.(108) The lack of clinical management clarity is due to barriers in obtaining high-quality evidence on the relative efficacy between common initial treatment options in the current era (i.e., RP and RT).(109) For instance, strong patient preferences surrounding RP and RT have resulted in numerous RCTs failing to accrue to completion, especially in North America.(110)(111) This occurs as RP is associated with increased rates of certain adverse functional outcomes, including erectile dysfunction,(88) and urinary incontinence,(112) while RT is typically associated with increased rates of urinary obstruction and irritation and bowel dysfunction and rectal bleeding.(113) Since most patients have strong preferences with regards to such outcomes, they generally decline randomization.

In the absence of RCT data, multiple investigations have been performed to compare common treatment options using non-randomized data, which has its own challenges/limitations when trying to generate credible evidence to inform clinical practice. As patients are not randomized, treatment assignment is influenced by both observed and unobserved factors that also influence relevant clinical outcomes of interest. For instance, candidates for RP compared to RT generally have less aggressive tumor characteristics, are younger and with fewer comorbidities.(114) This occurs as patients who are older, especially those with multiple comorbidities, are less suitable for RP because of peri-operative risks and delayed recovery.(115) As a result, these patients are more likely to be treated with RT. In addition, almost all patients are diagnosed with PCa by their urologists, and many are not seen by a radiation oncologist first to discuss management options. Since each specialist is more likely to recommend the treatment that they provide, candidates eligible for both RP and RT (i.e., younger patients with fewer comorbidities and less aggressive disease) are more likely to receive RP.(116) Age, comorbidity status and PCa aggressiveness are known to influence outcomes of interest negatively. As a result, patients undergoing RP are more likely to have better outcomes than those undergoing RT independent of treatment assignment. Crude categorization of confounders and/or not including all confounders in adjusted analyses and limitations in statistical adjustment can lead to improved CSS and OS among patients receiving RP compared to RT independent of treatment status.(117–120)

Common endpoints used to compare treatment effectiveness in PCa also have potential for bias. As mentioned before, these include BPFS, metastatic progression-free survival (MPFS), CSS and OS. As mentioned previously, the definition of biochemical failure among patients treated with RP and RT differ. Since RP removes the whole prostate gland, it is anticipated that the prostate specific biomarker (used to define biochemical failure) maintain a non-significant value of <0.2ng/ml.(121,122) Since RT does not remove the gland, it is anticipated that some prostate specific biomarker remains at significant levels of >2ng/ml above nadir.(61) The biochemical failure definition post-RP is intended to indicate cure whereas the definition post-RT is sensitive and specific for future clinical outcomes of interest (e.g., distant failure).(123,124) These definitions represent different disease kinetics and outcomes and should not be compared in the context of comparative effectiveness research. Evidence of metastasis depends on the presence of prompts to image such as biochemical failure and symptoms. Prompts may differ depending on whether the patient underwent RP or RT and the frequency of follow-up, which is dependent on characteristics of the patient, physician, and treatment centre among other factors. Finally, PCa-specific mortality as ascertained by death certificates is not immune to bias. Death is sometimes misattributed to PCa among PCa patients who die of other causes.(125) This misattribution bias is more likely to occur among those with multiple comorbidities, as deciphering cause of death among multiple causes can be difficult. Since RT patients are more likely to have multiple comorbidities, this would negatively impact survival outcomes when compared with patients undergoing RP independent of treatment status. OS poses little concern for measurement bias, as causes of death do not have to be ascertained.

Ascertainment bias can also influence treatment effect estimates. Consider a patient diagnosed with PCa and treated with RT. Diagnosis of PCa is accomplished through obtaining prostate tissue samples through needle biopsy to confirm the histopathological presence of cancer. In the case of a patient diagnosed with PCa and treated with RP, surgical resection of the prostate would allow for more thorough examination to detect any missed pathology on biopsy, resulting in upstaging of Grade Group.(126) As a result, patients undergoing RT may harbor more aggressive cancer identified at diagnosis compared to patients undergoing RP. Consider comparing survival outcomes between a group of PCa patients treated with RT who received a diagnosis of Grade Group 1 (low-risk PCa phenotype) on biopsy with a group of PCa patients treated with RP who received a diagnosis of Grade Group 1 based on surgical pathology. Some patients in the RT group will likely harbor Grade Group 2 disease, which is associated with lower rates of survival while individuals in the RP group are unlikely to harbor Grade Group 2 cancer. As a result, patients treated with RP will appear to have better survival outcomes independent of their treatment.

Finally, RT technology has evolved rapidly in recent years, such that observational data tends to reflect outdated regimens deemed less effective.(127)(128) Many authors have therefore discounted the results from such studies on the basis of irresolvable bias and technological drift.(108) Without relevant high-quality evidence, treatment decisions become more dependent on physician biases, differences in knowledge and experience as well as educational background. As such, it is important to address and account for identified sources of bias in observational studies and investigate modern RT approaches in relation to RP to improve evidence quality and credibility in guiding treatment decisions.

1.7 Transition to Chapter Two

This chapter provided a review on the nature of PCa, basic Epidemiological information and approaches to identification of PCa through screening and diagnosis, categorization of the risk of PCa through risk-stratification and subsequent management options. The main thrust of this thesis, as per the title, is mitigating bias in PCa comparative effectiveness research. As such, a focus was placed on the bias associated with comparing the effectiveness of RT and RP as an initial treatment for men diagnosed with unfavorable-risk non-metastatic PCa (i.e., unfavorable intermediate- to very high-risk non-metastatic PCa), as this is a substantial concern in the field of PCa comparative effectiveness research. The following chapter provided a review on popular methods in the management of confounding (i.e., both prevention through matching and control/adjustment through regression analysis) when performing comparative effectiveness research using non-experimental data.

Chapter 2: Background and Objectives

2.1 The Value of Non-Experimental Data in Comparative Effectiveness Research

Non-experimental comparative effectiveness research aims to generate evidence on the relative effectiveness and safety of different treatment approaches based on observations from routine clinical practice. This evidence can be used to identify more suitable treatment approaches for particular patients. Since the implementation of electronic patient health records, the proliferation of medical record and administrative claims databases has led to non-experimental research occupying a large proportion of comparative effectiveness research.(129) Benefits attributable to large non-experimental databases include potentially increased power for statistical analysis, a broader range of patient characteristics (which can help enhance the applicability of results), longer follow-up periods (allowing for study of longer-term outcomes), and the ability to address research questions that are impractical in the setting of a RCT.(129) Evidence from comparative effectiveness research using non-experimental data is also becoming increasingly implemented to guide clinical and policy decision-making based on the effectiveness and safety of treatments.(129)

2.2 The Issue of Confounding in Non-Experimental Datasets

Unlike in RCTs, subjects in non-experimental studies are not randomized to treatment groups; rather, treatment decisions are influenced by factors such as age, health status, disease severity, income, education, and patient and physician preferences, among other factors. These factors may also influence the occurrence/level of the outcomes of interest. As such, crude comparisons of the occurrence/level of the outcomes between treatment groups may not reflect differences in effectiveness in the treatments under comparison. Factors that influence both treatment decisions and the occurrence or level of the outcome are known as confounders and must be accounted for when estimating treatment effects using non-experimental data.

2.3 The Counterfactual Theory for Valid Causal Inference

The counterfactual theory for causal contrasts is a popular and useful framework for defining and addressing confounding in order to accurately estimate treatment effects.(130) The theory requires that in order to estimate a causal effect, we need to set up a valid causal contrast.(130) The ideal causal contrast is one where individuals from a population of interest that are exposed to a treatment of interest (index-treatment) are identical to those from the same population of interest but are unexposed or exposed to an alternative therapy used for comparison (reference-treatment). If this requirement is met, the observed variation in the outcome of interest is due to the difference in index- and reference-treatments. This relationship can be represented by the equation $E\left[Y^i\right] = \sum(Y^i - Y^r)$ wherein $E\left[Y^i\right]$ denotes the expected mean index-treatment effect at the population level and Y^i denotes the outcome in those who received the index-treatment and Y^r denotes the outcome in those who received the index-treatment had they been exposed to the reference-treatment. Since the ideal causal contrast is not observed in reality, it is counter to what is factual, and must be estimated from available data.(131) The goal of comparative effectiveness research using non-experimental data under this framework then becomes estimating the average outcome among the reference-population to accurately estimate the counterfactual contrast in order to validly estimate index-treatment effects.

The gold standard for estimating the counterfactual contrast is through a RCT. By randomizing individuals from a population of interest to index- and reference-treatments, each group contains individuals that, on average, share a similar distribution of baseline characteristics. As calendar time progresses, any variation in the presence or level of the outcome of interest observed between the index- and reference-treatment groups can be attributed to the difference in treatments. This assumes that drop-out between groups is non-random/informative with regard to the outcome risk. The index-treatment effect can then be calculated through comparing the average outcome level between index- and reference-treatment groups. This differs from the ideal measure in that no two individuals from different treatment groups will be identical, thus preventing us from calculating the individual index-treatment effect. However, we expect that, on average, the distribution of characteristics between the two groups will be approximately the same, so most of the observed outcome variation that results between the groups can be attributed to the index-treatment. This enables us to estimate the average index-treatment effect rather than the individual index-treatment effect (i.e., $E\left[Y^i\right] = E\left[Y^i\right] - E[Y^r]$). However, the average index-treatment effect is equal to the average of individual index-treatment effects.

Hernàn and Robins outline three key conditions required to identify and accurately estimate the index-treatment effects under the counterfactual framework: exchangeability, positivity and consistency.(130) Exchangeability implies that if individuals from the reference-treatment group were instead exposed to the index-treatment, we would observe the same population effect as those in the index-treatment group and vice versa. That is, prognostically relevant characteristics (i.e., potential confounders) from the reference-treatment and index-treatment groups are the same. Although in RCTs, random variability may introduce some imbalance in baseline characteristics between exchangeable groups, we anticipate that as the sample size grows larger, this imbalance dissipates and becomes less consequential. Positivity indicates that the probabilities of an individual receiving the index- or reference-treatment are both positive given their baseline characteristics (e.g., age, income, education, etc.) so that baseline characteristics that influence the outcome overlap between treatment groups. This is apparent in RCTs as the probability of receiving either treatment does not depend on individual characteristics but on a random process wherein each individual has an equal probability (e.g., 0.5) of being assigned to either treatment. Finally, consistency indicates that the index- and reference-treatments, explicitly and clearly defined, are the same between individuals so that one individual does not receive a different intensity or timing of either treatment, thereby reducing outcome heterogeneity as a result of one index- or reference-treatment rather than multiple subtypes of either treatment. This is approximated in RCTs, to varying degrees, as each participant is expected to, thought does not always, adhere to the same treatment protocol.

Since well-performed RCTs closely approximate these three conditions, they are regarded as the gold standard when making causal inferences. Unfortunately, RCTs are costly, time-intensive and often pose ethical constraints.(111) As a result, clinicians and policy makers often rely on evidence generated from non-experimental data, which tends to deviate substantially from all three principles of the counterfactual framework. For example, when trying to estimate index-treatment effectiveness using non-experimental datasets, those receiving the index-treatment likely differ in important prognostic variables compared with those receiving the reference-treatment, thus deviating from exchangeability. Also, the reason some participants do not receive the index- or reference-treatment could be due to absolute or relative contraindications, thus deviating from positivity. Finally, the index- and reference-treatments may vary in how they are delivered in terms of intensity, timing, and adjunctive therapies, thus deviating from the consistency condition. Issues of positivity and consistency in this scenario can be mitigated through restricting the analysis to only those eligible for both index- and reference-treatments who received similar exposure intensities, at similar times, and with similar adjunctive therapies. Exchangeability, however, remains as one of the most vexing issues in causal inference when using non-experimental datasets.(132) Lack of exchangeability can introduce confounding, since baseline characteristics that differ between treatment groups might also influence the outcome, making effect estimates inaccurate.

2.4 Adjusting for Bias in Non-Experimental Data Using Regression Modeling

Typically, when estimating the relative treatment effect, the occurrence/level of the outcome of interest would be modelled as a function of treatment received and potential-confounding variables. Thus, a multivariable regression model would generally be fitted to estimate the regression coefficients for treatment received and potentially confounding variables through algorithms such as maximum likelihood estimation, among others.(133) The model fit could then be assessed through examination of residual plots and goodness-of-fit tests.(134) If the model fit was poor, modifications in the functional form of relation being modelled could be made via adding higher-order terms for some characteristics, and/or interaction terms combining some independent variables.

However, the occurrence or level of the outcome of interest could vary as a function of the confounding variables 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. Inability to accurately quantify outcome variation attributable to confounding through appropriate modeling leads to residual confounding and biased effect-estimates.(135) Furthermore, since the functional form of the association of the study outcome with the treatment received and potential confounders can be specified to yield multiple models with reasonable fit, the model that is most consistent with the scientist's hypothesis can be chosen. This is known as model dependence and increases the likelihood of falsely rejecting the null hypothesis, increasing the frequency of a type I error.(136)

Issues of residual confounding and model dependence are further exacerbated with decreasing balance and/or overlap in the distribution of baseline covariates between treatment groups, as accurate estimation of regression coefficients becomes more reliant on model specification.(137) To explain, increasing imbalance in the multivariable distribution of baseline covariates increases the potential for confounding of effect estimates as treatment groups are not exchangeable. This can only be remedied through accurate model specification, which is dependent on overlap in the distribution, as the full range of values for each baseline covariate and combination of values for multiple baseline covariates must be observed in each treatment group (i.e., positivity) to accurately estimate regression coefficients. Otherwise, estimates for regression coefficients rely on interpolation and extrapolation, which can lead to bias in effect estimates.(130)

2.5 Preprocessing Non-Experimental Data to Improve Exchangeability and Positivity

To overcome these issues, data preprocessing techniques can increase the overlap and balance in the distribution of baseline covariates between treatment groups.(138,139)

2.5.1 Preprocessing Non-Experimental Data Using Propensity Score Matching

Propensity score matching (PSM) is an example of a data preprocessing technique that has become increasingly popular in recent years.(140) The propensity score (PS) can be defined as the probability of receiving the index- treatment given a subject's baseline covariates.(141) Issues of missing data may prevent an accurate estimation of the PS. Matching subjects between treatment groups on the PS has the potential to balance observed baseline covariates between treatment groups and can thereby reduce, or even eliminate, confounding by those covariates.(141) This reduces reliance on model specification to control for confounding and thus reduces the potential model dependence.(138,139) However, unmatched subjects may systematically differ from those who remain in the matched sample. This has implications that limit the representativeness of the study population and the generalizability of the overall study findings. Further, confounding control through PSM is limited to observed covariates and may, in some circumstances, exacerbate hidden bias through furthering imbalance in unobserved confounders. Methods exist to overcome issues of unobserved confounding, including instrumental variable approaches; however, are only effective when a valid instrumental variable exists for the causal contrast being considered.(134)

Many studies demonstrate strong comparability in effect estimates produced from regression modeling with and without preprocessing by PSM.(142,143) However, PSM procedures are generally not performed optimally.(140) That is, systematic identification of PSM strategies that optimize balance in baseline covariates between treatment groups while retaining a sufficient sample size is not commonly done.(143,144) In one review, data were obtained from large RCTs where violations of positivity and exchangeability are not of concern,(143) thus limiting the ability of PSM to further reduce bias in effect estimates. Moreover, many authors do not control for residual confounding through multivariable regression modelling after PSM, leading to biased effect estimates.(140,142)

Although PSM has many benefits when applied before regression modeling in the realm of comparative effectiveness research, King and Nielsen have recently identified a major issue in PSM that threatens the validity of treatment effect estimates.(136) Specifically, the “PSM paradox” is that as the strictness of the match (i.e., smaller allowed distance in the PS between matched individuals) increases, imbalance in baseline variables decreases until a certain point where imbalance begins to increase. This phenomenon occurs as matching on the PS does not use all the information from baseline covariates provided but rather takes an aggregate score, so it becomes incapable of discriminating between individuals who differ in individual or specific combinations of baseline covariates that are not captured in their PS. Once the PS extends beyond its means, increasing the strictness of the match increases imbalance through random elimination of individuals from the final matched cohort. Although concerning when considering the increasing number of studies using PSM,(140) Ripollone et al rigorously reviewed the pharmacoepidemiology literature and did not find an empirical instance of the PSM paradox, but rather found all studies that used PS matching to have improved balance in baseline covariates between treatment groups.(145)

2.5.2 Preprocessing Non-Experimental Data Using Coarsened Exact Matching

King and Nielsen recommend coarsened exact matching (CEM) as a superior alternative to PSM that uses information from all baseline covariates to decrease imbalance further than with PSM.(136) In this approach, continuous and ordinal characteristics are categorized, while some categories of inherently nominal characteristics get 'collapsed', resulting in fewer categories. In other words, CEM involves 'coarsening' of (at least some of) the potential confounders to facilitate the matching process. After variables are coarsened, multivariable strata containing observations from both treatment groups (i.e., areas of positivity) are retained, while the remaining strata are discarded. Index- and reference-treatments are defined, and a binary variable is adopted to represent the treatment group. Weights are then applied to each observation within each stratum to estimate the average treatment effect in the index-treatment group. This is accomplished with a weight of one for each observation remaining in the index-treatment group after matching. Weights for observations in the reference-treatment group are calculated as the proportion of total observations from the matched index-treatment group in a particular stratum (i.e., n_{i_s}/n_i) divided by the proportion of total observations from the matched reference-treatment group in that same stratum (i.e., n_{r_s}/n_r).(146)

A disadvantage to using CEM compared to PSM is that as the number of parameters increases, fewer matches become available unless the coarsening of parameters increases, which increases imbalance. Although PSM is not immune to this issue,(136) Elze and colleagues have demonstrated ability to balance baseline covariates between groups with as many as 17 baseline covariates, while still retaining a substantial portion of the original population.(143) However, this was shown in an RCT, so it may have little applicability to non-experimental datasets where less overlap in the distribution of baseline characteristics is expected. Since much of the research on PSM and CEM has surfaced only recently, little evidence exists to support the use of either approach over the other in specific situations.

Fullerton et al. examined the performance of different matching strategies in preprocessing non-experimental data, including PSM and CEM.(147) They found that although CEM improved balance in baseline covariates according to several measures of imbalance, it resulted in smaller matched subsamples that were not generalizable to the original cohort. In contrast, PSM led to improvements in only two of the three measures of imbalance but maintained the characteristics of the original cohort to a greater degree. Ripollone et al. also found that CEM was superior to PSM with regard to providing balanced covariate datasets according to the Mahalanobis distance measure.(148) However, CEM produced the least precise estimates due to lower levels of data retention after matching. Both comparisons used high-dimensional datasets with the smallest dataset involving 19 continuous and binary covariates and the largest having >100 covariates. Upon simulation analyses, Ripollone et al found effect estimates obtained through CEM maintained comparable precision to PSM in lower dimensional datasets involving 8 covariates, while providing more balance in baseline covariates between comparison groups.

Major limitations in these studies that hinder the validity and applicability of their results should be noted. First, neither study used a systematic approach to identifying and evaluating PS models that optimize data retention and balance. They also did not evaluate different matching ratios or caliper widths when using PSM, which has a potential to further improve balance between groups and precision in the effect estimate.(144) Third, both comparisons used high-dimensional empirical datasets with the smallest dataset involving 19 continuous and binary covariates and the largest having >100 covariates. Fourth, they often relied upon quantile-based rules for coarsening continuous variables, such as Sturges' rule.(149) In contrast, in many areas of comparative effectiveness research, the number of baseline covariates that consistently influence treatment decisions and are associated with important outcomes is relatively small;(144,150) furthermore, there often is *a priori* information on the prognostic value of continuous and ordinal variables that can allow one to create more prognostically meaningful strata rather than strata formed from quantile-based rules, which may not align. This can lead to retention of observations from the index-treatment and reference-treatment groups in the same strata that have distinguishable clinical prognoses. Moreover, index-treatment and/or reference-treatment observations with similar prognoses that do not fall into quantile-based strata might be lost. Overall, retention of observations with distinguishable clinical prognoses and loss of observations with similar prognoses reduces potential balance and levels of data retention, respectively, that could be achieved if ranges of baseline covariate values were informed by prior evidence on prognosis. Finally, Fullerton et al and Ripollone et al did not control for residual confounding through multivariable regression modelling after matching, which is recommended to control for remaining bias after matching.(140,142)

2.6 Transition to Chapter Three

This chapter provided a review of the value of evidence produced from non-experimental data in informing treatment decision making. Importance was placed on the concern for bias through confounding in evidence produced from non-experimental data. The counterfactual theory was reviewed to provide a theoretical basis from which to discuss the ability of popular statistical techniques in preventing (i.e., PSM and CEM) and adjusting for (i.e., regression modeling) confounding when estimating treatment effects and their associated shortcomings. Although previous studies have compared PSM and CEM in the prevention of confounding when estimating treatment effects using non-experimental data, several shortcomings in these comparisons were noted. In particular, there was a lack of a systematic approach to developing and evaluating PSM and CEM strategies that optimize data retention and balance.

Chapter three provides a review of diagnostics used to assess balance in the multivariable distribution of baseline covariates when comparing treatment outcomes using non-experimental data. A sufficient set that captures important differences in the distribution of baseline covariates with potential to introduce confounding, according to the counterfactual theory, is recommended to evaluate the success of matching in preventing confounding. This groundwork was necessary to inform the systematic approach to developing and evaluating

PSM and CEM strategies that optimizes data retention and balance between treatment groups obtained from non-experimental data covered in chapter four.

Chapter 3: Assessing the Potential for Confounding through Balance Diagnostics

To compare the performance of matching strategies in the ability to balance treatment groups, the appropriate metric should be used. Many methods have been developed to assess comparability between baseline covariates between exposure groups after matching. These methods are generally termed “balance diagnostics”.(151) All balance diagnostics have a similar goal: to quantify the difference in the multivariate distribution of baseline covariates between treatment groups in order to measure the degree of exchangeability. A common approach to quantifying balance is to compare means and/or medians of continuous variables and the distribution of categorical variables in index- and reference-treatment groups.(152) This approach is in line with the CONSORT statement, which requires that authors provide a summary table of the baseline characteristics in different treatment arms.(153) Comparison of continuous variables using a t-test and dichotomous variables using a chi-square test have been proposed and are most commonly utilized in non-experimental comparative effectiveness research.(153) However, these are not appropriate when performing adjustment techniques that reduce the sample size (i.e., matching). Since these tests are dependent on sample size, as the number of subjects who are not eligible for matching increases, the sample size in each group decreases, increasing the likelihood of a non-statistically-significant difference in baseline covariates, irrespective of whether balance actually improves (Figure 3.1).(154)

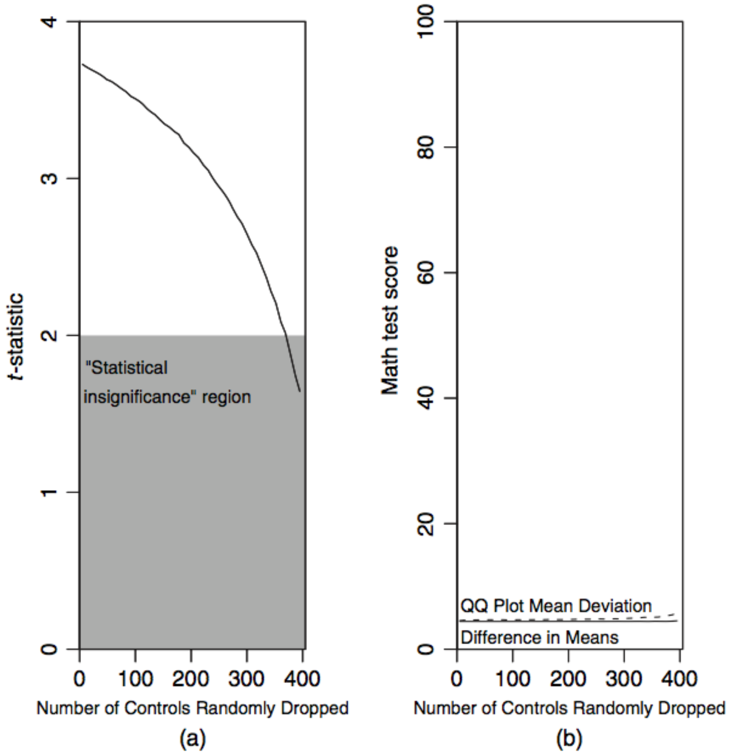


Figure 3.1 (a) the t-statistic comparing difference in means between treatment and control groups decreases as control units are randomly dropped; (b) indicates a constant difference in means and quantile–quantile plot mean deviation between treatment and control groups as control units are randomly dropped.(154)

3.1 Balance Diagnostics for the Central Tendency of Single Variables

Another approach to comparing the difference in means and proportions between treatment groups involves calculating the standardized mean difference (SMD).(152) This has become the standard approach, as the SMD is easy to compute and understand,(155) while being independent of sample size. Moreover, it allows for the comparison between covariates with different units since the SMD is unitless. For continuous variables, this can be calculated as the quotient of the difference in means between groups in the numerator and the root of the average of the sample variance in each group in the denominator. For categorical variables, this can be calculated as the quotient of the difference in proportions between groups in the numerator and the average of the variance in sample proportions in each group in the denominator.

Peter Austin shows that when the SMD <0.1, the amount of non-overlap in the baseline covariate between groups is <7.7%, indicating a high-degree of comparability.(152) However, this is based on several assumptions including that the covariate under investigation is normally distributed with equal variance between the two treatment groups being compared and that groups have similar sample sizes. Although such a difference might seem negligible, it is important to consider the prognostic association of the baseline variable. For example, if the association is non-linear then confounding can still manifest when SMD <0.1 since a greater proportion of one group might occupy greater levels of the confounding variable with greater prognostic value, while the other group clusters around the mean. Stuart et al. performed multiple data simulations with varying confounder relationships between a binary treatment and continuous outcome.(156) They found that both the correlation between the mean SMD of all confounders or proportion of confounders with SMD <0.1 and the true bias was stronger among simulations limited to continuous confounding variables with linear main effects compared with simulations involving both continuous and categorical confounding variables with non-linear effects. This indicates that small values of SMD do not necessarily indicate comparability, especially among covariate structures involving categorical and non-linear continuous variables, which are common in the medical literature.

3.2 Balance Diagnostics for the Variance of Single Variables

Imai et al. suggest that higher order moments of the baseline covariate distributions between treatment groups be compared in order to address this potential issue.(154) After the mean, which indicates the central tendency of the distribution, the variance, which indicates the amount of deviation, is the second moment in the covariate distribution. As such, estimating the ratio of variances in baseline covariates between groups in addition to SMD has been recommended.(152) Peter Austin demonstrated that under the null hypothesis, the 95% confidence interval for equality in variances between two independent groups amounts to a lower bound of approximately 0.92 and an upper bound of 1.08.(152) This was derived from a data simulation study of 2430 matched pairs and an *F*-distribution with 2429 and 2429 degrees of freedom so has limited applicability in different settings. Imai et al., have suggested quantile-quantile plots for comparison of the distribution of continuous covariates for assessing balance between two groups.(154) As such, comparing both the variance ratios and quantile-quantile plots can assist in identifying any notable differences between exposure groups.

3.3 Limitations of Balance Diagnostics for Single Confounding Variables

Thus far, we have discussed imbalance in single covariates, which has only the potential to assess main effects of confounding. However, most comparative effectiveness research includes multiple baseline covariates that have potential for confounding. As the number of baseline covariates increases, the amount of potential bias increases even when SMD <0.1 and the variance ratio is between 0.92 and 1.08. Moreover, other aspects of multivariable distributions between comparison groups could differ on the basis of covariance between multiple confounding variables, which might indicate different levels of effect modification at the level of confounding between comparison groups. Small differences in SMD and variance ratios in many baseline covariates coupled with differences in covariance have potential to bias treatment effect estimates in the presence of “balanced” baseline covariate distributions as commonly defined by the “acceptable” ranges of SMD and variance ratios.(155,156) For instance, Stuart et al found that the correlation between true bias and measures of imbalance based on SMD, as described above, became weaker when their data simulations involved an interaction effect between confounding variables compared to main effects alone. This issue can be overcome through calculating the SMD for appropriate interaction terms;(152) however, such interaction is not always easily identifiable in real datasets. Moreover, interactions between two variables can vary according to a third,(157) which makes identification of appropriate interaction terms for assessing balance more complicated.

3.4 Prognostic Scores

Stuart et al propose checking balance by examining SMD in a prognostic score.(156) This is accomplished by first identifying prognostic factors through a thorough examination of the literature and expert consultation. The outcome of interest is then modelled as a function of the prognostic variables using the referent category under comparison in order to obtain the estimated baseline prognosis in the absence of the index-treatment. Using this model, the predicted prognosis is estimated for each individual in the index- and reference-treatment groups to derive their “prognostic scores”. The SMD in prognostic scores between groups is then calculated to obtain estimates of balance. Compared to a number of other balance diagnostics, including mean SMD and proportion of SMD <0.1, the SMD in prognostic score consistently obtained stronger correlations with true bias in most situations (>0.90). However, this was dependent on whether the true prognostic score was appropriately specified. In situations where the prognostic score was not appropriately specified, the correlation between the SMD in prognostic score with the true bias were as low as 0.31. Since the true prognostic score might involve many regression coefficients, estimation in smaller datasets may be unreliable, leading to improperly estimated prognostic scores and thus unreliable measures of imbalance. One way to overcome this is to use validated prognostic scores. For example, in PCa research, as mentioned in chapter one, risk-groups, reflecting different prognoses post-treatment, have been identified as one of the most reliable predictors for biochemical progression, which is an important oncological outcome.(158) The SMD in the proportion of observations occupying different risk-groups are used in chapter four as the SMD in prognostic scores.

3.5 Global Imbalance Measure

One way to overcome the issues of assessing balance in multiple covariates and their relationships between one another as well as model dependence would be to calculate the distance between multivariate distributions between comparison groups. This can be accomplished using the global imbalance measure (L_1) developed by Iacus et al.(159) In this approach, each variable whether continuous or categorical are stratified using bounds that define acceptable levels of variation based on previous research.

The stratified covariates can be cross tabulated to create multidimensional histograms. The absolute difference in multidimensional histograms divided by two provides an estimate of global imbalance between groups (L_1), as demonstrated below:

$$\mathcal{L}_1(f, g; H) = \frac{1}{2} \sum_{l_1 \dots l_k \in H(\boldsymbol{X})} |f_{l_1 \dots l_k} - g_{l_1 \dots l_k}|$$

Wherein *f* and *g* denote the relative empirical multivariable frequency distributions for the index- and referent-units, respectively, and $f_{l_1 \dots l_k}$ and $g_{l_1 \dots l_k}$denote the relative frequency for observations belonging to the cell with coordinates l_1 … l_k of the multivariable cross-tabulation.(159) This can be interpreted in that if two empirical distributions are completely separated, then L_1=1 and there is 0% overlap between the two groups with regard to the multivariate distribution in baseline covariates. If the two distributions are the same, L_1=0 and there is 100% overlap in multivariable distribution between the two groups.

In theory, the global imbalance measure appears superior to other balance diagnostics in its ability to adequately capture all bias due to differences between groups in single covariates and any interaction among two or more covariates. However, data simulations have demonstrated that decreasing bias associated with improvements in covariate balance are not strongly correlated with reduction in the global imbalance measure.(160) These findings are consistent when trying to balance covariates using cohort data in that improvements in balance in prognostic covariates are not strongly correlated with reduction in the global imbalance measure.(147) This might occur as it does not weigh covariates and their interactions according to their prognostic value but rather measures all imbalances in the multivariable distribution equally important. This is further emphasized by Belitser et al. who demonstrated that when weights were applied to balance diagnostics to represent their prognostic value, the negative correlation between balance and bias became stronger.(161)

The absence of a strong correlation between the global imbalance measure and bias limits the utility of it as an imbalance measure. For instance, when optimizing matching algorithms to improve baseline covariate balance between exposure groups and thus reduce bias, changes in balance diagnostics sensitive to improvements in covariate balance are helpful to identify the matching algorithm capable of reducing the most imbalance given a level of data retention. As such, the global imbalance measure has limited utility in assessing bias due to covariate imbalance in the context of causal inference using nonrandomized data.

3.6 Other balance diagnostics

Although other measures of imbalance exist such as the Kolmogorov-Smirnov distance, the Lévy distance, and the overlapping coefficient, previous studies have demonstrated weaker correlation with bias in addition to theoretical limitations that make them less informative and suitable.(152,160,161) As such, these imbalance measures are not considered.

3.7 Recommended use of balance diagnostics for development and evaluation of the PS model in PSM

Herein, we recommend a set of balance diagnostics to measure all imbalances in the multivariable distribution with prognostic value when comparing treatment effectiveness from non-experimental data. Since the average absolute SMD of all baseline covariates reflects balance in typical values that are deemed most important, it should be used to evaluate general improvements in balance as the width of the PSM caliper decreases. However, since the average absolute SMD might decrease, while the absolute SMD of individual covariates might increase, we also propose that the number of individual covariates with |SMD| <0.1 be measured to ensure no gross imbalance in any single variable. Likewise, to capture general improvements in balance for higher-order moments, the average of variance ratios for continuous variables can be monitored in conjunction with the number of variables falling below 1.08 (if ratio is <1.00 then take the reciprocal) to ensure no gross violations in the second moment of individual covariate distributions.(152) Finally, a validated prognostic score should be used to capture important combinations of covariate values that might be missed. If no validated prognostic scores are available, one can estimate a prognostic score from their dataset using methods established by Stuart et al,(156) provided there are a sufficient number of events and non-events available.

3.8 Transition to Chapter Four

This chapter provided a review of balance diagnostics and their relative informativeness with regard to assessing the potential for confounding when comparing treatment outcomes using non-experimental data. A recommended set of diagnostics was identified that comprehensively evaluates differences in the multivariable distribution of baseline covariates with confounding potential. The following chapter demonstrates how to use the recommended set of balance diagnostics to guide the development of a propensity score model for matching that optimizes efficiency with respect to balance achieved in the multivariable distribution as well as retention of observations from the original dataset.

Chapter 4: A Systematic Approach to Developing and Evaluating Propensity Score Models for Matching

Despite the existence of multiple guidelines on how to perform PSM in comparative effectiveness research,(162–164) information surrounding the use of balance diagnostics in guiding the development and evaluation of PSM strategies remains elusive and subject to criticism. For example, most guidelines for clinical researchers emphasize the use of a standardized caliper (e.g., 0.2 of the standard deviation of the logit of the PS) for matching strategies and recommend iteratively exploring balance achieved after matching as a diagnostic tool for adequacy of the PS model.(162–166) However, the performance of different PS models used in matching depends on the balance diagnostic used for assessment (e.g. standardized mean difference (SMD), variance ratio, etc.) and caliper size. For instance, one PS model might lead to better balance than some alternative model given certain combinations of caliper sizes and balance diagnostics but to worse balance when compared to the same alternative given different balance diagnostics and caliper sizes.

Moreover, the balance diagnostics recommended by such guidelines often do not capture all differences in the multivariable distribution of baseline covariates between treatment groups. Often, the SMD is recommended to evaluate the differences in the central moment and, sometimes, the variance ratio to evaluate higher-order moment of imbalance in the distribution of individual covariates. As mentioned in chapter three, these balance diagnostics neglect important interactions that hold prognostic value about the outcome of interest and thus have potential to confound associations even in the presence of balanced individual covariates.

In this chapter, we offer a systematic approach to developing and evaluating PS models used in matching that maximize balance in all key aspects of the multivariable distribution to mitigate potential for confounding, while also maximizing retention of observations from the original dataset. A working example is provided from the PCa literature.

4.1 Working Example: Comparing the Rates of Biochemical Failure between Different Radiotherapy Approaches for Prostate Cancer

4.1.1 Background and Data Source

Different approaches to RT are available for the treatment of PCa.(2) BT involves the insertion of a radioactive isotope into the prostate gland.(3) Compared with the EBRT, BT is capable of delivering greater doses of radiation to the prostate gland while sparing adjacent structures such as the rectum and bladder.(4) As such, it is generally reserved for the monotherapy of tumors that maintain a lower-risk of extraprostatic extension unless it is combined with EBRT for higher risk situations.(4) Risk of extraprostatic extension is estimated by consideration of PSA, GS, and cT stage.(5) These characteristics also hold considerable prognostic value regarding important oncological outcomes such as BPFS, MPFS, CSS and OS.(6) Thus, when comparing BT with EBRT, it is important to adjust for all of these factors, among other baseline patient characteristics.

For this demonstration, we abstracted data from the Prostate Cancer Risk Stratification (ProCaRS) database. This database contains data on 7974 patients diagnosed with PCa and treated with different forms of primary RT between 1994 and 2010 from four Canadian institutions in Toronto, Quebec City, Montreal and Vancouver.(144) Details regarding ethics approval, database construction and quality assurance have been previously described.(20) The example comparison for our approach was informed by a RCT performed by Morris and colleagues, which compared the rate of biochemical failure among men diagnosed with intermediate-risk PCa according to the National Comprehensive Network and treated with either BT and hormone therapy (BT) or EBRT and hormone therapy (EBRT).(167) In this trial, the authors found an increased incidence of biochemical failure in the EBRT relative to the BT group (hazard ratio [95% confidence interval]: 2.04 [1.25, 3.33]).

4.1.2 Descriptive Measures

We begin with an initial examination of characteristics in the unmatched samples from the ProCaRS dataset. Table 4.1 shows that the BT group was, on average, treated at earlier dates than the EBRT group (median treatment year of 2002 vs 2003, respectively), and had less advanced tumor characteristics, as expected.

Table 4.1 Descriptive statistics and balance diagnostics for unmatched BT and EBRT samples

	BT (n=433)		EBRT (n=132)		|SMD|	Variance Ratio
RT Start Year						
Median	2002		2003		0.3029	1.49
IQR	2001, 2004		2002, 2004			
PSA (ng/ml)						
Median	7.60		9.02		0.3468	1.60
IQR	5.70, 10.50		5.88, 12.60			
Clinical T-Stage						
T1a-2a	373	86.14%	122	92.42%	0.2041	
T2b-c	60	13.86%	10	7.58%		
Gleason Grade						
1	126	29.10%	23	17.42%	0.2790	
2	249	57.51%	64	48.48%	0.1815	
3	58	13.39%	45	34.09%	0.5014	
PROCARS Risk-Group						

Low-intermediate	404	93.30%	99	75%	0.5177	
High-intermediate	29	6.70%	33	25%		
				Average	0.2936	1.29

Table 4.1 also shows that the two groups are imbalanced with regard to the SMD for individual covariates, and risk-group as well as the average absolute SMD for all covariates. The average variance ratio and individual variance ratios were also greater than the accepted threshold of 1.08.

4.2 Model Development

In developing PS models, we propose beginning with more general models (i.e., using simple linear terms for continuous covariates, original categories for categorical variables and without interactions) and increasing the complexity to better fit the relation between treatment predictors and treatment received by adding higher-order and interaction terms. Here, we use a logistic model to specify our PS. Since power is driven by the smaller comparison group, we would like to optimize retention of the smaller group. In this case, EBRT is the smaller group, so treatment status is coded as EBRT = 1 and BT = 0.

4.2.1 Model One

The first model involved simple linear terms for continuous characteristics and 'dummy' variables to represent categorical characteristics. The reader is referred to an article by Brookhart et al for an in-depth discussion on covariate selection in PS models.(168) In brief, it is recommended that any variable notably associated with the outcome be included to enhance precision and accuracy in effect estimates. The term '*PSA*' in model one represents the regression coefficient for baseline PSA (ng/ml), '*GS*' represents the regression coefficient for Gleason score, '*TS*' represents the regression coefficient for clinical T-stage, '*TxYr*' represents the regression coefficient for year that RT was initiated. We specify a logistic regression model with treatment as the dependent variable using the '*glm'* (or generalized linear model) command with '*family=binomial*' option in RStudio.(169)

```
> model1 <- glm(Tx ~ PSA + GS + TS + TxYr, data = PROCARS, family = "binomial")
```

The general form for the linear logistic model is given below:

$$log\left[\frac{P(x)}{1-P(x)}\right] = \beta_0 + \beta_{PSA}\mathrm{x}_{PSA} + \beta_{GS}\mathrm{x}_{GS} + \beta_{TS}\mathrm{x}_{TS} + \beta_{TxYr}\mathrm{x}_{TxYr}$$

4.2.2 Model two: Identifying departures from linearity and improving functional form in the relation between continuous predictors and treatment received

The second model attempts to improve the accuracy of the PS in predicting the treatment received through improving model fit of the functional form of the relation between continuous covariates and treatment status. To identify departures from linearity in the relationship between continuous predictors and the logit of the probability for receiving the treatment of interest, locally weighted scatterplot smoothers can be used. Evaluation of whether proposed transformations using higher-order terms improve model fit can be verified using the likelihood-ratio test for nested models and the pseudo-R^2 for both nested and non-nested models.(134) In our example, we found that a restricted cubic spline for baseline PSA ('*rPSA*') and categorization of treatment start date into two-year increments (*'TxYr2'*) improved model fit,(134) which was confirmed using the likelihood ratio test (p=0.00052) and pseudo-R^2 (model one = 0.13 vs model two = 0.17). As such, we added these terms to our second model:

```
> model2 <- glm(Tx ~ PSA + rPSA + GS + TS + TxYr2, data = PROCARS, family = "binomial")
```

4.2.3 Model Three: Identifying interaction terms

Finally, the third model attempts to improve accuracy of the PS in predicting the treatment received through identifying interaction terms between independent variables that improve model fit. Again, improvements in model fit can be verified using the likelihood-ratio test and pseudo-R^2 value. In the example provided, all two-way interactions among independent variables were assessed and a notable improvement in model fit relative to model two (likelihood ratio test p=0.0057 and pseudo-R^2 = 0.18) was found with the addition of an interaction term between baseline PSA and Gleason score, represented by the '*PSA*GS*' term, which was added to our third model:

```
> model3 <- glm(Tx ~ PSA + rPSA + PSA*GS + GS + TS + TxYr2, data = PROCARS, family = "binomial")
```

4.3 Propensity Score Matching Characteristics

Now that multiple candidate PS models have been specified, decisions regarding matching algorithm, matching ratio, and caliper size must be made. The most common combination involves the use of nearest-neighbor matching with select caliper width and without replacement.(170) That is, a random exposed subject is matched to an unexposed subject with the most similar propensity score (i.e. nearest-neighbor) who has a predicted PS within a pre-specified range of the exposed subject's PS (i.e. caliper). After the unexposed subject is matched, they are removed as a candidate for further matching (i.e., without replacement). This is often done in a one-to-one fashion; however, depending on the proportion of exposed to unexposed subjects, many to one or one to many might be ideal. In our example, there are approximately 3.3 BT observations for every one EBRT observation. As such, a matching ratio of 3:1 BT to EBRT observations might be favorable to maximize retention of BT observations. Other methods of matching involve nearest-neighbor matching without calipers, optimal matching wherein exposed and unexposed subjects are matched so as to minimize the total within-matched PS difference, and full-matching. The reader is referred to another guideline that reviews these matching algorithms thoroughly.(163)

Caliper size impacts variance and bias in the effect estimate as well as how baseline variables are balanced in the final model. A smaller caliper size will reduce the number of observations matched, thus eliminating more subjects from the final matched sample, while larger calipers will have the opposite effect. This ultimately impacts precision of the effect estimate with larger and smaller calipers generally leading to more and less precise effect estimates, respectively. Bias increases with caliper size as larger calipers enable more dissimilar subjects to be matched. Thus, when selecting caliper widths, a balance must be achieved between precision and validity of the effect estimate.

Popular calipers include fixed widths of the PS as well as a function of the logit of the propensity score (Logit(PS)), as the logit is more likely to approximate a normal distribution than the probability metric. Cochrane and Rubin have demonstrated that matching on a normally distributed confounding variable with caliper widths of 0.6 and 0.2 standard deviations (SD) of the Logit(PS) can remove 90% to 99% of confounding for that particular variable, respectively.(171) However, the caliper width that optimizes the balance between bias and precision will depend on the characteristics of the dataset (i.e., the multivariable distributions of baseline covariates in the treatment groups). As such, we propose that analysts start with wider calipers and explore progressively more narrow calipers to identify a 'plateau' in the association between improvements in balance and retention of the original dataset. The plateau is operationally defined as the point where further decreasing the PSM caliper leads to negligible improvements in balance, while leading to further decreases in data retention. The concept of a plateau will be demonstrated in the section 4.4.

We used the 'MatchIt' package in R to perform matching strategies with nearest-neighbor matching (R input option: *method="nearest"*) as the algorithm,(146) a matching ratio of 3:1 for BT:EBRT (R input option: *ratio = 3*), matching BT to EBRT observations, without replacement (R input option: *replace = FALSE*), and with progressively more narrow calipers (R input option: *caliper = 0.1*). An example is given below:

```
psmatch<-matchit(Tx ~ BasePSA + cGS + cTS + TxYr, data=PROCARS, method="nearest", ratio = 3, caliper = 0.1,  replace = FALSE)
```

4.4 Evaluating Model Performance

To evaluate the performance among candidate PS models used in matching, we propose that the efficiency of the candidate models be compared. Efficiency, as defined here, is the amount of balance improvement for a given level of the original population retained. Measures of SMD, and variance ratios can be obtained through the cobalt package in R.(172) Afterward, measures can be examined graphically using plots with the balance achieved through PSM for each candidate PS model for a given percentage of the original population retained. In other words, the PS model that leads to better balance among all diagnostics with the same or greater percentage of the original population retained is deemed the most efficient PS model and will be the selected model based on this performance.

4.4.1 Identifying the most efficient caliper width to compare efficiency between PS models

A standardized rule for identifying a specific caliper width is required to compare the efficiency of each candidate model. We propose that the point at which further narrowing PSM calipers for each PS model leads to a 'plateau' be used for this purpose. That is, when further narrowing the matching caliper leads to negligible improvements in or worsening balance at the expense of a decrease in the percentage of the original population retained. In our example, we explored balance achieved after matching without a caliper wherein the nearest-neighbor referent-unit of a randomly selected index-unit is matched and this process is repeated until each index-unit is matched with the specified number of referent-units. After, we matched with caliper limits of 2.0, 1.5, 1.0, 0.8, 0.6, 0.4, 0.3, 0.2, 0.1, 0.05, 0.025, 0.02, 0.015 and 0.01 SD(logit(PS) to capture the plateau. Using model one from our example and looking from right to left, we see that the average SMD plateaus at approximately 70% retention (Figure 4.1). This corresponds to a plateau in the average variance ratio (Figure 4.2) and risk-group SMD (Figure 4.3). We can also see at approximately 40-50% data retention, what King and Nielsen have termed, the PSM paradox where decreasing caliper size paradoxically leads to increasing imbalance.(136) The caliper width of 0.1 standard deviations of the logit of the PS, which leads to approximately 67% retention of the original cohort and plateau in all balance diagnostics appears to be the most efficient PSM strategy using model one. Likewise, a caliper width of 0.1 standard deviations of the logit of the PS for models two and three, which lead to approximately 66% and 63% retention of the original cohort, respectively, lead to a plateau in balance reduction for all balance diagnostics.

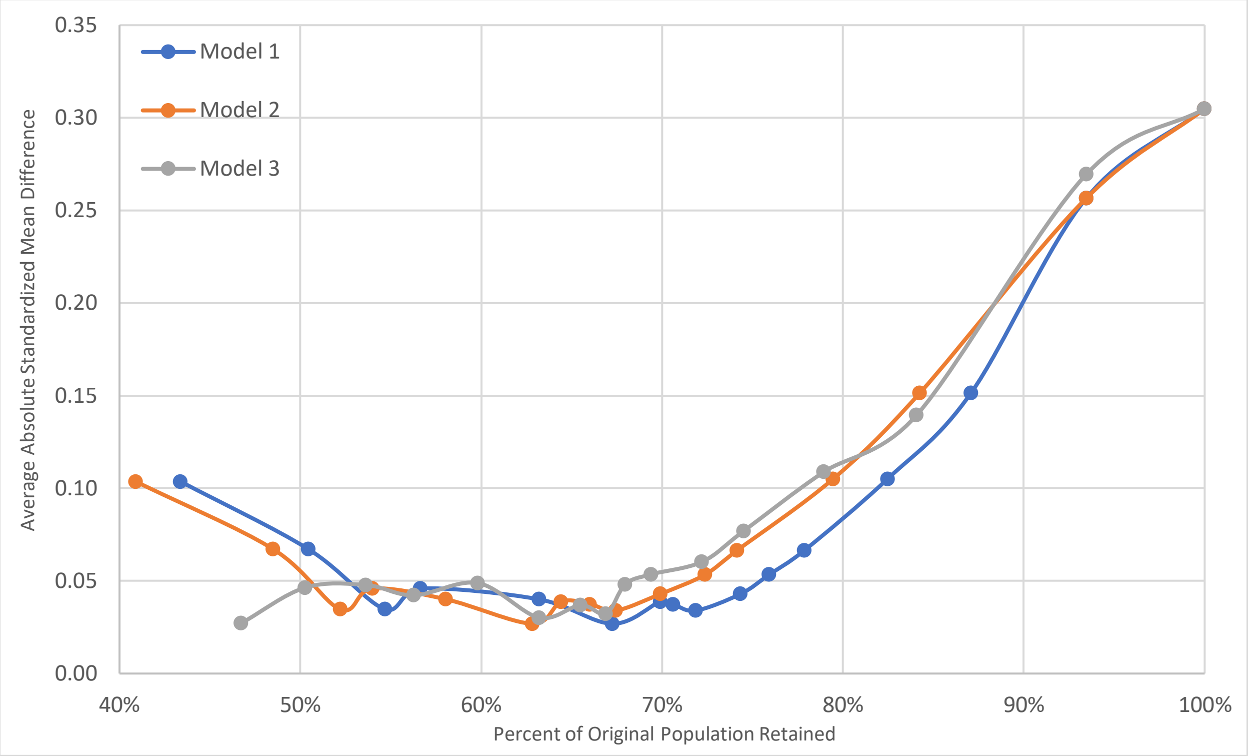


Figure 4.1 Association between percent of original population retained and average absolute standardized mean difference

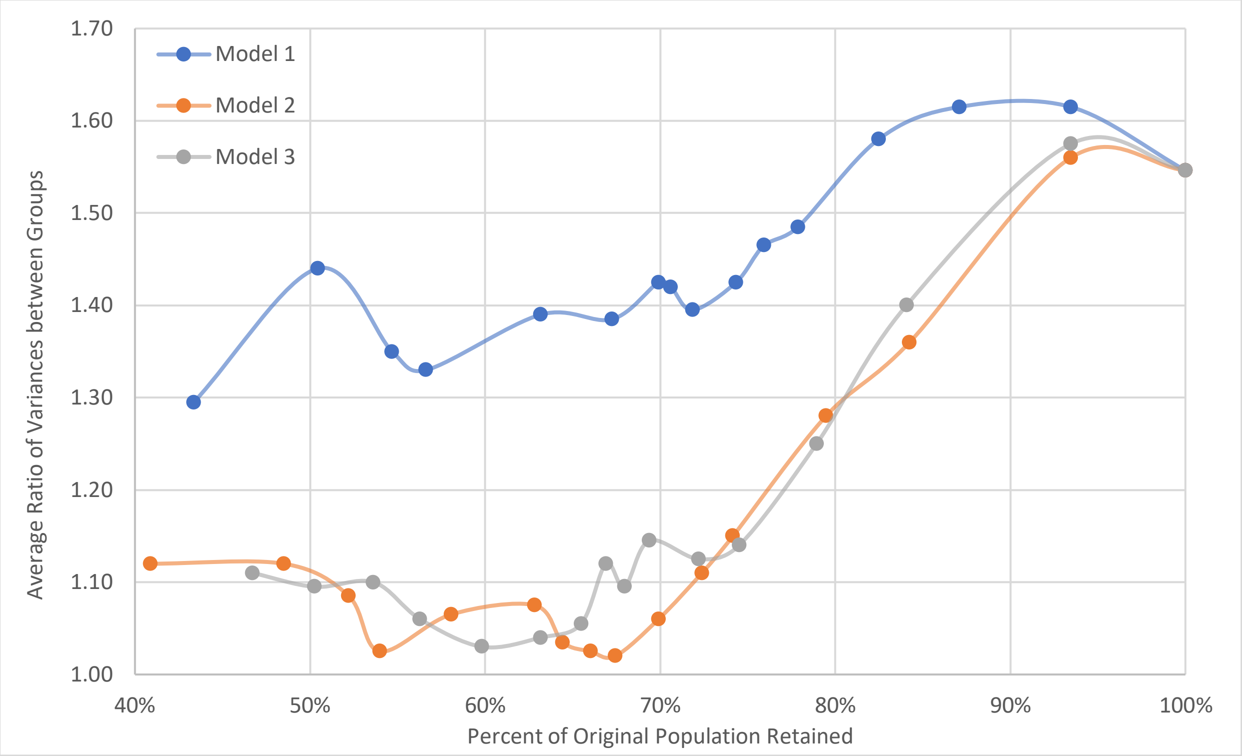


Figure 4.2 Association between percent of original population retained and the average variance ratio for continuous baseline covariates between treatment groups

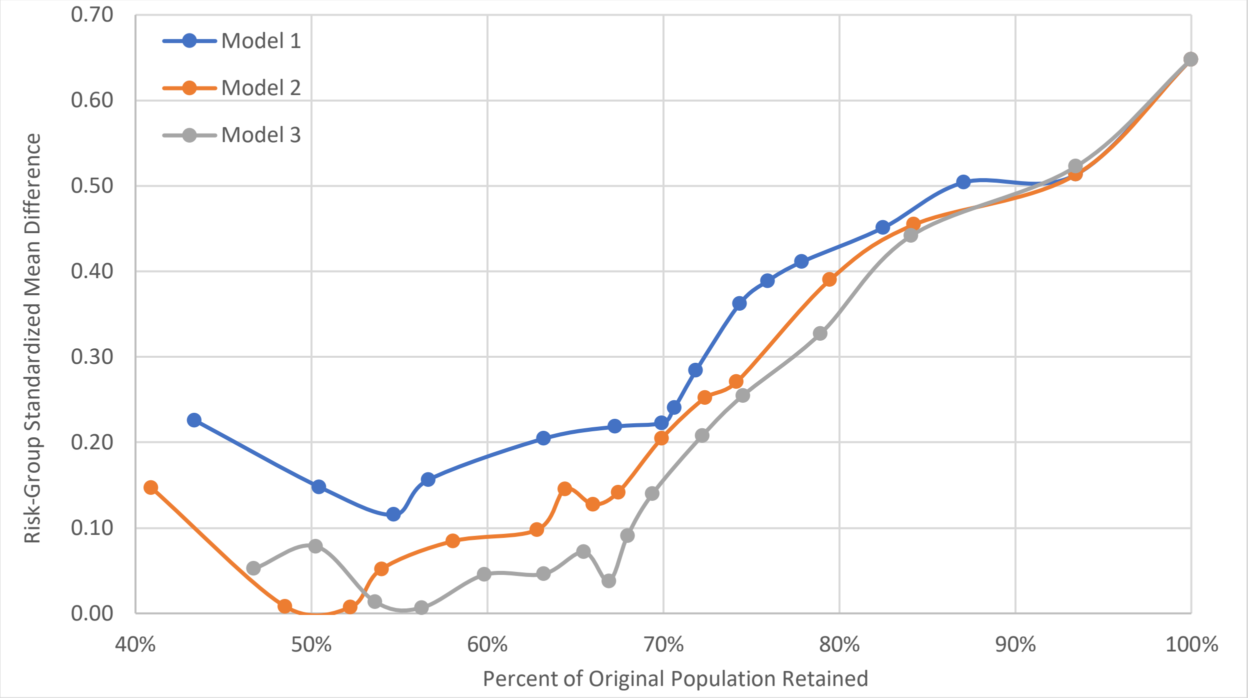


Figure 4.3 Association between percent of original population retained and the risk-group standardized mean difference.

4.4.2 Comparing balance after matching on candidate PS models

Given the similar percentage of retention among all matching strategies at the previously selected caliper width, we can compare the balance achieved in our five balance diagnostics between our candidate models (Table 4.2). The first model leads to an average absolute SMD of 0.027 with no covariate SMD>0.1, average variance ratio of 1.39 with both continuous variable variance ratios >1.08, and a risk-group SMD of 0.22. The second model leads to an average absolute SMD of 0.027 with no covariate SMD>0.1, average variance ratio of 1.08 with one continuous variable variance ratio >1.08, and risk-group SMD of 0.098. Finally, the third model leads to an average absolute SMD of 0.030 with no covariate SMD>0.1, average variance ratio of 1.04 without either continuous variable variance ratio >1.08, and risk-group SMD of 0.046. The second and third model led to better overall balance with the third outperforming the second. As a result, the third model provided the most efficient matching strategy in comparison with models one and two.

Table 4.2 Comparison of the percentage retained in the original cohort and balance in baseline covariates after matching using the most efficient caliper width for each candidate PS model

Model	Retention (%)	Average Absolute SMD	Number of covariates with SMD>0.1	Average Variance Ratio	Number of covariates with a variance ratio >1.08	Risk-group SMD
1	67.3	0.027	0	1.39	2	0.22
2	66.0	0.027	0	1.08	1	0.098
3	63.2	0.030	0	1.04	0	0.046

4.5 Strengths

In this section, we review the strengths of this approach in light of previous guidelines. The standardization in developing and evaluating PS models used for PSM in comparative effectiveness research helps reduce researcher bias. To explain, since multiple combinations of

PS models and caliper widths might lead to “reasonable” balance, and effect estimates randomly vary, the combination of a particular PS model and caliper width that led to an effect estimate more in-line with the scientist’s hypothesis can be chosen. Although this is still possible with our approach, it is considerably more limited since there will be fewer available options. Second, the set of balance diagnostics proposed is comprehensive in evaluating multiple characteristics of the multivariable distribution in baseline covariates that have potential to confound effect estimates based on the counterfactual theory of causal contrasts. Obtaining balance in this set of diagnostics offers more convincing conclusions regarding relative treatment effectiveness from which to base patient- and policy-level decisions. Compared to previous suggestions of iteratively exploring random combinations of PS models, and calipers, our systematic approach to identifying the most efficient caliper width associated with each PS model after matching allows the most efficient approach to identifying the PS model and caliper width that leads to the least imbalance for a given level of data retention.

4.6 Limitations

In this section, we discuss some limitations to the proposed approach that might arise with different datasets and offer potential solutions. One concern is if a plateau occurs in one balance diagnostic at a different level of data retention than another balance diagnostic (e.g., stabilization in the average absolute SMD occurs at approximately 80% data retention but occurs at 70% for the average variance ratio and 60% with the prognostic score based SMD). In this scenario, if the level of balance remains stable with progressively narrower calipers in the balance diagnostics that plateau first, the point at which the other balance diagnostics plateau should be used as the most efficient caliper width for comparison. If the level of balance worsens with progressively narrower calipers in the balance diagnostics that plateau first, multiple calipers consistent with each plateau in each balance diagnostic should be used for comparison. This might lead to two sets of matched samples for comparison generated from two caliper widths if the plateau occurs at a different data retention level for only one balance diagnostic or three sets of matched samples for comparison generated from three caliper widths if the plateau occurs at different data retention levels for all balance diagnostics. A reasonable alternative to this approach might include a caliper that leads to a mid-level of balance in all diagnostics for the same level of data retention. Inevitably, there will always be some exceptions where these rules will not apply, and the analyst will need to formulate their own decision. It is recommended that the rationale for such a decision be transparent to allow the reader sufficient information to evaluate the reasonableness of the decision. In addition, the author should report results from other reasonable matching strategies to demonstrate consistency in effect estimates.

4.7 Summary

In summary, we propose that a set of balance diagnostics that sufficiently capture important differences in the multivariable distribution be used in a systematic approach to developing and evaluating PS models used for PSM. We invite criticism and commentary surrounding the approach developed and presented here to improve the conduct and reporting of PSM for observational data in comparative effectiveness research. Since RCTs are becoming increasingly more difficult to perform due to multiple available standards of care with varying characteristics that preclude randomization for ethical and financial reasons, the accuracy of evidence produced from observational datasets is becoming increasingly important to guide patient- and policy-level decisions. As such, improvements in this field hold considerable value to researchers, clinicians, and patients alike.

4.8 Transition to Chapter Five

This chapter provided a systematic approach to developing and evaluating propensity score models for matching that optimizes balance in the multivariable distribution of baseline covariates and data retention. Improvements in balance per amount of data retained were achieved in models involving higher order terms for continuous covariates and interaction terms between some terms relative to simpler models. This shows that previous guidelines for developing models used in propensity score matching are inadequate and highlights the need to assess model performance using multiple balance diagnostics in a systematic fashion that also evaluates data retention.

The systematic approach to developing matching strategies is used in the following chapter to compare the performance of propensity score matching and coarsened exact matching in ability to balance the multivariable distribution of baseline covariates per level of data retention.

Chapter 5: Comparing the performance of coarsened exact matching and propensity score matching in non-experimental prostate cancer comparative effectiveness research

5.1 Objective

In this chapter, the performance of CEM and PSM in preprocessing data from a non-experimental database containing information on men diagnosed with intermediate-risk PCa and treated with different combinations of RT and ADT was compared.

5.2 Methodology

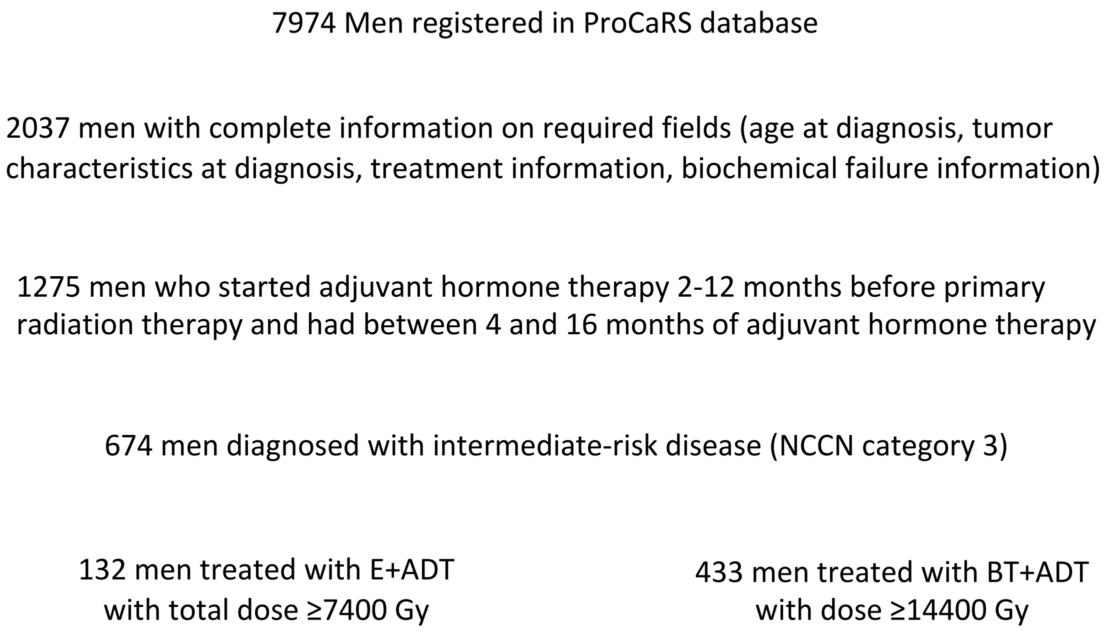
5.2.1 Data source

Data were abstracted from the Prostate Cancer Risk Stratification (ProCaRS) database. This database contains data on 7974 patients diagnosed with prostate cancer and treated with different forms of primary RT between 1994 and 2010 from four Canadian institutions in Toronto, Quebec City, Montreal and Vancouver.(144) Median follow-up was 79 months, and a total of 1442 (19%) patients developed biochemical failure. Details regarding ethics approval, database construction and quality assurance have been previously described.(20)

5.2.2 Comparison one: BT and ADT versus EBRT and ADT

The first comparison was based on a RCT performed by Morris and colleagues, which compared the rate of biochemical failure among men diagnosed with intermediate-risk PCa according to the National Comprehensive Network and treated with EBRT and either low-dose rate BT boost therapy and ADT (BT+ADT) or dose escalated EBRT and ADT (E+ADT).(167)

Patients from the ProCaRS database were included in the comparison if they met the PCa-specific eligibility criteria specified by Morris et al., with two modifications. First, patients who received low-dose rate BT as a monotherapy and without EBRT were included in the BT+ADT. Second, a range for ADT duration (4 to 16 months) was allowed rather than that specified by Morris et al (12 months) to accommodate a greater number of patients for analysis, as few patients underwent approximately 12 months of ADT in both treatment groups. Second, instead of specific dose-escalation protocols as investigated by Morris et al, patients undergoing E+ADT with a dose of ≥74 Gy or BT with a dose of ≥144 Gy were eligible for comparison. The final sizes of the BT+ADT and E+ADT groups were 433 and 132, respectively. The patient selection process is outlined in Figure 5.1a.



7974 Men registered in ProCaRS database

2037 men with complete information on required fields (age at diagnosis, tumor characteristics at diagnosis, treatment information, biochemical failure information)

1275 men who started adjuvant hormone therapy 2-12 months before primary radiation therapy and had between 4 and 16 months of adjuvant hormone therapy

674 men diagnosed with intermediate-risk disease (NCCN category 3)

132 men treated with E+ADT with total dose ≥7400 Gy

433 men treated with BT+ADT with dose ≥14400 Gy

Figure 5.1a Selection process for comparison one.

5.2.3 Comparison two: EBRT with vs without ADT

As shown in chapter four, issues of confounding arose when comparing BPFS between men diagnosed with intermediate-risk non-metastatic PCa and treated with BT or EBRT due to differences in the risk of extraprostatic extention as defined by baseline PSA, cTS, and GS. Confounding is also a concern when comparing the occurrence of oncological outcomes among men diagnosed with PCa and treated with RT alone or in combination with androgen deprivation therapy (ADT). Administration of ADT has demonstrated improvements in oncological outcomes attributed to a radiosensitization effect that improves response to RT while targeting occult micrometastases and extraprostatic extension.(173,174) However, since ADT leads to side effects and increases the risk of non-PCa death,(175) administration is reserved for those with higher PSA, Gleason score and clinical stage who are at an increased risk of biochemical failure, and PCa-specific death.

The second treatment comparison was based on a RCT performed by Jones and colleagues, which sought to compare, among other outcomes, rate of biochemical failure among men diagnosed with localized PCa and treated with External Beam RT alone (EBRT) or in combination with short-term ADT (E+ADT).(91) Patients from the ProCaRS database were included in the comparison if they adhered to PCa-specific eligibility criteria as specified by Jones et al. This involved those with histologically confirmed prostate adenocarcinoma who had PSA levels of ≤20 ng/ml, a clinical T-stage ≤2, and without nodal or metastatic involvement at the time of diagnosis. Since the results of the study by Jones et al. suggested effect modification by risk-group, and the majority of men selected for this comparison (64.4%) from the ProCaRs database had intermediate-risk PCa, we further limited the source population to those diagnosed with intermediate-risk PCa, using the definition provided by Jones et al. EBRT dose in both groups was limited to a total of ≥66 Gy, which slightly varies from the 66.6 Gy implemented by Jones and colleagues but was done to increase the number of participants for our comparison. Further, those who received 3-6 months of ADT before EBRT were included in the E+ADT group, which varies slightly from the four-month duration implemented by Jones et al. but was done for the purpose of increasing the sample size. The definition of biochemical failure was defined as an increase in the PSA level post-treatment of >2ng/ml above the nadir.(61) The final sample size included 126 and 579 men in the E+ADT and EBRT-only groups, respectively. A flowchart of the patient selection process is shown in Figure 5.1b.

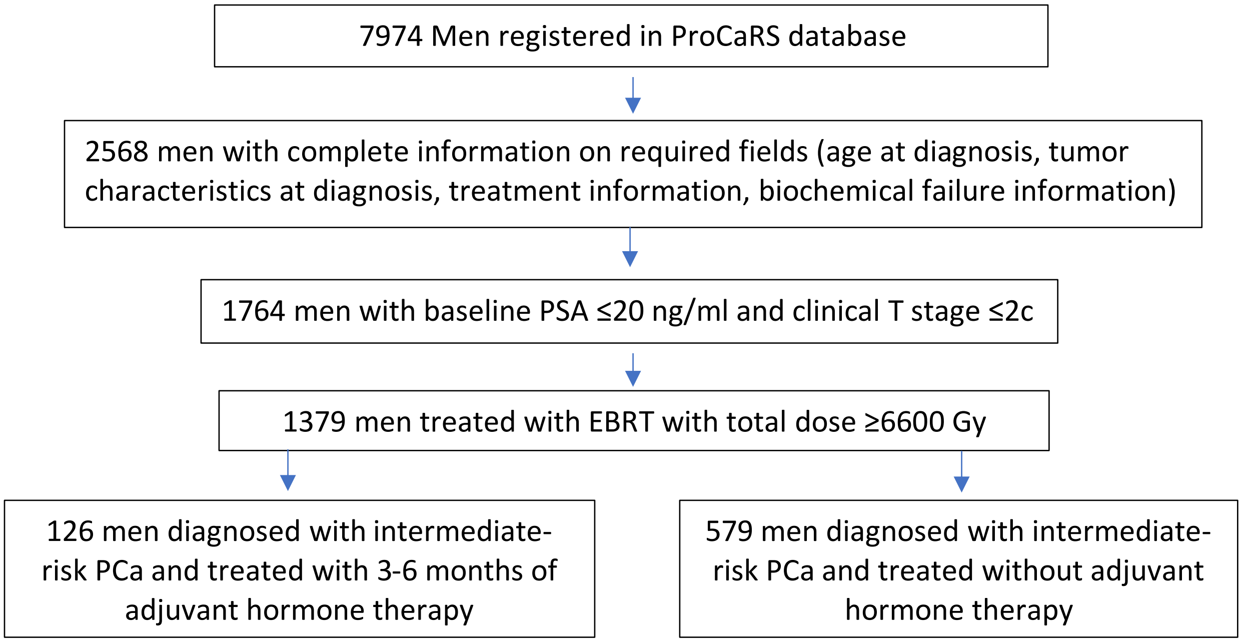

7974 Men registered in ProCaRS database
2568 men with complete information on required fields (age at diagnosis, tumor characteristics at diagnosis, treatment information, biochemical failure information)
1764 men with baseline PSA ≤20 ng/ml and clinical T stage ≤2c
1379 men treated with EBRT with total dose ≥6600 Gy
126 men diagnosed with intermediate-risk PCa and treated with 3-6 months of adjuvant hormone therapy
579 men diagnosed with intermediate-risk PCa and treated without adjuvant hormone therapy

Figure 5.1b Selection process for comparison two.

5.2.4 Covariate selection

We explored the potential for confounding through examining differences between treatment groups in distributions of baseline covariates that have demonstrated a prognostic role in relation to the rate of biochemical failure in previous literature.(19) Covariates included tumor characteristics (i.e. pre-biopsy PSA level, clinical T stage, and Gleason score), EBRT dose and treatment start date. Age was not included as a covariate since it has not demonstrated a consistent association with rate of biochemical failure in previous literature,(176,177) and did not demonstrate a notable association with the rate of biochemical failure in either comparison. Further, age was strongly associated with treatment choice, which would bias effect estimates if adjusted for.(168)

5.2.5 Propensity score matching

The PS model was built according to the systematic approach given in chapter four using the set of balance diagnostics recommended in chapter three to guide development and evaluation. Briefly, the PS model was a logistic model with prognostic characteristics as independent variables and type of treatment received as a binary dependent variable.(168) We explored the possibility of interactions and non-linearity for baseline covariates when developing the PS model, as appropriate specification of interaction and non-linear terms has demonstrated ability to achieve greater balance in baseline covariates between treatment groups.(152,178) Locally weighted scatterplot smoothers were used to assess for departures from linearity in the relationship between continuous predictors and the log odds of the probability for receiving E+ADT. Improvements in the model fit were assessed using the likelihood ratio test and pseudo-R^2. Baseline PSA was modeled as a restricted cubic spline with four knots, treatment start year was treated as a discrete variable with 2-year categories, and an interaction term between baseline PSA and Gleason score was added, as the model specifications improved the predictive value. The Hosmer-Lemeshow goodness-of-fit statistic was examined to test for model adequacy. Further, a plot of DFBETA statistics revealed one outlier wherein a subject received E+ADT instead of BT+ADT despite a very low PSA, clinical T-stage, and Gleason score. This patient was retained, as he did not have any relative or absolute contraindications to receiving E+ADT. The MatchIt package in R was used to match participants between treatment groups on the PS using progressively smaller calipers to identify the optimal balance to sample size trade-off.(146) Specifically, ratios of 1:3 and 1:4 were used for comparison one and two, respectively, given the ratio of index to reference observations available. We also examined matching ratios of 1:1 and 1:2 but did not find any meaningful difference in balance or effect estimates other than decreased precision compared to 1:3 and 1:4 matching ratios. Caliper widths included a range of 0.5 to 0.005 standard deviations of the logit of the PS (S Tables 5.1a and 5.1b). Nearest-neighbour matching was used without replacement. Compared to other matching approaches, nearest-neighbor matching with calipers and without replacement has been found to be less computationally burdensome and produce matches resulting in similar or increased balance, and similar or decreased bias and variance.(179)

5.2.6 Coarsened exact matching

Coarsening of baseline covariates used in CEM were informed by previous evidence surrounding the risk of PCa-specific death over 15 years after diagnosis.(180) In patients with a PSA of <4, 4 to 10, 10.1 to 20 and 20.1 to 50 ng/ml, the risk of PCa-specific death has been estimated to be 4%, 9%, 11%, and 22%, respectively.(180) Moreover, risk evaluation used to guide treatment decisions heavily relies on such thresholds.(19) As such, progressive coarsening for PSA was based on these ranges. Clinical stage, as determined through physical examination of the prostate or imaging, is an ordinal characteristic that takes on values of stages 1-4 with substages ‘a’ through ‘d’. Stages 1a-1c at diagnosis are assigned when there is no palpable tumor detected through digital rectal examination. Along with palpable tumors that occupy one side of the prostate (i.e., T2a and T2b, depending on year of classification), the risk of PCa-specific death over 15 years has been estimated to be between 6% and 7%, so were collapsed into one category.(180) In contrast, patients with bilateral disease not felt to extend outside the prostate upon digital rectal examination (T2b depending on year of classification and T2c) have been estimated to have approximately twice the risk of PCa-specific death over 15 years post-diagnosis as those with unilateral disease, so formed another category for matching. Gleason score was divided into 6 (3+3), 7 (3+4) and 7 (4+3), as these values are associated with notable differences in prognosis.

Patients were matched directly on ordinal covariates (i.e., Gleason score, and collapsed categories of clinical T-stage) and progressively coarsened continuous variables (i.e., PSA, year of RT, EBRT dose (if applicable)). Progressive coarsening for year of treatment was accomplished by dividing the range of values approximately evenly (i.e., halves, thirds, etc.) in the E+ADT group. EBRT dose was split into low (≥6600 Gy and <7300 Gy) and high (≥7300 Gy to <7980 Gy) dosage. Coarsening ranges are presented in S Tables 5.2a and 5.2b.

5.2.7 Balance diagnostics

Many balance diagnostics exist and have been rigorously assessed using various empirical and simulation datasets that represent a broad range of data characteristics. We chose three balance measures that consider different data characteristics in order to monitor improvements in balance when further restricting matching strategies (i.e., using finer ranges for continuous variables in CEM and smaller caliper widths in PSM), while enabling a comprehensive comparison of improvements in balance between PSM and CEM. The standardized mean difference (SMD) in proportion of observations having high-intermediate risk versus low-intermediate risk PCa as defined by the ProCaRS system was used as a prognostic score-based balance measure.(19) 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. Stuart et al. have demonstrated that a prognostic score-based imbalance measure strongly correlated with bias in effect estimates.(156) Since our prognostic score-based balance measure only involved two risk-groups, it was limited in capturing subtle differences in individual variables. As such, we also examined the absolute SMD for individual variables to improve sensitivity in identifying violations of balance in individual variables and the average absolute SMD for baseline covariates to monitor average reduction in absolute SMD when restricting matching strategies to improve balance. Both the absolute and average absolute SMD for baseline covariates have demonstrated a strong correlation with bias in effect estimates in simulation studies.(156,160,161) In addition to these three measures, we also examined the overlap in continuous baseline covariates through overlying density plots and variance ratios.

5.2.8 Descriptive statistics and multivariable regression analysis

All statistical analyses were performed using RStudio version 3.6.0.(169) Descriptive statistics were calculated for each treatment group before and after matching. The median and interquartile range are presented for continuous variables and proportions for categorical variables. Cox proportional-hazards regression analyses for estimating the effect of treatment group on the hazard of biochemical failure were performed using the Survival package.(181) 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. Improvements in model fit were examined through informally comparing the model log likelihoods after incorporating higher order terms and transformations for continuous covariates. Examination of a plot of DBETA statistics did not identify any influential observations. Hazard ratios and 95% confidence intervals were estimated from unmatched data both without and with adjustment for the natural logarithm of PSA, clinical stage, Gleason score, RT start year, and EBRT dose (if applicable). For matched data, we employed Cox models clustered by the matched sets with associated weights to account for variable matching ratios, using robust variance estimators to generate confidence intervals.(181,182) For the CEM strategies, the continuous covariates were included in the model to control for possible residual confounding. For the PSM strategies, all covariates were included in the Cox model.

5.3 Results

5.3.1 Comparison one: BT+ADT versus E+ADT

5.3.1.1 Descriptive statistics

Descriptive statistics for the unmatched treatment groups in comparison one are reported in Table 5.1a. Men treated with BT+ADT were, on average, younger (median age of 68 vs 72 years, respectively) and were treated at earlier dates than men in the E+ADT group (median treatment start year of 2002 vs 2003, respectively). Tumor characteristics were generally less advanced in the BT+ADT group than in the E+ADT group (median PSA: 7.5 vs 9.0 ng/ml, respectively; percentage of Gleason score 7 (4+3): 13% vs 34%) other than clinical stage wherein a greater proportion of BT+ADT had clinical T stage of T2b-c (14% vs 8%, respectively). This paralleled the smaller percentage of BT+ADT occupying the high-intermediate risk strata (7% vs 25%, respectively). Finally, the median duration of ADT was similar between groups (6.0 vs 5.4 months).

Table 5.1a Descriptive statistics for comparison one

	BT+ADT (n=433)		E+ADT (n=132)		SMD	Variance Ratio
Age						
Median	68		72		0.7644	0.55
IQR	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
IQR	5.70, 10.50		5.88, 12.60			
Gleason Grade						
1	126	29.10%	23	17.42%	0.2790	
2	249	57.51%	64	48.48%	0.1815	
3	58	13.39%	45	34.09%	0.5014	
RT Start Year						
Median	2002		2003		0.3029	0.67
IQR	2001, 2004		2002, 2004			
ADT Duration (Months)						
Median	5.98		5.45		0.1285	2.84
IQR	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%		

5.3.1.2 Performance of matching strategies

The number of patients and events retained for each CEM and PSM strategy were examined in comparison one and are presented in S Tables 5.1a and 5.3a, respectively. PSM strategy 10 and CEM strategy eight led to optimal balance to sample size trade-off so were used for further analysis.

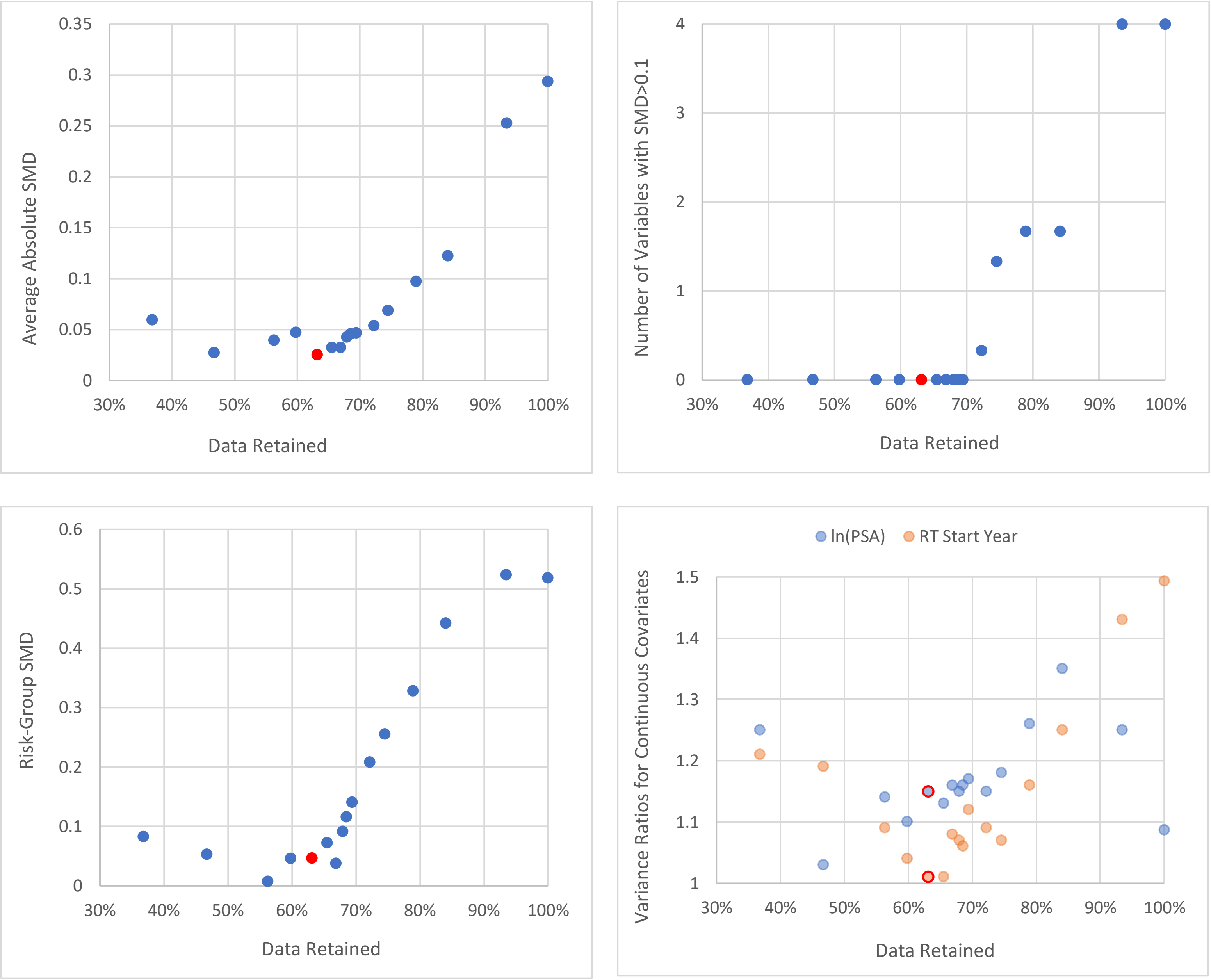


Figure 5.2a Balance achieved with each PSM strategy by percent of data retained for comparison one.

*The red dot indicates the chosen matching strategy

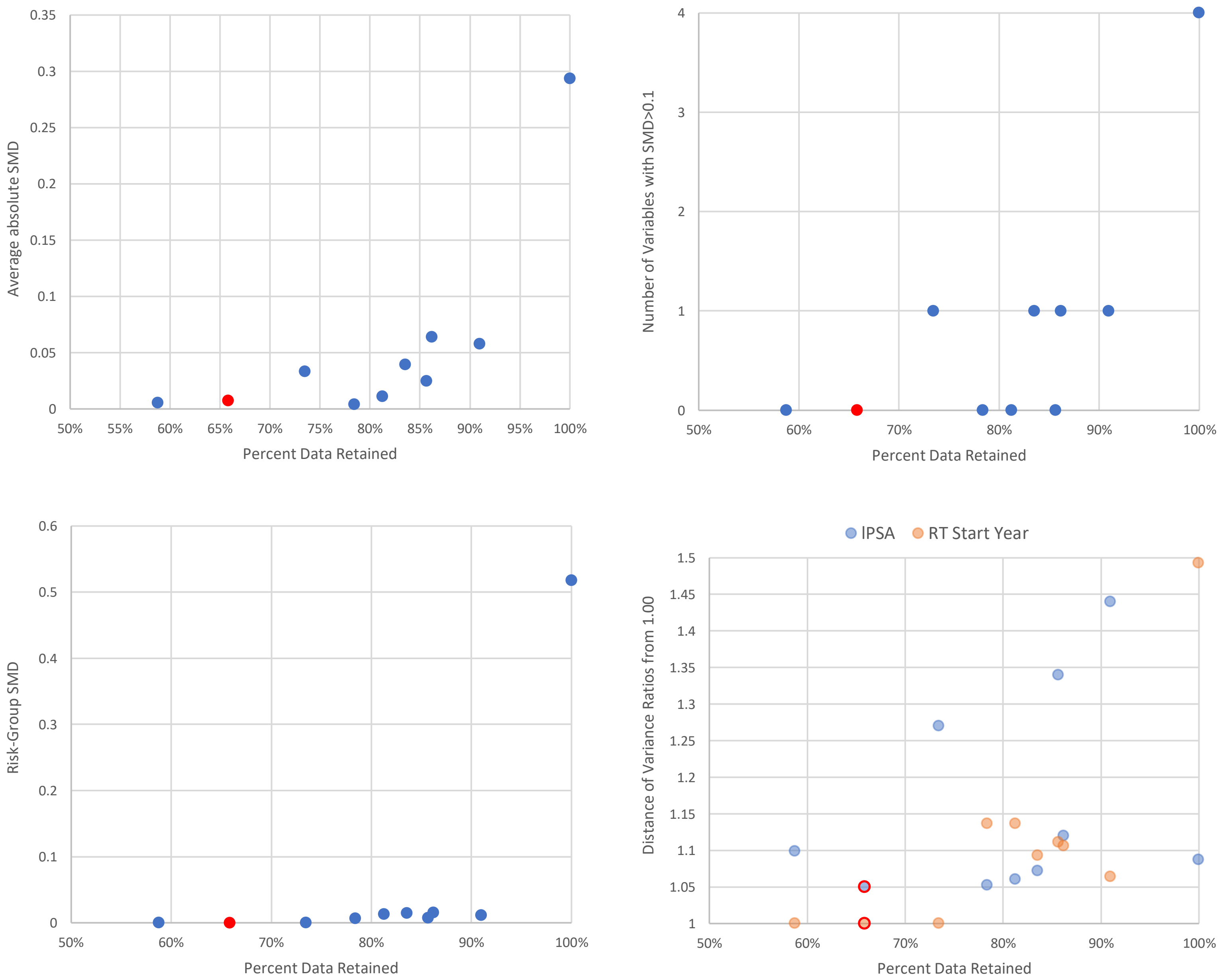


Figure 5.2b Balance achieved with each CEM strategy by percent of data retained for comparison one.

*The red dot indicates the chosen matching strategy

Median values for continuous covariates and proportions for categorical covariates according to matching strategy are presented in S Figures 5.1a and 5.1b. As matching approaches became stricter, the BT+ADT group characteristics tended toward those of the

E+ADT until a certain point wherein characteristics in both groups tended toward those of the BT+ADT group. In the matching strategy chosen, characteristics for both groups represented an average of both groups, as would be expected in areas of common support.

Density plots for continuous covariates before and after matching are presented in S Figures 5.2a and 5.2b. Overlap in treatment start date and baseline ln(PSA) improved after both matching strategies.

Descriptive statistics for the matched groups from the selected PSM and CEM strategies are presented in Table 5.2a. The distribution of baseline covariate values in the matched samples represents an average of the distribution of covariates from both groups.

Table 5.2a Descriptive statistics for PSM strategy 10 and CEM strategy 8 in comparison one

	PSM 10				CEM 8			
	BT+ADT (n=248)		E+ADT (n=109)		BT+ADT (n=276)		E+ADT (n=96)	
Age (years)								
Median	69		72		69.5		72	
IQR	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	
IQR	5.65, 10.60		5.63, 10.70		6.08, 11.00		5.72, 11.78	
Gleason Grade								
1	52	21.10%	23	21.10%	63	22.92%	22	22.92%
2	119	47.86%	52	47.71%	152	55.21%	53	55.21%
3	77	31.04%	34	31.19%	60	21.88%	21	21.88%
RT Start Year								
Median	2003		2003		2003.5		2003.5	
IQR	2002, 2004		2002, 2004		2002, 2005		2002, 2005	
ADT Duration (Months)								
Median	5.98		5.49		5.95		5.67	
IQR	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%

SMDs of individual covariates after PSM and CEM relative to the source population are presented in Figure 5.3. PSM and CEM improved balance in the absolute SMD relative to the unmatched sample. CEM achieved similar or more balance in the absolute SMD relative to PSM.

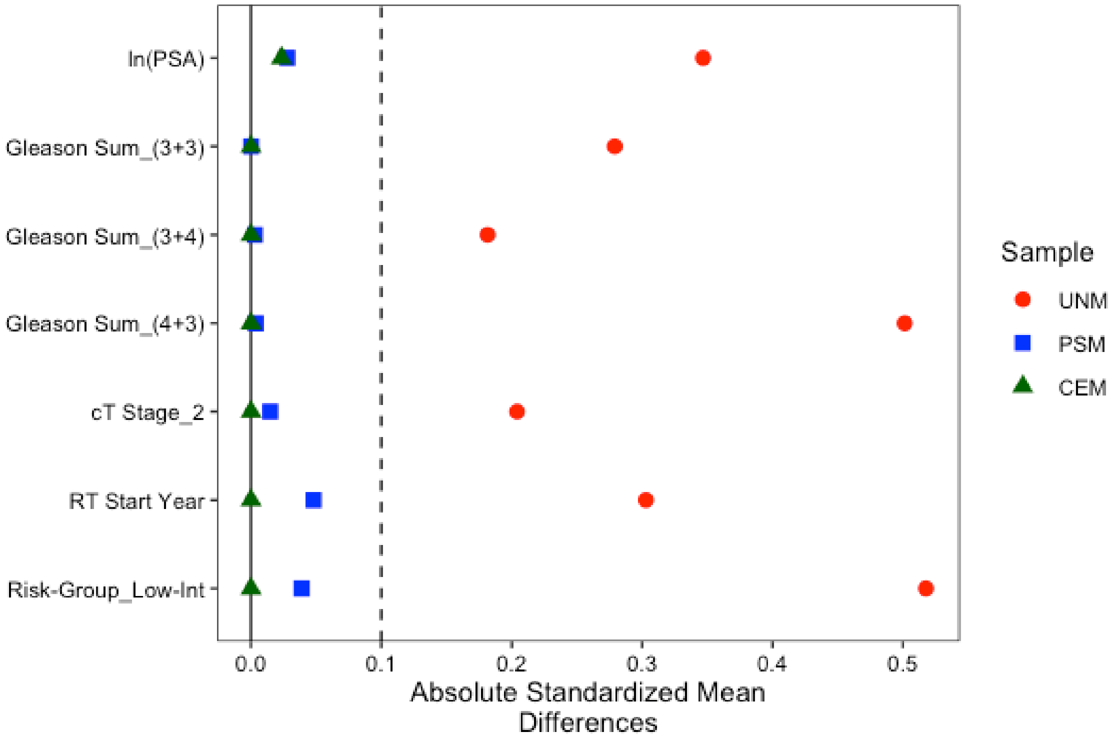


Figure 5.3 Love plot of the absolute SMD for individual baseline covariates before matching and after PSM and CEM in comparison one.

The effect estimates are presented in Table 5.3a. For the benchmark RCT hazard ratio estimate (95% confidence interval) of 2.04 [1.25, 3.33], the corresponding unadjusted effect hazard-ratio estimate (95% CI) was 6.55 [3.82, 11.26], while adjusting for relevant baseline covariates, the hazard-ratio estimate (95% CI) decreased to 4.48 [2.44, 8.22]. The unadjusted and multivariable adjusted hazard-ratio estimates (95% CI) 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 demonstrated similar improvements in imbalance led to similar point estimates and confidence intervals (Table 5.3a).

Table 5.3a 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

5.3.2 Comparison two: EBRT versus E+ADT

5.3.2.1 Descriptive Statistics

Descriptive statistics for unmatched treatment groups in analysis two are reported in Table 5.1b. Treatment groups were similar (SMD<0.1) with respect to age, PSA, and proportion of low- vs high-intermediate risk-group status. The E+ADT group had a slightly greater proportion of men diagnosed with clinical T1a-2a disease than the EBRT group. Those in the E+ADT group also had a greater proportion of Gleason sum 7 (3+4) and 7 (4+3) disease and received higher doses of EBRT, on average. Men from the E+ADT group also received treatment, on average, later than those from the EBRT group in calendar time.

Table 5.1b Descriptive statistics for PSM strategy 4 and CEM strategy 7 in comparison two

	PSM 6				CEM 7			
	E+ADT (n=126)		EBRT (n=347)		E+ADT (n=118)		BT+ADT (n=377)	
Age (years)								
Median	72		72		72		72	
IQR	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	
IQR	5.71, 12.46		5.87, 12.05		5.75, 12.15		6.10, 12.30	
Gleason Grade								
1	23	18.25%	68	19.64%	22	18.64%	70	18.64%
2	67	53.18%	179	51.46%	66	55.93%	211	55.93%
3	36	28.57%	100	28.90%	30	25.42%	96	25.42%
RT Start Year								
Median	2001		2001		2001		2001	
IQR	2000, 2004		2000, 2004		2000, 2004		2000, 2004	
EBRT Dose (Gy)								
Median	7560		7560		7560		7560	
IQR	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%

5.3.2.2 Performance of matching strategies

Data characteristics for PSM and CEM strategies examined in comparison two are presented in S Tables 5.1b and 5.3b, respectively. PSM strategy six and CEM strategy seven led to optimal balance to sample size trade-off so were used for further analysis. Figures 5.4a and 5.4b show 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-eight percent 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. Both strategies maintained SMD for all individual covariates under <0.1, and variance ratios for continuous covariates within the acceptable range of 0.92 to 1.08.(152)

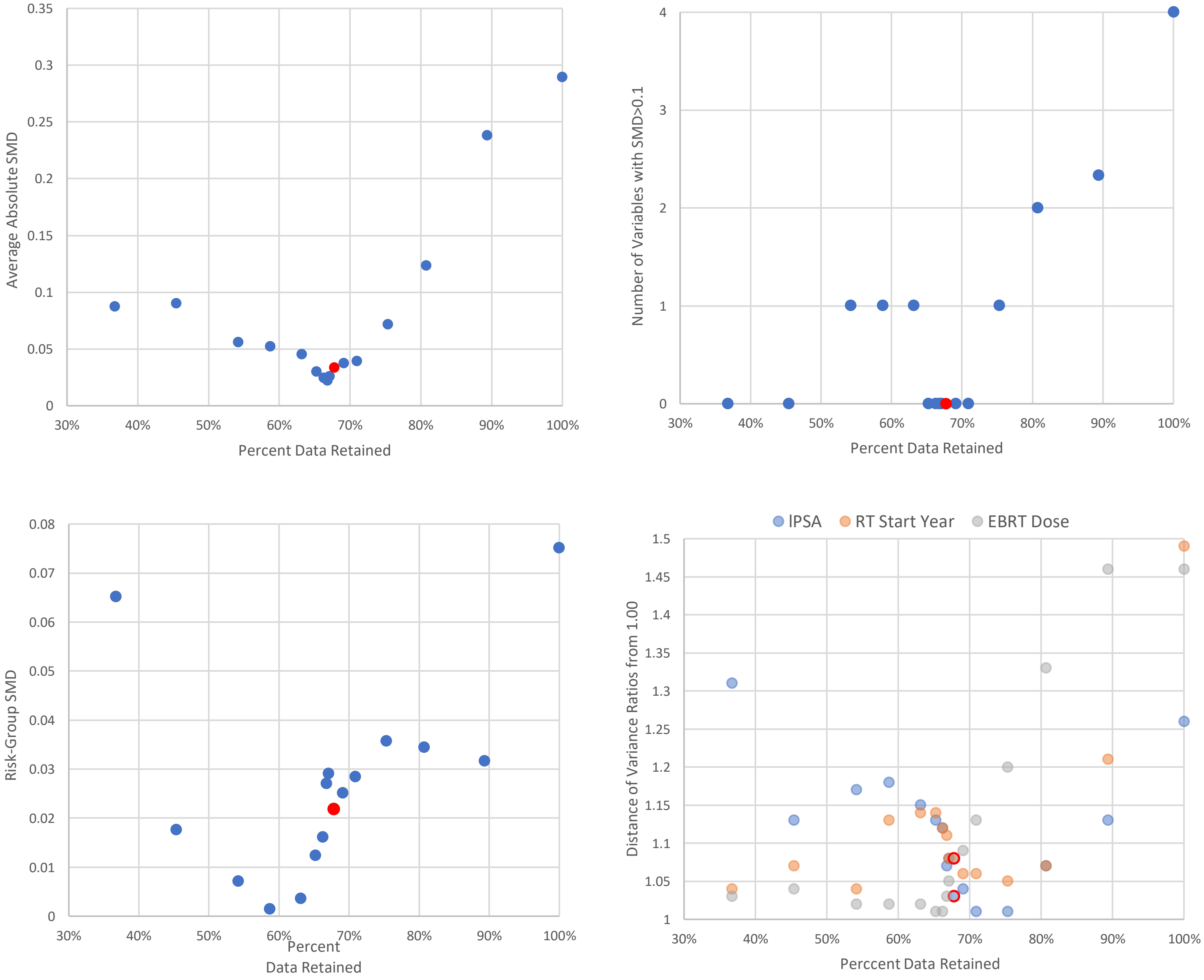


Figure 5.4a Balance achieved with each PSM strategy by percent of data retained for comparison two.

*The red dot indicates the chosen matching strategy

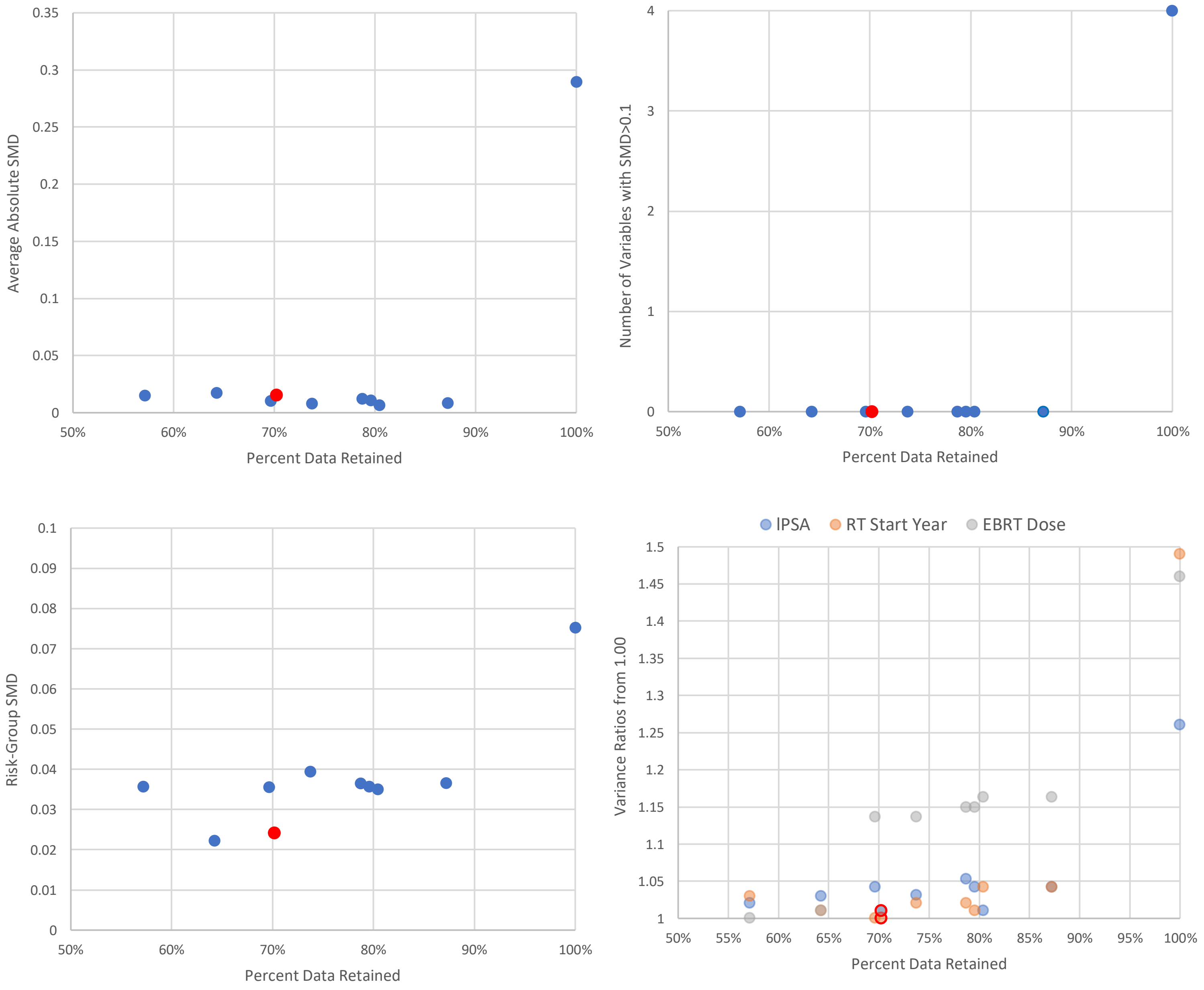


Figure 5.4b Balance achieved with each CEM strategy by percent of data retained for comparison two.

*The red dot indicates the chosen matching strategy

Descriptive statistics for the matched groups from the selected CEM and PSM strategies are presented in Table 5.2b. The distribution of baseline covariate values in the matched samples represents an average of both groups where common support exists.

Table 5.2b Descriptive statistics for PSM strategy 4 and CEM strategy 7 in comparison two

	PSM 6				CEM 7			
	E+ADT (n=126)		EBRT (n=347)		E+ADT (n=118)		BT+ADT (n=377)	
Age (years)								
Median	72		72		72		72	
IQR	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	
IQR	5.71, 12.46		5.87, 12.05		5.75, 12.15		6.10, 12.30	
Gleason Grade								
1	23	18.25%	68	19.64%	22	18.64%	70	18.64%
2	67	53.18%	179	51.46%	66	55.93%	211	55.93%
3	36	28.57%	100	28.90%	30	25.42%	96	25.42%
RT Start Year								
Median	2001		2001		2001		2001	
IQR	2000, 2004		2000, 2004		2000, 2004		2000, 2004	
EBRT Dose (Gy)								
Median	7560		7560		7560		7560	
IQR	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%

Median values for continuous covariates and proportions for categorical covariates according to matching strategy are presented in S Figures 5.3a and 5.3b. As matching approaches became stricter, the EBRT group characteristics tended toward those of the E+ADT until a certain point wherein characteristics in both groups tended toward those of the EBRT group. In the matching strategy chosen, characteristics for both groups represented an average of both groups, as would be expected in areas of common support.

Density plots for continuous covariates before and after matching are presented in S Figures 5.4a and 5.4b. Similar to comparison one, overlap in treatment start date and baseline ln(PSA) improved after both matching strategies compared to the unmatched sample.

SMDs of individual covariates after PSM and CEM relative to the unmatched sample are presented in Figure 5.5. Both matching strategies improved balance in the absolute SMD in all covariates relative to the unmatched sample except baseline ln(PSA), as this variables was already balanced between groups.

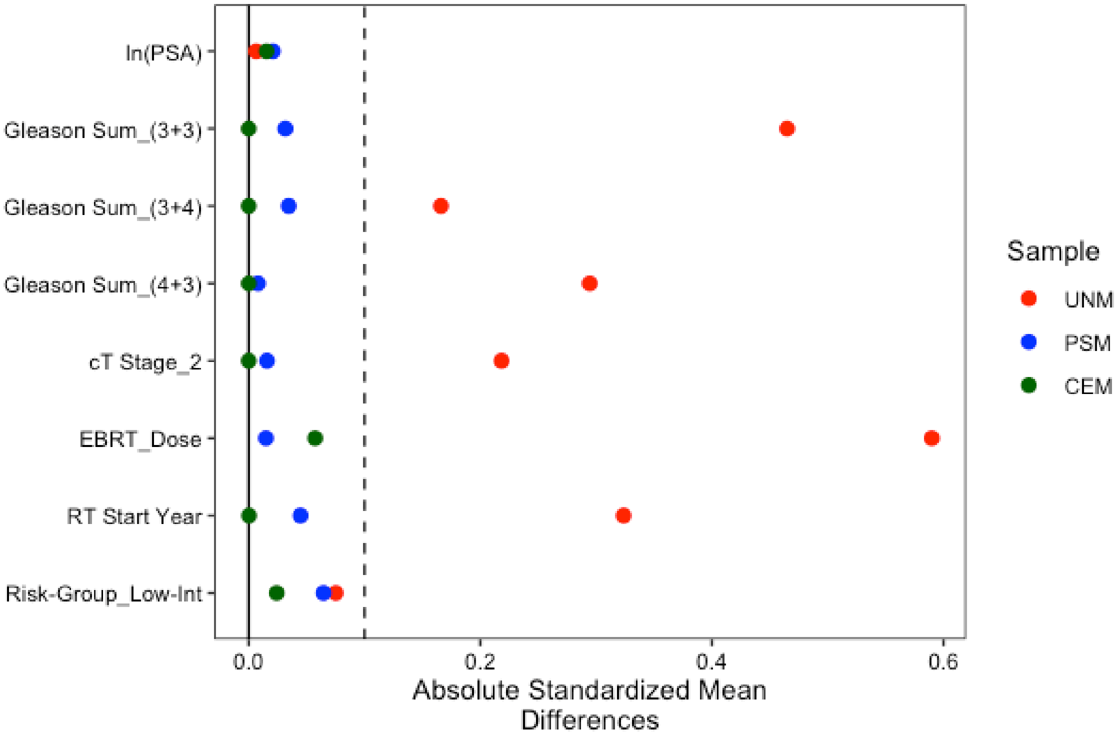


Figure 5.5 Love plot of the absolute SMD for individual baseline covariates before matching and after PSM and CEM in comparison two.

The effect estimates are presented in Table 5.3b. Compared to the benchmark RCT hazard ratio (95% CI) of 1.79 [1.45, 2.21], the unadjusted effect estimate (95% CI) was 1.40 [0.99, 1.98]. After adjusting for relevant baseline covariates, the hazard ratio estimate (95% CI) increased to 1.52 [1.06, 2.16]. The unadjusted and multivariable adjusted hazard-ratio estimates (95% CI) after PSM were 1.39 [0.97, 1.99] and 1.44 [1.00, 2.05]. CEM provided similar effect estimates (1.53 [0.95, 2.46] without adjustment and 1.55 [0.98, 2.45] after multivariable adjustment). Other candidate matching strategies for both PSM and CEM that demonstrated similar improvements in imbalance led to similar point estimates and confidence intervals (Table 5.3b).

Table 5.3b 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

5.4 Discussion

The purpose of this study was to compare the performance of two popular data-preprocessing techniques in the context of non-experimental datasets, using two examples from PCa CER. 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 proportions of retention of observations from the original dataset. Furthermore, the risk-group SMD, which reflects imbalance in prognostic score between treatment groups 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 than PSM for baseline PSA and treatment start date; however, PSM led to a variance-variance ratio closer to one for radiation treatment dosing between treatment groups. These findings are consistent with other studies wherein large improvements in balance were observed after CEM compared to PSM using balance diagnostics based on the comparison of multivariable distributions between treatment groups.(147,148)

In the first comparison, the rate of biochemical progression was elevated in the E+ADT compared with the BT+ADT group. Some of the difference in the rate of biochemical progression between treatment groups can be attributed to differences in the risk of extraprostatic extension, which is reflected in the baseline measures of PSA, clinical stage, and Gleason sum in addition to changing clinical practices and technology that occur overtime, which are reflected, in part, by the calendar year of treatment. After adjusting for such variables, the hazard ratio was attenuated (6.55 vs 4.48). The adjusted effect estimate is more consistent with that of the benchmark RCT (2.04). After both PSM and CEM, however, the effect estimate after multivariable modeling was, on average, closer to that of the benchmark RCT (3.84 and 3.84, respectively). Attenuation of the effect estimate after matching might be due to the limitations of multivariable regression modeling to adequately control for confounding. To clarify, appropriate model specification would require adequate representation of the functional forms of the relations between the study 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 modelling of effect estimates rests on the assumption of positivity, with even small violations of which potentially resulting in biased effect estimates.(183)

Even without further adjustment for confounding, matching led to a stronger attenuation in the effect estimate than multivariable matching (4.06 and 4.04 for PSM and CEM, respectively, versus 4.48 after multivariable modeling alone) that was closer to the benchmark RCT. This might demonstrate the bias reduction potential offered through PSM and CEM even without further multivariable adjustment. However, the further attenuation in the effect estimate afforded through multivariable adjustment after matching demonstrates the remaining confounding not entirely managed by through matching strategies performed.

In the second comparison, the rate of biochemical progression was elevated in the EBRT compared to the E+ADT group; however, the difference was likely underestimated since, as mentioned in the background, ADT is generally reserved for those with a greater risk of extraprostatic extension. This is reflected in the differences in baseline characteristics between treatment groups wherein those undergoing E+ADT had, on average, worse prognosis than those in the EBRT group. However, such differences in prognosis were not as substantial as differences in prognosis between treatment groups in the first comparison, as demonstrated by the smaller risk-group SMD in comparison two relative to one (0.075 versus 0.52, respectively). Multivariable adjustment led to an increased effect estimate (1.52 versus 1.40) after adjusting for potential confounding variables. The relative difference in unadjusted and adjusted effect estimate compared with the first comparison was much smaller. This is likely attributable to the greater balance observed between treatment groups in comparison two relative to that in comparison one. This notion is further supported in that matching did not substantially change effect estimates (1.44 and 1.55 after PSM and CEM, respectively).

An alternative explanation for the observed differences in the effect estimates might be the differences in treatments implemented in the randomized trial relative to those administered in the ProCaRS database. Since BT monotherapy and EBRT without dose escalation are likely to have different impacts on the rate of biochemical failure compared to EBRT with BT boost and EBRT with dose-escalation, the ASCENDE-trial can serve only as a loose guideline in interpreting effect estimates rather than a gold standard. Another explanation could be that each approach estimates a different parameter. Specifically, multivariable 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., E+ADT), which has been termed the average treatment effect among the treated. 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 termed by Iacus and colleagues as the feasible sample average treatment effect among the treated.(159) Random variation of the hazards ratio might also explain the findings, at least partly. However, effect estimates drawn from several candidate PSM and CEM strategies consistently estimated effects more in line with 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 multivariable modeling where imbalance was not as substantial. The comparison of results from the benchmark RCTs with those obtained after matching to support trends in bias reduction is limited since characteristics of the treatment groups and treatment approaches for each comparison varied notably from the chosen benchmark RCT.

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. This is also favorable since results are more 'generalizable' to patient groups with characteristics that are amenable for either treatment under comparison. 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.(147) This seeming discrepancy is likely explained by the difference in dimensionality between datasets used for matching. The covariate sets Fullerton et al. used to match between comparison groups involved over 80 variables, whereas we only used matching on four and five covariates. This explanation is supported by findings from Ripollone et al., who reported that smaller covariate sets of 8 covariates retained a substantially greater proportion of the original population compared to larger covariate sets with up to 119 covariates (32.5% vs 3.6%, respectively).(148)

The precision of effect estimates did not differ notably between PSM and for CEM. Although Fullerton et al. and Ripollone et al. found that greater precision was observed for PSM than for CEM in high-dimensional datasets, differences attenuated as the number of covariates decreased.(147,148) Since our datasets included a small number of covariates, similarity in precision achieved after PSM and CEM is to be expected.

The PSM paradox, as demonstrated by King and Nielsen,(136) became apparent when restricting matches to exceedingly smaller caliper ranges. In particular, measures of imbalance became very sporadic, increasing and decreasing with progressively smaller calipers ≤0.1 in analysis one and ≤0.2 in analysis two (Figures 5.2a and 5.4a). This could also be due to the fact that sample sizes of comparison groups became exceedingly smaller as caliper size progressively decreased. Austin noted that the standard deviation of SMD increases in smaller samples, so greater variation of SMD is expected with progressively smaller calipers.(152) This phenomenon was also observed in a study by Belitser et al., who found correlation of SMD with bias decreased in smaller sample sizes.(161) These findings have substantial implications for researchers who use standard caliper sizes instead of exploring progressively smaller caliper ranges to identify optimal balance before the PSM paradox kicks in or when estimation of balance becomes sporadic and unreliable. Thus, we recommend that analysts explore progressively smaller matching calipers when PSM in order to examine the trade-off between balance and sample size and identify a suitable matching strategy for their research purposes.

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 retention of the original study population with progressively stricter matching criteria (i.e., smaller caliper sizes in PSM and finer ranges in continuous variables in CEM). 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 cut-points for continuous variables rather than rely solely on quantile-based rules, as in previous studies.(145,147) This has the potential to optimize the efficiency of matching strategies by reducing imbalance while retaining a greater part of the original sample. 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.

5.5 Conclusions

In summary, both matching strategies appear to be effective in managing confounding. The use of multivariable adjustment 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 pre-processing of non-experimental data with a relatively small number of covariates that can result in retention of a large proportion of the original study sample from which to generate effect estimates with reasonable precision and utility to inform clinical practice in the absence of RCTs. CEM also has potential to further improve balance in the multivariable distribution of baseline covariates between comparison groups, compared with PSM.

5.6 Transition to Chapter Six

This chapter compared the performance of propensity score matching and coarsened exact matching in the ability to balance the multivariable distribution of baseline covariates per level of data retention using methods developed in chapter four. Two treatment comparisons of men diagnosed with PCa and treated with different combinations of RT and ADT in Canada were used to compare these matching strategies. RCTs that compared similar groups of men who were similarly managed were used to further inform how changes in balance led to changes in effect estimates. The results from this comparison thus add to the real-world evidence on the performance of matching strategies and show the potential that matching can have in improving the validity of effect estimates in PCa comparative effectiveness research. As such, the methods developed and evaluated in chapters four and five were used in a comparison of the rate of MPFS between men diagnosed with unfavorable-risk non-metastatic PCa who were initially treated with either RT or RP in chapter seven.

However, before chapter seven, a systematic review and meta-analysis of studies comparing the rate of CSS between men diagnosed with high-risk non-metastatic PCa who were initially treated with either RT or RP was performed in response to limitations in a previous meta-analysis on the topic and to provide context from the medical literature for the comparison done in chapter seven. Comparisons using records from men diagnosed with high-risk non-metastatic PCa instead of unfavorable-risk (i.e., unfavorable-intermediate-, high- and very-risk) non-metastatic PCa is performed, as the term unfavorable-risk is a recent change in risk-stratification terminology.(19) Therefore, most studies with sufficient follow-up for CSS that are available for systematic review and meta-analysis do not use this terminology.

Chapter 6: Characterizing Surgical and Radiotherapy Outcomes in Non-Metastatic High-Risk Prostate Cancer: A Systematic Review and Meta-analysis

Authors: David Guy (MD/PhD candidate),[1,2] Hanbo Chen (MD),[3] Gabriel Boldt (MLIS),[2] Joseph Chin (MD),[1,2] George Rodrigues (MD, PhD)[1,2]

Affiliations:

[1]Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

[2]London Health Sciences Centre, London, ON, Canada

[3]Sunnybrook Odette Cancer Centre, Toronto, ON, Canada

Abstract:

Background

Identifying the optimal management of high-risk non-metastatic prostate cancer (PCa) is an important public health concern given the large burden of this disease. We performed a meta-analysis of studies comparing PCa-specific mortality (CSM) and all-cause mortality (ACM) among men diagnosed with high-risk non-metastatic PCa who were treated with primary radiotherapy (RT) and radical prostatectomy (RP).

Methods

Medline and EMBASE were queried for articles between 2005 to 2020. After title and abstract screening, two authors independently reviewed full-text articles for inclusion. Data were abstracted and a modified version of the Newcastle-Ottawa Scale, involving a comprehensive list of confounding variables, was used to assess risk of bias.

Results

Fourteen studies involving 88,543 patients were included. No difference in adjusted CSM in RT relative to RP was shown (hazard ratio, 1.02 [95% confidence interval: 0.84, 1.25]). Increased CSM was found in a moderator analysis comparing external beam radiation therapy (EBRT) with RP (1.35 [1.10, 1.68]) whereas EBRT combined with brachytherapy (BT) versus RP showed lower CSM (0.68 [0.48, 0.95]). All studies demonstrated a high risk of bias, as none fully adjusted for all confounding variables.

Conclusion

We found no difference in CSM and ACM between men diagnosed with non-metastatic high-risk PCa treated with RP or RT; however, this is likely explained by increased CSM in men

treated with EBRT and decreased CSM in men treated with EBRT+BT studies relative to RP. High-risk of bias in all studies identifies the need for better data collection and confounding control in PCa research.

6.1 Rationale and Objectives

Prostate cancer (PCa) was the second most frequently diagnosed cancer and fifth leading cause of cancer death worldwide as of 2018.(3) High-risk PCa, as defined by a clinical stage ≥T3, Gleason score 8-10 or prostate-specific antigen >20ng/ml at the time of diagnosis,(184) accounts for approximately one quarter of all PCa diagnoses but was responsible for a disproportionately larger share of PCa-specific mortality (CSM).(185) Optimal selection and sequencing of therapy for high-risk non-metastatic PCa, such as the choice between radical prostatectomy (RP) and radical radiotherapy (RT), remains an area of intense academic and clinical debate.(108) Unfortunately, no RCTs on this topic have been completed due to low patient and provider equipoise surrounding RP and RT, especially in North America.(110)(111) As such, investigations comparing RP and RT outcomes have mostly been performed using non-randomized data. In the absence of RCTs, meta-analyses that summarize non-randomized data can inform treatment decisions for physicians and policymakers.

Previous meta-analyses that have compared mortality outcomes between patients diagnosed with PCa and treated with RP or RT involved studies that compared older treatment approaches, which greatly differ from current standards of care.(186) Publications included in these meta-analyses have since been updated to include longer follow-up periods of more contemporary RT approaches such as dose-escalation protocols for external beam radiation therapy (EBRT), use of brachytherapy boost (BT) and adjuvant androgen deprivation therapy (ADT),(101,120,187,188) which may lead to better oncological outcomes for men diagnosed with high-risk non-metastatic PCa.(91,167,189) Although a more recent meta-analyses has been conducted,(190) numerous errors were made, limiting the utility of the aggregated effect estimates for use in clinical practice. For instance, multiple effect estimates were generated from overlapping data,(99,120,191–196) leading to some patient data overinfluencing aggregate effect estimates as well as inclusion of a study investigating low-risk PCa.(187) Moreover, the authors aggregated studies involving patients diagnosed with non-metastatic and nodal metastatic high-risk PCa,(193) which have heterogenous disease trajectories and ultimately call for different management approaches that are not comparable.(21)

The objective of this study was to compare the rates of CSM and ACM between men diagnosed with high-risk non-metastatic PCa and treated with RP or RT as their primary treatment modality.

6.2 Methods

6.2.1 Research question

The primary and secondary objectives of the study were to summarize the relative CSM and ACM, respectively, of patients diagnosed with non-metastatic high-risk PCa treated primarily with either RP or RT.

Common endpoints used to compare treatment effectiveness in PCa also have potential for bias. These include BFFS, metastatic progression-free survival (MPFS), CSS and OS. The definition of biochemical failure among patients treated with RP and RT differ. Since RP removes the whole prostate gland, it is anticipated that the prostate specific biomarker (used to define biochemical failure) maintain a non-significant value of <0.2ng/ml.(121,122) Since RT does not remove the gland, it is anticipated that some prostate specific biomarker remains.(61) The biochemical failure definition post-RP intended to indicate cure whereas the definition post-RT is sensitive and specific for future clinical outcomes of interest (e.g. distant failure, PSA >25 ng/ml, etc.).(123,124) These definitions represent different disease kinetics and outcomes and should not be compared in the context of comparative effectiveness research. Evidence of metastasis depends on the presence of prompts to image such as biochemical failure and symptoms. Prompts may differ depending on whether the patient underwent RP or RT and the frequency of follow-up, which is dependent on characteristics of the patient, physician, and treatment centre among other factors. Finally, PCa-specific mortality as ascertained by death certificates is not immune to bias. Death is sometimes misattributed to PCa among PCa patients who die of other causes. This misattribution bias is more likely to occur among those with multiple comorbidities, as deciphering cause of death among multiple causes can be difficult. Since RT patients are more likely to have multiple comorbidities, this would negatively impact survival outcomes when compared with patients undergoing RP independent of treatment status. OS poses little concern for measurement bias, as causes do not have to be ascertained.

6.2.2 Protocol and search strategy

The systematic review was conducted in accordance with the PRISMA guidelines.(197) The review protocol has been registered with PROSPERO (registration number: CRD42020150710). The search strategy is provided in Appendix E1. Studies were included in our analysis if they were published after 2005 to limit attention to analyses of more contemporary treatment periods up to February 11, 2020. Only full-text articles published in English in a peer-reviewed journal were considered.

We included only cohort studies in our review since case-control studies typically do not evaluate hazard ratios. Furthermore, previous RCTs were excluded since, due to insufficient numbers of men diagnosed with nonmetastatic high-risk PCa, hazard ratios for this risk-group were not provided.(198) Editorials, letters to the editor, commentaries, guidelines and review articles were also excluded.

We included studies that reported on men of any age diagnosed with non-metastatic high-risk PCa, according to the National Comprehensive Cancer Network (clinical stage ≥ T3 or Gleason score 8-10 or prostate specific antigen > 20 ng/ml),(184) or D'Amico criteria (clinical stage ≥ T2c or Gleason score 8-10 or prostate specific antigen > 20 ng/ml)(25) who were treated with either primary RP or RT. All common forms of RP (e.g., open retropubic, laparoscopic, and robotic) and RT (e.g., conformal external beam, intensity-modulated, brachytherapy or combination of radiotherapy modalities with curative intent) were considered. Studies assessing adjuvant or salvage therapies as the primary objective were excluded. We included only studies that provided a hazard ratio for CSM or ACM, both having managed confounding (i.e., through prevention or adjustment). Studies reporting on surrogate outcome measures such as biochemical progression were excluded, since definitions for RP and RT differ.

6.2.3 Article review

The first phase of the project involved title and abstract review by DG to discard non-relevant citations and duplications. Full-text reviews of remaining studies were examined in the second phase by DG and HC to determine eligibility for inclusion based on pre-determined criteria. Afterward, DG and HC independently reviewed the records, and GBR settled discrepancies on inclusion/exclusion of certain records. Where more than one publication existed using the same patient population, the most relevant, updated, and complete publication was selected. A diagram describing the study flow is outlined in Figure 6.1.

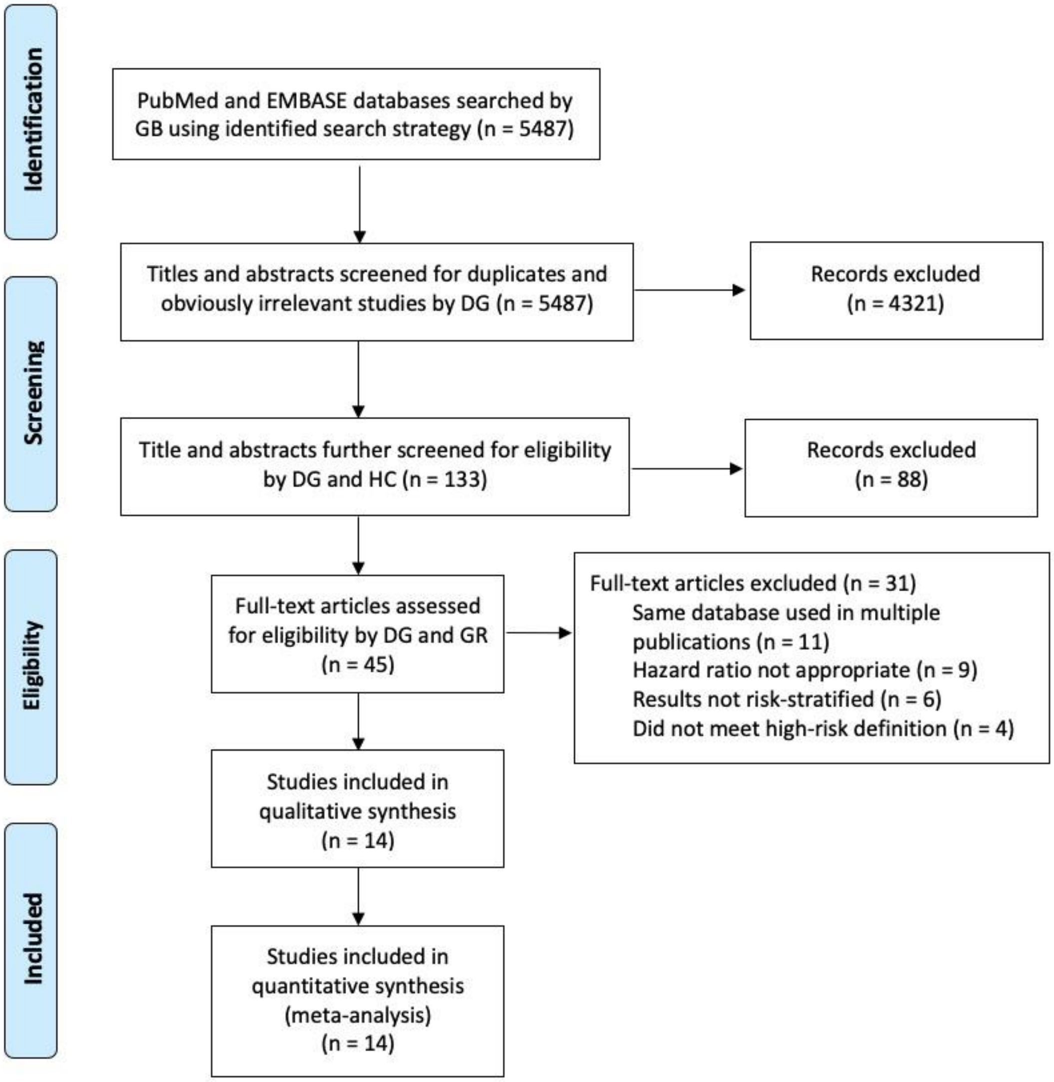

Identification
Screening
Eligibility
Included
PubMed and EMBASE databases searched by GB using identified search strategy (n = 5487)
Titles and abstracts screened for duplicates and obviously irrelevant studies by DG (n = 5487)
Records excluded (n = 4321)
Title and abstracts further screened for eligibility by DG and HC (n = 133)
Records excluded (n = 88)
Full-text articles assessed for eligibility by DG and GR (n = 45)
Full-text articles excluded (n = 31)
Same database used in multiple publications (n = 11)
Hazard ratio not appropriate (n = 9)
Results not risk-stratified (n = 6)
Did not meet high-risk definition (n = 4)
Studies included in qualitative synthesis (n = 14)
Studies included in quantitative synthesis (meta-analysis) (n = 14)

Figure 6.1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram outlining search strategy and final included and excluded studies.

6.2.4 Data extraction and risk of bias assessment

A data extraction form was completed for each study as outlined in Appendix E2. We used a modified version of the Newcastle Ottawa Scale to include a comprehensive list of items identifying confounding variables (see Appendix E3). Confounding variables included those relating to tumor characteristics (baseline PSA, Gleason score, and clinical stage), age, comorbidity status, year of diagnosis or treatment, study center (if multiple), and at least one demographic characteristic (e.g., education, income, rural or urban residence). This list was reviewed and approved by both a radiation oncologist (GR) and uro-oncologist (JC).

6.2.5 Publication bias

We assessed publication bias using funnel plots and the Egger test. Hazard ratios from included studies were plotted as a function of their standard error in relation to the aggregate effect estimate generated through random-effects models. Residual values were also estimated using mixed-effects models to account for heterogeneity due to moderator variables (RT approach for CSM and ACM, and age for ACM) in order to improve interpretation of funnel plots for the assessment of publication bias.

6.2.6 Assessment of heterogeneity

The Q-test was performed to identify significant heterogeneity in treatment effect estimates, using the Dersimonian-Laird method, and quantified through the I^2 statistic.(199)

6.2.7 Statistical analysis

General study information, PCa treatment and endpoint information, and methodological information were categorized into tables using frequency or proportions for categorical variables, medians or means for continuous variables, and descriptive terms for other variables where appropriate.

The meta-analysis was performed in R (x64, version 3.3.2; R Foundation for Statistical Computing) with the "metafor" package (version 1.9-9).(200) The primary meta-analysis with CSM as the outcome and initial treatment received (i.e., RP or RT) as the only independent variable was carried out using inverse variance-weighted random effects models. We then performed a series of univariable meta-regressions to explore sources of heterogeneity. Input variables included treatment era (examined as a binary variable with values of 1 and 0 for values above and below the median year of diagnosis, respectively), approach to RT (external beam radiation therapy with or without brachytherapy boost), length of follow-up (examined as a binary variable with values of 1 and 0 for values above and below the median, respectively), geographical location (United States versus other) and age (examined as a binary variable with values of 1 and 0 for values above and below the median, respectively). Insufficient data were available to explore the effect of RT dose, RP approach (i.e., open, laparoscopic, robotic), proportion receiving systemic therapy (i.e., androgen deprivation therapy, chemotherapy, and adjuvant RT), and type of EBRT (i.e., 3D conformal, IMRT, etc.). All statistical tests were two-sided with significance levels of <0.05.

6.3 Results

Fourteen studies involving 89,167 total patients were identified for inclusion. The article selection flowchart is outlined in Figure 6.1.

6.3.1 Study characteristics

Table 6.1 shows characteristics for individual studies. Four studies compared treatment groups from a single institution, another four studies compared groups from different institutions, another four studies used national registries to compare treatment groups and two studies made comparisons across multiple institutions. Patient characteristics varied across studies due to variations in inclusion and exclusion criteria. In general, RT patients were, on average, older, had a greater number of comorbidities and poorer prognostic characteristics. Median follow-up varied substantially between studies and between treatment groups. Treatment details were scarcely reported for the RP group, while details regarding RT dose, proportion receiving ADT and whether EBRT was performed in conjunction with BT were provided in most studies.

Table 6.1 General characteristics of included studies

Author	Year	Treatment Comparison	Data Source (study interval)	Median follow-up duration (RP/RT), months	RP (n)	RT (n)	Median (IQR) age (RP/RT), years	Median RT Dose (Gy)	Adjuvant Therapy
Yin	2019	EBRT+BT±ADT v RP	SEER 21 (2004, 2015)	58/87	59540	355	63.8/66.1	na	ADT: RT: "majority" RP: na
		EBRT±ADT v RP		/62		2638	/69.4		
Jayadevappa	2019	EBRT+BT±ADT v RP	SEER-Medicare (1996, 2003)	≥120	677	4141	71.7/73.1	na	not reported
		EBRT+ADT v RP				1478	/75.5		ADT: RT: 100% RP: na
Gunnarsson	2019	EBRT±BT±ADT v RP+RT+ADT	Kalmar County Hospital, Sweden (RP); The National Prostate Cancer Register (RT) (1995, 2010)	na	153	702	65/65	EBRT≤78 and EBRT+BT 20/50	ADT: RT: "preferred" RP: 100% aRT: 64%
Cano-Velasco	2019	EBRT+ADT v RP+ADT	Hospital General Universitario Gregorio Maran˜ón, Madrid, Spain (1996, 2008)	152/97	145	141	65/71	EBRT 74	ADT: RT: 100% RP: 100%
Tilki	2018	EBRT+BT+ADT v RP	Chicago Prostate Cancer Centre (RT); Martini-Klinik Prostate Cancer Center (RP) (1992, 2013)	58.7/66.1	372	80	66.4/70.3	EBRT 45 BT (I125 Pd103 and Cs131) 108/90/100	ADT: RT: 100% RP: 0% aRT: 0%
		v RP+ADT		46.4/	88		66.6/		RP: 100% aRT: 0%
		v RP+aRT		58.6/	49		66/		RP: 0% aRT: 100%
		v RP+ADT+aRT		57.4/	50		66.4/		RP: 100% aRT: 100%
Robinson	2018	EBRT±BT±ADT v RP	Swedish National Prostate Cancer Registry (1998, 2012)	75.6/70.8	3761	6462	63.1/67	na	not reported
Ciezki	2017	EBRT v RP	Cleveland Clinic (1996, 2012)	55.6/94.6	1308	734	62/68.5	(52%) at 78 (2 Gy fraction) & (48%) at 70 (2.5 Gy fraction)	ADT: RT: 93% RP: 19%
		EBRT+BT v RP		/48.9		515	/70		ADT: RT: 53% RP: 19%
Kishan	2017	EBRT v RP	Multi-institutional (12 centres) (2000, 2013)	50.4/61.2	639	734	61.2/68	EBRT 74.3	ADT: RT: 89.5% RP: 39% aRT: 34%
		EBRT+BT v RP		50.4/75.6		436	/68		ADT: RT: 92.4% RP: 39% aRT: 34%
Greenberg	2015	EBRT+ADT v RP	Anglia Cancer Network, UK (2000, 2010)	na/na	na	na	na/na	na	ADT: RT: 88.2% RP: na
Lee	2014	EBRT±ADT v RP	Severance Hospital, Seoul, Korea (1990, 2009)	74/85.5	251	125	67.5/68.6	EBRT (range) 74-79	not reported
Yamamoto	2014	EBRT±ADT v RP	Cancer Institute Hospital in Tokyo, Japan (1994, 2005)	93/85	112	119	67/72	EBRT 70	ADT: RT: 95.8% RP: 76.8%
Westover	2012	EBRT+BT+ADT v RP	Duke University (RP) (1988, 2008); Chicago Prostate Cancer Centre /21st Century Oncology Establishment (RT) (1991, 2005)	91.2/43.2	285	372	65/70	EBRT 45 BT I125/Pd103 108/90	ADT: RT: 100% RP: 0%

Kibel	2012	EBRT v RP	Barnes-Jewish Hospital and Cleveland Clinic (1995, 2005)	(59 to 72)/(70 to 74)	525	676	60.4/69.4	EBRT (median) 74 to 78 BT na	ADT: RT: 82% RP: na
		EBRT+BT v RP		(59 to 72)/(51 to 70)		33	/68.4		
Boorjian	2011	EBRT v RP	Mayo Clinic Prostatectomy Registry (RP) and the Fox Chase Cancer Centre (RT) (1988, 2004)	122.4/87.6	1238	344	66/69.3	EBRT 72	ADT: RT: 0% RP: 40.6%
		EBRT+ADT v RP		/72		265	/68.8		ADT: RT: 100% RP: 40.6%

Abbreviations: RP = radical prostatectomy; RT = radiation therapy; Gy = Gray; EBRT = external beam radiation therapy; BT = brachytherapy; SEER = Surveillance, Epidemiology, and End Results Program; ADT = androgen deprivation therapy; I125 = Iodine-125; Pd103 = Palladium-103; Cs131 = Cesium-131

6.3.2 Risk of bias assessment

The overall risk of bias was high for all studies (Table 6.2), as none examined all potential confounders and applied adjustments as appropriate. Most studies had a low risk of bias for the 'selection' section other than those comparing treatment groups from tertiary centers. The 'comparability' section varied due to variation in covariate control. All studies controlled for age, most studies provided adequate control for tumor characteristics (i.e., PSA, clinical stage, and Gleason scores) (13/14), while fewer studies controlled for comorbidities (7/14), demographic characteristics (4/14) and study center (7/14). Finally, most studies did not have a sufficient median follow-up, leading to a score of 2/3 for the 'outcome' section for 12/14 studies. There was no indication of publication bias. The Egger test for publication bias was not statistically significant (p = 0.22 for CSM and 0.92 for ACM; Figure 6.2).

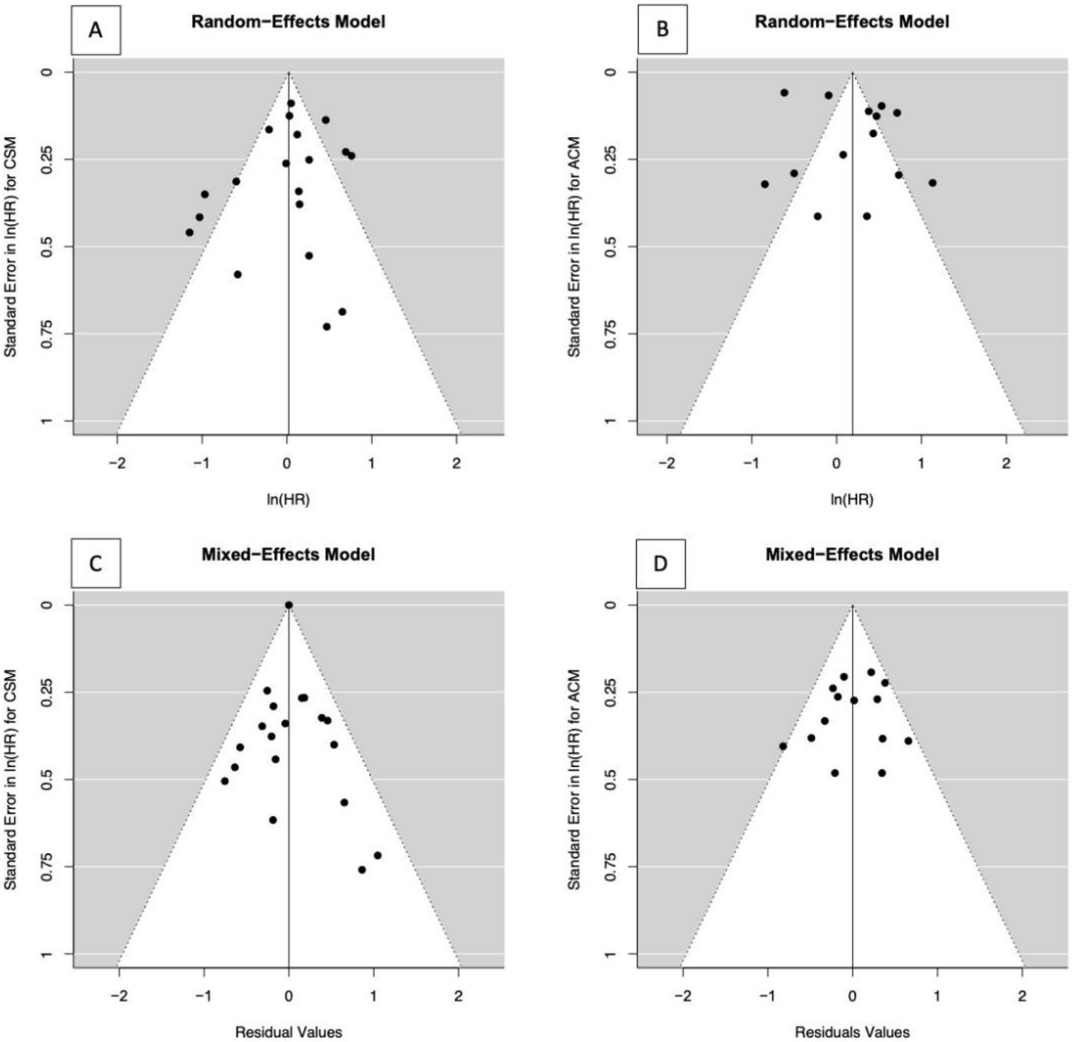


Figure 6.2 Funnel plots of meta-analysis for (a) prostate cancer-specific mortality, and (b) all-cause mortality using random-effects models. Mixed-effects models with moderators to reduce heterogeneity in effect estimates and improve symmetry in funnel plots for assessment of publication bias are shown in (c) for CSM (adjusted for receipt of BT) and (d) for all-cause mortality (adjusted for receipt of BT and age).

Abbreviations: HR = hazard ratio

Table 6.2 Modified Newcastle-Ottawa Scale for risk of bias assessment of studies included in the meta-analysis

Study Information	Selection				
Author (Year)	Representativeness of the exposed cohort (RT)	Representativeness of the non-exposed cohort (RP)	Ascertainment of exposure	Demonstration that outcome of interest was not present at start	Total
Yin (2019)	1	1	1	1	4
Jayadevappa (2019)	1	1	1	1	4
Gunnarsson (2019)	1	0.5	1	1	3.5
Cano-Velasco (2019)	0.5	0.5	1	1	3
Tilki (2018)	0.5	0.5	1	1	3
Robinson (2018)	1	1	1	1	4
Ciezki (2017)	0.5	0.5	1	1	3
Kishan (2017)	1	1	1	1	2
Greenberg (2015)	1	1	1	1	4
Lee (2014)	1	1	1	1	4
Yamamoto (2014)	0.5	0.5	1	1	3
Westover (2012)	0.5	0.5	1	1	3
Kibel (2012)	1	1	1	1	4
Boorjian (2011)	0.5	0.5	1	1	4

	Comparability								
Author (Year)	cT	GS	PSA	Age	Comorbidity	Demographic characteristic	Year of diagnosis or treatment	Study center (if applicable)	Total
Yin (2019)	1	1	1	1	0	1	1	1	3.5
Jayadevappa (2019)	0	0	0	1	1	1	1	1	2.5
Gunnarsson (2019)	0.5	0.5	0.5	1	0	1	0	0	1.75
Cano-Velasco (2019)	1	1	1	1	0	0	0	1	2.5
Tilki (2018)	1	1	1	1	1	1	0	0	3
Robinson (2018)	1	1	1	1	1	1	1	0	3.5
Ciezki (2017)	1	1	1	1	0	0	0	1	2.5
Kishan (2017)	1	1	1	1	0	0	0	1	2.5
Greenberg (2015)	1	1	1	1	0	0	0	0	2
Lee (2014)	1	1	1	1	1	1	0	1	3.5
Yamamoto (2014)	1	1	1	1	1	0	0	1	3
Westover (2012)	1	1	1	1	1	1	0	0	3
Kibel (2012)	1	1	1	1	1	0	1	0	3
Boorjian (2011)	1	1	1	1	0	1	0	0	2.5

	Outcome				
Author (Year)	Ascertainment of outcome	Adequate cohort follow-up intensity	Sufficient follow-up duration?	Total	Risk of Bias
Yin (2019)	1	1	0	2	High
Jayadevappa (2019)	1	1	0	2	High
Gunnarsson (2019)	1	1	0	2	High
Cano-Velasco (2019)	0	1	1	2	High
Tilki (2018)	1	1	0	2	High
Robinson (2018)	1	1	0	2	High
Ciezki (2017)	1	1	0	2	High
Kishan (2017)	1	1	0	2	High
Greenberg (2015)	1	1	0	2	High
Lee (2014)	1	0	0	1	High
Yamamoto (2014)	1	1	0	2	High
Westover (2012)	1	1	0	2	High
Kibel (2012)	1	1	0	2	High
Boorjian (2011)	1	1	1	3	High

Abbreviations: RP = radical prostatectomy; RT = radiation therapy; cT = clinical stage; GS = Gleason score; PSA = prostate-specific antigen

Table 6.3. Subgroup analyses assessing risk of prostate cancer-specific mortality and all-cause mortality following radiotherapy and surgery for prostate cancer

	Prostate cancer-specific mortality		Overall mortality	
	Adjusted HR (95% CI; p-value)	I^2	Adjusted HR (95% CI; p-value)	I^2
Radiotherapy modality				
EBRT±ADT	1.35 (1.10, 1.67; p=0.0048)	59%	1.54 (1.14, 2.09; p=0.0054)	91%
EBRT+BT±ADT	0.68 (0.48, 0.95; p=0.024)	47%	0.89 (0.55, 1.44; p=0.64)	86%
Treatment Era				
Before 2002	1.03 (0.75, 1.42; p=0.84)	69%	1.45 (0.92, 2.30; p=0.11)	97%
After 2002	1.00 (0.76, 1.30; p=0.98)	71%	1.00 (0.71, 1.42; p=0.99)	73%
Age				
≤67.4 years	1.04 (0.84, 1.29; p=0.72)	59%	1.58 (1.36, 1.85; p<0.0001)	39%
>67.4 years	0.97 (0.63, 1.47; p=0.87)	77%	0.94 (0.62, 1.43; p=0.78)	95%
Median follow-up				
≤67 months	1.04 (0.82, 1.32; p=0.73)	63%	1.06 (0.62, 1.81; p=0.83)	83%
>67 months	0.98 (0.67, 1.41; p=0.90)	74%	1.28 (0.83, 1.98; p=0.27)	97%
Geographic region				
United States	1.10 (0.87, 1.38; p=0.42)	71%	1.35 (0.90, 2.01; p=0.14)	97%
Other	0.81 (0.49, 1.32; p=0.40)	62%	1.01 (0.61, 1.66; p=0.98)	77%

Abbreviations: HR = hazard ratio; CI = confidence interval; EBRT = external beam radiation therapy; BT = brachytherapy; ADT = androgen deprivation therapy

6.3.4 Prostate cancer-specific mortality

Ten studies with 88,026 patients were included in the primary meta-analysis for CSM. The resulting adjusted hazard ratio [95% confidence interval] was 1.02 [0.84, 1.25] with substantial heterogeneity (I^2=69%) as shown in Figure 6.3a. Moderator analysis revealed a statistically significant effect by RT approach (p<0.0001). Specifically, CSM was increased among EBRT±ADT compared to RP (1.35 [1.10, 1.68]; p=0.0048), but decreased among EBRT+BT±ADT compared to RP (0.68 [0.48, 0.95]; p=0.024) (Table 6.3; Figure E6.1a). Including this variable in the moderator analysis was also associated with decreased, though still substantial, heterogeneity (I^2=59% and 47%, respectively). The remaining moderator analyses did not differ notably from the primary analysis.

6.3.5 All-cause mortality

Seven studies with 74,210 patients were included in the secondary meta-analysis for ACM. The resulting adjusted HR [95%CI] was 1.21 [0.89, 1.65] with substantial heterogeneity (I^2=95%) as shown in Figure 6.3b. Moderator analysis revealed a statistically significant effect by RT approach (p=0.03). Specifically, ACM was increased among EBRT±ADT compared to RP (1.54 (1.14, 2.09; p=0.0054)), but no statistically significant difference among those treated with EBRT+BT±ADT relative to RP (0.89 (0.55, 1.44; p=0.64)) (Table 6.3; Figure E6.1b). Both moderator analyses were associated with substantial heterogeneity (I^2=91% and 86%, respectively). Moderator analysis by median age also revealed a significant effect (p=0.0001). A statistically significantly higher rate of ACM among RT relative to RP was observed among studies with younger patient groups (1.58 [1.36, 1.85]; p<0.0001; I^2=39%) compared to those with older patient groups (0.94 [0.62, 1.43]; p=0.78; I^2=95%) (Table 6.3; Fig E6.2b). Effect estimates also varied from the main analysis among moderator analyses of studies assessing men diagnosed and/or treated before 2002 (1.45 [0.92, 2.30]), but not after 2002 (1.00 [0.71, 1.42]) median follow-up of >67 months (1.28 [0.83, 1.98]), but not ≤67 months (1.06 (0.62, 1.81), and studies performed in the United States (1.35 [0.90, 2.01]) versus other geographic locations (1.01 [0.61, 1.66]).

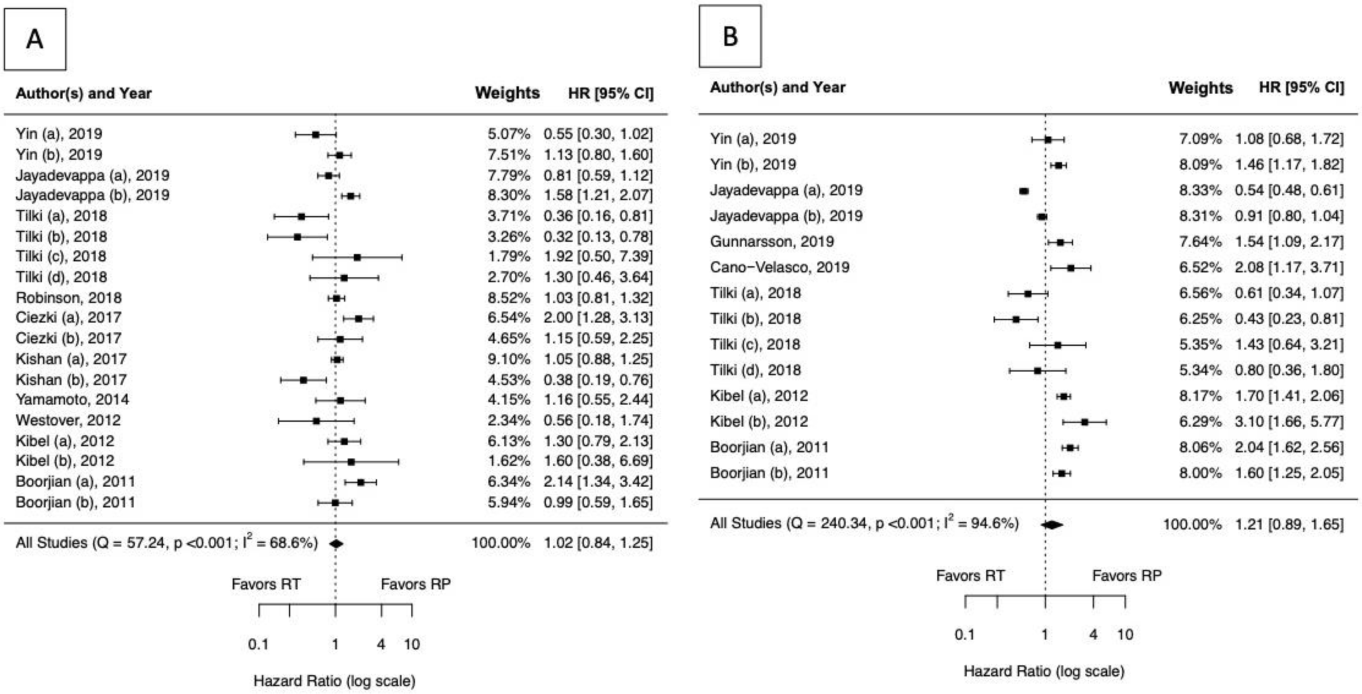


Figure 6.3 Forest plot assessing the risk of (a) prostate cancer-specific mortality and (b) all-cause mortality following radiotherapy and surgery for prostate cancer

Abbreviations: HR = hazard ratio; CI = confidence interval; RT = radiation therapy; RP = radical prostatectomy

6.4 Discussion

Our aggregate effect estimates for adjusted CSM showed no statistically significant differences between RP and RT for high-risk non-metastatic PCa patients. Moderator analysis revealed a significant increased incidence of CSM among men treated with EBRT±ADT relative to the RP group and a decreased incidence of CSM among men treated with EBRT+BT±ADT relative to the RP group. This is consistent with results from the ASCENDE-RT trial wherein an increased incidence of biochemical failure was found among men diagnosed with intermediate- and high-risk non-metastatic PCa and treated with dose-escalation RT protocols using EBRT alone compared with those using combination EBRT+BT (HR [95%CI]: 2.04 [1.25, 3.33]).(167) Remaining moderator analyses did not differ from the primary analysis.

Multiple reports indicate that, since the early 2000's, the use of BT boost in high-risk patients has declined in use in the United States,(201) and other geographic regions.(202) Interestingly, however, the use of prostate BT boost has increased since the early 2000's in certain European centers and in Canada.(203,204) This discrepancy may be attributable to differences in resident exposure in providing sufficient training opportunities given the steep learning curve associated with administering BT,(205–207) and unfavorable reimbursement relative to EBRT in the United States relative to publicly funded healthcare systems.(203,208) Given the CSM benefit associated with BT boost among high-risk patients reported in RCTs and estimated here, we encourage investment in overcoming the aforementioned obstacles through increasing resident exposure, and improving reimbursement models to encourage use BT boost.

The HR comparing relative incidence of CSM between EBRT±ADT and RP groups was smaller compared to that in a previous meta-analysis performed in 2016 (1.35 [1.10, 1.68] versus 1.83 [1.51–2.22]).(118) These differences might be explained by more recent changes in treatment approaches including the increasing use of dose-escalation protocols and adjuvant ADT paired with RT,(201,202) which have both demonstrated improvements in oncological outcomes, though only the addition neoadjuvant ADT to RT has demonstrated improvements in CSM.(91,186,202)

The analysis of relative ACM between RT and RP also revealed no statistically significant difference between the treatment groups. However, moderator analysis revealed a statistically significantly increased incidence of ACM among the EBRT±ADT relative to the RP group, while there was an insignificant decrease in ACM between the EBRT+BT±ADT and RP groups. In addition to the CSM benefit afforded through RP and EBRT+BT±ADT relative to EBRT±ADT, differences in cardiopulmonary health requirements before undergoing general anesthetic that is required for RP and BT, and lack of control for comorbidities in many of the included studies might contribute to the observed differences. Studies conducted among younger age groups demonstrated increased incidence of ACM in the RT relative to the RP group. Finally, a tendency toward increased incidence of ACM in the RT relative to the RP group was also noted when restricting analyses among studies conducted in earlier treatment eras, longer follow-up periods and among studies conducted only in the United States. However, this is likely explained by the greater proportion of comparisons with RP involving EBRT±ADT instead of EBRT+BT±ADT among studies conducted in earlier treatments eras, which were associated with longer follow-up periods and were mostly performed in the United States versus other geographic locations.

Overall, the risk of bias was deemed high for all studies due to the partial control of confounding variables. This stands in contrast with a previous meta-analysis performed by Wallis et al who found a low to moderate risk of bias for all studies included in their meta-analysis comparing the rate of ACM and CSM between patients who underwent RT and RP. Interestingly, four studies used in both analyses indicated perfect comparability between RT and RP groups by Wallis et al., yet some of these studies did not control for study center,(92,209,210) year of diagnosis,(100,209,210) or demographic characteristics.(92) Since patients undergoing RT are more likely to be older, have poorer prognostic characteristics, and sociodemographic characteristics that are associated with poorer CSM and ACM,(211–213) we anticipate the influence of these unaccounted-for biases to overestimate CSM and ACM in the RT group relative to the RP group. However, the discrepancy in such baseline characteristics appears more prominent among those undergoing EBRT±ADT rather than EBRT+BT±ADT wherein patients are more similar to those undergoing RP.(211,212) As such, collecting information on these variables and properly controlling for them is crucial when estimating relative treatment effects between groups to more accurately inform treatment decisions.

Our study has certain limitations. There was a high level of heterogeneity in effect estimates. This was substantially reduced through moderator analyses comparing RP with EBRT±ADT and EBRT+BT±ADT, and among comparisons involving younger populations, though heterogeneity still remained high and was unaccounted for through additional moderator analyses. Unfortunately, information surrounding treatment details such as RT dose, type of EBRT (i.e., 3D conformal, IMRT, etc.), use of adjunct therapies and surgeon experience, which might account for a large proportion of this heterogeneity, was missing in many of the studies.

Given the high risk of bias in all studies, the aggregated effect estimates provided in this study are limited in informing clinical decisions. In light of this and considering the relatively small difference in CSM between treatment approaches, other factors such as patient preferences, patient health (i.e., comorbidities), and treatment factors (e.g., operative risk and prostate volume for BT) should be considered when forming treatment decisions. This should occur through a shared decision-making process, involving the patient and providing urologist and radiation oncologist to optimize satisfaction in patient outcomes.

6.5 Conclusions

We identified no statistically significant difference in the rate of CSM between patients diagnosed with high-risk non-metastatic PCa and treated with RP relative to RT. However, there was significant subgroup effect with the use of EBRT+BT±ADT, highlighting the necessity of differentiating RT with or without BT in future comparative effectiveness studies. The high risk of bias in all studies reviewed emphasizes the need for better control of all potentially confounding variables to provide higher quality non-randomized evidence. This is exceedingly important when RCTs are unlikely to be feasible in this patient population.(110,111)

6.6 Transition to Chapter Seven

In this chapter, we found no difference in the rate of CSM between men diagnosed with high-risk non-metastatic PCa who were initially treated with RT or RP. Although differences in outcomes seemed to depend on RT type and age. Moreover, the risk of bias for all included studies was high. This was mainly due to inadequate control for all observable confounding variables, which might have been due to limitations in data collection and/or inadequate/inappropriate management of confounding. The results from this systematic review and meta-analysis provide context for the following chapter, which compared MPFS, a validated surrogate for CSM, between men diagnosed with unfavorable-risk non-metastatic PCa who were initially treated with RT or RP.

Chapter 7: Characterization of Outcomes of Multimodal Approaches in Unfavorable-Risk Prostate Cancer

Authors: David Guy (MD/PhD candidate)[1], Rachel M. Glicksman (MSc, MD)[2], Andrew Loblaw (MD, MSc)[2], Joseph Chin (MD)[1], George Rodrigues (MD/PhD)[1]

Affiliations:

[1]London Health Sciences Centre, London, ON, Canada

[2]Sunnybrook Health Sciences Centre, Toronto, ON, Canada

Abstract

Background

Identifying the optimal management of unfavorable-risk non-metastatic prostate cancer (PCa) is an important public health concern given the large burden of this disease. We compared the rate of metastatic progression-free survival among men diagnosed with unfavorable-risk non-metastatic PCa and treated with radiation therapy (RT) or radical prostatectomy (RP).

Methods

We reviewed medical records obtained from two academic centers in Toronto and London, Ontario, Canada of men diagnosed with unfavorable-risk non-metastatic PCa and treated with primary RT or RP. Patients were matched on prognostic covariates using two matching techniques. Multivariable Cox proportional hazards models were used to estimate the hazard ratios and confidence intervals for metastatic progression-free survival between groups.

Results

A total of 164 and 169 men were included in the RT and RP treatment groups, respectively. After a median follow-up of 83.9 and 96.9 months of men in the RT and RP groups, respectively, no difference in the rate of metastatic progression-free survival was found between groups (unadjusted HR [95%CI]: 1.29 [0.74, 2.26]; p=0.37 and adjusted: 1.16 [0.63, 2.13]; p=0.64). Effect estimates did not change notably after matching.

Conclusion

The rate of metastatic progression-free survival did not differ between men diagnosed with unfavorable-risk non-metastatic PCa who were treated with either RT or RP.

7.1 Background and Rationale

Prostate cancer (PCa) was the second leading cancer diagnosis and fifth leading cause of cancer death globally in 2018.(3) Unfavorable-risk non-metastatic disease, including high-intermediate, high and extremely high-risk disease,(19) accounts for approximately one third of all PCa diagnoses, but a disproportionate amount of morbidity and mortality.(180,202,214) Optimizing the efficacy and safety of treatments for this disease is thus a major public health concern. Common definitive management options include surgical resection of the prostate (radical prostatectomy [RP]) and irradiation of the prostate through radiation therapy (RT).(21) Compared to watchful waiting, definitive management with RT or RP among men diagnosed with localized PCa has been shown in RCTs to decrease the rate of metastatic progression, PCa-specific mortality and overall mortality.(87,215)

The selective use of adjuvant and salvage therapies alongside definitive management has also been shown to further improve outcomes. For instance, the use of adjuvant RT in the context of adverse pathological findings post-RP has been found to decrease rates of biochemical recurrence and local relapse.(216–218) The addition of androgen deprivation therapy (ADT) to RT post-RP has been shown to reduce rates of metastatic progression and PCa-specific mortality among those with adverse pathological features relative to RT alone. For patients with unfavorable-risk non-metastatic PCa who undergo RT, decreased risk of metastatic progression and PCa-specific death has been observed with the use of adjuvant ADT.(105,219) Results from the ASCENDE-RT trial have also shown improvements in biochemical control from combination external beam RT (EBRT) with BT compared to EBRT alone.(167) Finally, RT dose-escalation protocols have demonstrated improvements in biochemical control such that traditional regimens of <70 Gy are no longer standard.(21)

Despite the progress made in the selection and sequencing of adjuvant and salvage therapies and refinements in RT approaches, optimal local control has not been adequately evaluated through a RCT for this patient population. In turn, clinicians and patients rely on evidence generated from observational data to guide treatment decisions, which have limitations due to confounding and comparisons involving outdated treatment regimens. For example, candidates for RP compared to RT generally have less aggressive tumor characteristics, are younger and with fewer comorbidities.(114) Vast disparities in these baseline characteristics make the assumptions of positivity required for valid estimation of treatment effects through regression modeling questionable.(183) As such, identifying patients with similar baseline characteristics who are treated with RP and RT and who have undergone more contemporary forms of treatment is necessary to improve the internal and external validity of evidence on this topic.

In this study, we compared the rate of metastatic-progression between men diagnosed with unfavorable-risk non-metastatic PCa and treated with RT and RP as definitive local therapies. Issues of non-positivity are mitigated through the use of data obtained from a multidisciplinary clinic wherein RT patients were also eligible for RP. Furthermore, we take advantage of data preprocessing techniques that have been developed to improve the degree of comparability between treatment groups obtained from observational data.(138,139) Established techniques include propensity score matching and coarsened exact matching.(141,159)

7.2 Methodology

7.2.1 Data source

Ethics approval was provided from both institutional review boards at Sunnybrook Health Sciences Centre and London Health Science Centre (LHSC). We identified the records of men diagnosed between 2007-2012 with high-intermediate to extremely high-risk non-metastatic PCa in the multi-disciplinary diagnostic assessment program in the Gale and Graham Wright Prostate Centre (GGWPC) at North York General Hospital in Toronto, Ontario, Canada. Patients in the RT group included those who had undergone EBRT with or without brachytherapy boost (BT) (low-or high dose rate) and with or without ADT. Patients in the RP group included those who had undergone RP as their primary treatment modality. Due to limited RP observations from the GGWPC, we also included men diagnosed between 2007-2012 with high-intermediate to extremely high-risk non-metastatic PCa who were treated with primary RP at LHSC in London, Ontario, Canada.

7.2.2 Data Collection

We reviewed electronic medical records from identified patients. Information regarding patient age at diagnosis, biopsy date, prognostic factors at diagnosis (pre-biopsy PSA level, TNM stage, Gleason Score (GS), and percentage of biopsy cores containing tumor), initial treatment decision, treatment date, and treatment details were obtained. Patients were eligible for the study if they met the following criteria:

1. Diagnosed with high-intermediate, high or extremely high-risk PCa according to the Prostate Cancer Risk Stratification (ProCaRS) database.(19)
2. No evidence of regional or metastatic disease at the time of diagnosis and staging
3. Consulting radiation oncologist offered the RT
4. Consulting urologist offered patient RP
5. Diagnosed between July 2007 and December 2012
6. Had at least one year of follow-up

ProCaRS high intermediate-risk disease (HIR) is defined as having a GS=7 and one or both of PSA 10-20 ng/mL and/or bilateral clinical disease. High-risk disease (HR) is defined as having a PSA > 20 ng/mL, cT stage = 3-4 or GS = 8-10, while extremely high-risk disease (EHR) is defined as having a PSA > 30 ng/mL or high-volume disease, defined as > 87.5% biopsy core involvement. Information on patient comorbidities, socioeconomic and demographic characteristics was not available for the majority of patients, and consequently, was not used in analysis.

7.2.3 Outcomes

We analyzed the rate of metastatic progression-free survival between treatment groups. Metastatic progression was confirmed through imaging reports. Progression-free survival time was defined as the interval between the date of PCa treatment initiation (i.e., RT or RP) and the date of metastatic progression or last documented encounter with their providing oncologist. Patients who were event-free at the end of the study period were censored at that point and contributed the time interval from their date of treatment to the end of the study in the survival analysis.

7.2.4 Covariate selection

We explored the potential for confounding through examining differences between treatment groups in distributions of baseline covariates that have demonstrated a prognostic role in relation to the rate of treatment failure in previous literature.(19) Covariates included tumor characteristics (i.e. pre-biopsy PSA level, clinical T (cT) stage, and GS). Age was not included as a covariate since it has not demonstrated a predictable association with the outcome examined in previous research,(176,177) nor did it demonstrate an association in either of the comparisons performed here (adjusted for treatment received: HR [95%CI]: 0.98 [0.93, 1.03]; p-value=0.50 and HR [95%CI]: 1.00 [0.98, 1.02]; p-value=0.99 for datasets used in comparison one and two, respectively). Further, age was strongly associated with treatment choice, which would bias effect estimates if adjusted for.(168) An insufficient number of patients had specific information related to the percentage of tumor containing biopsy cores. Thus, this variable is only reported in the descriptive statistics and not used for adjustment.

7.2.5 Propensity score matching

The propensity score model was a logistic model with prognostic characteristics as independent variables and type of treatment received as a binary dependent variable.(168) We explored the possibility of interactions and non-linearity for baseline covariates when fitting the propensity score model.(152,178) Locally weighted scatterplot smoothers were used to assess for departures from linearity in the relationship between baseline PSA and the logit of the probability of receiving RP. Improvements in the model fit were assessed using the likelihood ratio test and pseudo-R^2. DFBETA statistics did not reveal any outliers. Model one involved baseline PSA as a linear term and cT stage (1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4a, and 4b) and GS (6, 7, 8, 9, 10) as categorical variables. A restricted cubic spline with four knots was found to improve model fit for baseline PSA, and thus was included in model two. The MatchIt package in R was used to match participants with a ratio of 1:1 between treatment groups.(146) Although chapters four and five used many-to-one and one-to-many matching ratios, the number of participants in those comparisons varied substantially between treatment groups. Here, the number of participants in each treatment group is approximately the same so a 1:1 matching ratio was used. We explored a range of caliper widths between 1.0 and 0.01 standard deviations of the logit of the propensity score. Nearest-neighbour matching was used without replacement.(179)

7.2.6 Coarsened exact matching

Patients were matched on progressively coarsened covariates (i.e., GS, cT stage, and baseline PSA). GS was first dichotomized into ≤7 or 8-10 and then using each category (6, 7, 8, 9, and 10). Clinical T stage was first dichotomized into ≤2 and 3-4 and then using each category (1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4a and 4b). Due to missing information, categorizing GS in more detail (i.e., GS=7 as 3+4 vs 4+3) was not feasible without substantial data loss. Progressive coarsening for PSA involved cut points from 0 ng/ml to 300 ng/ml first at 20 and 100 ng/ml, with additional cut points at 30 and 50 ng/ml and further at 6 and 10 ng/ml. Coarsening ranges are presented in S Table 1.

7.2.7 Balance diagnostics

Many balance diagnostics exist and have been rigorously assessed using various empirical and simulation datasets that represent a broad range of data characteristics. We chose four balance measures that considered different data characteristics in order to monitor improvements in balance when further restricting matching strategies (i.e., using finer ranges for covariates in coarsened exact matching and smaller caliper widths in propensity score matching). This enabled systematic identification of matching strategies that optimized balance in the multivariable distribution of baseline covariates as a function of data retention. This process is shown in Figures E1&2 for both matching strategies.

7.2.8 Descriptive statistics and multivariable regression analysis

All statistical analyses were performed using RStudio version 3.6.0.(169) Descriptive statistics were calculated for each treatment group before and after matching. The mean and standard deviation are presented for continuous variables and proportions for categorical variables. MPFS was plotted for each group using Kaplan-Meier curves and the log-rank test was used to calculate if MPFS significantly differed between groups. Cox proportional-hazards regression analyses were performed for estimating the effect of treatment group on the hazard of metastatic progression using the Survival package.(181) The proportional-hazards assumption was confirmed using log-minus-log survival plots and scaled Schoenfeld residuals.

Improvements in model fit were examined through comparing the model log likelihoods after incorporating interaction terms and higher order terms and transformations for continuous covariates. Examination of a plot of DBETA statistics did not identify any influential observations. Hazard ratios and 95% confidence intervals were calculated from unmatched data both without and with adjustment for baseline PSA, cT stage, and GS as well as interactions between baseline PSA and GS and baseline PSA and cT stage. For matched data, we employed Cox models clustered by the matched sets, using robust variance estimators to generate confidence intervals.(181,182) Both unadjusted and adjusted estimates were calculated.

7.3 Results

Descriptive characteristics are displayed in Table 1. At diagnosis, men treated with RT relative to RP were older, had higher PSA levels, a greater percentage of tumor containing biopsy cores, less advanced tumor staging, and comparable GS. A greater proportion of men treated with RT presented with high-intermediate risk, and a smaller proportion of high-risk disease than those treated with RP, while similar proportions of each treatment group were considered extremely high-risk disease.

Sixty-seven (40.9%) men treated with RT received neoadjuvant ADT compared to only 40 (23.7%) men treated with RP. Twenty-eight (17.1%) of men treated with RT received adjuvant ADT, while over 43% of the RP group did. Local and systemic salvage therapy was initiated among 30.8% and 28.4%, respectively, among men treated with RP, while only one man (0.6%) treated with RT received systemic salvage therapy aside from adjuvant ADT. The most common form of RT was EBRT without BT boost for which 84.6% of men received. The median dose for men receiving EBRT without BT boost was 78 Gy. Of the 19 (11.2%) of men who received BT boost, the vast majority received HDR, while only one (0.6%) man received LDR. Finally, two men (1.2%) treated with RT received SBRT boost. The median dose for this group was 113.57 Gy.

Descriptive characteristics after matching are displayed in Table 2. After propensity score matching, 117 subjects were retained in each group, while coarsened exact matching led to retention of 138 and 141 patients from the RT and RP groups, respectively. Both matching strategies led to balance in the multivariable covariate structure according to conventional thresholds for balance (i.e., SMD<0.1 and variance ratio between 0.92 to 1.08).(152) Mean percent of tumor containing biopsy cores remained imbalanced.

Kaplan-Meier curves showing the probability of metastatic progression-free survival over time stratified by treatment group are shown in Figure 2. Overall, both treatment groups demonstrated similar rates of metastatic progression-free survival over time. Unadjusted and adjusted hazards ratios and 95% confidence intervals are presented in Table 3. The unadjusted HR [95%CI] estimated before matching was 1.29 [0.74, 2.26], which attenuated to 1.16 [0.63, 2.13] upon adjustment. The HR [95%CI] was 1.06 [0.50, 2.26] after propensity score matching and 1.55 [0.60, 3.98] after coarsened exact matching. Given the small number of events, variation in the effect estimates might be attributable to random error. Moreover, changes in the effect estimates are expected with sample changes due to matching.

Table 7.1 Descriptive patient and treatment characteristics

Treatment Group	Radiation Therapy	Radical Prostatectomy	|SMD|	Variance Ratio	GGWPC RP	LHSC RP	|SMD|	Variance Ratio
	N=164	N=169			N=75	N=94		
Follow-up time (months) Median (Q1, Q3)	83.9 (58.8, 106.3)	96.9 (67.8, 118.4)			98.8 (69.3, 124.0)	94.5 (65.2, 113.1)		
Metastatic Events n (%)	20 (12.2)	33 (19.5)	0.20		11 (14.7)	22 (23.4)	0.22	
Age (years) at Diagnosis, Mean (SD)	72.5 (7.5)	62.6 (6.4)	1.42	0.74	62.1 (6.7)	63.6 (5.9)	0.15	0.83
Missing n (%)	3 (1.8)	0 (0)			0 (0)	0 (0)		
Baseline PSA (ng/ml) Mean (SD)	19.7 (21.9)	16.4 (15.3)	0.18	0.49	14.4 (10.9)	17.9 (18.0)	0.24	2.70
Missing n (%)	1 (0.6)	0 (0)			0 (0)	0 (0)		
Clinical T Stage								
1	75 (48.4)	65 (40.1)	0.17		43 (62.3)	22 (23.7)	0.85	
2	67 (43.2)	57 (35.2)	0.17		20 (29.0)	37 (39.8)	0.23	
3	13 (8.4)	37 (22.8)	0.41		6 (8.7)	31 (33.3)	0.63	
4	0 (0)	3 (1.9)	0.19		0 (0)	3 (3.2)	0.26	
Missing n (%)	9 (5.5)	7 (4.1)			6 (8.0)	1 (1.1)		
Gleason Score								
≤6	4 (2.4)	5 (3.0)	0.03		2 (2.9)	3 (3.2)	0.03	
7	108 (65.9)	103 (61.0)	0.10		46 (66.7)	55 (58.5)	0.13	
8	20 (12.2)	31 (18.3)	0.17		7 (10.1)	22 (23.4)	0.30	
9	32 (19.5)	28 (16.6)	0.08		14 (20.3)	13 (13.8)	0.17	
10	0 (0)	2 (1.2)	0.15		0 (0)	1 (1.1)	0.02	
Missing n (%)	0 (0)	0 (0)			0 (0)	0 (0)		
(%) Core Positivity Mean (SD)	56.4 (27.8)	51.0 (25)	0.20	0.81	50.5 (24.2)	51.7 (25.6)	0.06	1.14
≥50%	97 (59.2)	93 (59.2)	0.00		39 (56.5)	48 (63.2)	0.18	
Missing n (%)	0 (0)	12 (7.1)			0 (0)	12 (12.8)		
ProCaRS Risk-Groups								
High-Intermediate	72 (47.4)	58 (40.6)	0.14		29 (45.7)	29 (38.7)	0.08	
High	48 (31.6)	59 (41.3)	0.20		27 (39.7)	32 (42.7)	0.06	
Extremely High	32 (21.1)	26 (18.2)	0.07		12 (17.7)	14 (18.7)	0.03	
Missing n (%)	12 (7.3)	26 (18.4)			7 (10.4)	19 (20.2)		
Treatment Characteristics								
Radiotherapy Patients								
EQD2 for EBRT Median (min, max)	78 (70, 108.5)							
EQD2 for EBRT+BT Median (min, max)	113.57 (113.1, 116.7)							
ADT n (%)	95 (57.9)							
Initial ADT n (%)	67 (40.9)							
Duration ADT Median (min, max)	22.1 (2.5, 43.3)							
Brachytherapy boost type								
Low-dose rate	1 (0.6)							
High-dose rate	18 (10.6)							
Prostatectomy patients								

Neoadjuvant systemic therapy		40 (23.7)			2 (2.7)	38 (40.4)		
Adjuvant radiotherapy		57 (33.7)			9 (12)	48 (51.1)		
Adjuvant systemic		32 (18.9)			8 (10.7)	24 (25.5)		
All patients								
Local salvage	0 (0)	52 (30.8)			36 (48.0)	16 (17.0)		
Systemic salvage	1 (0.6)	48 (28.4)			20 (26.7)	28 (29.8)		

Abbreviations: |SMD| = absolute standardized mean difference; LHSC = London Health Sciences Centre; RP = Radical Prostatectomy; ADT = androgen deprivation therapy; RT = radiation therapy; EQD2 = Equivalent dose in 2-Gy fractions

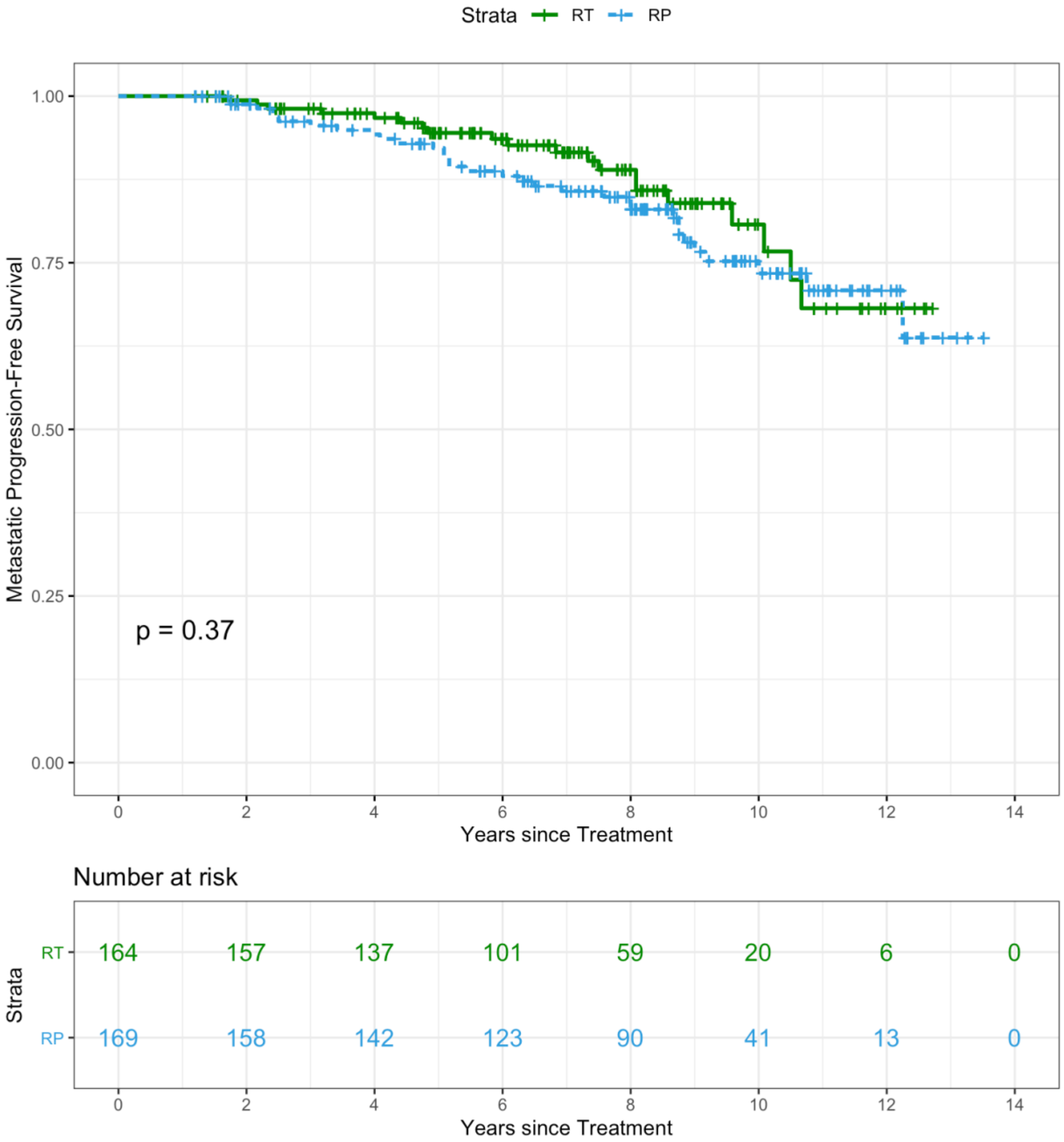


Figure 7.2 Kaplan-Meier curves showing the probability of metastatic progression-free survival over time stratified by treatment group.

Table 7.2 Descriptive patient characteristics in matched samples

	Propensity Score Matching				Coarsened Exact Matching			
Treatment Group	Radiation Therapy	Radical Prostatectomy	|SMD|	Variance Ratio	Radiation Therapy	Radical Prostatectomy	|SMD|	Variance Ratio
	N=117	N=117			N=138	N=141		
Age at Diagnosis Mean (SD)	72.3 (7.4)	62.1 (6.5)	1.46	1.28	72.4 (7.3)	62.6 (6.3)	1.45	1.39
Missing n (%)	0 (0)	0 (0)						
Baseline PSA (ng/ml) Mean (SD)	17.3 (13.2)	17.1 (13.6)	0.02	1.07	16.4 (10.5)	16.0 (11.9)	0.06	1.08
Missing n (%)	0 (0)	0 (0)			0 (0)	0 (0)		
Clinical T Stage								
1	58 (49.6)	63 (53.9)	0.09		71 (51.5)	73 (51.5)	0	
2	47 (40.2)	44 (37.6)	0.05		58 (42.0)	59 (42.0)	0	
3	12 (10.3)	10 (8.6)	0.06		9 (6.5)	9 (6.5)	0	
4	0 (0)	0 (0)	0		0 (0)	0 (0)	0	
Missing n (%)	0 (0)	0 (0)						
Gleason Score								
≤6	3 (2.6)	3 (2.6)	0		2 (1.4)	2 (1.4)	0	
7	80 (68.4)	80 (68.4)	0		98 (71.0)	100 (71.0)	0	
8	17 (14.5)	15 (12.8)	0.05		15 (10.9)	15 (10.9)	0	
9	17 (14.5)	19 (16.2)	0.05		23 (16.7)	24 (16.7)	0	
10	0 (0)	0 (0)	0		0 (0)	0 (0)	0	
Missing n (%)	0 (0)	0 (0)						
(%) Core Positivity Mean (SD)	57.7 (27.8)	49.2 (24.2)	0.33	1.33	56.0 (27.0)	49.0 (22.0)	0.26	1.51
≥50%	75 (64.1)	63 (54.8)	0.19		83 (60.1)	80 (58.6)	0.03	
Missing n (%)	0 (0)	2 (1.7)			0 (0)	9 (6.4)		
ProCaRS Risk-Groups								
High-Intermediate	57 (50.9)	54 (48.2)	0.05		72 (54.5)	76 (57.9)	0.07	
High	35 (31.3)	40 (35.7)	0.09		41 (31.1)	41 (31.2)	0.00	
Extremely High	20 (17.9)	18 (16.1)	0.05		19 (14.4)	14 (10.9)	<0.10	
Missing n (%)	5 (4.3)	5 (4.3)			6 (4.3)	9 (6.4)		

Abbreviations: |SMD| = absolute standardized mean difference; LHSC = London Health Sciences Centre; RP = Radical Prostatectomy; ADT = androgen deprivation therapy; RT = radiation therapy

Table 7.3 Hazards ratios and confidence intervals for metastatic progression in RP relative to RT

	Unadjusted			*Adjusted		
	Hazard ratio	95% CI	p-value	Hazard ratio	95% CI	p-value
Unmatched	1.29	0.74, 2.26	0.37	1.16	0.63, 2.13	0.64
PSM	1.01	0.50, 2.05	0.64	1.06	0.50, 2.26	0.87
CEM	1.32	0.62, 2.82	0.47	1.55	0.60, 3.98	0.37

Abbreviations: RP = radical prostatectomy; RT = radiation therapy; PSM = propensity score matched; CEM = coarsened exact matched; CI = confidence interval

*Adjusted model includes baseline PSA, clinical T stage, and Gleason score as continuous linear variables with interactions between baseline PSA and clinical T stage and baseline PSA and Gleason score

7.4 Discussion

We compared the rate of metastatic progression between men diagnosed with unfavorable-risk non-metastatic PCa who were treated with RT or RP. No significant difference was observed in the rate of metastatic progression between treatment groups. Previous reports have demonstrated reduced rates of metastatic progression among men diagnosed with unfavorable-risk non-metastatic PCa who were initially treated with RT relative to RP.(220–222) This includes a very similar study of men diagnosed with high- and very high-risk PCa in a multidisciplinary clinic and treated with RT or RP wherein rates of distant metastasis were elevated among those treated with RP relative to those treated with RT (HR [95%CI]: 2.5 [0.8, 7.8]; p=0.11); however, this analysis might not have been adequately powered, given the point estimate, as only 35 events were observed.(222) These findings are consistent with another comparison of rates of metastatic progression among men diagnosed with unfavorable-risk PCa and treated with EBRT+ADT relative to RP.(221) The attenuated effect estimate in our study relative to previous studies might be attributable to non-consistency in ADT administration, as only 57.9% of men in our study received any neoadjuvant or adjuvant ADT whereas ADT use in the aforementioned studies among those treated with EBRT was approximately 100%.

Substantial variation in oncological outcomes has also been observed with the use of combination EBRT+BT relative to EBRT alone. For instance, lower rates of metastatic progression have been found among men diagnosed with unfavorable-risk non-metastatic PCa and treated with combination EBRT+BT relative to RP (0.27 [0.17, 0.43]). This finding is consistent with other reports demonstrating improved PCa-specific survival among combination EBRT+BT relative to EBRT alone.(211,212) Only 19 (11.2%) men in our cohort received combination EBRT+BT so subgroup comparisons were not feasible. As such, our findings likely represent a combination of two different effect estimates for combination EBRT+BT and EBRT alone.

The rate of salvage therapy post-RP was much higher than that post-RT. Local and systemic salvage therapy were administered to approximately 30% of men post-RP, while only one man treated with RT received salvage therapy. These observations are consistent with previous investigations by Kishan et al who found similar rates of local and systemic salvage therapy post-RP,(220) and Markovina et al who found salvage much more common post-RP than post-RT.(221) This can, in part, be explained by the increased rates of biochemical-failure among men diagnosed with unfavorable-risk non-metastatic PCa who undergo RP relative to RT. Administration of salvage therapy is also less likely among men undergoing RT. Since men who undergo RT are generally older, with poorer health and lower life expectancies, the benefits of salvage therapy are limited, while side effects from it can adversely affect quality of life. Moreover, the rate of salvage therapy post-RT might be hampered due to limited awareness and availability of modalities such as cryotherapy and high intensity focused ultra-sound.

The median follow-up time was 13 months shorter in the RT relative to the RP group. This might be explained by increased rates of competing events that would increase losses to follow-up. To explain, those receiving RT were also approximately 10 years older than those receiving RP and likely had increased comorbidities. During later phases of PCa management, competing illnesses that decrease life-expectancy may take priority and patients might stop attending follow-up appointments for their PCa if it poses less threat to their survival. Unfortunately, data from other clinics indicating development of metastasis was not available, preventing competing risks analyses. This missing data issue can bias effect estimates either through limiting the contribution of event-free follow-up time or limiting the identification metastatic progression.

Other missing data issues involved 17 subjects for clinical tumor stage and 23 subjects for percent core positivity, which prevented these observations from contributing to regression and post-matching effect estimation. However, due to the limited number of subjects missing information on these variables, inferences regarding the distribution of missing data is limited and data imputation methods, such as multiple imputation, are unlikely to lead to notably different effect estimates or provide additional information.

The strengths of our study include the comparison of men treated with RT who were also eligible for RP thereby mitigating violations of positivity required for the conduct of regression analysis.(183) Moreover, systematic identification of comparable treatment groups through propensity score matching and coarsened exact matching has potential to reduce reliance on model specification,(138) thereby improving the robustness of confounding control. Finally, since men were diagnosed between 2007 and 2012 from two large academic centers, treatment approaches are expected to be more consistent with contemporary treatment approaches.

The findings of this study are subject to limitations. First, the proportion of men treated with RT who received ADT was much lower than other similar investigations. Since ADT has been shown to decrease the rate of metastatic progression, the rates observed among men treated with RT in our cohort may exceed those achievable through the current standard of care, which recommends RT and ADT for men diagnosed with unfavorable-risk PCa.(90) In addition, the series of men treated at LHSC may not have been comparable to men treated at GGWPC so there is potential for confounding of effect estimates by treatment center. Finally, due to data limitations, percent of tumor containing biopsy cores, comorbidities, and socioeconomic and demographic characteristics could not be controlled for, potentially biasing effect estimates. Based on the risk of bias assessment tool used in the chapter seven, this study would be given a high-risk of bias.

7.5 Conclusion

The results from our study support findings from previous analyses that more contemporary forms of treatment involving RT as an initial strategy may be, at least, comparable to those involving initial RP for men diagnosed with unfavorable-risk PCa. Furthermore, the decreased use of salvage therapies among men treated with RT relative to RP may have benefits with regard to fewer side effects in the long-term management of this patient population. Given the aforementioned limitations of this study, the results provided here must be interpreted with caution. However, since evidence from RCTs is unlikely to surface within the next decade,(109) the value of observational research holds great value in informing treatment decisions.

8.0 Integrated Discussion

In this thesis, we developed a systematic approach to developing and evaluating matching strategies that improve balance in the multivariable distribution of baseline covariates while optimizing data retention. In our analyses, both propensity score matching, and coarsened exact matching performed similarly. However, due to the small number of covariates, the performance of coarsened exact matching might decrease with larger covariate sets, as shown in previous studies.

One concern with propensity score matching is its limited ability to balance past the propensity score matching paradox. A hybrid matching approach, with elements of both propensity score matching and coarsened exact matching, could be used to overcome the limitations of propensity score, and coarsened exact matching. Specifically, matching on latent variables would reduce the number of variables used in coarsened exact matching, while providing more granular data than the scalar value of the propensity score from which to inform matches to improve balance potential past the propensity score matching paradox. For example, a latent variable for risk of extraprostatic extension could be represented by PSA, Gleason Grade, clinical tumor stage, and percent of positive biopsy cores. Another latent variable for how estimated life expectancy impacts treatment decisions could be modeled using comorbidity information and age at diagnosis.

Moving forward, I am currently working with a PhD candidate in Statistics at the University of Waterloo to develop an RStudio package to improve upon the statistical sophistication and automatization of the systematic matching algorithm developed in chapter four. This work will provide analysts with an open access to a more user-friendly version of our systematic approach to matching and hopefully improve the quality of matching in the realm comparative effectiveness research using non-experimental data. We also aim to develop a latent variable matching strategy and assess its performance relative to propensity score matching and coarsened exact matching, using the approach demonstrated in chapter five.

The matching methods developed and presented in this thesis (specifically in chapter four) were used in chapter seven to compare the rate of metastatic progression between men diagnosed with unfavorable-risk non-metastatic prostate cancer who were initially treated with radiation therapy or radical prostatectomy. Despite the high-level of sophistication in the management of confounding through matching and regression modeling, the risk of bias in this study was still deemed to be high. This was the case for each study included in the systematic review and meta-analysis performed in chapter six as well. There are two reasons for this. Most studies, including that performed in chapter seven, did not involve all observable confounding variables. The second reason pertains to the inappropriate use of matching and regression strategies.

The reason for not including all potentially confounding variables could be due to limitations in obtaining information on comorbidities, demographic, socioeconomic, geographic, diagnostic, treatment, and outcome information. Institutional databases included in our systematic review and meta-analysis generally provided sufficient detail on diagnostic, treatment, and outcome related information. However, information pertaining to comorbidities, geographic location, demographic, and socioeconomic characteristics were generally unavailable. On the other hand, national registry databases provided this information, but in turn lacked detailed diagnostic, treatment and/or outcome related information. In our case, primary care medical records of patients diagnosed and treated at the GGWPC and LHSC were not readily obtainable due to concerns with privacy and confidentiality as well as practicality in collating information from paper-based documents and electronic medical records that likely used different software platforms. This resulted in inconsistently reported and incomplete information on comorbidities, demographic, socioeconomic, and geographic location, limiting our ability to effectively control for any potential confounding these variables might have introduced.

A concerted effort to improve the standardization in reporting diagnostic, treatment, and clinical follow-up information in detail and that has potential for linkage with medical records from primary and tertiary care providers (to provide information on comorbidities) and with administrative databases (that have demographic, and socioeconomic information) would have great value in the realm of comparative effectiveness research. This has been demonstrated, to some extent, with the National Prostate Cancer Register in Sweden, which collects detailed screening, diagnostic, risk-stratification, treatment and follow-up information and has potential for linkage to other national registries with information on comorbidities, demographic, socioeconomic, and geographic location.(223) However, follow-up information on margin status post-RP, biochemical progression, and metastatic progression and subsequent adjuvant and salvage therapy is still lacking.

To overcome the inappropriate use of regression techniques used in comparative effectiveness research, research ethics boards, grant reviewing bodies, and other organizations should ensure that projects meet standards for statistical analyses. Additionally, academic journals, such as the Journal of Urology, should continue to mandate that authors report specifics related to their statistical analyses, ensuring standard recommendations are met. This approach should be emulated by all journals to motivate the proper use and reporting of statistical analyses. An explanation for the subpar use of matching techniques could be due to the relatively recent development of any standardized guideline.(140) Further, such guidelines, as shown in chapter four, are incomplete in that no set of balance diagnostics are recommended to guide the development and evaluation of matching strategies.

It is my goal that through the publication of the recommended set of balance diagnostics presented in chapter three and the proposed systematic approach to developing and evaluating propensity score models for matching presented in chapter four, that the conduct of propensity score matching might improve in the field of prostate cancer and other fields of comparative effectiveness research.

References

1. Epstein J. Male Genital System and Lower Urinary Tract. In: Robbins basic pathology. 2013. p. 657–80.
2. Crawford D. Understanding the Epidemiology, Natural History, and Key Pathways Involved in Prostate Cancer. J Urol. 2009;73(5A):4–10.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.
4. Canadian Cancer Society. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Can Cancer Stat 2020. 2020;
5. Fradet Y, Klotz L, Trachtenberg J, Zlotta A. The burden of prostate cancer in Canada. Can Urol Assoc J. 2009;3(3 Suppl 2):S92.
6. Canadian Cancer Society. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Can Cancer Stat 2017 [Internet]. 2017;2016:1–132. Available from: cancer.ca/Canadian-CancerStatistics-2017-EN.pdf%0A
7. Bostwick D, Burke H, Djakiew D, Euling S, Ho S, Landolph J, et al. Human prostate cancer risk factors. Cancer. 2004;15(101):2371–490.
8. Bratt O. Hereditary prostate cancer: clinical aspects. J Urol. 2002;168(3):903–13.
9. Hoffman R, Gilliland F, Eley J, Harlan L, Stephenson R, Stanford J, et al. Racial and ethnic differences in advanced-stage prostate cancer: the Prostate Cancer Outcomes Study. J Natl Cancer Inst. 2001;93(5):388–95.
10. Noone A, Howlader N, Krapcho M, Miller D, Brest A, Yu M, et al. SEER Cancer Statistics Review, 1975-2015. National Cancer Institute. 2017.
11. Waldron N, Chowdhury S. Prostate cancer. Med (United Kingdom) [Internet]. 2020;48(2):119–22. Available from: http://www.elsevier.com/wps/find/journaldescription.cws_home/709606/description#description
12. Guy D, Ghanem G, Loblaw A, Buckley R, Persaud B, Cheung P, et al. Diagnosis, referral, and primary treatment decisions in newly diagnosed prostate cancer patients in a multidisciplinary diagnostic assessment program. J Can Urol Assoc. 2016;10(3-4April):120–5.
13. Balk S, Ko Y-J, Bubley G. Biology of Prostate-Specific Antigen. Biol Neoplasia. 2003;21(2):383–91.
14. Kim E, Andriole G. Prostate-specific antigen-based screening: controversy and guidelines. BMC Med. 2015;13(61):1–4.
15. Rendon R, Mason R, Marzouk K, Finelli A, Saad F, So A, et al. Canadian Urological Association recommendations on prostate cancer screening and early diagnosis. Can Urol Assoc J. 2017;11(10):298–309.
16. Cancer Care Ontario. Prostate Cancer Diagnosis Pathway [Internet]. 2018. p. 1–5. Available from: https://www.cancercareontario.ca/sites/ccocancercare/files/assets/DPMProstateDiagnosis.pdf
17. Epstein J, Egevad L, Amin M, Delahunt B, Srigley J, Humphrey P. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. Am J Surg Pathol. 2016;40(2):244–52.

18. Saad F, McCormack M. Prostate Cancer. 2019. 104–109 p.
19. Rodrigues G, Lukka H, Warde P, Brundage M, Souhami L, Crook J, et al. The prostate cancer risk stratification (ProCaRS) project: Recursive partitioning risk stratification analysis. Radiother Oncol [Internet]. 2013;109(2):204–10. Available from: http://dx.doi.org/10.1016/j.radonc.2013.07.020
20. Rodrigues G, Gonzalez-Maldonado S, Lukka H, Warde P, Brundage M, Souhami L, et al. The Prostate Cancer Risk Stratification (ProCaRS) Project: Database Construction and Outcome Analysis. Int J Radiat Oncol [Internet]. 2012;84(3):S57. Available from: http://dx.doi.org/10.1016/j.ijrobp.2012.07.356
21. Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically Localized Prostate Cancer: AUA/ASTRO/SUO Guideline. Part I: Risk Stratification, Shared Decision Making, and Care Options. J Urol [Internet]. 2018;199(3):683–90. Available from: https://doi.org/10.1016/j.juro.2017.11.095
22. Shekarriz B, Upadhyay J, Bianco Jr F, Tefilli M, Tiguert R, Gheiler E, et al. Impact of preoperative serum PSA level from 0 to 10 ng/ml on pathological findings and disease-free survival after radical prostatectomy. Prostate. 2001;48(3):136–43.
23. Aleman M, Karakiewicz P, Kupelian P, Kattan M, Graefen M, Cagiannos I, et al. Age and PSA predict likelihood of organ-confined disease in men presenting with PSA less than 10 ng/mL: implications for screening. Urology. 2003;62(1):70–4.
24. Partin A, Kattan M, Subong E, Walsh P, Wojno K, Oesterling J, et al. Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer: a multi-institutional update. JAMA. 1997;277(18):1445–51.
25. D’Amico A, Whittington R, Malkowicz S, Schultz D, Blank K, Broderick G. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. JAMA. 1998;280(11):969–74.
26. Bickley L. Bates’ Guide to Physical Examination and History Taking. Phlidelphia: Lippincott Williams and Wilkins; 2003.
27. Edge S, Compton C. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann surical Oncol. 2010;17(6):1471–4.
28. Cheng L, Montironi R, Bostwick D, Lopez-Beltran A, Berney D. Staging of prostate cancer. Histopathology. 2012;60(1):87–117.
29. Douglas T, McLeod D, Mostofi F, Mooneyhan R, Connelly R, Moul J, et al. Prostate-specific antigen-detected prostate cancer (Stage T1c): An analysis of whole-mount prostatectomy specimens. Prostate. 1997;32(1):59–64.
30. Armatys S, Koch M, Bihrle R, Gardner T, Cheng L. Is it necessary to separate clinical stage T1c from T2 prostate adenocarcinoma? BJU Int. 2005;96(6):777–80.
31. Ramos C, Carvalhal G, Smith D, Mager D, Catalona W. Clinical and pathological characteristics, and recurrence rates of stage T1c versus T2a or T2b prostate cancer. J Urol. 1999;161(5):1525–9.
32. Freedland S, Presti JR J, Terris M, Kane C, Aronson W, Dorey F, et al. Improved clinical staging system combining biopsy laterality and TNM stage for men with T1c and T2 prostate cancer: results from the SEARCH database. J Urol. 2003;169(6):2129–35.
33. Lerner S, Seay T, Blute M, Bergstralh E, Barrett D, Zincke H. Prostate specific antigen detected prostate cancer (clinical stage T1c): an interim analysis. J Urol. 1996;155(3):821–6.

34. Ghavamian R, Blute M, Bergstralh E, Slezak J, Zincke H. Comparison of clinically nonpalpable prostate-specific antigen-detected (cT1c) versus palpable (cT2) prostate cancers in patients undergoing radical retropubic prostatectomy. Urology. 1999;54(1):105–10.
35. Obek C, Louis P, Civantos F, Soloway M. Comparison of digital rectal examination and biopsy results with the radical prostatectomy specimen. J Urol. 1999;161(2):494–9.
36. Buyyounouski M, Horwitz E, Hanlon A, Uzzo R, Hanks G, Pollack A. Positive prostate biopsy laterality and implications for staging. Urology. 2003;62(2):298–303.
37. Cantrell B, DeKlerk D, Eggleston J, Boitnott J, Walsh P. Cantrell BB, DeKlerk DP, Eggleston JC, Boitnott JK, Walsh PC. Pathological factors that influence prognosis in stage A prostatic cancer: the influence of extent versus grade. J Urol. 1981;125(4):516–20.
38. Mathieu R, Moschini M, Beyer B, Gust K, Seisen T, Briganti A, et al. Prognostic value of the new grade groups in prostate cancer: a multi-institutional European validation study. Prostate Cancer Prostatic Dis. 2017;20(2):197.
39. Macleod L, Ellis W, Newcomb L, Zheng Y, Brooks J, Carroll P, et al. Timing of adverse prostate cancer reclassification on first surveillance biopsy: results from the Canary Prostate Cancer Active Surveillance Study. J Urol. 2017;197(4):1026–33.
40. Rajab R, Fisher G, Kattan M, Foster C, Møller H, Oliver T, et al. An improved prognostic model for stage T1a and T1b prostate cancer by assessments of cancer extent. Mod Pathol. 2011;24(1):58.
41. Epstein J, Walsh P, Carmichael M, Brendler C. Pathologic and clinical findings to predict tumor extent of nonpalpable (stage t1 c) prostate cancer. JAMA1. 1994;271(5):368–74.
42. Tosoian J, Mamawala M, Epstein J, Landis P, Wolf S, Trock B, et al. Intermediate and longer-term outcomes from a prospective active-surveillance program for favorable-risk prostate cancer. J Clin Oncol. 2015;33(30):3379.
43. Iremashvili V, Pelaez L, Manoharan M, Jorda M, Rosenberg D, Soloway M. Pathologic prostate cancer characteristics in patients eligible for active surveillance: a head-to-head comparison of contemporary protocols. Eur Urol. 2012;62(3):462–8.
44. Kasperzyk J, Shappley III W, Kenfield S, Mucci L, Kurth T, Ma J, et al. Watchful waiting and quality of life among prostate cancer survivors in the Physicians' Health Study. J Urol2. 2011;186(5):1862–7.
45. Shappley III W, Kenfield S, Kasperzyk J, Qiu W, Stampfer M, Sanda M, et al. Prospective study of determinants and outcomes of deferred treatment or watchful waiting among men with prostate cancer in a nationwide cohort. J Clin Oncol. 2009;27(30):4980.
46. Sartor O. Endpoints in prostate cancer clinical trials. Urology [Internet]. 2002;60(3 Suppl 1):101–7; discussion 107-8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12231062
47. Pound CR, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural History of Progression After PSA. J Am Med Assoc. 1999;281(17):1591–7.
48. Freedland S, Humphreys E, Mangold L, Eisenberger M, Dorey F, Walsh P, et al. Risk of prostate cancer–specific mortality following biochemical recurrence after radical prostatectomy. JAMA. 2005;294(4):433–9.
49. Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: A proposal for a standardized definition. J Clin Oncol. 2006;24(24):3973–8.
50. Djavan B, Moul J, Zlotta A, Remzi M, Ravery V. PSA progression following radical

prostatectomy and radiation therapy: new standards in the new millenium. Eur Urol. 2003;43(1):12–27.

51. Villers A. PSA in a follow-up after radical prostatectomy: a review. In: Proceedings from the First International Consultation on Prostate Cancer. 1997.

52. Han M, Partin A, Pound C, Epstein J, Walsh P. Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy: the 15-year Johns Hopkins experience. Urol Clin. 2001;28(3):555–6.

53. Zincke H, Oesterling J, Blute M, Bergstralh E, Myers R, Barrett D. Long-term (15 years) results after radical prostatectomy for clinically localized (stage T2c or lower) prostate cancer. J Urol. 1994;152(5):1850–7.

54. Foster L, Jajodia P, Fournier G, Shinohara K, Caroll P, Narrayan P. The value of PSA and TRUS-guided biopsy in detecting prostate fossa recurrence following radical prostatectomy. J Urol. 1993;149:1024.

55. Lange P, Ercole C, Lightner D, Fraley E, Vessella R. The value of serum prostate specific antigen determinations before and after radical prostatectomy. J Urol. 1989;141(4):873–9.

56. Amling C, Bergstralh E, Blute M, Slezak J, Zincke H. Defining prostate specific antigen progression after radical prostatectomy: what is the most appropriate cut point? J Urol. 2001;165(4):1146–51.

57. Eggener S, Scardino P, Walsh P, Han M, Partin A, Trock B, et al. Predicting 15-year prostate cancer specific mortality after radical prostatectomy. J Urol. 2011;185(3):869–75.

58. Zagars G, Pollack A. The fall and rise of prostate-specific antigen: Kinetics of serum prostate-specific antigen levels after radiation therapy for prostate cancer. Cancer. 1993;72(3):832–42.

59. Critz F, Williams W, Benton J, Levinson A, Holladay C, Holladay D. Prostate specific antigen bounce after radioactive seed implantation followed by external beam radiation for prostate cancer. J Urol. 2000;163(4):1085–9.

60. ASTRO. Consensus Statement. Guidelines for PSA Following Radiation Therapy. Int J Radiat Oncol Biol Phys. 1997;37:1035–41.

61. Roach M, Hanks G, Thames H, Schellhammer P, Shipley WU, Sokol GH, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: Recommendations of the RTOG-ASTRO Phoenix Consensus Conference. Int J Radiat Oncol Biol Phys. 2006;65(4):965–74.

62. Thompson A, Keyes M, Pickles T, Palma D, Moravan V, Spadinger I, et al. Evaluating the Phoenix definition of biochemical failure after 125I prostate brachytherapy: Can PSA kinetics distinguish PSA failures from PSA bounces? Int J Radiat Oncol Biol Phys. 2010;78(2):415–21.

63. Critz F, Williams W, Levinson A, Benton J, Schnell F, Holladay C, et al. Prostate specific antigen bounce after simultaneous irradiation for prostate cancer: the relationship to patient age. J Urol. 2003;170(5):1864–7.

64. Stock R, Stone N, Cesaretti J. Prostate-specific antigen bounce after prostate seed implantation for localized prostate cancer: descriptions and implications. Int J Radiat Oncol Biol Phys. 2003;56(2):448–53.

65. Crook J, Gillan C, Yeung I, Austen L, McLean M, Lockwood G. PSA kinetics and PSA bounce following permanent seed prostate brachytherapy. Int J Radiat Oncol Biol Phys. 2007;69(2):426–33.

66. Bostancic C, Merrick G, Butler W, Wallner K, Allen Z, Galbreath R, et al. Isotope and

patient age predict for PSA spikes after permanent prostate brachytherapy. Int J Radiat Oncol Biol Phys. 2007;68(5):1431–7.

67. Das P, Chen M, Valentine K, Lopes L, Cormack R, Renshaw A, et al. Using the magnitude of PSA bounce after MRI-guided prostate brachytherapy to distinguish recurrence, benign precipitating factors, and idiopathic bounce. Int J Radiat Oncol Biol Phys. 2002;54(3):698–702.

68. Romesser P, Pei X, Shi W, Zhang Z, Kollmeier M, McBride S, et al. Prostate-Specific Antigen (PSA) Bounce After Dose-Escalated External Beam Radiation Therapy Is an Independent Predictor of PSA Recurrence, Metastasis, and Survival in Prostate Adenocarcinoma Patients. Int J Radiat Oncol Biol Phys. 2018;100(1):59–67.

69. Wennberg J. Unwarranted variations in healthcare delivery: implications for academic medical centres. J Natl Cancer Inst. 2006;98(5):355–7.

70. O'Connor A, Llewellyn-Thomas H, Flood A. Modifying unwarranted variations in health care: shared decision making using patient decision aids. Health Aff. 2004;63.

71. Morash C, Tey R, Agbassi C, Klotz L, McGowan T, Srigley J, et al. Active surveillance for the management of localized prostate cancer: Guideline recommendations. Can Urol Assoc J. 2015;9(5–6):171.

72. Dall'Era M, Cowan J, Simko J, Shinohara K, Davies B, Konety B, et al. Surgical management after active surveillance for low-risk prostate cancer: pathological outcomes compared with men undergoing immediate treatment. BJU Int. 2011;107(8):1232–7.

73. Iremashvili V, Manoharan M, Rosenberg D, Acosta K, Soloway M. Pathological findings at radical prostatectomy in patients initially managed by active surveillance: a comparative analysis. Prostate. 2012;72(14):1573–9.

74. Sugimoto M, Shiraishi T, Tsunemori H, Demura T, Saito Y, Kamoto T, et al. Pathological findings at radical prostatectomy in Japanese prospective active surveillance cohort. Jpn J Clin Oncol. 2010;68(12):1257–62.

75. Warlick C, Trock B, Landis P, Epstein J, Carter H. Delayed versus immediate surgical intervention and prostate cancer outcome. J Natl Cancer Inst. 2006;98(5):355–7.

76. Radomski L, Gani J, Trottier G, Finelli A. Active surveillance failure for prostate cancer: does the delay in treatment increase the risk of urinary incontinence? Can J Urol. 2012;19(3):6287–92.

77. Kravchick S, Peled R, Cytron S. Watchful waiting and active surveillance approach in patients with low risk localized prostatic cancer: An experience of out-patients clinic with 12-year follow-up. Pathol Oncol Res. 2011;17(4):893–7.

78. Ischia J, Pang C, Tay Y, Suen L, Christopher F, Aw H, et al. Active surveillance for prostate cancer: an Australian experience. BJU Int. 2012;109(s3):40–3.

79. Ercole B, Marietti S, Fine J, Albertsen P. Outcomes following active surveillance of men with localized prostate cancer diagnosed in the prostate specific antigen era. J Urol. 2008;180(4):1336–41.

80. Bul M, Zhu X, Valdagni R, Pickles T, Kakehi Y, Rannikko A, et al. Active surveillance for low-risk prostate cancer worldwide: the PRIAS study. Eur Urol. 2013;63(4):597–603.

81. Klotz L. Active surveillance for prostate cancer: overview and update. Curr Treat Options Oncol. 2013;14(1):97–108.

82. Stattin P, Holmberg E, Johansson J, Holmberg L, Adolfsson J, Hugosson J. Outcomes in localized prostate cancer: National Prostate Cancer Register of Sweden follow-up study. J Natl Cancer Inst. 2010;102(13):950–8.

83. Abd-Alazeez M, Ahmed H, Arya M, Allen C, Dikaios N, Freeman A, et al. Can multiparametric magnetic resonance imaging predict upgrading of transrectal ultrasound biopsy results at more definitive histology? Urol Oncol Semin Orig Investig. 2014;32(6):741–7.

84. Stamatakis L, Siddiqui M, Nix J, Logan J, Rais-Bahrami S, Walton-Diaz A, et al. Accuracy of multiparametric magnetic resonance imaging in confirming eligibility for active surveillance for men with prostate cancer. Cancer. 2013;119(18):3359–66.

85. Margel D, Yap S, Lawrentschuk N, Klotz L, Haider M, Hersey K, et al. Impact of multiparametric endorectal coil prostate magnetic resonance imaging on disease reclassification among active surveillance candidates: a prospective cohort study. J Urol. 2012;187(4):1247–52.

86. Wilt T, Brawer M, Jones K, Barry M, Aronson W, Fox S, et al. Radical prostatectomy versus observation for localized prostate cancer. N Engl J Med. 2012;367(3):203–13.

87. Bill-Axelson A, Holmberg L, Ruutu M, Garmo H, Stark J, Busch C, et al. Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med. 2011;364(18):1708–17.

88. Donovan J, Hamdy F, Lane J, Mason M, Metcalfe C, Walsh E, et al. Patient-Reported Outcomes after Monitoring, Surgery, or Radiotherapy for Prostate Cancer. N Engl J Med. 2016;375(15):1425–37.

89. Hamdy F, Donovan J, Lane J, Mason M, Metcalfe C, Holding P, et al. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. N Engl J Med. 2016;375:1415–24.

90. Sanda MG, Cadeddu JA, Kirkby E, Chen RC, Crispino T, Fontanarosa J, et al. Clinically Localized Prostate Cancer: AUA/ASTRO/SUO Guideline. Part II: Recommended Approaches and Details of Specific Care Options. J Urol [Internet]. 2018;199(4):990–7. Available from: https://doi.org/10.1016/j.juro.2018.01.002

91. Jones C, Hunt D, McGowan D, Amin M, Chetner M, Bruner D, et al. Radiotherapy and Short-Term Androgen Deprivation for Localized Prostate Cancer. N Engl J Med. 2011;365(2):107–18.

92. Kibel AS, Ciezki JP, Klein EA, Reddy CA, Lubahn JD, Haslag-Minoff J, et al. Survival among men with clinically localized prostate cancer treated with radical prostatectomy or radiation therapy in the prostate specific antigen era. J Urol. 2012;187(4):1259–65.

93. Tewari A, Divine G, Chang P, Shemtov M, Milowsky M, Nanus D, et al. Long-term survival in men with high grade prostate cancer: a comparison between conservative treatment, radiation therapy and radical prostatectomy—a propensity scoring approach. J Urol. 2007;177(3):911–5.

94. Cooperberg M, Vickers A, Broering J, Carroll P, Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) Investigators. Comparative risk-adjusted mortality outcomes after primary surgery, radiotherapy, or androgen-deprivation therapy for localized prostate cancer. Cancer. 2010;116(22):5226–34.

95. Albertsen P, Hanley J, Penson D, Barrows G, Fine J. 13-year outcomes following treatment for clinically localized prostate cancer in a population-based cohort. J Urol. 2007;177(3):932–6.

96. Merglen A, Schmidlin F, Fioretta G, Verkooijen H, Rapiti E, Zanetti R, et al. Short-and long-term mortality with localized prostate cancer. Arch Intern Med. 2007;167(18):1944–50.

97. Zelefsky M, Eastham JA, Cronin A, Fuks Z, Zhang Z, Yamada Y, et al. Metastasis After Radical Prostatectomy or External Beam Radiotherapy for Patients With Clinically Localized Prostate Cancer: A Comparison of Clinical Cohorts Adjusted for Case Mix. J Clin Oncol. 2010;28(9):1508–13.
98. Abdollah F, Schmitges J, Sun M, Jeldres C, Tian Z, Briganti A, et al. Comparison of mortality outcomes after radical prostatectomy versus radiotherapy in patients with localized prostate cancer: A population-based analysis. Int J Urol. 2012;19(8):836–44.
99. Hoffman RM, Koyama T, Fan KH, Albertsen PC, Barry MJ, Goodman M, et al. Mortality after radical prostatectomy or external beam radiotherapy for localized prostate cancer. J Natl Cancer Inst. 2013;105(10):711–8.
100. Lee JY, Cho KS, Kwon JK, Jeh SU, Kang HW, Diaz RR, et al. A Competing Risk Analysis of Cancer-Specific Mortality of Initial Treatment with Radical Prostatectomy versus Radiation Therapy in Clinically Localized High-Risk Prostate Cancer. Ann Surg Oncol. 2014;21(12):4026–33.
101. Sooriakumaran P, Nyberg T, Akre O, Haendler L, Heus I, Olsson M, et al. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: Observational study of mortality outcomes. BMJ [Internet]. 2014;348:1–13. Available from: http://dx.doi.org/doi:10.1136/bmj.g1502
102. DeGroot J, Brundage M, Lam M, Rohland S, Heaton J, Mackillop W, et al. Prostate cancer-specific survival differences in patients treated by radical prostatectomy versus curative radiotherapy. Can Urol Assoc J. 2013;7(5–6):E299.
103. Bolla M, Collette L, Blank L, Warde P, Dubois J, Mirimanoff R, et al. Long-term results with immediate androgen suppression and external irradiation in patients with locally advanced prostate cancer (an EORTC study): a phase III randomised trial. Lancet. 2002;360(9327):103–8.
104. Bolla M, De Reijke T, Van Tienhoven G, Van den Bergh A, Oddens J, Poortmans P, et al. Duration of androgen suppression in the treatment of prostate cancer. N Engl J Med. 2009;360(24):2516–27.
105. Horwitz E, Bae K, Hanks G, Porter A, Grignon D, Brereton H, et al. Ten-year follow-up of radiation therapy oncology group protocol 92-02: a phase III trial of the duration of elective androgen deprivation in locally advanced prostate cancer. J Clin Oncol. 2008;26(15):2497–504.
106. Donnelly B, Saliken J, Brasher P, Ernst S, Rewcastle J, Lau H, et al. A randomized trial of external beam radiotherapy versus cryoablation in patients with localized prostate cancer. Cancer. 2010;116(2):323–30.
107. Lukka H, Waldron T, Chin J, Mayhew L, Warde P, Winquist E, et al. High-intensity focused ultrasound for prostate cancer: a systematic review. Clin Oncol. 2011;23(2):117–27.
108. Roach M, Ceron Lizarraga TL, Lazar AA. Radical prostatectomy versus radiation and androgen deprivation therapy for clinically localized prostate cancer: How good is the evidence? Int J Radiat Oncol Biol Phys [Internet]. 2015;93(5):1064–70. Available from: http://dx.doi.org/10.1016/j.ijrobp.2015.08.005
109. Greenberger B, Chen V, Den R. Combined Modality Therapies for High-Risk Prostate Cancer: Narrative Review of Current Understanding and New Directions. Front Oncol. 2019;9(1273):1–14.
110. Crook JM, Gomez-Iturriaga A, Wallace K, Ma C, Fleshner N, Jewett M. Comparison of

Health-Related Quality of Life 5 Years after Treatment for Men Who Either Chose or Were Randomized to Radical Prostatectomy or Brachytherapy after a SPIRIT (ACOSOG Z0070) Trial Education Session. Brachytherapy [Internet]. 2010;9:S23. Available from: http://dx.doi.org/10.1016/j.brachy.2010.02.004

111. Ritchie A, Verbaeys A, Fellows G, Hehir M, Mason M, Moffat L, et al. Early closure of a randomized controlled trial of three treatment approaches to early localised prostate cancer: The MRC PR06 trial [3]. BJU Int. 2004;94(9):1400–1.

112. Chen RC, Basak R, Meyer AM, Kuo TM, Carpenter WR, Agans RP, et al. Association between choice of radical prostatectomy, external beam radiotherapy, brachytherapy, or active surveillance and patient-reported quality of life among men with localized prostate cancer. JAMA. 2017;317(11):1141–50.

113. Buron C, Le Vu B, Cosset JM, Pommier P, Peiffert D, Delannes M, et al. Brachytherapy versus prostatectomy in localized prostate cancer: Results of a French multicenter prospective medico-economic study. Int J Radiat Oncol Biol Phys. 2007;67(3):812–22.

114. Petrelli F, Vavassori I, Coinu A, Borgonovo K, Sarti E, Barni S. Radical prostatectomy or radiotherapy in high-risk prostate cancer: A systematic review and metaanalysis. Clin Genitourin Cancer [Internet]. 2014;12(4):215–24. Available from: http://dx.doi.org/10.1016/j.clgc.2014.01.010

115. Alibhai SMH, Leach M, Tomlinson G, Krahn MD, Fleshner N, Holowaty E, et al. 30-day mortality and major complications after radical prostatectomy: Influence of age and comorbidity. J Natl Cancer Inst. 2005;97(20):1525–32.

116. Keyes M, Crook J, Morris WJ, Morton G, Pickles T, Usmani N, et al. Canadian prostate brachytherapy in 2012. J Can Urol Assoc. 2013;7(2):51–8.

117. Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. Cancer. 2008;112(11):2456–66.

118. Wallis CJD, Saskin R, Choo R, Herschorn S, Kodama RT, Satkunasivam R, et al. Surgery Versus Radiotherapy for Clinically-localized Prostate Cancer: A Systematic Review and Meta-analysis. Eur Urol [Internet]. 2016;70(1):21–30. Available from: http://dx.doi.org/10.1016/j.eururo.2015.11.010

119. Tree AC, van As NJ, Dearnaley DP. Re: Christopher J.D. Wallis, Refik Saskin, Richard Choo, et al. Surgery Versus Radiotherapy for Clinically-localized Prostate Cancer: A Systematic Review and Meta-analysis. Eur Urol 2016;70:21–30. Eur Urol [Internet]. 2016;70(1):e10. Available from: http://dx.doi.org/10.1016/j.eururo.2016.02.044

120. Robinson D, Garmo H, Lissbrant IF, Widmark A, Pettersson A, Gunnlaugsson A, et al. Prostate Cancer Death After Radiotherapy or Radical Prostatectomy: A Nationwide Population-based Observational Study. Eur Urol. 2018;73(4):502–11.

121. Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, Van Der Kwast T, et al. EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. Eur Urol [Internet]. 2014;65(2):467–79. Available from: http://dx.doi.org/10.1016/j.eururo.2013.11.002

122. Cookson MS, Aus G, Burnett AL, Canby-Hagino ED, D’Amico A V., Dmochowski RR, et al. Variation in the Definition of Biochemical Recurrence in Patients Treated for Localized Prostate Cancer: The American Urological Association Prostate Guidelines for Localized Prostate Cancer Update Panel Report and Recommendations for a Standard in the Re. J Urol. 2007;177(2):540–5.

123. Kuban D, Thames H, Levy L, Horwitz E, Kupelian P, Martinez A, et al. Failure definition-dependent differences in outcome following radiation for localized prostate cancer: Can one size fit all? Int J Radiat Oncol Biol Phys. 2005;61(2):409–14.
124. Nielsen ME, Partin AW. The impact of definitions of failure on the interpretation of biochemical recurrence following treatment of clinically localized prostate cancer. Rev Urol. 2007;9(2):57–62.
125. Hoffman RM, Stone SN, Hunt WC, Key CR, Gilliland FD. Effects of misattribution in assigning cause of death on prostate cancer mortality rates. Ann Epidemiol. 2003;13(6):450–4.
126. D'Elia C, Cerruto M, Cioffi A, Novella G, Cavalleri S, Artibani W. Upgrading and upstaging in prostate cancer: From prostate biopsy to radical prostatectomy. Mol Clin Oncol [Internet]. 2014;2(6):1145–9. Available from: https://www.spandidos-publications.com/10.3892/mco.2014.370
127. Chin J, Rumble RB, Loblaw DA. Brachytherapy for Patients With Prostate Cancer: American Society of Clinical Oncology/Cancer Care Ontario Joint Guideline Update. J Clin Oncol. 2017;35(15):1737–43.
128. Roach M, Alexander M. The prognostic significance of race and survival from breast cancer: a model for assessing the reliability of reported survival differences. J Natl Med Assoc [Internet]. 1995;87(3):214–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7731072%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2607822
129. Schaumberg DA, Shah S, Nordstrom BL, McDonald L, Ramagopalan S V., Stokes M. Evaluation of comparative effectiveness research: A practical tool. J Comp Eff Res. 2018;7(5):503–15.
130. Hernán M, Robins J. Causal Inference [Internet]. Boca Raton: Chapman & Hall/CRC, forthcoming.; 2019. Available from: https://www.taylorfrancis.com/books/9781134881819/chapters/10.4324/9781315542287-6
131. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal Inference in Public Health. Annu Rev Public Health. 2013;34(1):61–75.
132. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci [Internet]. 2010;25(1):1–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20871802%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2943670
133. Etz A. Introduction to the Concept of Likelihood and Its Applications. Adv Methods Pract Psychol Sci. 2018;1(1):60–9.
134. Vittinghoff E, Glidden D V., Shiboski SC, McCulloch CE. Regression Methods in Biostatistics. 2nd ed. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors. New York City, NY: Springer; 2012. 1–527 p.
135. Greenland S, Schwartzbaum J, Finkle W. Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol. 2000;151(5):531–9.
136. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Polit Anal. 2019;
137. King G, Lucas C, Nielsen R. Optimizing balance and sample size in matching methods for causal inference [Internet]. 2013. p. 1–27. Available from: https://gking.harvard.edu/files/gking/files/frontier_0.pdf

138. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal. 2007;15(3):199–236.

139. Grijalva CG, Roumie CL, Murff HJ, Hung AM, Beck C, Liu X, et al. The role of matching when adjusting for baseline differences in the outcome variable of comparative effectiveness studies. J Comp Eff Res. 2015;4(4):341–9.

140. Yao XI, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. J Natl Cancer Inst. 2017;109(8):1–9.

141. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika [Internet]. 1983;70(1):41–55. Available from: https://www-jstor-org.ucsf.idm.oclc.org/stable/pdf/2335942.pdf?refreqid=excelsior%3Ab4e97f50c22058d74adcea6eff2b539f

142. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies : a systematic review. J Clin Epidemiol. 2005;58:550–9.

143. Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, et al. Comparison of Propensity Score Methods and Covariate Adjustment. J Am Coll Cardiol [Internet]. 2017;69(3):345–57. Available from: https://linkinghub.elsevier.com/retrieve/pii/S073510971637036X

144. Smith GD, Pickles T, Crook J, Martin AG, Vigneault E, Cury FL, 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 [Internet]. 2015;91(3):505–16. Available from: http://dx.doi.org/10.1016/j.ijrobp.2014.11.018

145. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin M. Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. Am J Epidemiol. 2018;187(9):1951–61.

146. Ho D, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J Stat Softw. 2011;42(8):1–43.

147. 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. 2016;51(5):1960–80.

148. 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. 2020;189(6):613–22.

149. Sturges H. The choice of a class interval. J Am Stat Assoc. 1926;21(153):65–6.

150. Chen H, Laba JM, Boldt G, Goodman CD, Palma DA, Senan S, et al. Stereotactic Ablative Radiotherapy Versus Surgery in Early Lung Cancer: A Meta-Analysis of Propensity Score-Adjusted Comparative Effectiveness Studies. Int J Radiat Oncol. 2017;99(2):E445.

151. Camillo F, D’Attoma I. %GI : A SAS Macro for Measuring and Testing Global Imbalance of Covariates within Subgroups . J Stat Softw. 2012;51(Code Snippet 1).

152. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;(July):3083–107.

153. Moher D, Schulz KF, Altman D. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. JAMA. 2001;285:1787–991.

154. Imai K, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc A. 2008;481–502.

155. Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Ann Transl Med. 2019;7(1):16–16.

156. Stuart E, Lee B, Leacy F. Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. J Clin Epidemiol. 2013;66(8 Suppl):S84–90.

157. Szklo M, Nieto J. EpidemiologyBeyond the Basics. 4th ed. Burlington, MA: Jones and Bartlett Learning; 2019. 3 p.

158. Raymond E, O'Callaghan M, Campbell J, Vincent A, Beckmann K, Roder D, et al. An appraisal of analytical tools used in predicting clinical outcomes following radiation therapy treatment of men with prostate cancer: a systematic review. Radiat Oncol. 2017;12(56):1–20.

159. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. J Am Stat Assoc. 2011;106(493):345–61.

160. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med. 2014;33:1685–99.

161. Belitser S V, Martens EP, Pestman WR, Groenwold RHH, Boer A De, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf. 2011;20:1115–29.

162. Kim DH, Pieper CF, Ahmed A, Colón-Emeric CS. Use and Interpretation of Propensity Scores in Aging Research: A Guide for Clinical Researchers. J Am Geriatr Soc. 2016;64(10):2065–73.

163. Lee J, Little TD. A practical guide to propensity score analysis for applied clinical research. Behav Res Ther [Internet]. 2017;98:76–90. Available from: http://dx.doi.org/10.1016/j.brat.2017.01.005

164. Williamson EJ, Forbes A. Introduction to propensity scores. Respirology. 2014;19(5):625–35.

165. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399–424.

166. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From naïve enthusiasm to intuitive understanding. Stat Methods Med Res. 2012;21(3):273–93.

167. Morris WJ, Tyldesley S, Rodda S, Halperin R, Pai H, McKenzie M, 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 f. Int J Radiat Oncol Biol Phys [Internet]. 2017;98(2):275–85. Available from: http://dx.doi.org/10.1016/j.ijrobp.2016.11.026

168. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149–56.

169. Team Rs. RStudio: Integrated Development Environment for R [Internet]. Boston, MA: RStudio, PBC; 2020. Available from: http://www.rstudio.com/

170. Austin PC. Some Methods of Propensity-Score Matching had Superior Performance to Others : Results of an Empirical Investigation and Monte Carlo simulations. Biometrical J. 2009;51:171–84.

171. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhyā Indian J Stat Ser A. 1973;417–46.

172. Greifer N. cobalt: Covariate Balance Tables and Plots. 2020.

173. Kent EC, Hussain MH. Neoadjuvant Therapy for Prostate Cancer: An Oncologist's Perspective. Rev Urol. 2003;5 Suppl 3:S28-37.

174. Pilepich M, Krall J, Al-Sarraf M, John M, Dogget R, Sause W, et al. Androgen deprivation with radiation therapy compared with radiation therapy alone for locally advanced prostatic carcinoma — a randomized comparative trial of the Radiation Therapy Oncology Group. Urology. 1995;45:616–23.

175. Higano C. Side effects of androgen deprivation therapy: monitoring and minimizing toxicity. Urology. 2003;61:32–8.

176. 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. 2015;15(190):1–8.

177. Khor R, Duchesne G, Tai K, Foroudi F, Chander S, Van Dyk S, 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. 2013;85(3):679–85.

178. Rubin DB, Rosenbaum PR. Reducing Bias in Observational Studies Using Score on the Propensity Subclassification. J Am Stat Assoc. 1984;79(387):516–24.

179. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014;33(6):1057–69.

180. Stephenson AJ, Kattan MW, Eastham JA, Bianco FJ, Yossepowitch O, Vickers AJ, et al. Prostate cancer-specific mortality after radical prostatectomy for patients treated in the prostate-specific antigen era. J Clin Oncol. 2009;27(26):4300–5.

181. Therneau TM, Lumley T, Atkinson E, Crowson C. Package 'survival' [Internet]. 2020. p. 1–176. Available from: https://github.com/therneau/survival

182. Gayat E, Resche-rigon M, Mary J. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. Pharm Stat. 2012;11(3):222–9.

183. Westreich D, Cole SR. Invited commentary: Positivity in practice. Am J Epidemiol. 2010;171(6):674–7.

184. National Comprehensive Cancer Network. NCCN clinical practice guidelines in oncology: prostate cancer v4.2018. 2018.

185. Punnen S, Cooperberg MR. The epidemiology of high-risk prostate cancer. Curr Opin Urol. 2013;23(4):331–6.

186. Podder T, Song D, Showalter T, Beaulieu L. Advances in Radiotherapy for Prostate Cancer Treatment. Prostate Cancer. 2016;2016:2–4.

187. Sun M, Sammon JD, Becker A, Roghmann F, Tian Z, Kim SP, et al. Radical prostatectomy vs radiotherapy vs observation among older patients with clinically localized prostate cancer: A comparative effectiveness evaluation. BJU Int.

2014;113(2):200–8.

188. Yin M, Zhao J, Monk P, Martin D, Folefac E, Joshi M, et al. Comparative effectiveness of surgery versus external beam radiation with/without brachytherapy in high-risk localized prostate cancer. Cancer Med. 2020;9:27–34.

189. Dearnaley DP, Jovic G, Syndikus I, Khoo V, Cowan RA, Graham JD, et al. Escalated-dose versus control-dose conformal radiotherapy for prostate cancer: Long-term results from the MRC RT01 randomised controlled trial. Lancet Oncol [Internet]. 2014;15(4):464–73. Available from: http://dx.doi.org/10.1016/S1470-2045(14)70040-3

190. Wang Z, Ni Y, Chen J, Sun G, Zhang X, Zhao J, et al. The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: A systematic review and meta-analysis. World J Surg Oncol. 2020;18(1):1–13.

191. Feldman AS, Meyer CP, Sanchez A, Krasnova A, Reznor G, Menon M, et al. Morbidity and Mortality of Locally Advanced Prostate Cancer: A Population Based Analysis Comparing Radical Prostatectomy versus External Beam Radiation. J Urol [Internet]. 2017;198(5):1061–8. Available from: https://doi.org/10.1016/j.juro.2017.05.073

192. Gu X, Gao X, Cui M, Xie M, Ma M, Qin S, et al. Survival outcomes of radical prostatectomy and external beam radiotherapy in clinically localized high-risk prostate cancer: A population-based, propensity score matched study. Cancer Manag Res. 2018;10:1061–7.

193. Jang T, Patel N, Faiena I, Radadia K, Moore D, Elsamra S, et al. Comparative effectiveness of radical prostatectomy with adjuvant radiotherapy versus radiotherapy plus androgen deprivation therapy for men with advanced prostate cancer. Cancer. 2018;124(20):4010–22.

194. Jayadevappa R, Chhatre S, Wong Y-N, Wittink MN, Cook R, Morales KH, et al. Comparative effectiveness of prostate cancer treatments for patient-centered outcomes. Medicine (Baltimore). 2017;96(18):e6790.

195. Muralidhar V, Mahal B, Butler S, Lamba N, Yang D, Leeman J, et al. Combined External Beam Radiation Therapy and Brachytherapy versus Radical Prostatectomy with Adjuvant Radiation Therapy of Gleason 9-10 Prostate Cancer. J Urol. 2019;202.

196. Gunnarsson O, Schelin S, Brudin L, Carlsson S, Damber JE. Triple treatment of high-risk prostate cancer. A matched cohort study with up to 19 years follow-up comparing survival outcomes after triple treatment and treatment with hormones and radiotherapy. Scand J Urol [Internet]. 2019;53(2–3):102–8. Available from: https://doi.org/10.1080/21681805.2019.1600580

197. Moher D, Liberati A, Tetzlaff J, Altman DG, Altman D, Antes G, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6(7):e1000097.

198. Akakura K, Suzuki H, Ichikawa T, Fujimoto H, Maeda O, Usami M, et al. A randomized trial comparing radical prostatectomy plus endocrine therapy versus external beam radiotherapy plus endocrine therapy for locally advanced prostate cancer: results at median follow-up of 102 months. Jpn J Clin Oncol. 2006;36(12):789–93.

199. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ [Internet]. 2003;327(7414):557–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12958120%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC192859

200. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package . J Stat Softw.

2015;36(3).

201. Cooperberg MR, Cowan J, Broering JM, Carroll PR. High-risk prostate cancer in the United States, 1990-2007. World J Urol. 2008;26(3):211–8.

202. Beckmann K, Garmo H, Nilsson P, Franck Lissbrant I, Widmark A, Stattin P. Radical radiotherapy for prostate cancer: patterns of care in Sweden 1998–2016. Acta Oncol (Madr) [Internet]. 2020;59(5):549–57. Available from: https://doi.org/10.1080/0284186X.2020.1730003

203. Corkum MT, Morton G, Louie A V., Bauman GS, Mendez LC, Chin J, et al. Is prostate brachytherapy a dying art? Trends and variation in the definitive management of prostate cancer in Ontario, Canada. Radiother Oncol [Internet]. 2020;152:42–8. Available from: https://doi.org/10.1016/j.radonc.2020.07.036

204. Guedea F, Venselaar J, Hoskin P, Hellebust T, Peiffert D, Londres B, et al. Patterns of care for brachytherapy in Europe: updated results. Radiother Oncol. 2010;97(3):514–20.

205. Orio P, Nguyen P, Buzurovic I, Cail D, Chen Y. Prostate brachytherapy case volumes by academic and nonacademic practices: implications for future residency training. Int J Radiat Oncol Biol Phys. 2016;96:624–8.

206. Bockholt N, DeRoo E, Nepple K, Modrick J, Smith M, Fallon B, et al. First 100 cases at a low volume prostate brachytherapy institution: learning curve and the importance of continuous quality improvement. Can J Urol. 2013;20(5):6907–12.

207. Liu H, Malkoske K, Sasaki D, Bews J, Alain D, Nugent Z, et al. The dosimetric quality of brachytherapy implants in patients with small prostate volume depends on the experience of the brachytherapy team. Brachytherapy. 2010;9(3):202–7.

208. Burt L, Shrieve D, Tward J. Factors influencing prostate cancer patterns of care: an analysis of treatment variation using the SEER database. Adv Radiat Oncol. 2018;3:170–80.

209. Boorjian S, Karnes J, Viterbo R, Rangel L, Bergstralh E, Horwitz E, et al. Long-Term Survival After Radical Prostatectomy Versus External Beam Radiotherapy for Patients with High-Risk Prostate Cancer Stephen. Cancer. 2011;117(13):2883–91.

210. Westover K, Chen MH, Moul J, Robertson C, Polascik T, Dosoretz D, et al. Radical prostatectomy vs radiation therapy and androgen-suppression therapy in high-risk prostate cancer. BJU Int. 2012;110(8):1116–21.

211. Jayadevappa R, Lee DI, Chhatre S, Guzzo TJ, Malkowicz SB. Comparative effectiveness of treatments for high-risk prostate cancer patients. Urol Oncol Semin Orig Investig [Internet]. 2019;37(9):574.e11-574.e18. Available from: https://doi.org/10.1016/j.urolonc.2019.06.005

212. Yin M, Zhao J, Monk P, Martin D, Folefac E, Joshi M, et al. Comparative effectiveness of surgery versus external beam radiation with/without brachytherapy in high-risk localized prostate cancer. Cancer Med. 2020;9(1):27–34.

213. Caño-Velasco J, Herranz-Amo F, Barbas-Bernardos G, Polanco-Pujol L, Verdú-Tartajo F, Lledo-Garcia E, et al. Oncological control in high-risk prostate cancer after radical prostatectomy and salvage radiotherapy compared to radiotherapy plus primary hormone therapy. Actas Urol Esp. 2019;43(4):190–7.

214. Van Hemelrijck M, Wigertz A, Sandin F, Garmo H, Hellström K, Fransson P, et al. Cohort profile: The national prostate cancer register of sweden and prostate cancer data base Sweden 2.0. Int J Epidemiol. 2013;42(4):956–67.

215. Zelefsky M, Reuter V, Fuks Z, Scardino P, Shippy A. Influence of local tumor control on

distant metastases and cancer related mortality after external beam radiotherapy for prostate cancer. J Urol. 2008;179(4):1368–73.

216. Thompson I, Tangen C, Paradelo J, Lucia M, Miller G, Troyer D, et al. Adjuvant radiotherapy for pathologically advanced prostate cancer: a randomized clinical trial. JAMA. 2006;296(19):2329–35.

217. Bolla M, van Poppel H, Tombal B, Vekemans K, Da Pozzo L, de Reijke T, et al. Postoperative radiotherapy after radical prostatectomy for high-risk prostate cancer: long-term results of a randomised controlled trial (EORTC trial 22911). Lancet. 2018;380(9858):2018–27.

218. Wiegel T, Bottke D, Steiner U, Siegmann A, Golz R, Storkel S, et al. Phase III postoperative adjuvant radiotherapy after radical prostatectomy compared with radical prostatectomy alone in pT3 prostate cancer with postoperative undetectable prostate-specific antigen: ARO 96-02/AUO AP 09/95. J Clin Oncol. 2009;27(18):2924–30.

219. Zumsteg Z, Spratt D, Daskivich T, Tighiouart M, Luu M, Rodgers J, et al. Effect of Androgen Deprivation on Long-term Outcomes of Intermediate-Risk Prostate Cancer Stratified as Favorable or Unfavorable. JAMA. 202AD;3(9).

220. Kishan A, Cook R, Ciezki J, Ross A, Pomerantz M, Nguyen P, et al. Radical Prostatectomy, External Beam Radiotherapy, or External Beam Radiotherapy With Brachytherapy Boost and Disease Progression and Mortality in Patients With Gleason Score 9-10 Prostate Cancer. JAMA. 2019;319(9):896–905.

221. Markovina S, Marshall W, Badiyan SN, Vetter J, Gay H, Paradis A, et al. Superior metastasis-free survival for patients with high-risk prostate cancer treated with definitive radiation therapy compared to radical prostatectomy: A propensity score-matched analysis. Adv Radiat Oncol. 2017;3:190–6.

222. Reichard C, Hoffman K, Tang C, Williams S, Allen P, Achim M, et al. Radical prostatectomy or radiotherapy for high- and very high-risk prostate cancer: a multidisciplinary prostate cancer clinic experience of patients eligible for either treatment. BJU Int. 2019;124:811–9.

223. Hagel E, Garmo H, Bill-Axelson A, Bratt O, Johansson J-E, Adolfsson J, et al. PCBaSe Sweden: A register-based resource for prostate cancer research. Scand J Urol Nephrol. 2009;43:342–9.

224. Loeb S, Smith ND, Roehl KA, Catalona WJ. Intermediate-Term Potency, Continence, and Survival Outcomes of Radical Prostatectomy for Clinically High-Risk or Locally Advanced Prostate Cancer. Urology. 2007;69(6):1170–5.

Supplementary Tables and Figures

S Table 5.1a Characteristics of PSM strategies for comparison one

PSM Strategy	Caliper Width SD of logit(PS)	BT+ADT (n)	E+ADT (n)	Events (n)
Unmatched	-	433	132	56
Nearest	no caliper	396	132	56
1	2.0	343	132	52
2	1.5	321	125	51
3	1.0	299	122	48
4	0.8	287	121	47
5	0.6	274	118	44
6	0.5	270	117	45
7	0.4	268	116	45
8	0.3	264	114	45
9	**0.2**	**258**	**112**	**44**
10	**0.1**	**248**	**109**	**41**
11	**0.05**	**234**	**104**	**37**
12	0.025	214	104	33
13	0.01	182	82	30
14	0.005	139	69	24

S Table 5.1b Characteristics of PSM strategies for comparison two

PSM Strategy	Caliper Width SD of logit(PS)	E+ADT (n)	EBRT (n)	Events (n)
Unmatched	-	579	126	256
Nearest	No caliper	504	126	237
1	2.0	443	126	210
2	1.5	405	126	192
3	1.0	374	126	180
4	**0.8**	**361**	**126**	**175**
5	**0.6**	**352**	**126**	**173**
6	**0.5**	**347**	**126**	**171**
7	0.4	345	126	170
8	0.3	343	124	167
9	0.2	339	121	164
10	0.1	330	115	156
11	0.05	303	111	138
12	0.025	277	105	130
13	0.01	230	90	109
14	0.005	181	78	89

S Table 5.2a Coarsening of covariates used in CEM for comparison one

Variable	Coarsening	Matching Range Boundaries
PSA (ng/ml)		
	1	0, 2, 4, 6, 8, 10, 13, 16, 20
	2	0, 2, 4, 7, 10, 14, 20
	3	0, 4, 10, 20
Treatment Year		
	1	Exact
	2	1997, 2000, 2002, 2004, 2007
	3	1997, 2002, 2007

S Table 5.2b Coarsening of covariates used in CEM for comparison two

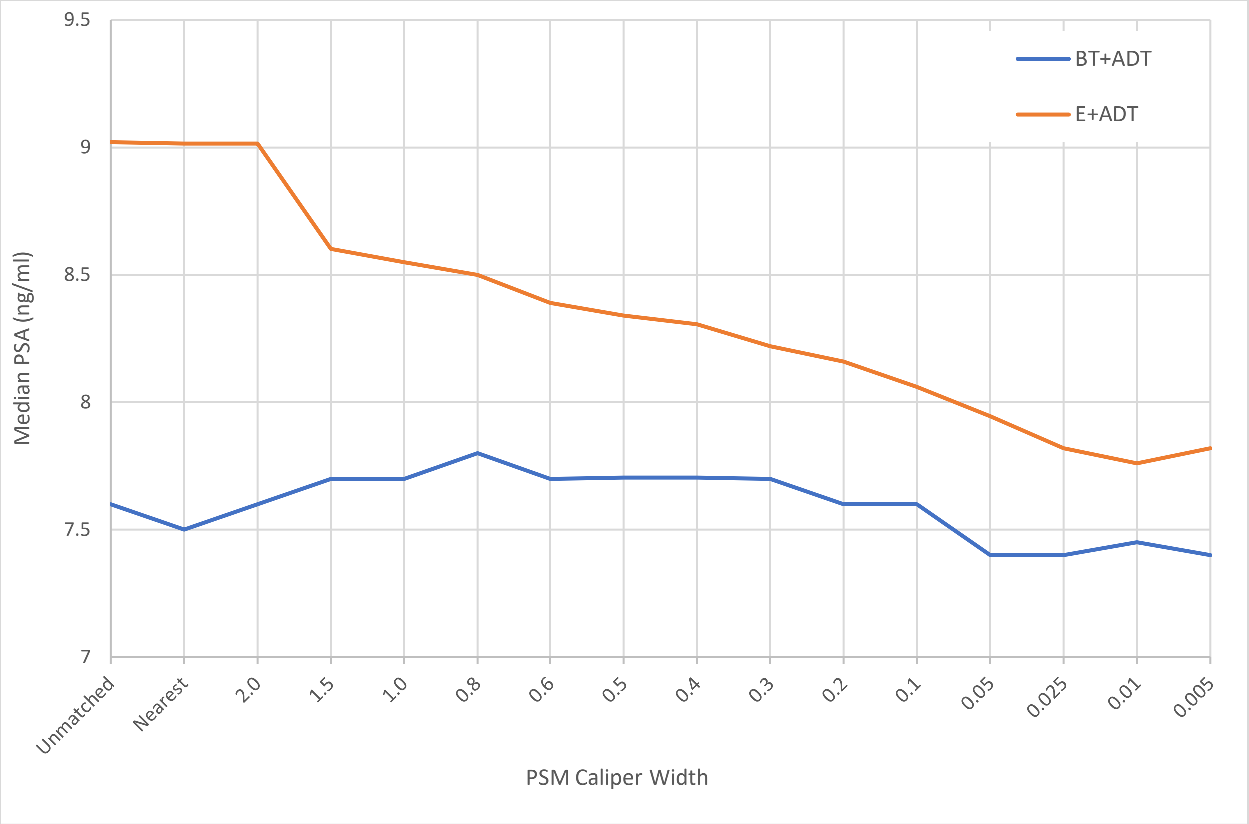
Variable	Coarsening	Matching Range Boundaries
PSA (ng/ml)		
	1	0, 2, 4, 6, 8, 10, 13, 16, 20
	2	0, 2, 4, 7, 10, 14, 20
	3	0, 4, 10, 20
Treatment Year		
	1	1995, 1997, 1999, 2001, 2003, 2006
	2	1995, 1999, 2002, 2006
	3	1993, 1999, 2006
EBRT Dose (Gy)		
	1	6600, 7300, 7980

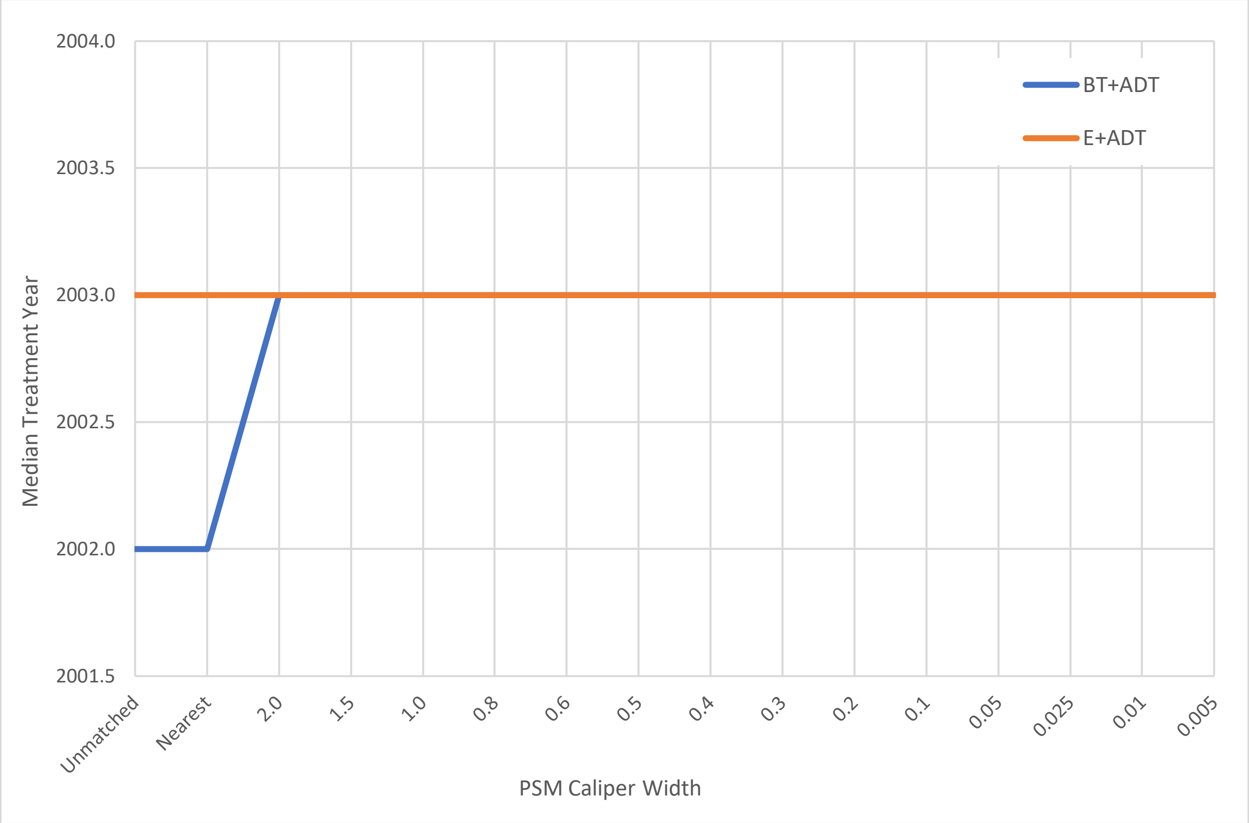
S Table 5.3a Characteristics of CEM strategies for comparison one

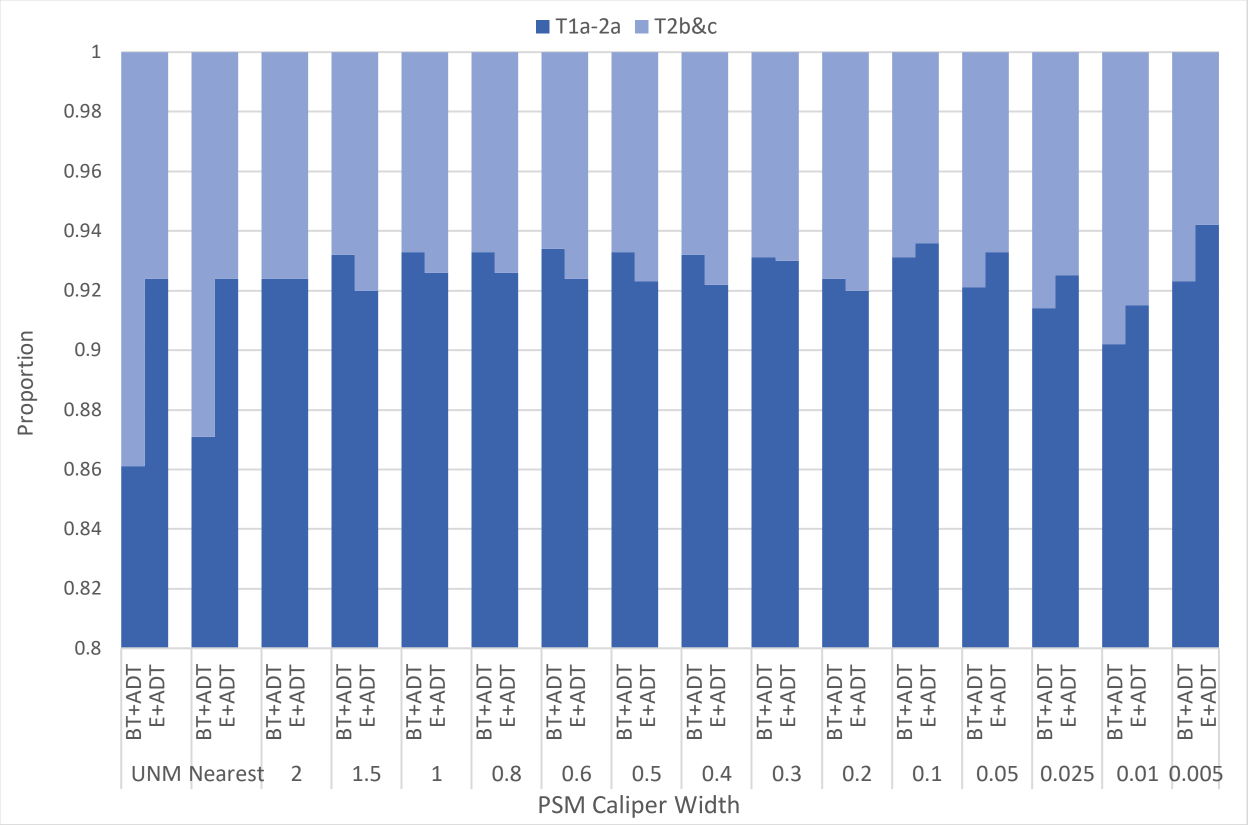
CEM Strategy	RT Year Coarsening	PSA Coarsening	BT+ADT (n)	E+ADT (n)	Events (n)
Unmatched	-	-	433	132	56
1	3	3	391	123	50
2	3	2	371	116	48
3	3	1	363	109	45
4	2	3	372	112	41
5	2	2	353	106	40
6	2	1	343	100	38
7	1	3	308	107	35
8	**1**	**2**	**276**	**96**	**34**
9	1	1	242	90	32

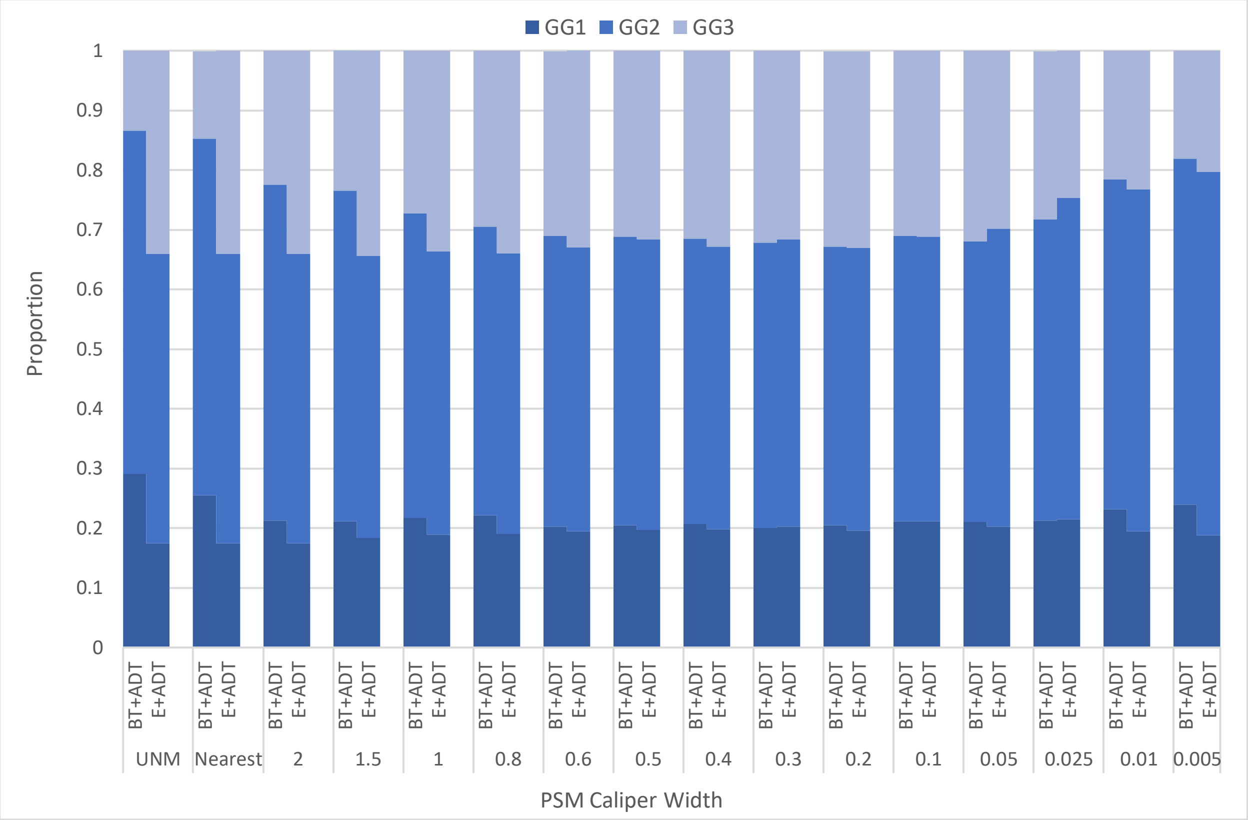
S Table 5.3b Characteristics of CEM strategies for comparison two

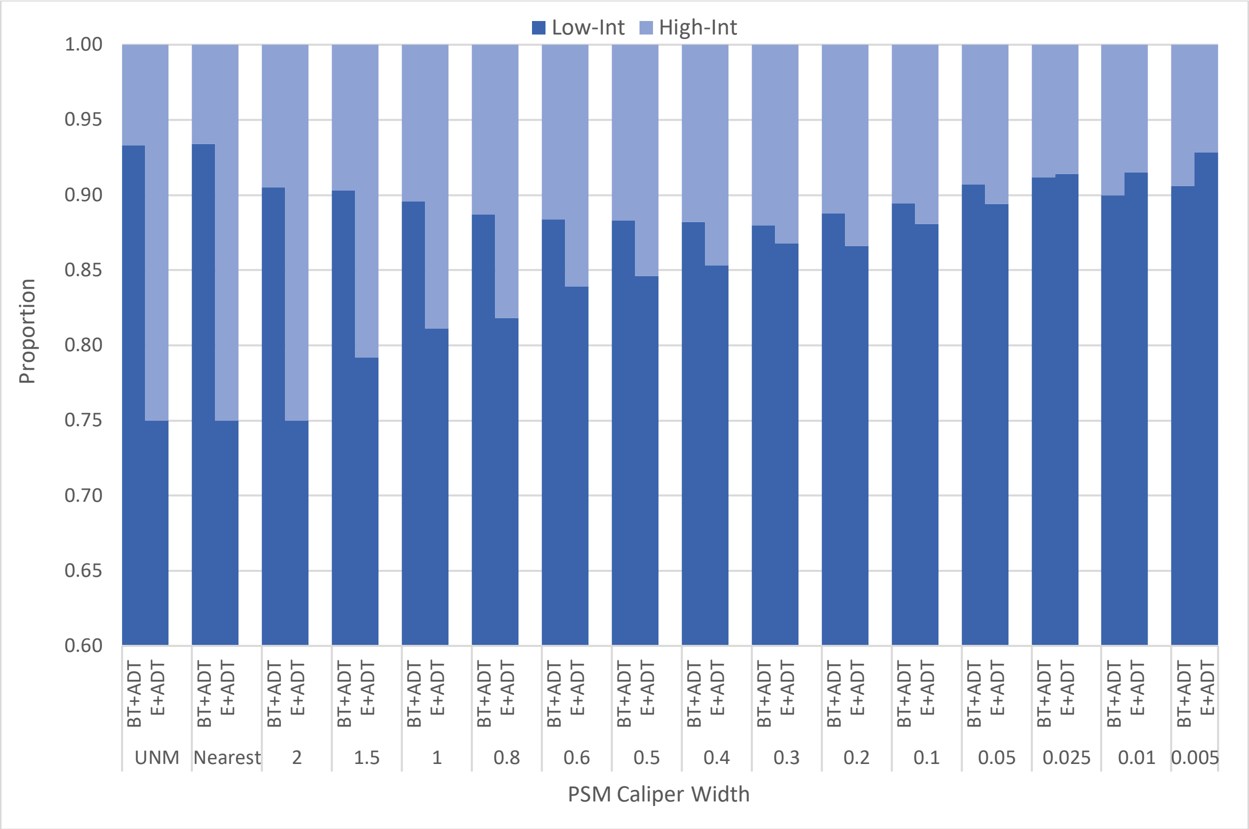
CEM Strategy	RT Year Coarsening	PSA Coarsening	EBRT Dose Coarsening	E+ADT (n)	EBRT (n)	Events (n)
Unmatched	-	-	-	579	126	256
1	**3**	**3**	**1**	**492**	**123**	**227**
2	3	2	1	445	122	209
3	3	1	1	436	119	203
4	3	3	1	439	122	211
5	3	2	1	398	122	194
6	3	1	1	373	118	180
7	2	3	1	377	118	176
8	2	2	1	337	116	160
9	2	1	1	295	108	138


9.5
9
8.5
8
7.5
7
Median PSA (ng/ml)
BT+ADT
E+ADT
Unmatched
Nearest
2.0
1.5
1.0
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

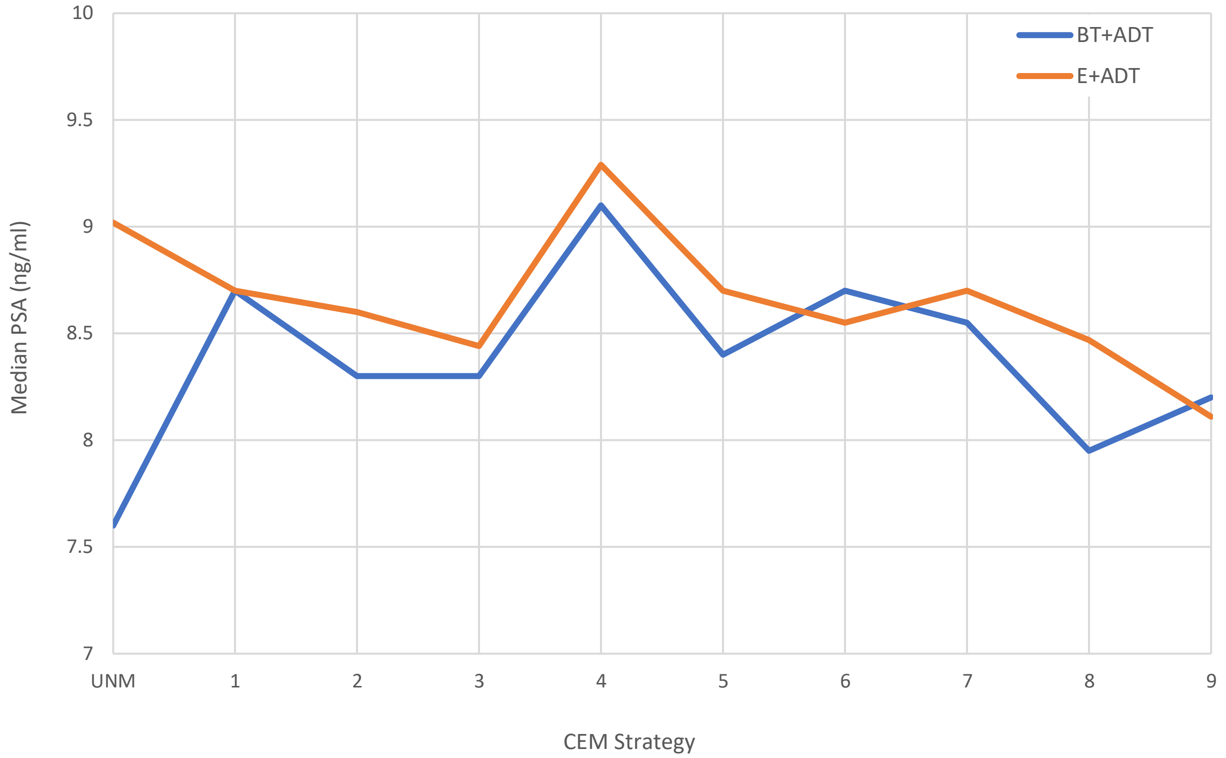

2004.0
2003.5
2003.0
2002.5
2002.0
2001.5
Median Treatment Year
BT+ADT
E+ADT
Unmatched
Nearest
2.0
1.5
1.0
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

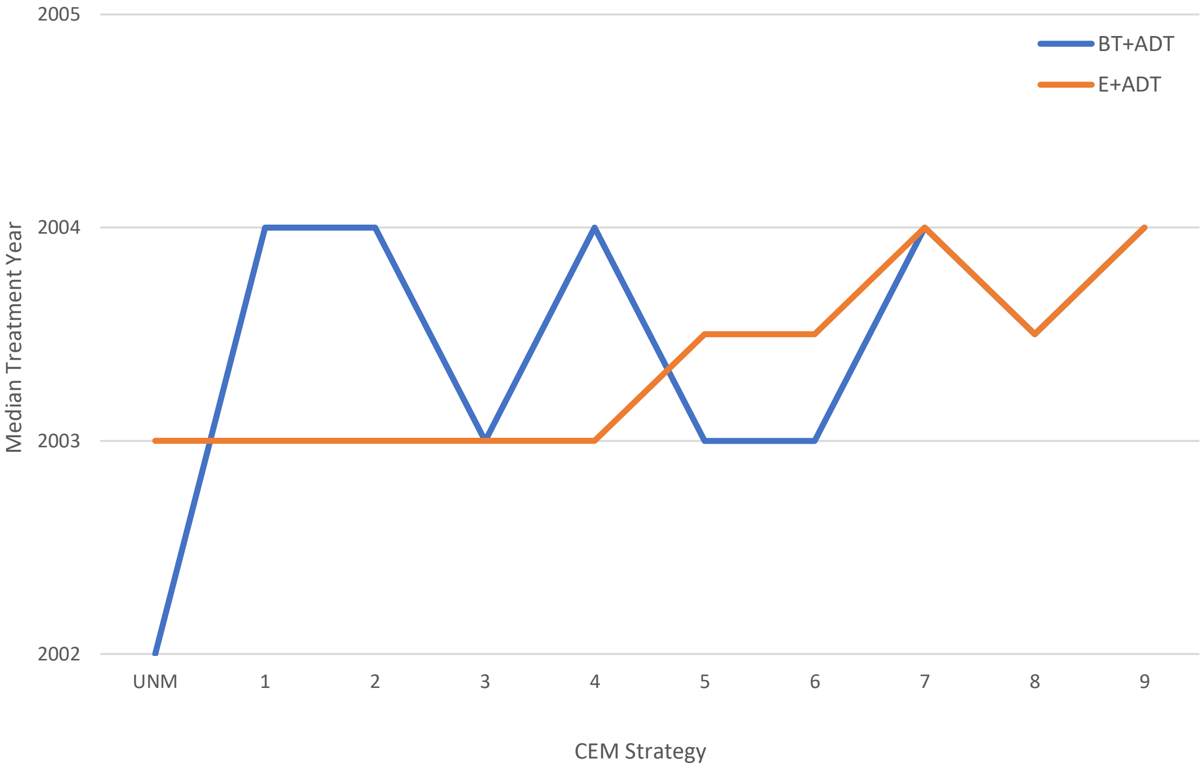

T1a-2a
T2b&c
Proportion
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
BT+ADT
E+ADT
UNM
Nearest
2
1.5
1
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

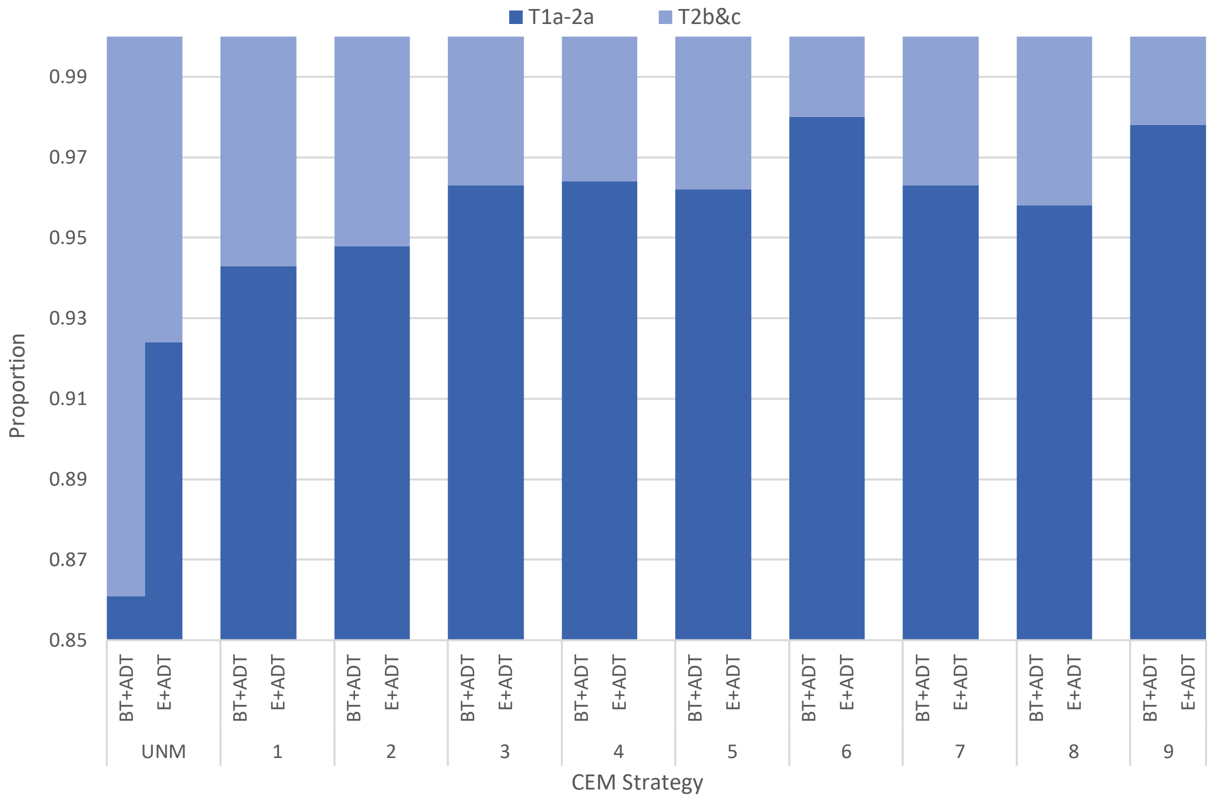

GG1
GG2
GG3
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BT+ADT
E+ADT
UNM
Nearest
2
1.5
1
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

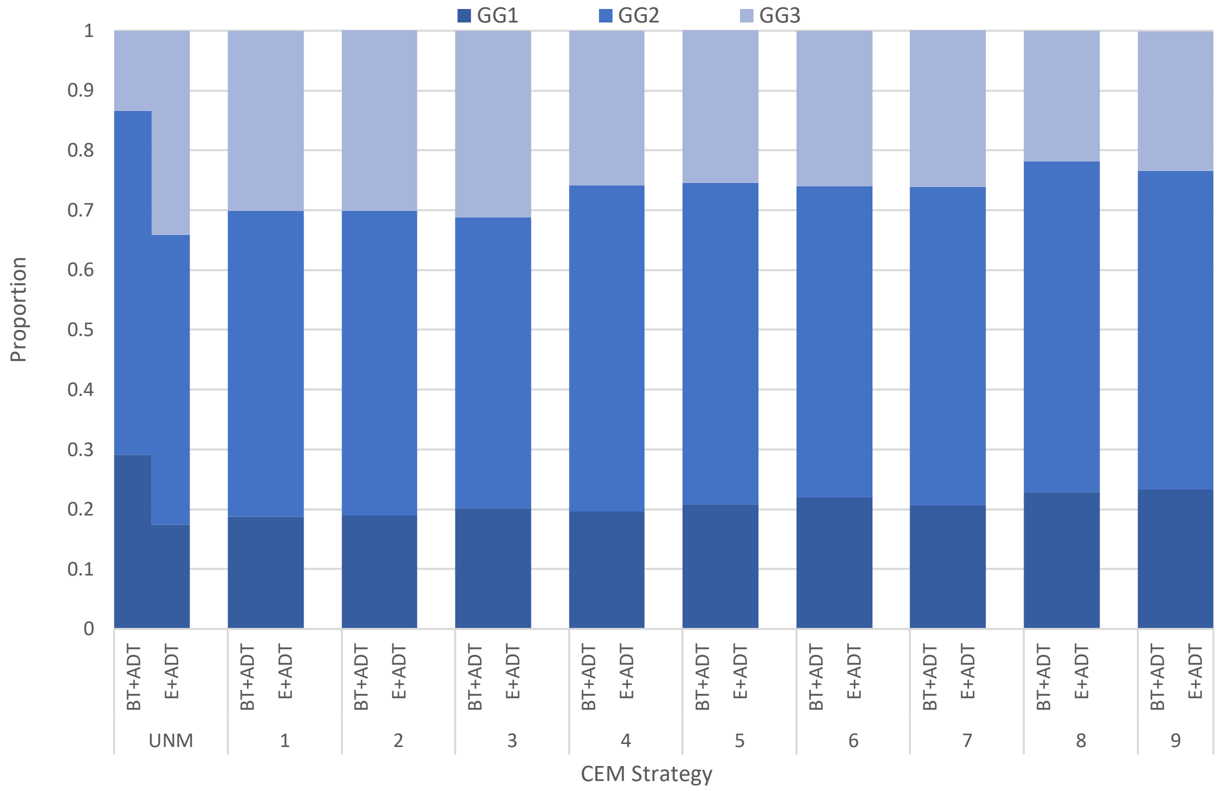


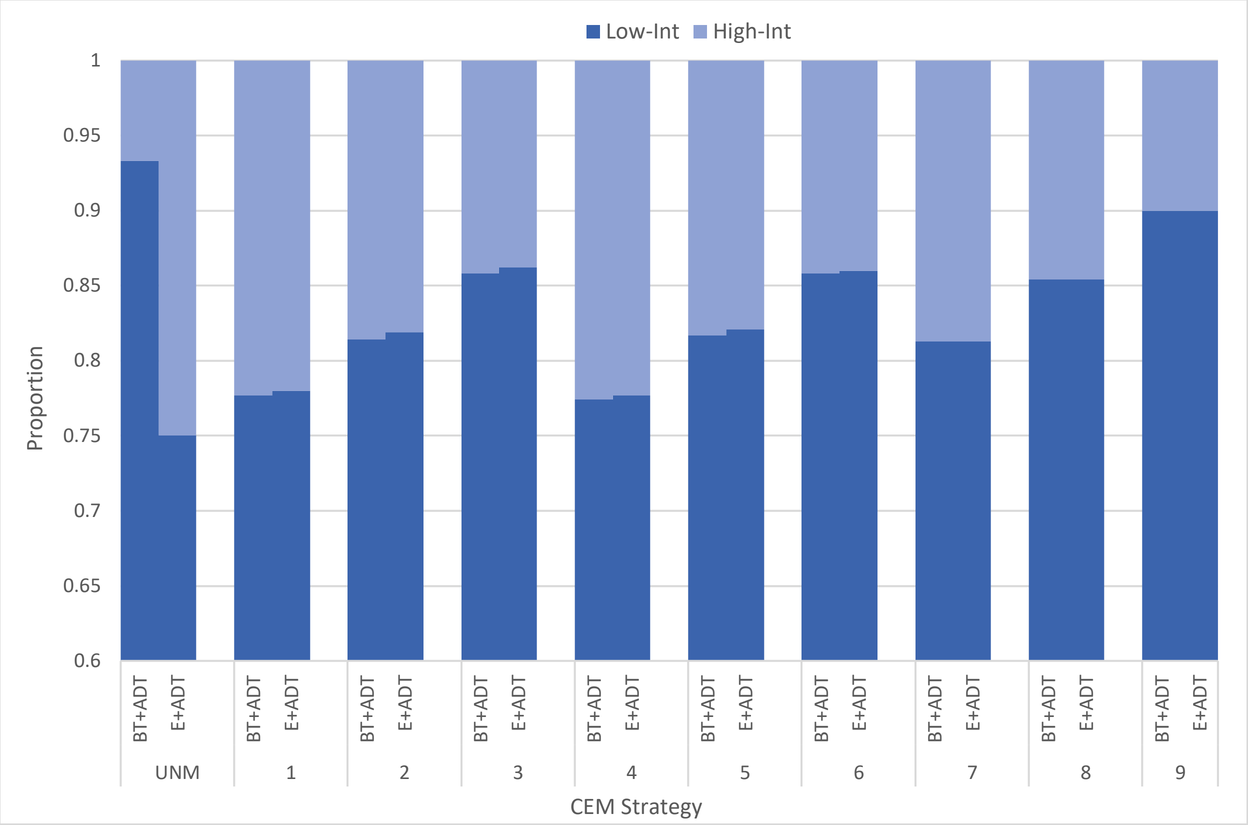
S Figure 5.1a Distribution of baseline covariates by PSM caliper width for comparison one


10
9.5
9
8.5
8
7.5
7
Median PSA (ng/ml)
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy
BT+ADT
E+ADT

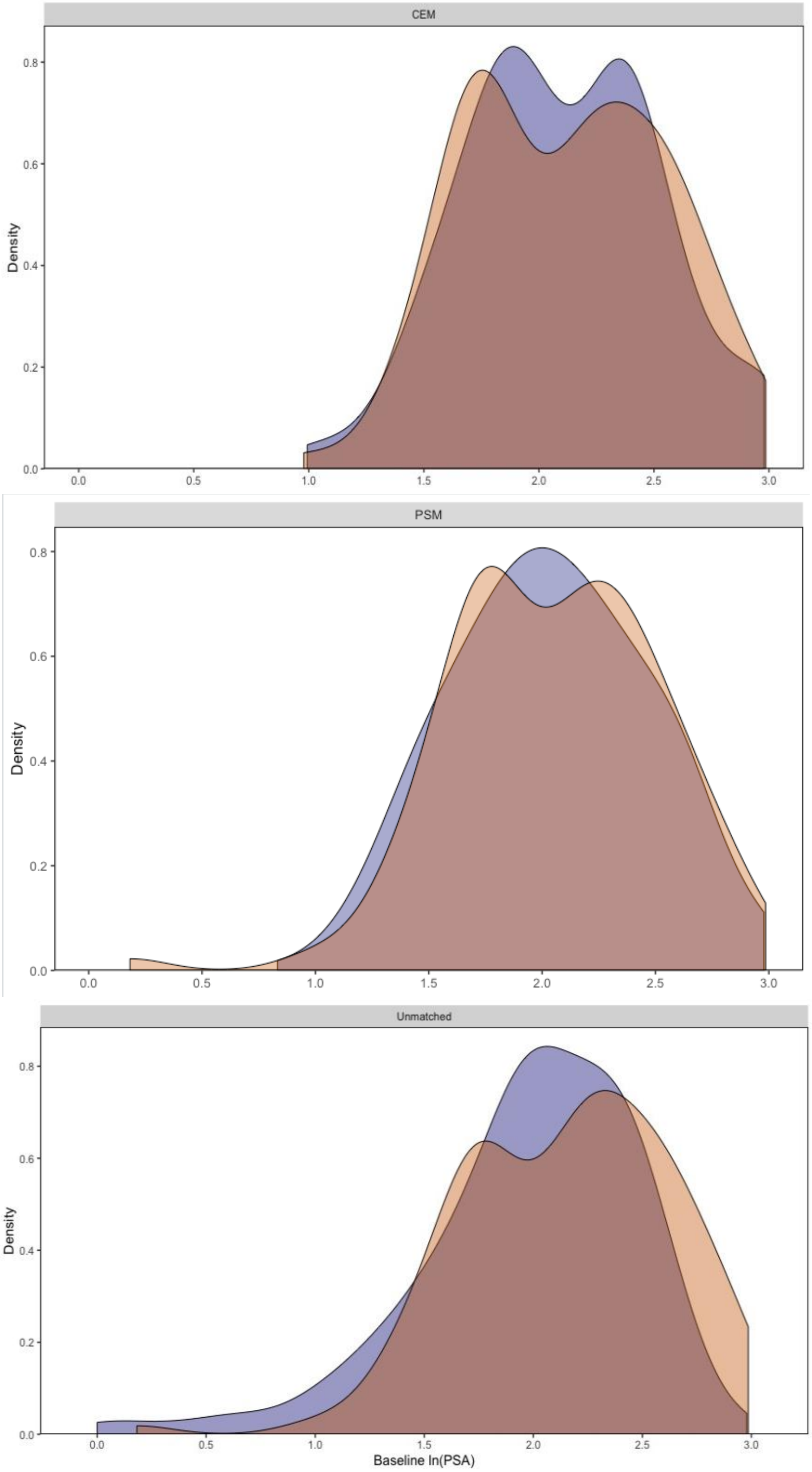

2005
2004
2003
2002
Median Treatment Year
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy
BT+ADT
E+ADT


T1a-2a
T2b&c
Proportion
0.99
0.97
0.95
0.93
0.91
0.89
0.87
0.85
BT+ADT
E+ADT
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy

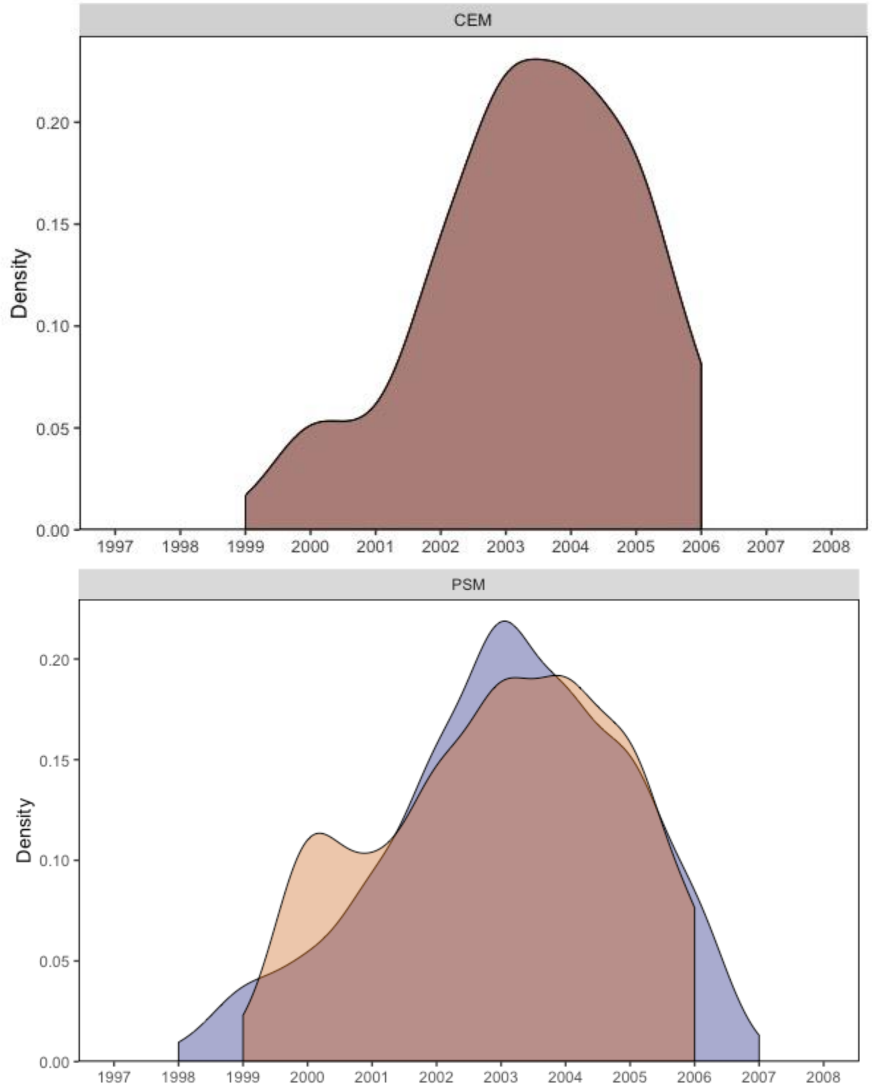

GG1
GG2
GG3
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BT+ADT
E+ADT
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy

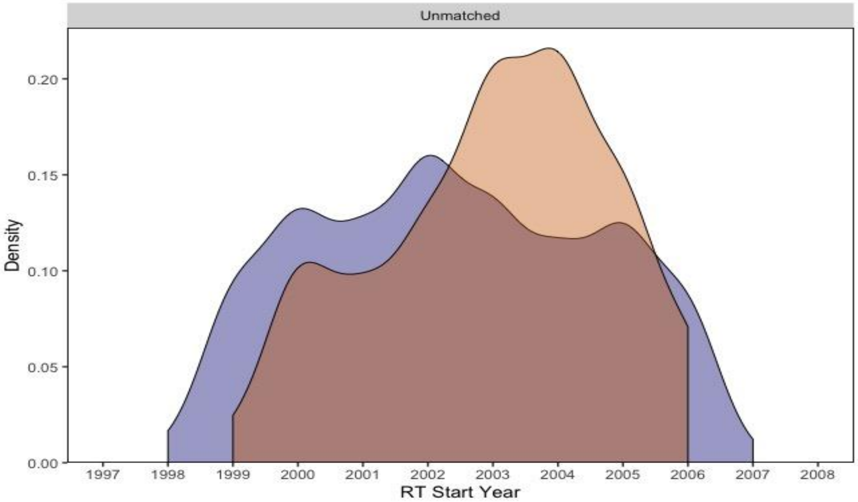


S Figure 5.1b Distribution of baseline covariates by CEM strategy for comparison one.

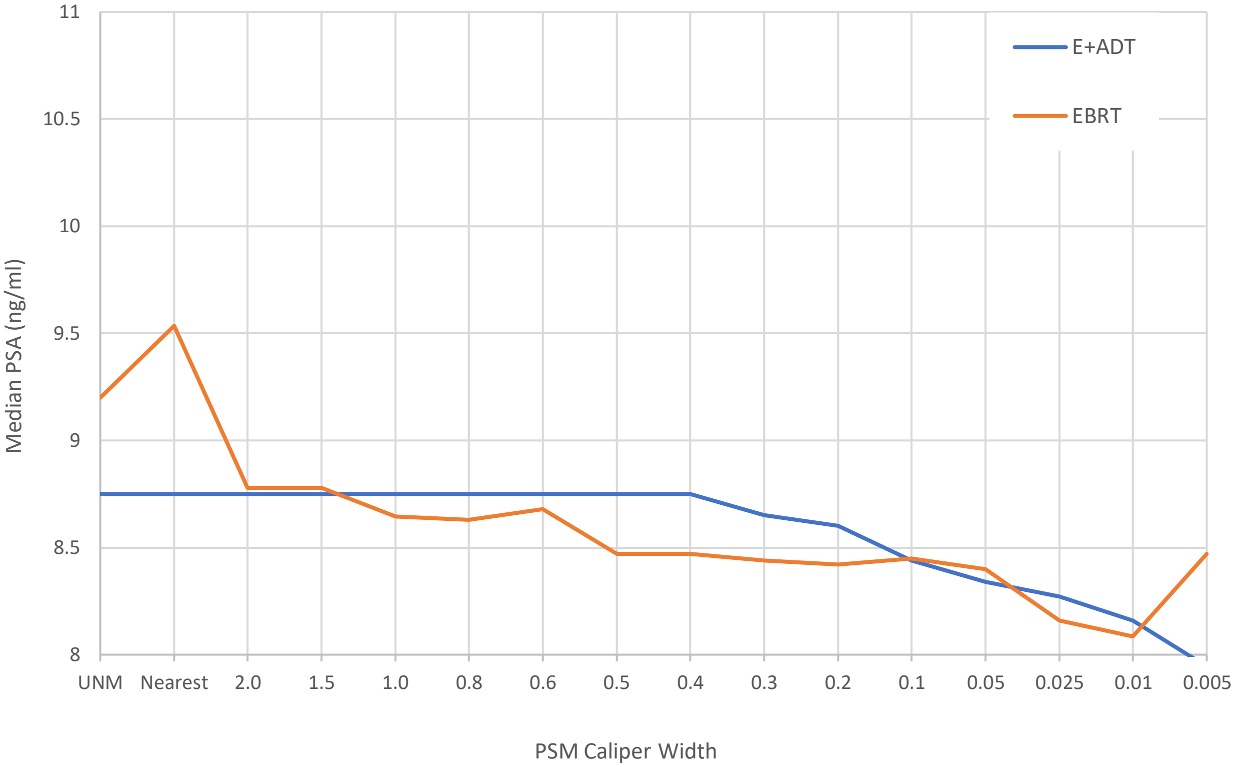

CEM
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
PSM
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Unmatched
Density
0.0
0.2
0.4
0.6
0.8
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Baseline ln(PSA)

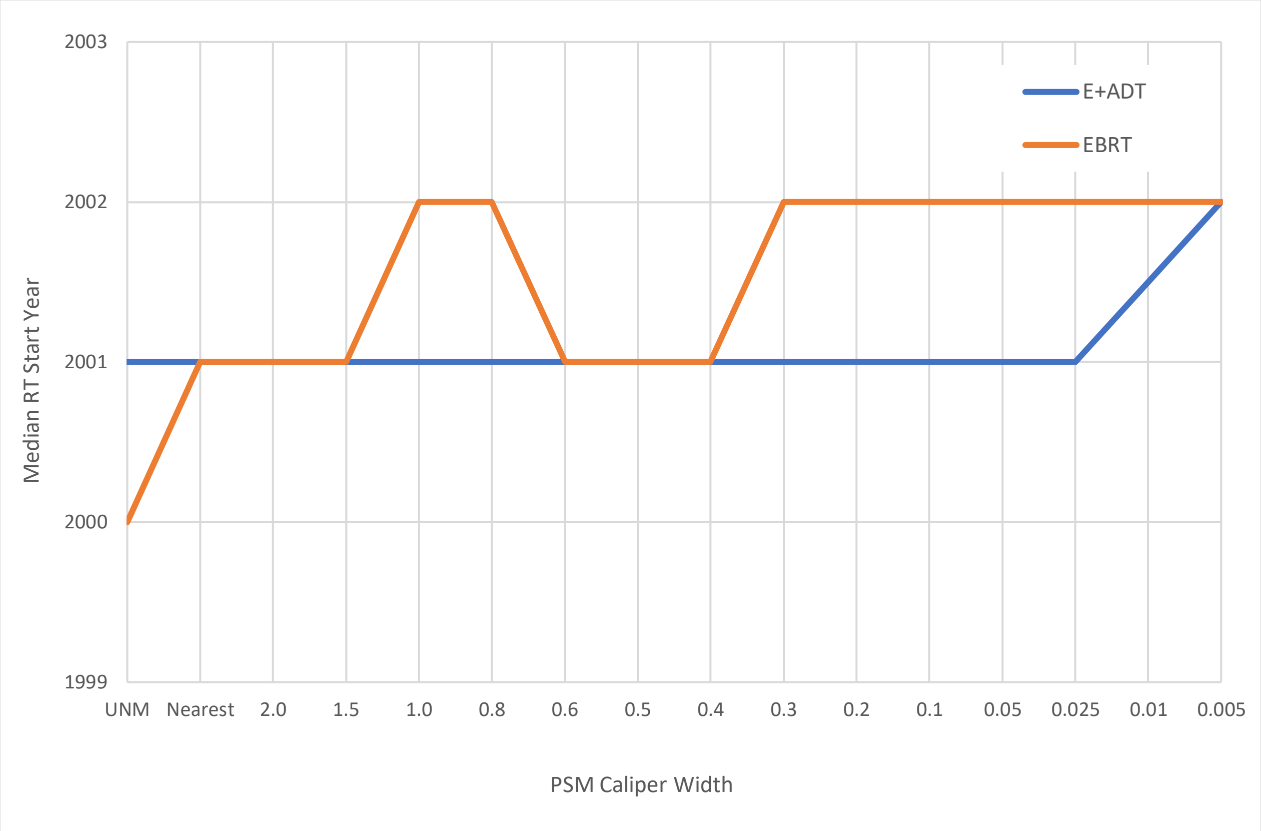
S Figure 5.2a Distribution of baseline ln(PSA) in BT+ADT (blue) and E+ADT (orange) groups for unmatched samples and for samples obtained after PSM and CEM.

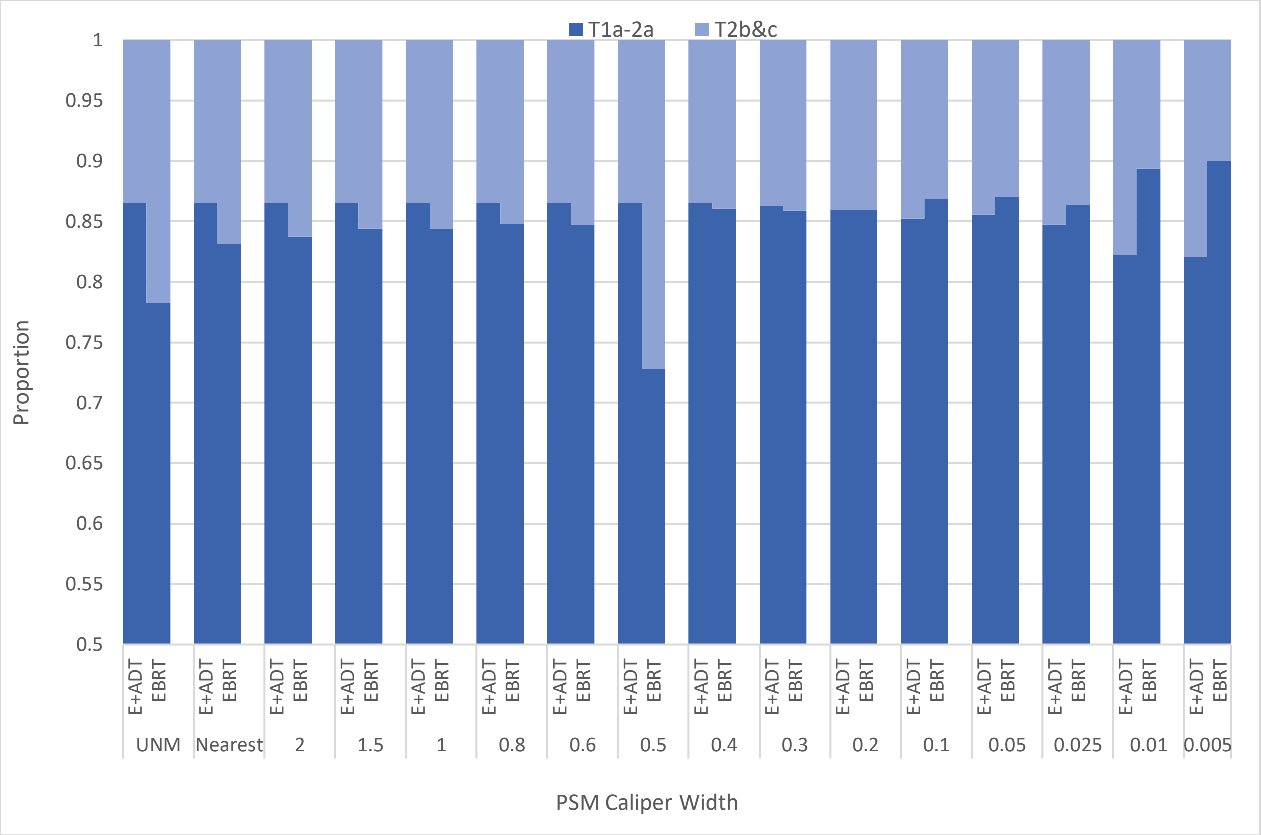


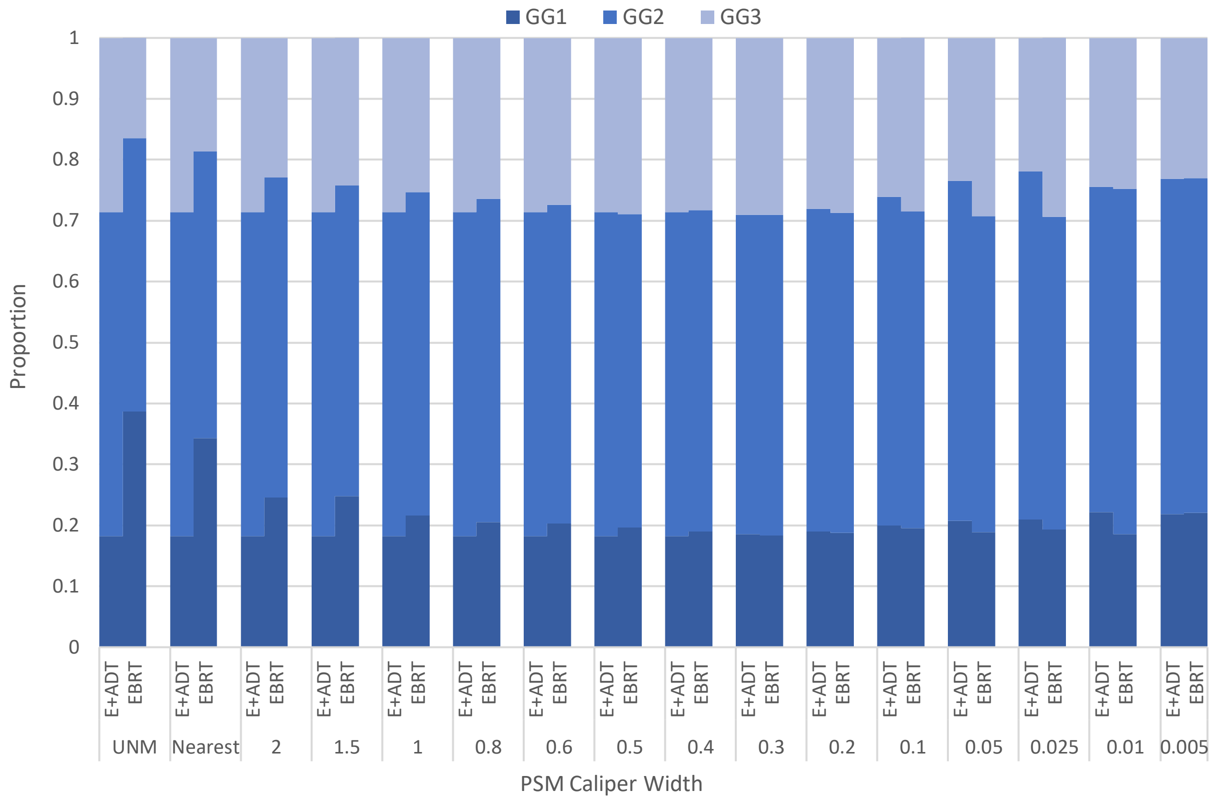


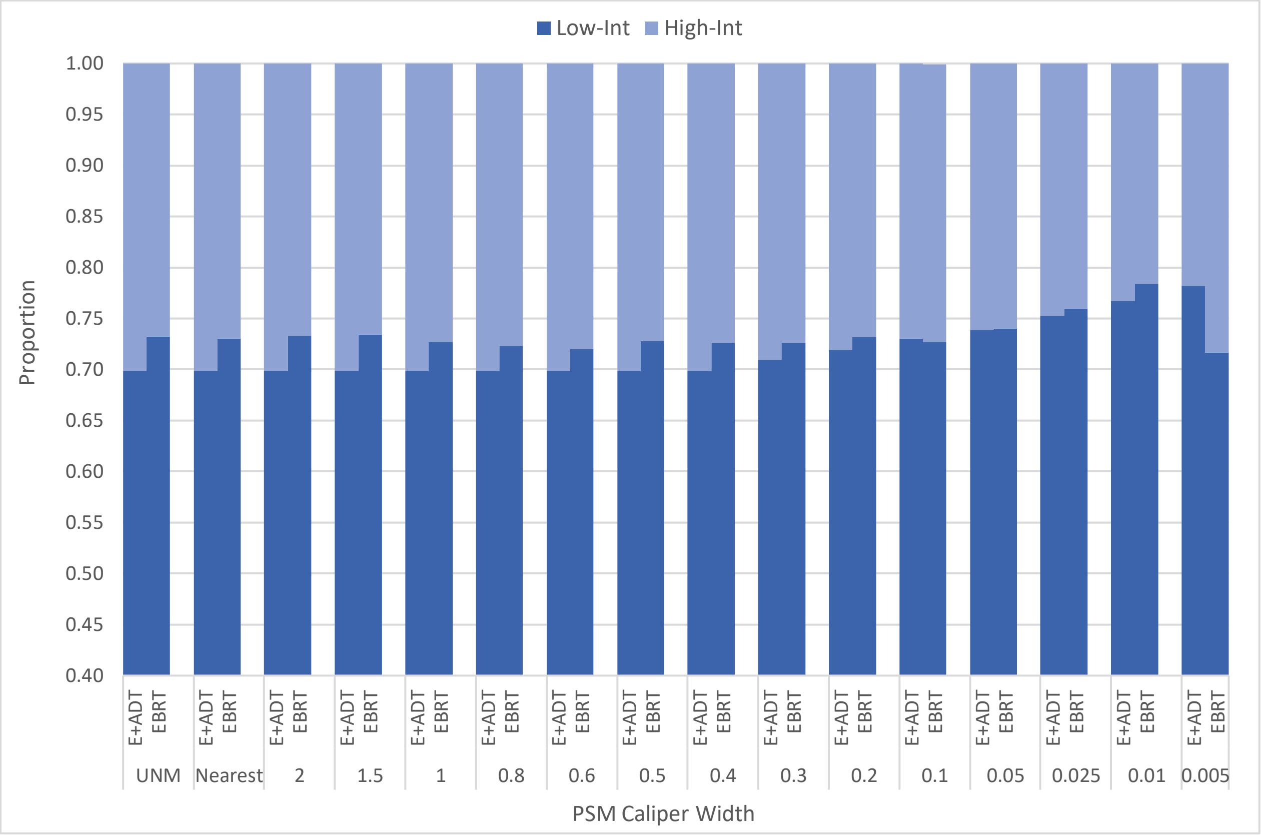
S Figure 5.2b Distribution of RT start year in BT+ADT (blue) and EBRT+ADT (orange) groups for unmatched samples and for samples obtained after PSM and CEM



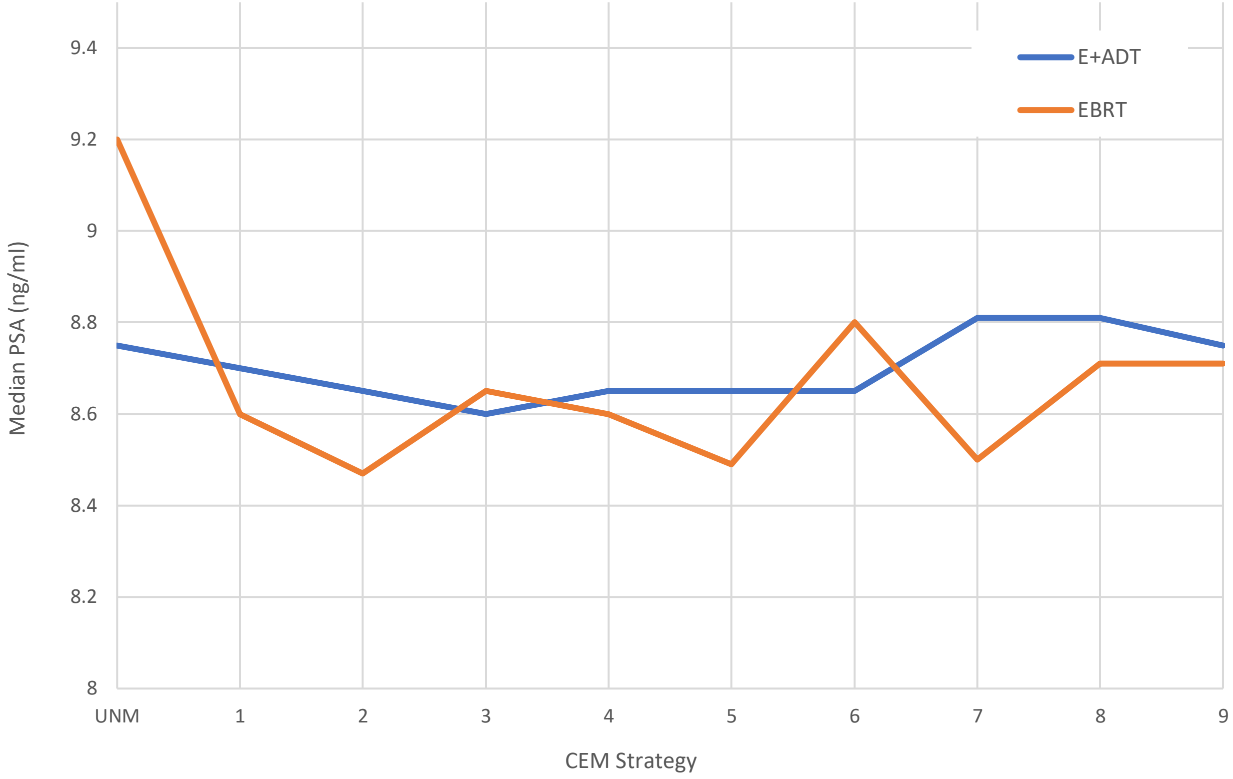

2003
2002
2001
2000
1999
Median RT Start Year
E+ADT
EBRT
UNM
Nearest
2.0
1.5
1.0
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

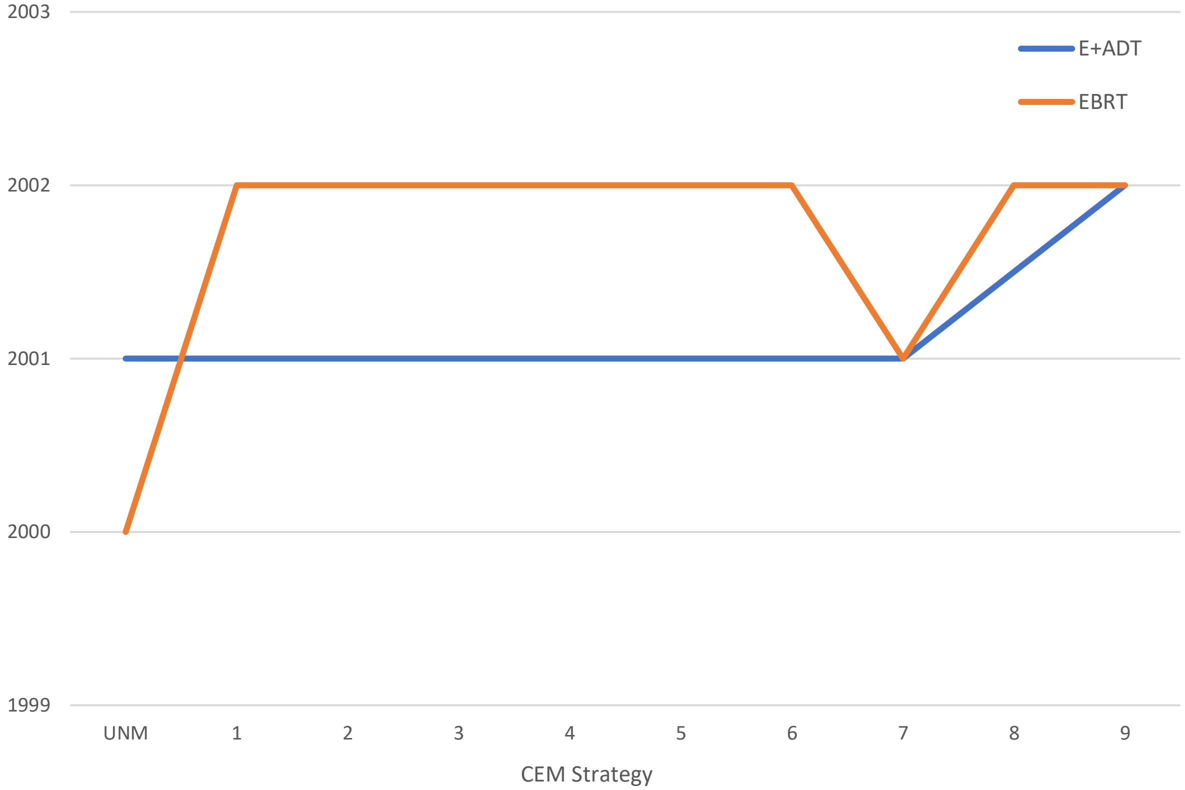

T1a-2a
T2b&c
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Proportion
E+ADT
EBRT
UNM
Nearest
2
1.5
1
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

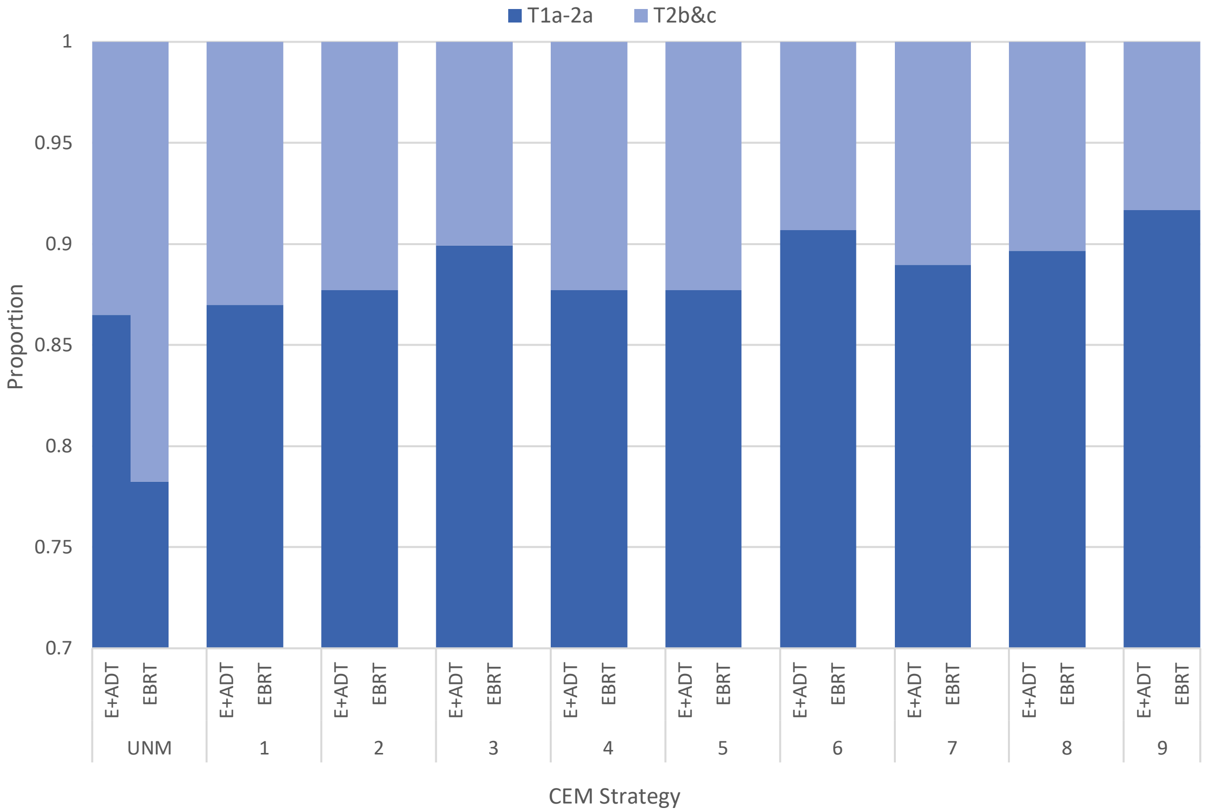

■ GG1 ■ GG2 ■ GG3
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
E+ADT
EBRT
UNM
Nearest
2
1.5
1
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

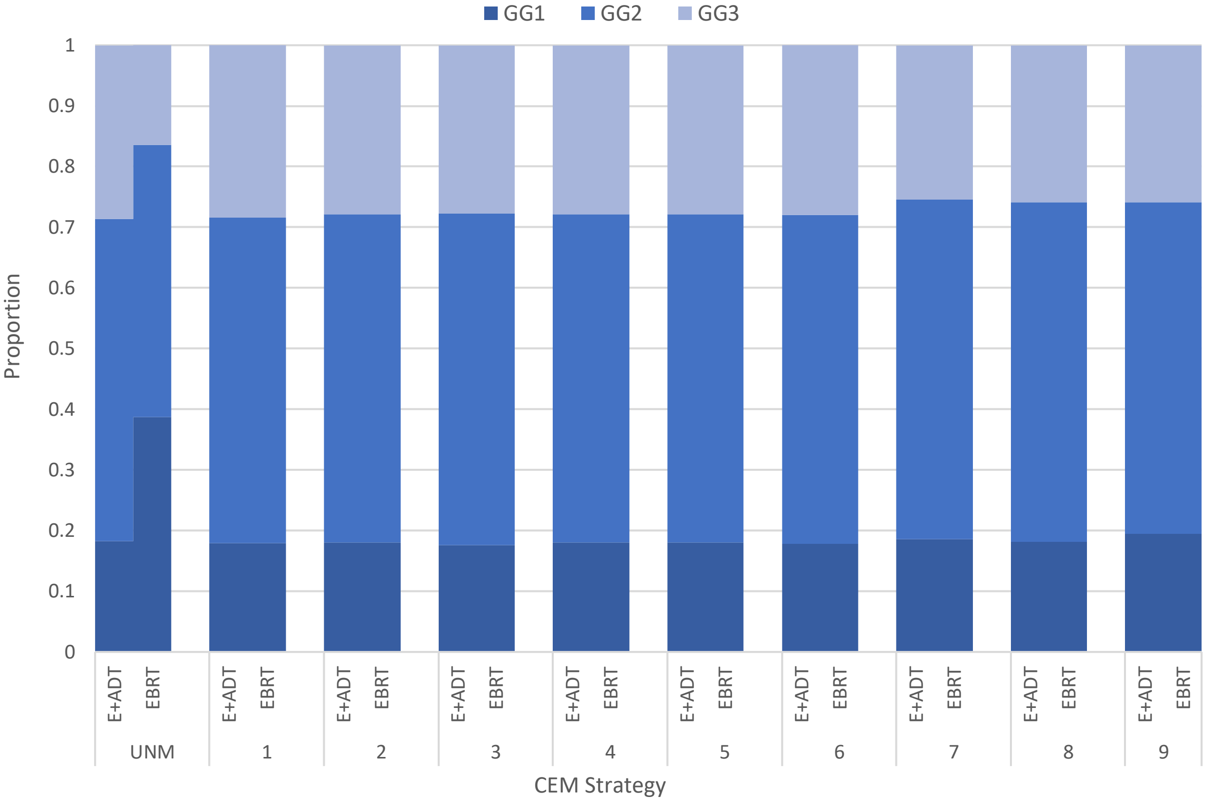

■ Low-Int ■ High-Int
Proportion
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
E+ADT
EBRT
UNM
Nearest
2
1.5
1
0.8
0.6
0.5
0.4
0.3
0.2
0.1
0.05
0.025
0.01
0.005
PSM Caliper Width

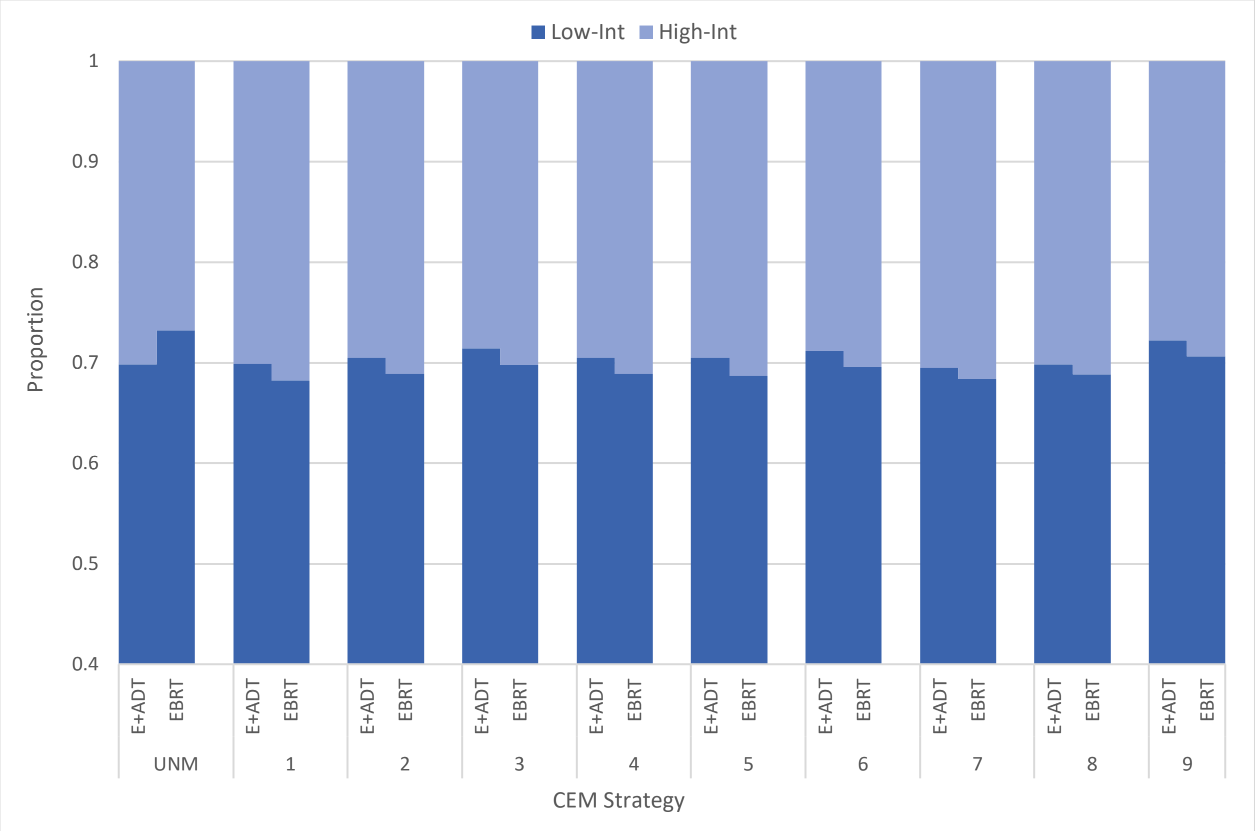
S Figure 5.3a Distribution of baseline covariates by PSM caliper width for comparison two.


Median PSA (ng/ml)
9.4
9.2
9
8.8
8.6
8.4
8.2
8
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy
E+ADT
EBRT

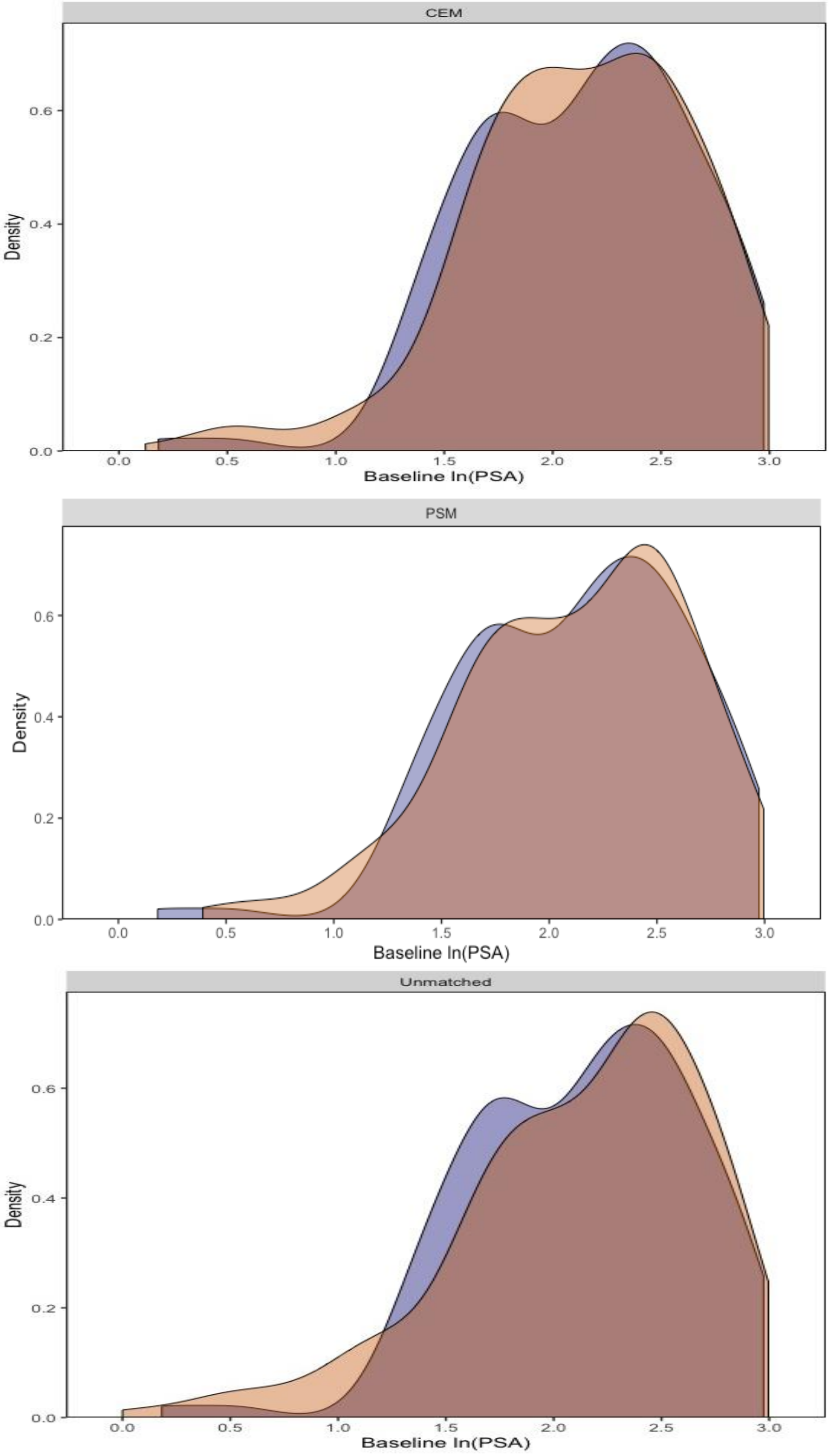

Median RT Start Year
2003
2002
2001
2000
1999
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy
E+ADT
EBRT


T1a-2a
T2b&c
Proportion
1
0.95
0.9
0.85
0.8
0.75
0.7
E+ADT
EBRT
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy

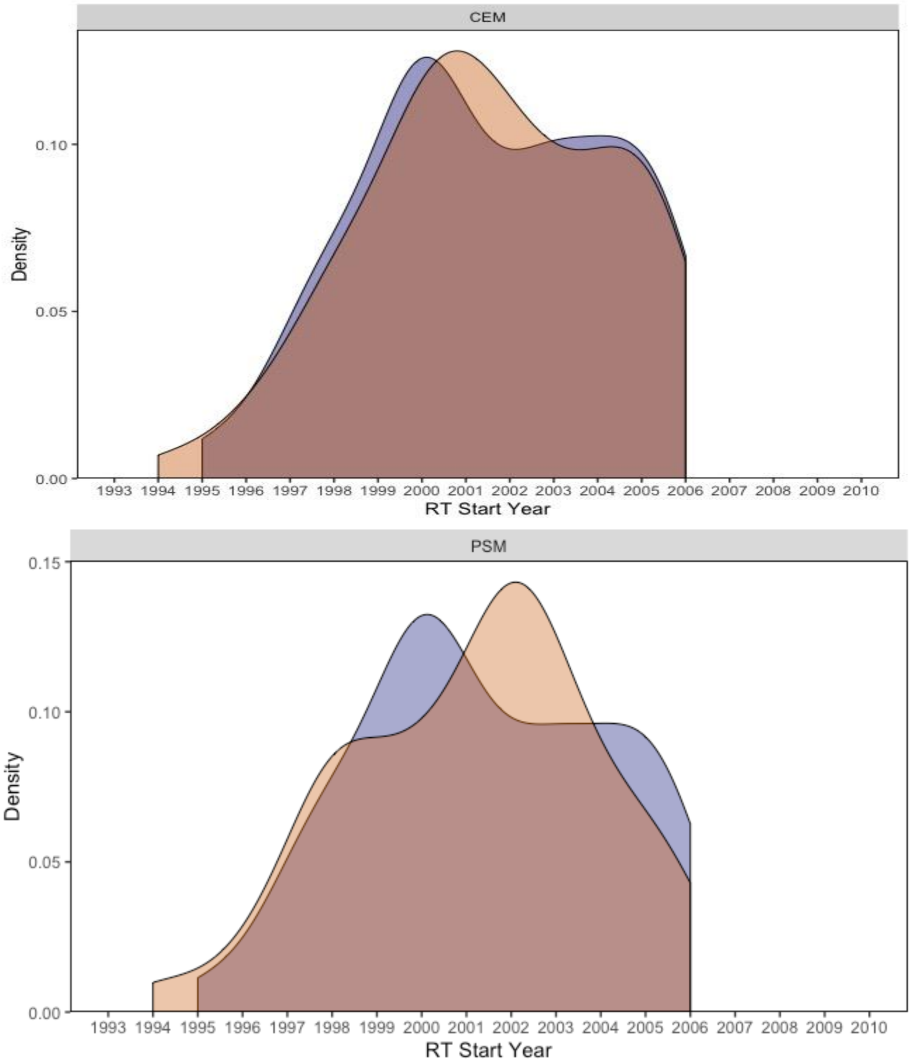

GG1
GG2
GG3
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
E+ADT
EBRT
UNM
1
2
3
4
5
6
7
8
9
CEM Strategy

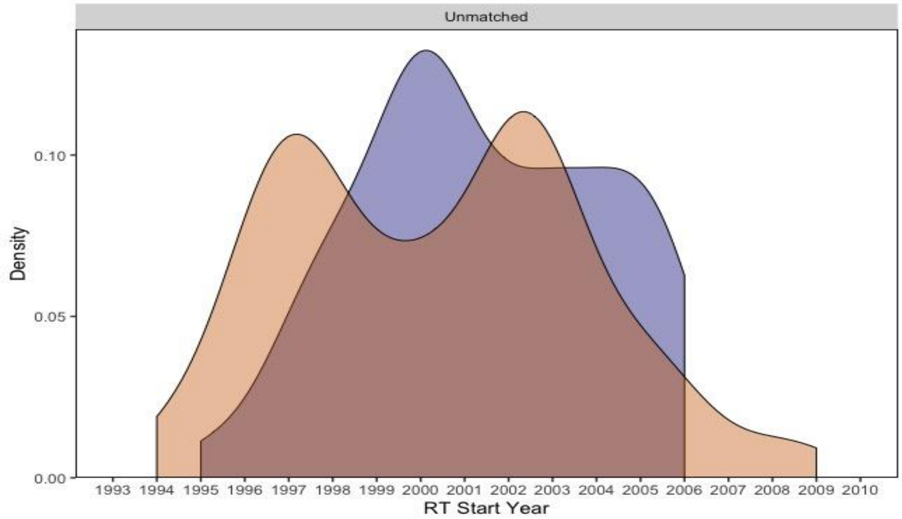


S Figure 5.3b Distribution of baseline covariates by CEM strategy for comparison two.

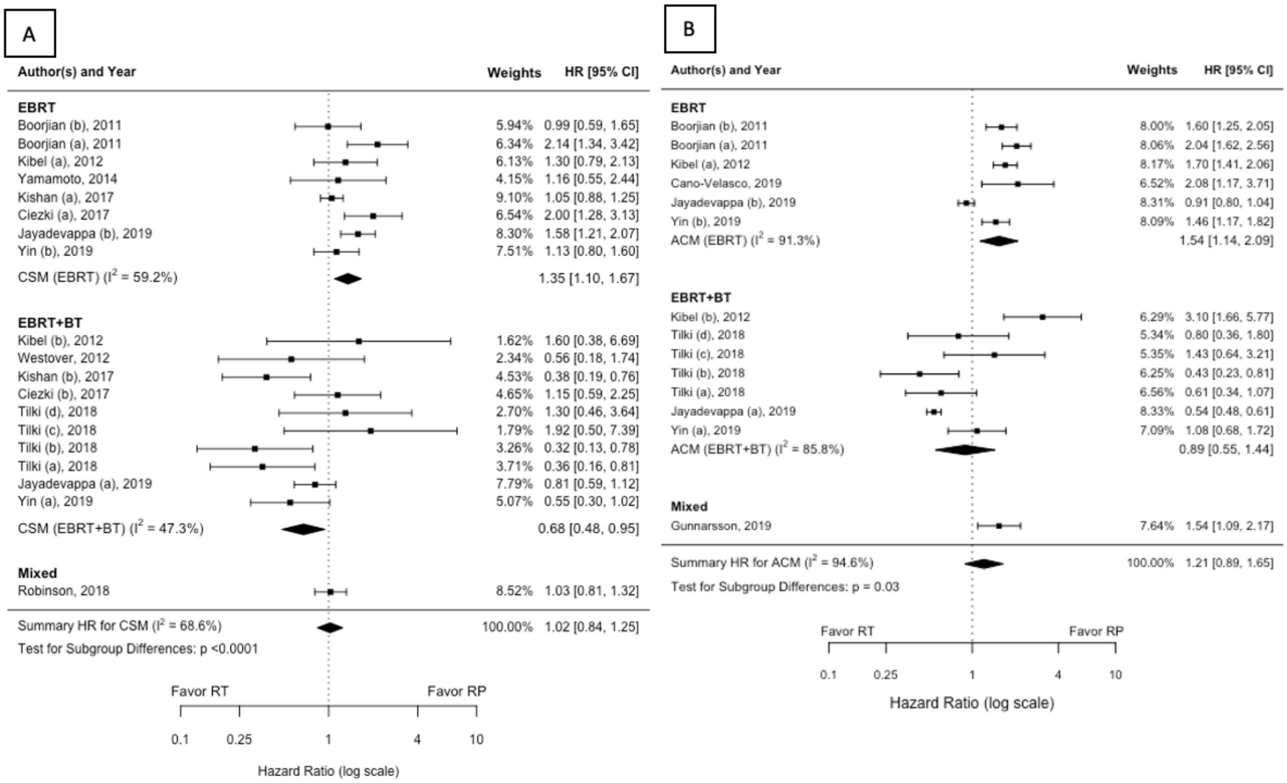

CEM
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Baseline ln(PSA)
PSM
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Baseline ln(PSA)
Unmatched
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Baseline ln(PSA)

S Figure 5.4a Distribution of baseline ln(PSA) in E+ADT (blue) and EBRT (orange) groups for unmatched samples and for samples obtained after PSM and CEM



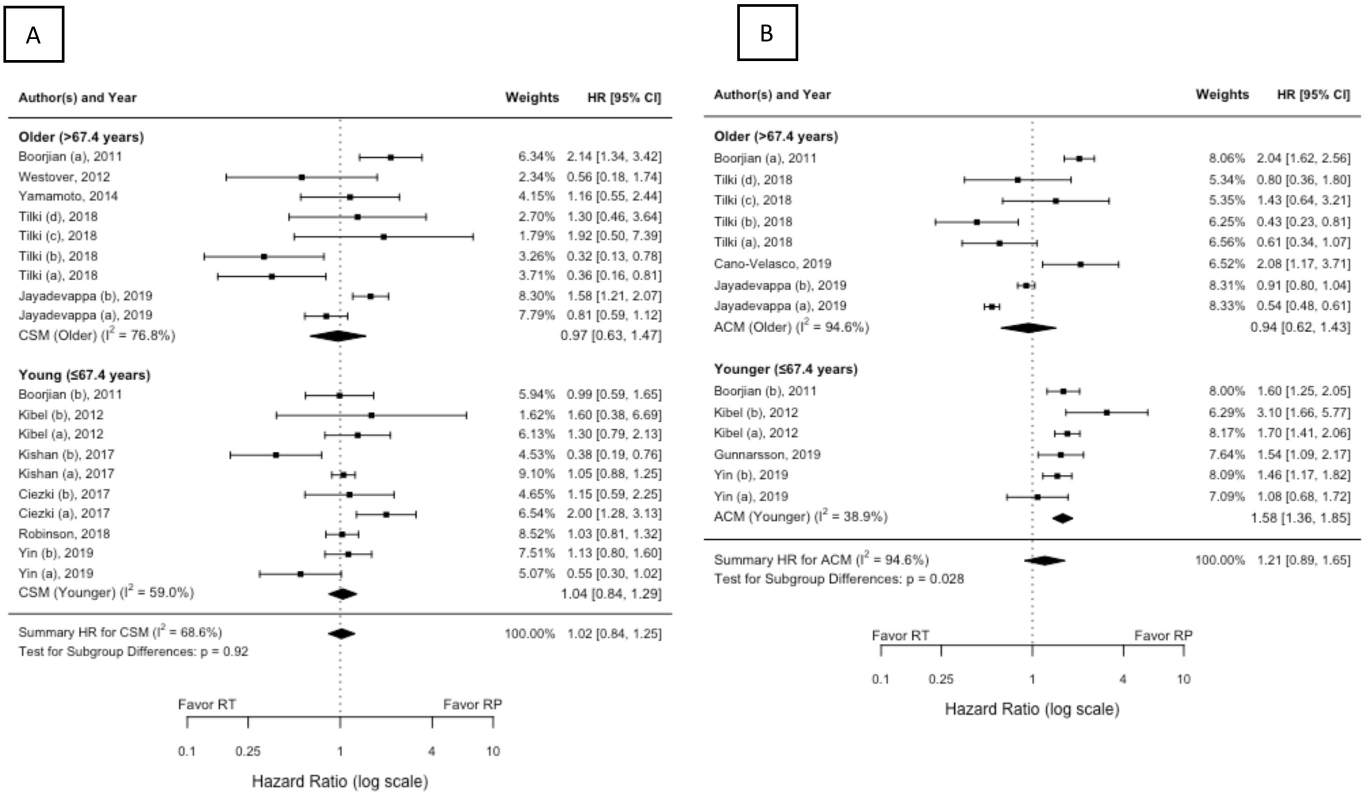


S Figure 5.4b Distribution of RT start year in E+ADT (blue) and EBRT (orange) groups for unmatched samples and for samples obtained after PSM and CEM



S Figure 6.1 Forest plot showing subgroup effects for EBRT and EBRT+BT assessing the risk of (a) prostate cancer-specific mortality and (b) all-cause mortality following radiotherapy and surgery for prostate cancer

Abbreviations: HR = hazard ratio; CI = confidence interval; RP = radical prostatectomy; RT = radiation therapy; EBRT = external beam radiation therapy; BT = brachytherapy; CSM = cancer-specific mortality; ACM = all-cause mortality

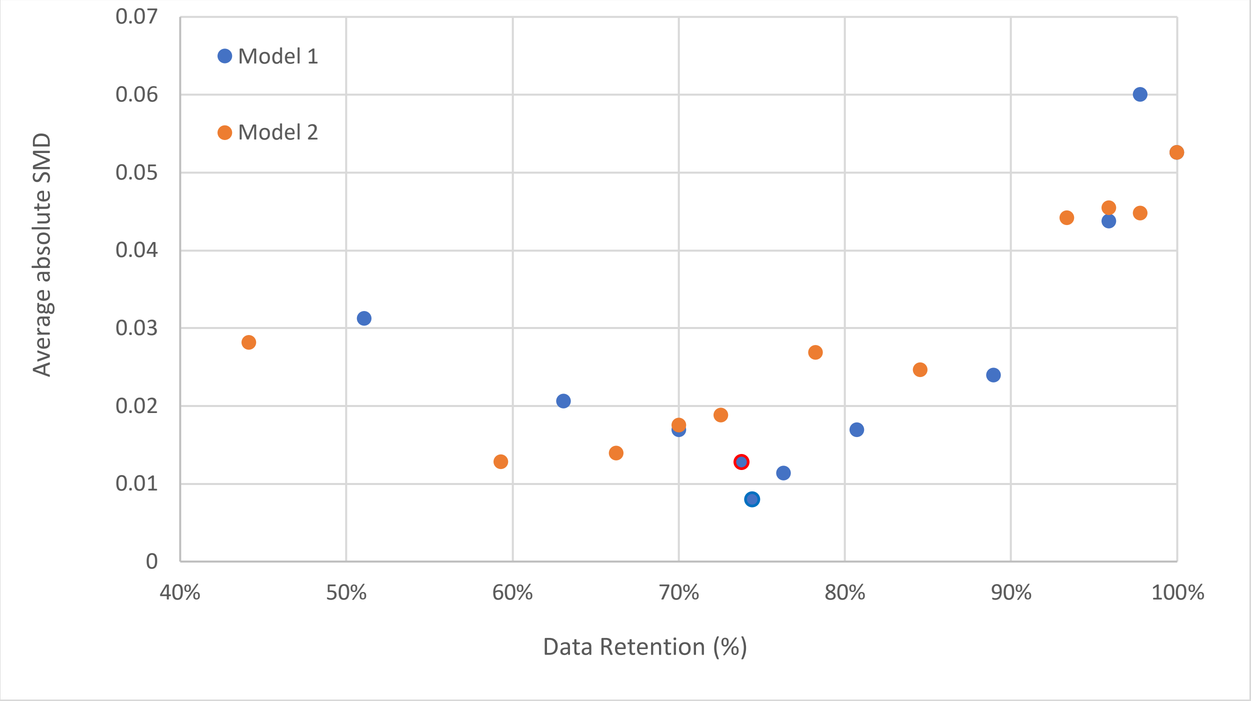


S Figure 6.2 Forest plot showing subgroup effects for studies conducted among younger and older patient groups assessing the risk of (a) prostate cancer-specific mortality and (b) all-cause mortality following radiotherapy and surgery for prostate cancer

Abbreviations: HR = hazard ratio; CI = confidence interval; RP = radical prostatectomy; RT = radiation therapy; CSM = cancer-specific mortality; ACM = all-cause mortality

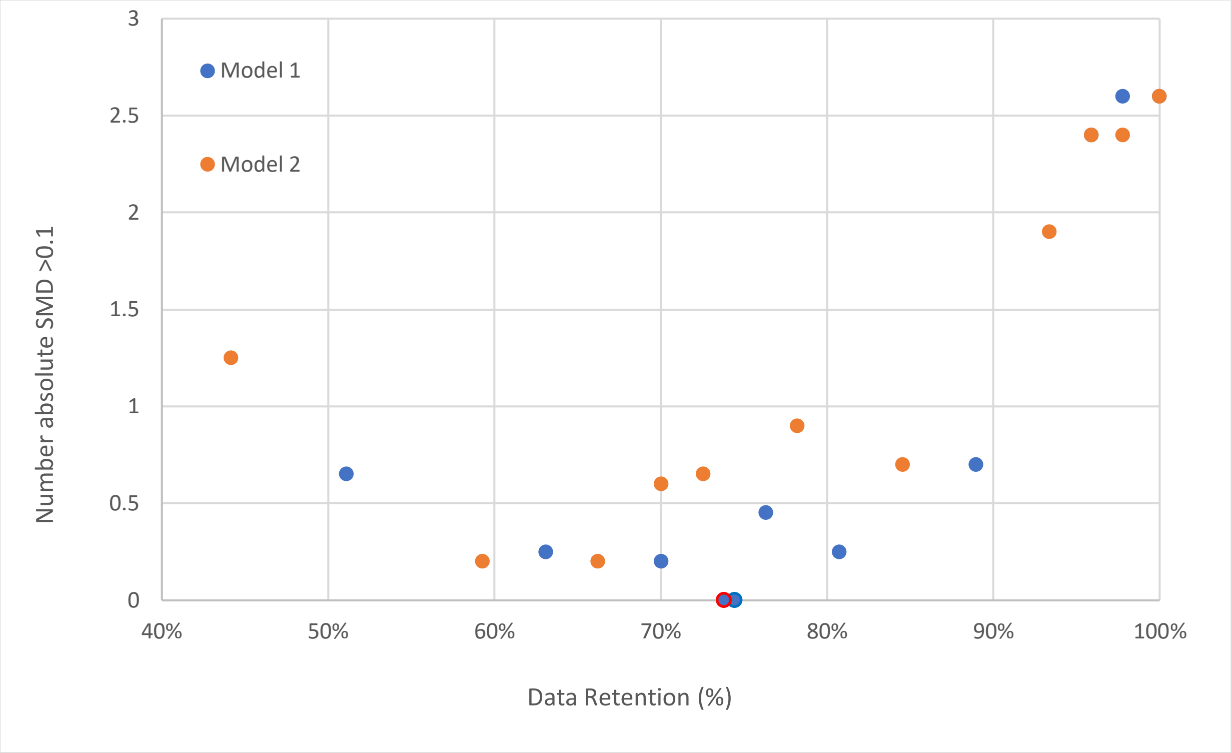
S Table 7.1 Ranges of coarsened variables

Variable	Ranges
PSA (ng/ml)	0, 20, 100, 300
	0, 20, 30, 50, 100, 300
	0, 6, 10, 20, 30, 50, 100, 300
Gleason Score	6, 8, 10
	6, 7, 8, 9, 10
Clinical Tumor Stage	1, 2, 4,
	1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4a and 4b



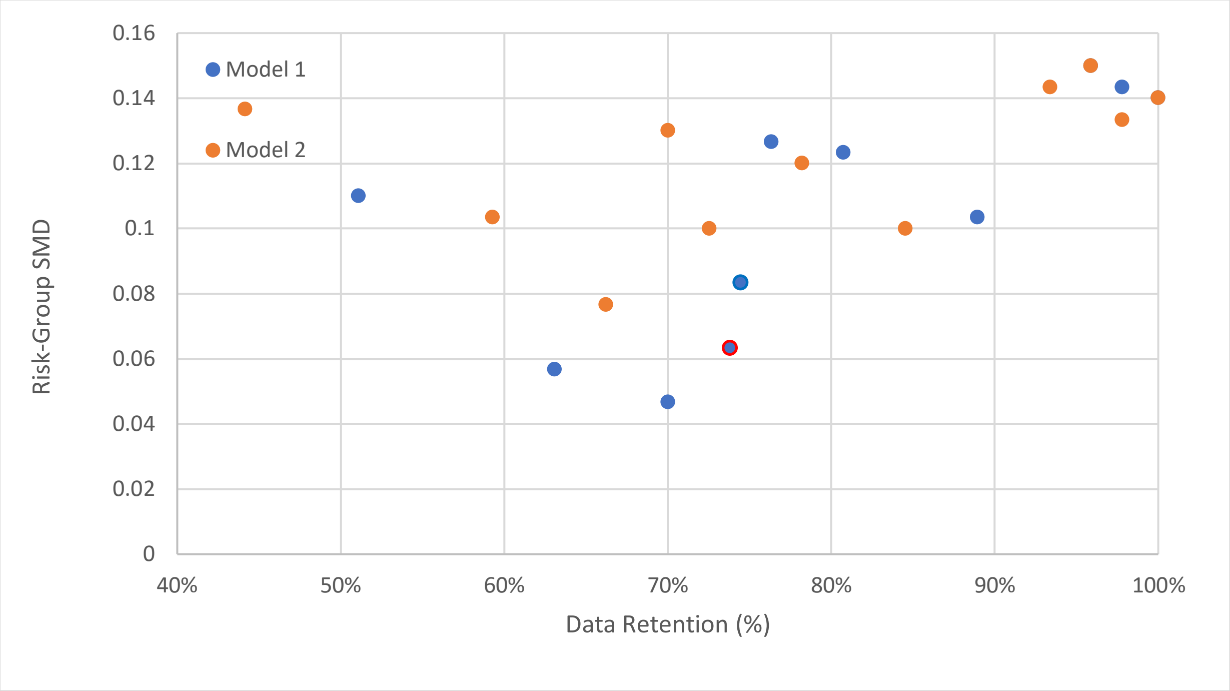
S Figure 7.1a The average absolute standardized mean difference (SMD) is plotted per level of data retention for each matching caliper used in propensity score matching.

*The red dot indicates the matching strategy chosen.



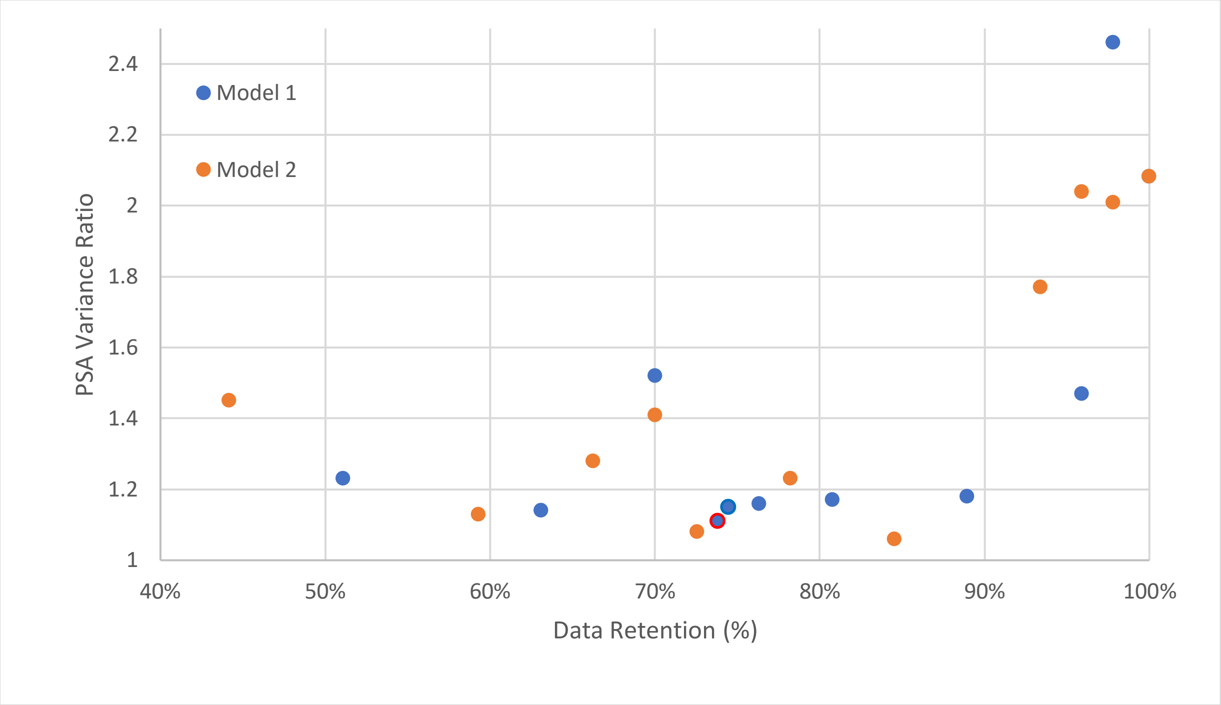
S Figure 7.1b The number of absolute standardized mean differences (SMD) that exceed the threshold of 0.1 is plotted per level of data retention for each matching caliper used in propensity score matching.

*The red dot indicates the matching strategy chosen.



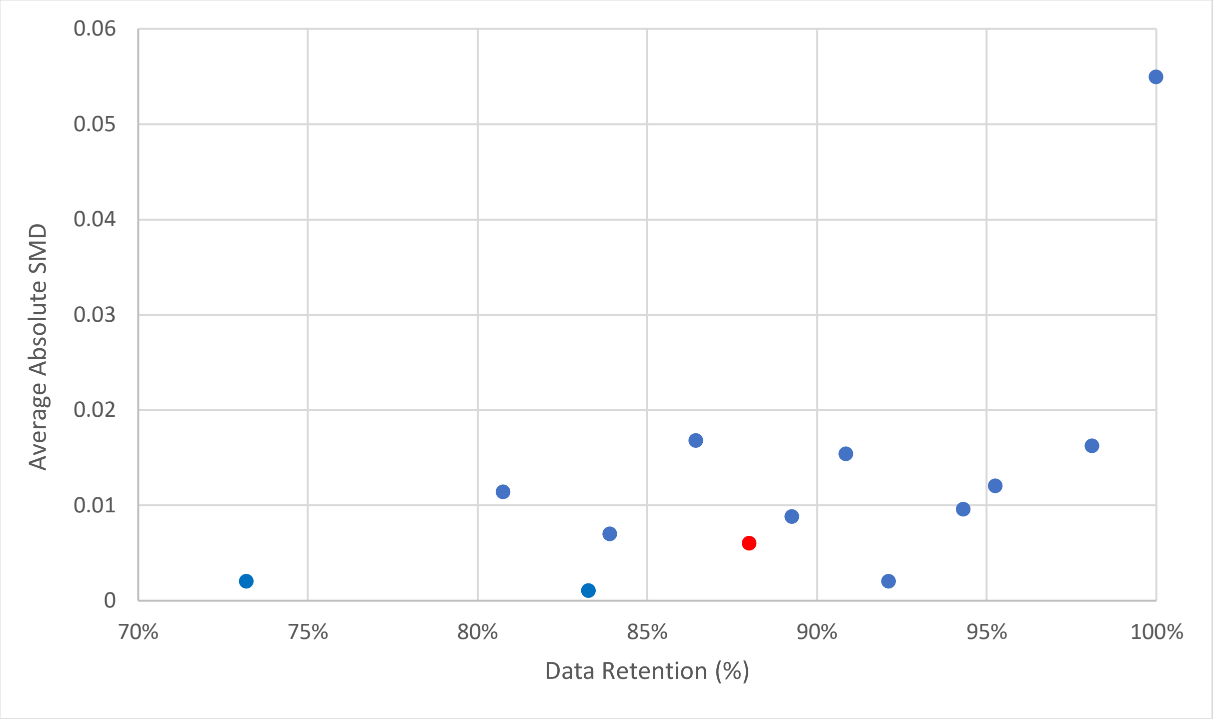
S Figure 7.1c The average absolute standardized mean difference (SMD) for the proportion of patients in each treatment group occupying each ProCaRS risk-group (i.e., high-intermediate, high, and extremely high) is plotted per level of data retention for each matching caliper used in propensity score matching.

*The red dot indicates the matching strategy chosen.



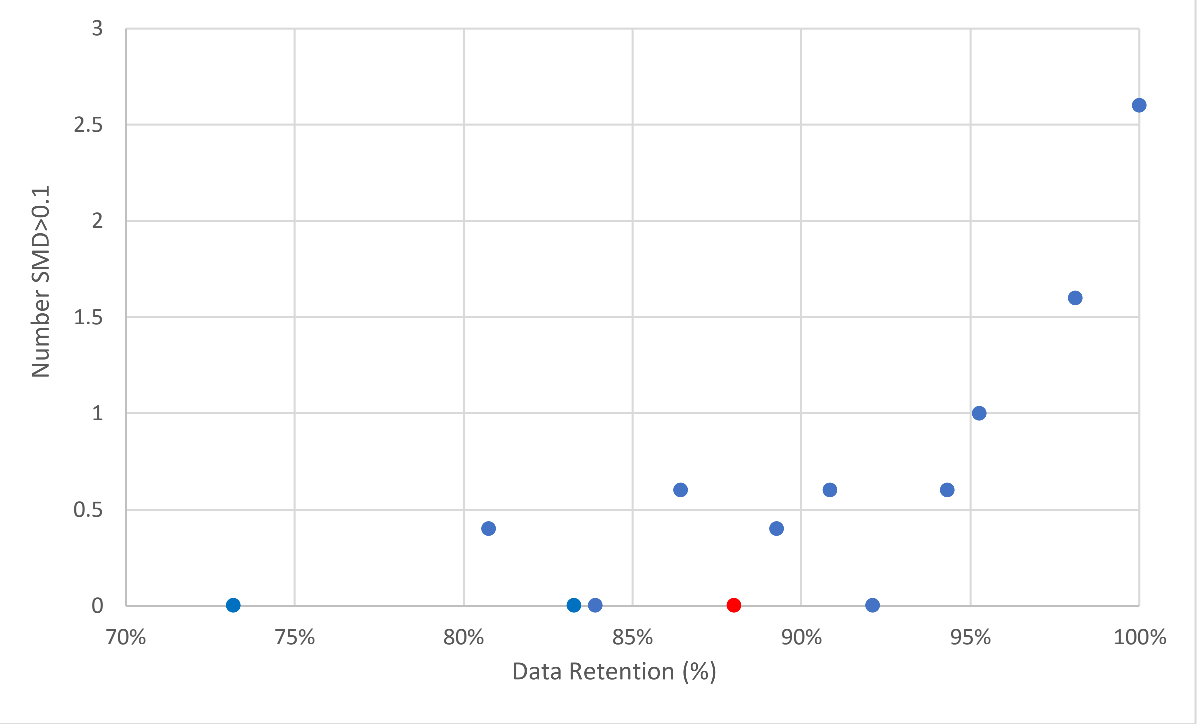
S Figure 7.1d The variance ratio for baseline PSA between treatment groups is plotted per level of data retention for each matching caliper used in propensity score matching.

*The red dot indicates the matching strategy chosen.



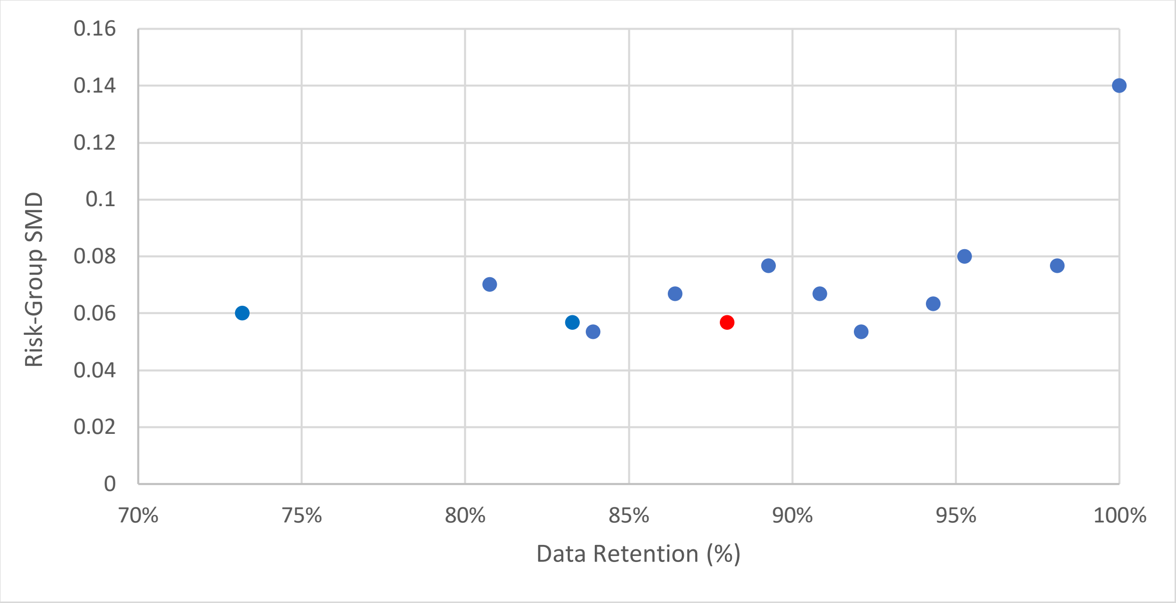
S Figure 7.2a The average absolute standardized mean difference (SMD) is plotted per level of data retention for each combination of coarsened variables used in coarsened exact matching.

*The red dot indicates the matching strategy chosen.



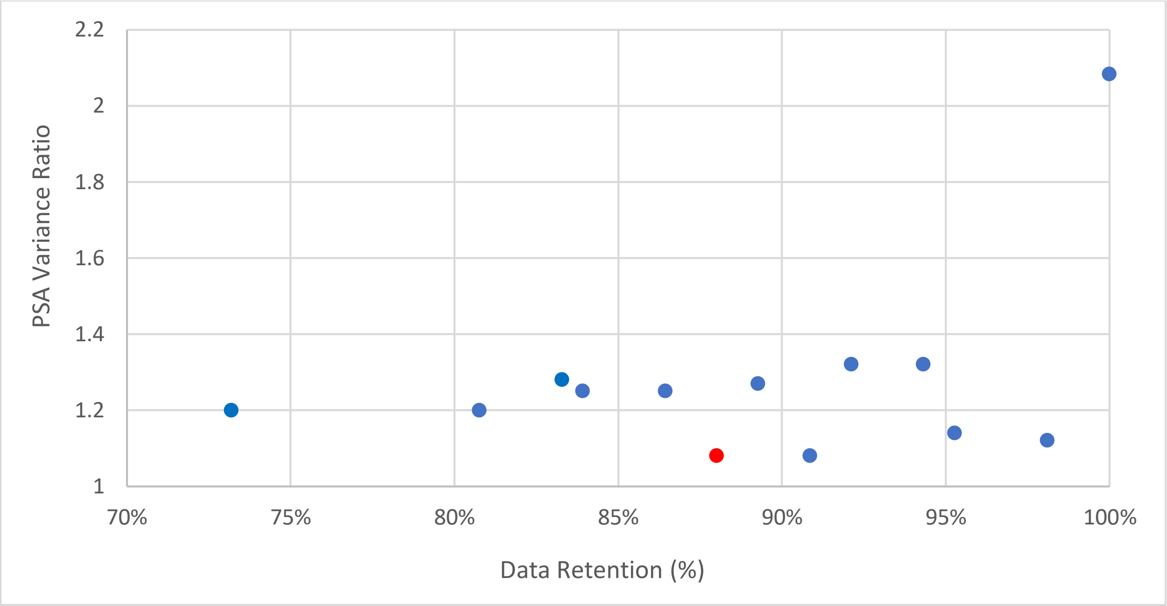
S Figure 7.2b The number of absolute standardized mean differences (SMD) that exceed the threshold of 0.1 is plotted per level of data retention for each combination of coarsened variables used in coarsened exact matching.

*The red dot indicates the matching strategy chosen.



S Figure 7.2c The average absolute standardized mean difference (SMD) for the proportion of patients in each treatment group occupying each ProCaRS risk-group is plotted per level of data retention for each combination of coarsened variables used in coarsened exact matching.

*The red dot indicates the matching strategy chosen.



S Figure 7.2d The variance ratio for baseline PSA between treatment groups is plotted per level of data retention for each combination of coarsened variables used in coarsened exact matching.

*The red dot indicates the matching strategy chosen.

Appendix A: Data Extraction Items

a) General study information:

- Title
- Authors
- Publication date
- Study design
    - Prospective vs retrospective
- Data source:
    - National-level databases
    - Single-institutional
    - Multi-institutional
    - Range of calendar years of diagnosis and treatment included
    - Geographical location

b) Prostate cancer, treatment and endpoint information:

- Dates of patient inclusion
- Follow-up duration
- Median age in each group
- Treatment information:
    - Number treated in each group
    - Approach to radiotherapy (e.g., dose, fractions, duration, 3D, IMRT, brachytherapy, dose-escalation, proton beam, SBRT, combination, etc.)
    - approach to radical prostatectomy (e.g., open- retropubic or perineal, laparoscopic or robotic)
    - use of neoadjuvant or adjuvant hormonal or chemotherapy and duration
- Adjusted HR for prostate cancer-specific mortality and all-cause mortality

Appendix B: Search Strategy

A search strategy was performed by Gabriel Boldt, a clinician librarian, and yielded a total of 5,487 articles total between PubMed and EMBASE databases before screening. Search strategies was completed as follows:

PubMed Strategy:

(radiotherapy[mh] OR radiation therapy[tw] OR radiotherapy[tw] OR surgery[mh] OR prostatectomy[tw] OR surgeries[tw])

AND

prostat*[tw]

AND

surviv*[tw]

AND

(high risk[tw] OR intermediate[tw] OR non-metastatic[tw] OR nonmetastatic[tw] OR localised[tw] OR localized[tw] OR locally[tw] OR local[tw])

NOT

review[pt]

Limits: Human, 2005-2020, English

Results 4325

EMBASE Strategy:

(radiotherapy.mp. or exp radiotherapy/ or radiation therapy.mp. or surgery.mp. or exp surgery/ or exp prostatectomy/ or prostatectomy.mp. or surgeries.mp.)

and

(prostate tumor/ or prostat*.mp. or exp prostate carcinoma/ or exp prostate cancer/ or exp prostate hypertrophy/)

and

(surviv*.mp. or exp survival/)

and

(high risk or intermediate or non-metastatic or nonmetastatic or localised or localized or locally or local).mp.

limit to (human and english language and exclude medline journals and yr="2005 - 2000")

Results 1162

Appendix C: Modified Newcastle-Ottawa Scale for Risk of Bias Assessment

Items having potential to bias the relationship between treatment modality (i.e. radical prostatectomy (RP) or radiation therapy (RT)) and outcomes of interest (i.e. cancer specific or overall survival)

Selection

1. Representativeness of the exposed cohort
    a. 1 point for data representing the general population (i.e. in terms of socioeconomic and demographic characteristics)
    b. 0 point if data is not representative or indicated (e.g. selected group of users like nurses, volunteers, insured, safety-net hospitals, secondary data from other clinical population, etc.)

2. Representativeness of the non-exposed cohort
    a. 1 point if drawn from the same community as the exposed cohort
    b. 0 points if drawn from a different source or not specified

3. Ascertainment of exposure
    a. 1 point if obtained from a secure record (e.g. surgical records) or self-report
    b. 0 points if no description

4. Demonstration that outcome of interest was not present at the start
    a. 1 point if yes
    b. 0 points if no

Comparability

5. Comparability of treatment groups after matching (if applicable) or accounted for in multivariable analysis. Maximum of 4 points awarded if the following factors are controlled for or not significantly different after matching as indicated by a standardized mean difference >0.10 or p>0.05:
    i. TNM
    ii. GS
    iii. PSA
    iv. Comorbidity status
    v. Age
    vi. ≥1 of year of diagnosis or treatment
    vii. ≥1 demographic characteristic (education, income, rural/urban)
    viii. Study center (if multiple)

    b. 0.5 point deducted for each variable not included in the model, unless tested and shown to have insignificant influence on the final results.

Outcome

6. Ascertainment of outcome
    a. 1 point if record linkage or blind assessment
    b. 0 points if assessment is not blinded or not reported

7. Adequacy of follow-up of cohorts
    a. 1 point if no subjects lost to follow up or those lost are unlikely to introduce bias (i.e. number lost ≤20% or description of those lost suggested no different from those followed)
    b. 0 points if follow up rate <80% and no description of those lost or if no statement was made

8. Was follow-up long enough for outcomes to occur?
    a. 1 point if median follow-up was ≥10 years, as 10-year cancer specific survival is estimated to be 88% in patients diagnosed with high-risk PCa undergoing multi-modal treatment.(224)

Thresholds for converting to low, moderate and high risks of bias:

Low risk of bias: ≥3 points in selection domain AND 4 points in comparability domain AND ≥2 points in outcome domain

Moderate risk of bias: 2 points in selection domain AND 4 points in comparability domain AND ≥2 points in outcome domain

High risk of bias: ≤1 point in selection domain OR ≤3 points in comparability domain OR ≤1 point in outcome domain

This scoring system is adapted from the Newcastle Ottawa Scale. We gave more weight to Item 5 as these confounding variables have demonstrated substantial impact on the comparison between RP and RT and overall and cause-specific mortality in prostate cancer research.(101)

Curriculum Vitae

Name: David Guy

Post-secondary Education and Degrees:

University of Waterloo
Waterloo, Ontario, Canada
2012-2017 BSc. (Kinesiology)

The University of Western Ontario
London, Ontario, Canada
2017-ongoing, MD/PhD (candidate)

Honors/Awards:

Faculty of Applied Health Sciences Kinesiology Scholarship, University of Waterloo, Waterloo, ON
2015-2016, 2016-2017

Charles J. Robson Research Day, People's Choice Award: Graduate Student/Surgeon Scientist Program, Division of Urology, Department of Surgery, University of Toronto, Toronto, ON.
2017

Jack Banham Hargreaves/Jessie Louisa Florence Hargreaves MD Award
2017, 2018, 2021

Province of Ontario Graduate Scholarship
2019-2020, 2020-2021

Physician's Services Incorporated Foundation
Research Trainee Fellowship
2020-2022

Related Work Experience

Clinical Research Assistant
Department of Radiation Oncology, Sunnybrook Health Sciences Centre
Toronto, ON
2013-2014

Student Researcher
Division of Urology, Department of Surgery, Sunnybrook Health Sciences Centre, Toronto, ON
2014-2017

Publications:

David Guy, Igor Karp, Piotr Wilk, Joseph Chin, George Rodrigues. Comparing the performance of coarsened exact matching and propensity score matching in non-experimental prostate cancer comparative effectiveness research. Submitted to J. Comp. Eff. Res. 2021 (accepted)

David Guy, Hanbo Chen, Gabriel Boldt, Joseph Chin. Characterizing Surgical and Radiotherapy Outcomes in Non-Metastatic High-Risk Prostate Cancer: A Systematic Review and Meta-analysis. Submitted to Prostate Cancer and Prostatic Dis. 2021 (under review)

David Guy, Rachel M. Glicksman, Roger Buckley, Patrick Cheung, Hans Chung, Stan Flax, David Hajek, Andrew Loblaw, Gerard Morton. Characterization of Outcomes of Multimodal Approaches in Unfavorable-Risk Prostate Cancer. 2021 J. Urol. (under review)

Kush M. Joshi, Arnon Lavi, Ray S. Jia, **David Guy**, Danielle Starcevic, Sophia M. Frost, Natan Veinberg, Shiva M. Nair, Joseph Chin. Preoperative neutrophil to lymphocyte ratio predicts adverse pathology at radical prostatectomy. Can Urol Assoc J. Abstracts. 2020. S102

David Guy, Avi D. Vandersluis, Laurence H. Klotz, Neil E. Fleshner, Alexander Kiss, Chris Parker, Vasundara Venkateswaran. Total energy expenditure and vigorous intensity physical activity are associated with reduced odds of reclassification among men on active surveillance. Prostate Cancer Prostatic Dis. 2017. Dec; doi: 10.1038/s41391-017-0010-0.

David Guy, Gabriella Ghanem, Andrew Loblaw, Roger Buckley, Beverly Persaud, Patrick Cheung, Hans Chung, Cyril Danjoux, Gerard Morton, Jeff Noakes, Lesley Spevack, Stanley Flax. Diagnosis, referral, and primary treatment decisions in newly diagnosed prostate cancer patients in a multidisciplinary diagnostic assessment program. Can Urol Assoc J. 2016. 10(3-4): 120–125.

Avi D. Vandersluis (co-first author), **David Guy (co-first author)**, Laurence H. Klotz, Neil E. Fleshner, Alexander Kiss, Chris Parker, Vasundara Venkateswaran. The role of lifestyle characteristics on prostate cancer progression in two active surveillance cohorts. Prostate Cancer Prostatic Dis. 2016. Sep;19(3):305-10. doi: 10.1038/pcan.2016.22. Epub 2016 June 28.

David Guy, Wade Wilson. Physical Activity and Chronic Disease in Rural Populations: Insights and Suggestions for Future Research. University of Waterloo. 2016. Graduate Leisure Research Symposium.