LOCALIZED PROSTATE CANCER PATIENTS' PREFERENCES FOR HYPOFRACTIONATED RADIOTHERAPY: A DISCRETE CHOICE EXPERIMENT

by

Samantha Sigurdson

A thesis submitted to the Department of Public Health Sciences

In conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(December, 2019)

Abstract

Introduction While external beam radiotherapy (EBRT) has always been a standard of care for localized prostate cancer patients, hypofractionated EBRT has now become a standard EBRT approach. We aimed to elicit the relative weight patients place on various aspects of EBRT and to compare these findings to clinical trials results, with a goal of informing clinical practice.

Methods A discrete choice experiment was designed in which men had to choose between two EBRT profiles that differed in five treatment attributes: effectiveness, acute toxicity, late toxicity, length, and fiducial marker implantation. We recruited prostate cancer patients with localized disease who had completed or were undergoing EBRT at our institution. Conditional (multinomial) logit regression (MNL), mixed multinomial logit (MXL), and latent class (LC) models were used to analyze the relative importance of treatment attributes, and the trade-offs this patient population were willing to make between attributes. These results were compared to explicit ranking and rating scores the participants completed for internal validation, as well as to clinical trials findings for interpretation. The MXL and LC models were used to compare realistic EBRT scenarios that corresponded to conventional fractionation, moderate hypofractionation, and stereotactic body radiotherapy, and determine what characteristics resulted in individuals preferring certain scenarios.

Results All treatment attributes' parameter estimates were significantly different from zero, and across the MNL and MXL models the most important attributes were risks of PSA recurrence and late toxicity. The rating exercises agreed with the quantitative results. Preference heterogeneity was present, largely due to age and distance patients traveled to the cancer centre. One latent class (the "convenience group") were not willing to lengthen EBRT to reduce risk of toxicity at any level within the model.

Conclusions Prostate cancer patients with localized disease place importance on convenience of EBRT. Older participants tend to be willing to accept a risk of increased toxicity to be treated with hypofractionated regimens. This risk is well within the expected risk based on clinical trial results. Participants had a good understanding of the DCE, and the responses were internally valid.

Co-Authorship

Harrison, M.[1,2], Richardson, H.[3], Brundage, M.[3,4]

1. University of British Columbia, Vancouver, British Columbia, Canada.
2. Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital, Vancouver, British Columbia, Canada.
3. Department of Public Health Sciences, Division of Cancer Care and Epidemiology, Queen's University Cancer Research Institute, Kingston, Ontario, Canada.
4. Department of Radiation Oncology, Cancer of Southeastern Ontario, Kingston, Ontario, Canada.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Michael Brundage, for his continuous support and guidance through each stage of this process. Thank you for encouraging me and allowing me to grow as a research scientist. I would like to acknowledge Dr. Mark Harrison for his invaluable expertise and encouragement, especially this past summer helping me develop STATA code for my analysis. I would like to thank Dr. Alison Pearce for her insightful comments and advice, as well as finding the time to video chat even with the difficulty of coordinating time zone differences from Australia. Thank you to the Radiation Oncologists who made time for my study in their busy clinics, and Dr. Harriet Richardson for your input throughout my degree. I would like to acknowledge the funding I received from the Clinical Investigator Program in support of my graduate studies.

I am grateful to my mom Dr. Karen Spalding who provided moral and emotional support throughout my life, as well as encouraged me to follow in her academic path. Thank you to my wonderful husband Andrew Surya for his encouragement throughout my academic studies. To my father and sister, Eric and McKenzie Sigurdson, thank you for believing in me and taking an interest in my thesis!

I would especially like to acknowledge the generosity of the participants who volunteered their valuable time and perspectives, ultimately making this study possible. This thesis is dedicated to them.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Acute	Attribute: Risk of moderate acute GI and GU toxicity
ADT	Androgen deprivation therapy
AIC	Akaike information criterion
ASCO	American Society for Clinical Oncology
ASTRO	American Society for Radiation Oncology
AUA	American Urological Association
β	Part-worth estimate or coefficient
BIC	Bayesian information criterion
CAIC	Consistent Akaike information criterion
CCO	Cancer Care Ontario
CFRT	Conventionally fractionated radiotherapy
DCE	Discrete choice experiment
df	Degrees of freedom
DFS	Disease free survival
EBRT	External beam radiotherapy
GI	Gastrointestinal
Gy	Gray
GU	Genitourinary
IGRT	Image guided radiotherapy
IMRT	Intensity modulated radiotherapy
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
Late	Attribute: Risk of moderate late GI and GU toxicity
LCM	Latent class model
Length	Attribute: Length of EBRT

Markers	Attribute: Fiducial marker implantation
MH	Moderate hypofractionation
MRS	Marginal rate(s) of substitution
MNL	Multinomial mixed logit
MXL	Multinomial mixed logit
NCCN	National Comprehensive Cancer Network
OAR	Organ(s) at risk
PFS	Progression free survival
PREFS	Purpose, Respondents, Explanation, Findings, Significance
PSA	Prostatic specific antigen
	Attribute: Risk of PSA recurrence within 5-years of completing treatment
RCT	Randomized controlled trials
RT	Radiotherapy
RTOG	Radiation Therapy Oncology Group
RUT	Random utility theory
SBRT	Stereotactic body radiotherapy
V	Total utility

Chapter 1

Introduction

This chapter briefly reviews the two main concepts key to my thesis: (1) radiotherapy to treat localized prostate cancer; and (2) quantifying patient preferences via a discrete choice experiment. I then provide my three objectives and in the final section an overview of my thesis.

Radiotherapy and hypofractionation

In North America prostate cancer is the most commonly diagnosed non-cutaneous cancer in men, and approximately 250 000 men were diagnosed in 2017 (1). Over 90% of men with prostate cancer are diagnosed with localized (non-metastatic) disease, and one of the primary treatment methods for these patients is radiotherapy (1, 2). Based on disease factors and with patient input, physicians initially prescribe one (or a combination) of several treatment methods that include surgery, radiotherapy, androgen deprivation therapy (ADT), or active surveillance (3). External beam radiotherapy (EBRT) has always been a standard of care for localized prostate cancer patients, and up until recently has consisted of radiotherapy doses of 76-80 Gray (Gy) delivered in 2 Gy per fraction. A fraction is the amount of radiation delivered each day, and typically patients receive radiotherapy at their local cancer centre daily until the complete dose of radiotherapy has been delivered. This means conventional fractionated radiotherapy (CFRT) results in patients travelling to the cancer centre daily Monday to Friday for up to 8 weeks.

Clinical studies suggest prostate cancer has biologic characteristics that make it more sensitive to larger doses of daily radiotherapy, known as hypofractionation (4). Hypofractionated EBRT has now become a standard EBRT approach. Hypofractionation also has the benefit of increased patient convenience and reduced treatment burden due to a shorter radiotherapy course. Recently, moderate hypofractionation (MH) regimens (2.4-4 Gy per fraction) have been investigated in several international multi-centre Phase III clinical trials comparing MH to CFRT (5-8). These trials have demonstrated that disease-related outcomes are quite similar between approaches, but toxicity outcomes differ among the studies. Three key uncertainties are: (1) long-term treatment efficacy; (2) long-term toxicity rates; and (3) acute toxicity rates. The trials do, however, provide evidence regarding the plausible range of outcomes (5-8). Two articles review the two main goals of these prostate hypofractionation RCTs; superiority trials hypothesize greater efficacy of hypofractionation with equivalent toxicity, and non-inferiority studies aim to demonstrate equivalent efficacy with reduced or similar toxicity (9, 10). The review papers concluded that MH can now be considered standard of care for localized prostate cancer (9).

Stereotactic body radiotherapy (SBRT) is a technique that delivers extremely conformal high dose radiotherapy in five or fewer fractions of greater or equal to 6.5 Gy per fraction. This modality is only safe in settings with precise, image-guided technology and implantation of fiducial markers (tiny permanent gold markers) into the prostate gland (1). Studies have investigated regimens of 35-36.25 Gy in five fractions (7-7.25 Gy per fraction) delivered every other day. The studies concluded these regimens are safe and appear effective, however there are no published phase III RCTs comparing CFRT and SBRT at this time (1, 11). A Scandinavian phase III trial has recently been published comparing ultra-fractionated radiotherapy of 42.7 Gy in seven fractions (6.1 Gy per fraction) compared to CFRT (78 Gy in 39 fractions) (12). This is similar to SBRT with slightly more fractions. With an average of five year follow up there was no significant difference of failure-free survival between the two groups, however there was significantly more acute urinary and bowel toxicity in the ultra-fractionated group (12).

A large proportion of Radiation Oncology department workload consists of radiotherapy to the prostate (in one study 27%) (13). With increased uptake of hypofractionation there should be a decrease in radiotherapy wait times in jurisdictions where waits for radiotherapy exist, and for all cancer patients since each prostate cancer patient would spend less time on treatment (13). The switch to hypofractionation could decrease the cost of treatment as SBRT is less expensive to the Canadian health care system than CFRT (14, 15). SBRT could reduce treatment costs by 27-55% compared with CFRT, and MH could account for a 22% decrease in cost (13). However, this is slightly offset by SBRT requiring more expensive technology (13, 16). Therefore, it follows that SBRT should be the radiotherapy modality of choice if studies are able to demonstrate non-inferiority to other radiotherapy techniques (1). Ideally patient preferences should be studied in this area to ensure they are in line with the evolving standard of care and changes due to cost considerations. The cost effectiveness of therapies can be increased by aligning care with patient preferences and utilities as this could lead to improved adherence and/or compliance, and therefore greater efficiency (17, 18).

The American Urological Association (AUA), the American Society for Radiation Oncology (ASTRO), and the American Society for Clinical Oncology (ASCO) published a consensus paper in 2019 that strongly recommended with high quality evidence that MH of 2.4-3.4 Gy per fraction be offered to prostate cancer patients with localized disease who are being treated with EBRT since MH provides similar prostate cancer control to CFRT (19). The paper notes that patients should be counseled about the small increased risk of acute gastrointestinal (GI) toxicity with MH (19). The consensus for SBRT is less clear due to the dearth of published RCTs examining SBRT. Currently the only SBRT phase III RCT, the gold standard of evidence in medicine, is a study at a centre in the United Kingdom (UK) has not yet reached maturity (11). The UK trial consists of two arms; non-operable patients randomized between 36.25 Gy in five fractions (SBRT) or CFRT, and operable patients randomized between SBRT or surgery (11). There is moderate consensus SBRT can be considered in a clinical trial setting in patients with lower disease burden (19). The National Comprehensive Cancer Network (NCCN) guidelines indicate that MH and SBRT can be considered as an alternative to CFRT for localized prostate cancer. However, longer follow up and prospective multi-institutional data are required to evaluate longer-term results since late toxicity could be worse in hypofractionated regimens compared to CFRT (20).

Some clinicians have expressed concern about the long-term risks of hypofractionation based on the variable trial findings and limited follow-up. However, there is a consensus in the literature that suggests a growing trend to prescribe hypofractionated regimens (14). There is a historical trend for Canadian clinicians to uptake hypofractionation at higher rates than their American colleagues. For example, hypofractionation regimens for breast cancer have been the standard of care for several years, but in 2011 only 35% of American breast cancer patients were prescribed hypofractionated regimens, while in Canada the rate was 70% (14). The difference between the two countries may reflect the variances in practice due to the varying models of health care, including reimbursement methods (14). Additionally, since there are no clear cut guidelines for clinicians to implement SBRT for the treatment of prostate cancer, this may contribute to lower rates of uptake (14). Therefore, a preference study in this patient population can help identify groups of patients willing to be treated with MH or SBRT.

Patient preferences

Discrete choice experiments

Patient choice has recently been an area of increasing interest, as addressed further below. Discrete choice experiments (DCEs) are an important method of quantitatively eliciting patient preferences that combine random utility theory, consumer theory, experimental design theory, and economic analysis (21). DCEs are surveys of hypothetical choices between two medical interventions (known as alternatives) and ask participants to choose their preferred option (21). DCE respondents are offered a series of choices (known as choice sets) and in each case are asked to choose between two or more alternatives (22). Each alternative is described by its characteristics (known as attributes) and the theoretical basis for DCEs makes the assumption that alternatives can be meaningfully described by a set number of attributes (21). Individuals’ valuations depend on the levels set for each of these attributes (21). For example, if the DCE presents two alternative drugs for relief of headaches, the attributes could be effectiveness in reducing headache pain, size of tablet, and side effects. The levels for the effectiveness attribute could include three probabilities of successful pain relief (e.g. 50%, 75%, and 100%); the levels for the size of tablet could be small, medium, and large; and the levels for the side effects could be minor,

	Drug A	Drug B
Chance of pain relief	50% of people will have no headache	75% of people will have no headache
Size of tablet	Small	Large
Risk of side effects	Minor: upset stomach	Intermediate: vomit
I choose:	☐	☐

Figure 1: Mock choice set of example in text for pain relief drug.

intermediate, or major GI side effects (Figure 1). By asking participants to choose between alternatives with different attribute levels multiple times, DCEs represent a stated preference survey method using conjoint analysis to systematically examine the factors that drive individuals choices' (22, 23). The relative importance of attributes, and the trade-offs that respondents make between them, can be modeled based on participants choices (21). A statistical model is used to analyze the influence of attribute levels on choice decisions, and then the relative importance of the attributes and the rate respondents are willing to trade between the attributes can be estimated (24, 25). The relative significance of attributes can be evaluated by taking a ratio of the coefficients of each possible pair of attributes, which provides the marginal rates of substitution (MRS) (26). For example, the MRS of attribute X for attribute Y is the amount of attribute X that an individual is willing to exchange for attribute Y.

Rationale

DCEs allow investigators to determine whether a health intervention or service is important to patients, the relative importance of the intervention's attributes, and the trade-offs individuals are willing to make between these attributes (21). Patients are both the potential beneficiaries of new treatments, but also the ones who bear the risks associated with these treatments, and their preferences warrant consideration (27). There has been recent emphasis on patient-centered healthcare in regulatory and policy decisions, and this has led to augmented interest in choice experiments for quantifying patient concerns and values (28). To date, no DCE has been conducted examining patient preferences' for SBRT or MH. Conducting a DCE to determine respondents' preferences for EBRT is ideal in the setting of localized prostate cancer because there are no clear guidelines for SBRT recommendations yet. Independent of health outcomes, cancer patients prefer treatments that are convenient (29), as there may be considerations such as family, work, or out-of-pocket expense. This suggests that patient outcomes can be optimized by considering how patients value disease and treatment-related outcomes, and other factors such as treatment duration and its associated implications (29, 30). Suitable prostate cancer patients with localized disease could be treated with CFRT, MH, or possibly SBRT. If the trade-offs prostate cancer patients with localized disease are willing to make for treatment convenience were quantified, then clinicians would be better able to counsel these patients regarding their treatment options when the phase III trial data become available. Therefore, a DCE is an ideal experimental design to determine patients' preferences for hypofractionation, and their willingness to trade off the possible increased risk of toxicity and extra intervention with fiducial marker implantation, for the convenience of shorter treatment time.

Objectives

The purpose of this study was to design and conduct a DCE to measure the trade-offs patients with localized prostate cancer were willing to make between five attributes of EBRT: treatment length; fiducial marker implantation; long-term PSA control (disease-specific outcome); risk of acute GI and genitourinary (GU) toxicity; and risk of late GI and GU toxicity. My intent was to compare the magnitude of these trade-offs with clinical trials results with a goal of informing clinical practice. The target population was men with prostate cancer who were undergoing or had previously completed EBRT for localized disease.

Objective 1

The first objective was to quantify the coefficient values, known as part-worth estimates, for each attribute to prioritize their relative importance to respondents. The null hypotheses were that the estimates for each attribute would not be significantly different from 0. If the null hypothesis was rejected for a given attribute, based on the literature that attribute could be concluded to statistically significantly play a role in respondents' choices across the studied population (31). The sign of a parameter coefficient indicated whether the attribute had a positive or negative impact on the participants' preferences (32). It was hypothesized *a priori* that the part-worth estimates for all the attributes would be negative as respondents should prefer: treatment alternatives that have a lower risk of not being effective; a lower risk of side effects; not to have a procedure to implant fiducial markers; and a shorter treatment time. Preference heterogeneity was expected to be present within the sample since one group of respondents may trade-off increased risk of toxicity for the benefit of shorter treatment length, while another group might not and prefer to minimize risk of treatment toxicity no matter the length of treatment. Several methods of analysis were explored to determine what respondent factors influence choice. For example, it was hypothesized that older respondents and those who lived further from the cancer centre would value length of EBRT more strongly than younger participants and those who lived closer to the cancer centre.

Objective 2

The second objective was to quantify how systematically varying the levels assigned to each attribute impacted the respondents' willingness to trade-off each attribute for the ease of shorter treatment time (30). This trade-off is known as the marginal rate of substitution (MRS). These values would be compared to the known risks for MH and SBRT from the RCT literature. Additionally, theoretical EBRT scenarios were created and the parameter estimates for significant attributes were used to compare the relative ranking of these scenarios with two statistical models.

Objective 3

The third objective was to assess the internal validity of the DCE via choice sets additional to the experimental design. Participants completed a post-DCE questionnaire to determine how they made choices in the DCE, explicit rankings, and a qualitative question would assess why certain attributes were important to participants. Participant-reported scores were compared to interviewer-reported scores for inter-rater consistency. Explicit ratings of the attributes were compared to the quantitative results.

Thesis overview

Following this introductory chapter, chapter 2 reviews the literature regarding the role of hypofractionation to treat localized prostate cancer, best methods to conduct a DCE, limitations of this type of study, critique of DCEs compared to other methods to elicit patient preference data, and review of selected published oncology DCEs. Chapter 3 is a detailed methods chapter, followed by the results in chapter 4, and my thesis concludes in chapter 5 with the discussion and implications of my results.

References

1. Loblaw A, Liu S, Cheung P. Stereotactic ablative body radiotherapy in patients with prostate cancer. Transl Androl Urol. 2018 Jun;7(3):330-40.
2. Zaorsky NG, Shaikh T, Murphy CT, Hallman MA, Hayes SB, Sobczak ML, et al. Comparison of outcomes and toxicities among radiation therapy treatment options for prostate cancer. Cancer Treat Rev. 2016 Jul;48:50-60.
3. 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. 2018 Mar;199(3):683-90.
4. Moon DH, Efstathiou JA, Chen RC. What is the best way to radiate the prostate in 2016? Urol Oncol. 2017 Feb;35(2):59-68.
5. Dearnaley D, Syndikus I, Mossop H, Khoo V, Birtle A, Bloomfield D, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. Lancet Oncol. 2016 Aug;17(8):1047-60.
6. Catton CN, Lukka H, Gu CS, Martin JM, Supiot S, Chung PWM, et al. Randomized trial of a hypofractionated radiation regimen for the treatment of localized prostate cancer. J Clin Oncol. 2017 Jun;35(17):1884-90.
7. Lee WR, Dignam JJ, Amin MB, Bruner DW, Low D, Swanson GP, et al. Randomized phase III noninferiority study comparing two radiotherapy fractionation schedules in patients with low-risk prostate cancer. J Clin Oncol. 2016 Jul;34(20):2325-32.
8. Incrocci L, Wortel RC, Alemayehu WG, Aluwini S, Schimmel E, Krol S, et al. Hypofractionated versus conventionally fractionated radiotherapy for patients with localised prostate cancer (HYPRO): final efficacy results from a randomised, multicentre, open-label, phase 3 trial. Lancet Oncol. 2016 Aug;17(8):1061-9.
9. Benjamin LC, Tree AC, Dearnaley DP. The role of hypofractionated radiotherapy in prostate cancer. Curr Oncol Rep. 2017 Apr;19(4):30.
10. Lee WR, Koontz BF. Moderate hypofractionation for prostate cancer. Transl Androl Urol. 2018 Jun;7(3):321-9.
11. Haque W, Butler EB, Teh BS. Stereotactic body radiation therapy for prostate cancer-a review. Chin Clin Oncol. 2017 Sep;6(Suppl 2):S10.
12. Widmark A, Gunnlaugsson A, Beckman L, Thellenberg-Karlsson C, Hoyer M, Lagerlund M, et al. Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial. Lancet. 2019 Aug;394(10196):385-95.
13. Sharieff W, Greenspoon JN, Dayes I, Chow T, Wright J, Lukka H. The technique, resources and costs of stereotactic body radiotherapy of prostate cancer: a comparison of dose regimens and delivery systems. Technol Cancer Res T. 2016 Feb;15(1):171-8.
14. Hunter D, Mauldon E, Anderson N. Cost-containment in hypofractionated radiation therapy: a literature review. J Med Radiat Sci. 2018 Jun;65(2):148-57.
15. Helou JA, Torres S, Musunuru HB, Cheung P, Vesprini D, Chung HT, et al. Low dose rate brachytherapy vs standard external beam radiotherapy vs stereotactic body radiotherapy for low risk prostate cancer: a cost-utility analysis. J Clin Oncol. 2017 Nov;29(11):718-31.
16. Muralidhar V, Nguyen PL. Maximizing resources in the local treatment of prostate cancer: a summary of cost-effectiveness studies. Urol Oncol. 2017 Feb;35(2):76-85.
17. Brazier JE, Dixon S, Ratcliffe J. The role of patient preferences in cost-effectiveness analysis: a conflict of values? Pharmacoeconomics. 2009 Sep;27(9):705-12.
18. Bien DR, Danner M, Vennedey V, Civello D, Evers SM, Hiligsmann M. Patients' preferences for outcome, process and cost attributes in cancer treatment: a systematic review of discrete choice experiments. Patient. 2017 Oct;10(5):553-65.

19. Morgan SC, Hoffman K, Loblaw DA, Buyyounouski MK, Patton C, Barocas D, et al. Hypofractionated radiation therapy for localized prostate cancer: executive summary of an ASTRO, ASCO and AUA evidence-based guideline. J Urol. 2019 Mar;201(3):528-34.
20. Carroll PH, Mohler JL. NCCN guidelines updates: prostate cancer and prostate cancer early detection. J Natl Compr Netw. 2018 May;16(Suppl 5):620-3.
21. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012 Feb;21(2):145-72.
22. Veldwijk J, Determann D, Lambooij MS, van Til JA, Korfage IJ, de Bekker-Grob EW, et al. Exploring how individuals complete the choice tasks in a discrete choice experiment: an interview study. BMC Med Res Methodol. 2016 Apr;16:45.
23. Havrilesky LJ, Alvarez Secord A, Ehrisman JA, Berchuck A, Valea FA, Lee PS, et al. Patient preferences in advanced or recurrent ovarian cancer. Cancer. 2014 Dec;120(23):3651-9.
24. Hauber AB, Gonzalez JM, Groothuis-Oudshoorn CG, Prior T, Marshall DA, Cunningham C, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR conjoint analysis good research practices task force. Value Health. 2016 Jun;19(4):300-15.
25. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics. 2008 Feb;26(8):661-77.
26. Ali S, Ronaldson S. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. Br Med Bull. 2012 Sep;103(1):21-44.
27. Hauber AB, Fairchild AO, Johnson RF. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. Appl Health Econ Health Pol. 2013 Aug;11(4):319-29.
28. Muhlbacher A, Johnson FR. Choice experiments to quantify preferences for health and healthcare: state of the practice. Appl Health Econ Health Pol. 2016 Jun;14(3):253-66.
29. Higgins A, Barnett J, Meads C, Singh J, Longworth L. Does convenience matter in health care delivery? A systematic review of convenience-based aspects of process utility. Value Health. 2014 Dec;17(8):877-87.
30. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014 Sep;32(9):883-902.
31. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health--a checklist: a report of the ISPOR good research practices for conjoint analysis task force. Value Health. 2011 Jun;14(4):403-13.
32. de Bekker-Grob EW, Niers EJ, van Lanschot JJ, Steyerberg EW, Wijnhoven BP. Patients' preferences for surgical management of esophageal cancer: a discrete choice experiment. World J Surg. 2015 Oct;39(10):2492-9.

Chapter 2

Literature Review

Introduction

This chapter has two main goals; to examine the current literature on the most effective external beam radiotherapy (EBRT) regimens for localized prostate cancer, and to review the papers that discuss best practices and current trends to conduct discrete choice experiments (DCEs), including specifically in the oncology patient population. Following the methods and results sections of this chapter, the discussion section reviews the recent randomized clinical trials (RCTs) underlying new guidelines to treat localized prostate cancer, and explains the theory, best methods for design and analysis, and limitations of DCEs. The final section reports on areas of future research and the justification for selecting a DCE to answer the study objectives.

Methods

Review articles published between 2016 to 2018 were searched through Pubmed using the terms; prostate cancer, hypofraction*, radiotherapy. Any articles identified would have their cited papers reviewed for phase III clinical trials and key papers. Updated guidelines by known leading organizations were searched in Pubmed (American Urological Association (AUA), American Society for Radiation Oncology (ASTRO), American Society for Clinical Oncology (ASCO), and National Comprehensive Cancer Network (NCCN)).

A literature search was conducted in Ovid Medline, CINAHL, and Ovid PsychInfo for review methodologies published from 2009 to 2018 with any of the following terms; discrete choice experiment(s), discrete choice modelling, discrete choice modeling, stated preference, part-worth utilities, functional measurement, paired comparisons, pairwise choices, conjoint analysis, conjoint measurement, conjoint studies, or conjoint choice experiments. These terms are from two older reviews, first Ryan and

Gerard in 2003 reviewed DCEs published between 1990 to 2000, and de Bekker-Grob et al. in 2012 reviewed DCEs published between 2001 to 2008 (1, 2). These two systematic reviews reference lists were reviewed to identify landmark papers for conducting DCEs.

A literature search was conducted in Pubmed using the terms; discrete choice experiment(s) and prostate cancer. The systematic review of oncology DCEs by Bien et al. (3) reference list was manually searched to identify any other relevant oncology DCEs.

Results

The literature search on prostate hypofractionation identified 11 review articles (Figure 2). Three guidelines were identified based on the targeted search, two published in 2017 and one each in 2018 and 2019 (Figure 2).

The literature search on DCEs found 153 articles with five duplicates, and of the 148 original articles 42 were included (Figure 3). The included articles were determined after screening abstracts for studies that focused on best current methods to design a DCE or that identified any issues with conducting a DCE and how best to resolve them. Additional papers were identified to supplement the original electronic search by manually reviewing the reference lists of articles retrieved from the databases.

Five DCEs conducted with prostate cancer patients were identified from this literature search. A further three DCEs conducted with cancer patients and deemed relevant to this study were included for the discussion of oncology DCEs in this chapter.

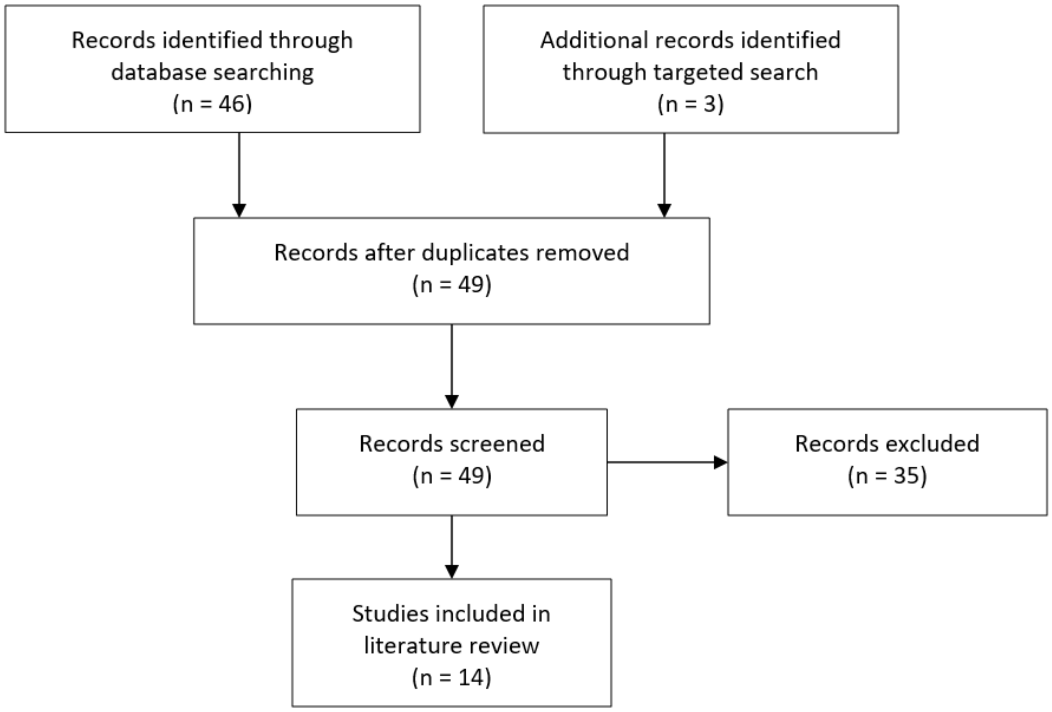


Figure 2: Flow chart of literature review for hypofractionated prostate cancer radiotherapy.

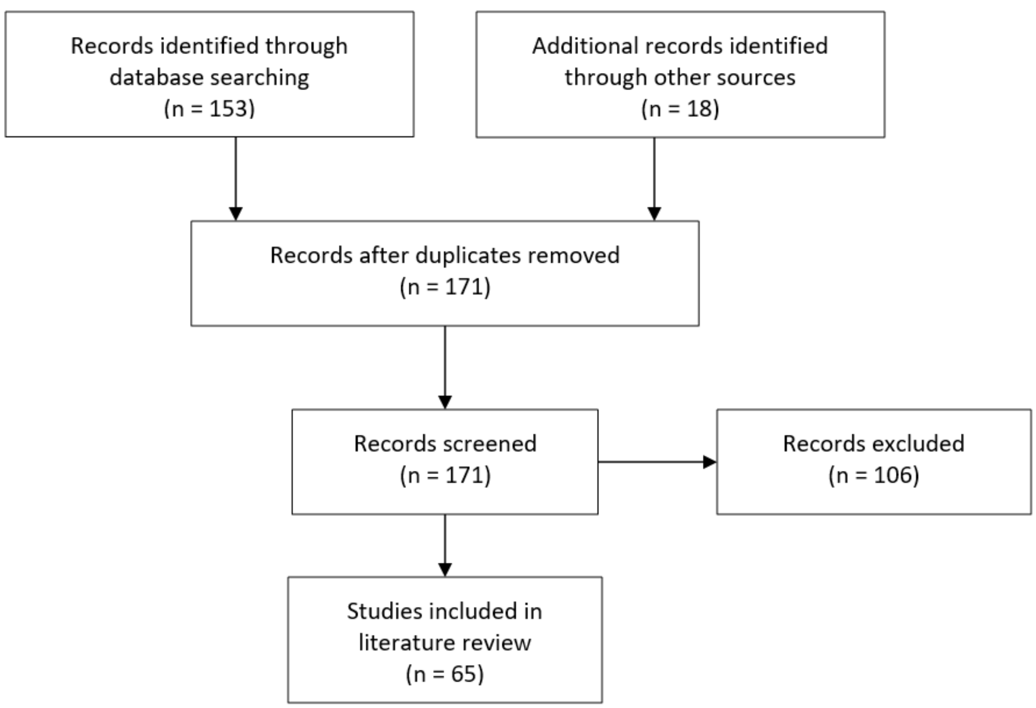


Figure 3: Flow chart of DCE literature review.

Discussion

Treating localized prostate cancer with radiotherapy

Localized prostate cancer is divided into three main risk groups; low, intermediate, and high. This is based on Prostate Specific Antigen (PSA; an enzyme that is secreted by prostatic cells) level, pathology findings, and tumour extent at time of diagnosis (4). The risk groups are linked to prognosis, and help health care practitioners determine the best treatment(s) to recommend to patients (4). For example, active surveillance is only recommended to low risk prostate cancer patients, while the combination of EBRT with ADT is usually only recommended to high risk prostate cancer patients (4). EBRT can be used as the sole treatment for low and intermediate risk disease. Recurrence is monitored for via a blood assay for PSA (4). If a biochemical recurrence has occurred testing for metastatic disease and/or further therapy is likely necessary (4).

EBRT is an effective method for treating localized prostate cancer, however there are risks of side effects from this therapy. This includes toxicity to organs near the prostate, known as ‘organs at risk’ (OARs), specifically the bladder and rectum. Early side effects are defined as acute toxicity that develops and resolves within three months of radiotherapy. Long term side effects are defined as late toxicity that develops at any point three months post EBRT, and may affect either organ (5). The Radiation Therapy Oncology Group (RTOG) defined the grades of toxicity for each OAR (6). Toxicity is graded on a scale of 0 (no toxicity) to 5 (death from the toxicity), with grade 2 toxicity considered moderate (6).

Intensity-modulated radiation therapy (IMRT) was developed in an attempt to decrease toxicity by delivering radiation from several beam angles and dynamically varying the shape and intensity of the beams to decrease dose delivered to OARs (7). Several comparative studies have reported a reduction in acute and late toxicities with IMRT compared to conventional 3-Dimensional conformal radiotherapy (3DCRT) (7). Image-guided radiation therapy (IGRT) has been incorporated to increase the accuracy of radiotherapy by enhancing the target definition. This results in better localization of the prostate with computed tomography (CT) imaging to verify patient position prior to delivery of radiotherapy (7).

Highly conformal IMRT with image guidance resulted in high doses of radiotherapy becoming the current standard of care by reducing radiation related side effects (7). However, dose escalation has been achieved through increasing the number of fractions, leading to prolongation of treatment time, increased hospital visits for patients, greater departmental workload, and higher costs (8).

A review article published in 2016 explored the history of conformal radiotherapy techniques enabling safer delivery of higher doses of radiation to the prostate (7). The current standard fractionation dose of up to 80 Gray (Gy) is known as ‘dose escalated’, delivered in daily 2 Gy fractions called conventionally fractionated radiotherapy (CFRT). Dose escalated radiotherapy increased grade 2 or greater acute and late GI and GU toxicity compared to the previous traditional doses of up to 70 Gy (7). However, dose-escalated radiotherapy was accepted as the new standard of care as the dose-escalated regimens consistently demonstrated improved disease free survival (DFS) in at least five RCTs (7).

Besides EBRT, another method of delivering radiotherapy exists called brachytherapy, which consists of radioactive implants inserted into and around the tumour. Brachytherapy is a second method of delivering a dose boost directly to the tumour after a patient receives a lower conventionally fractionated dose to the pelvis via EBRT, usually 46 Gy (9). In 2017 Cancer Care Ontario (CCO) and the American Society of Clinical Oncology (ASCO) updated their guidelines and indicated for suitable intermediate or high risk prostate cancer patients treated with EBRT, a boost with brachytherapy should be offered as standard of care, and eligible patients with low risk disease can be treated with brachytherapy or EBRT alone (10). In an RCT with 9 years of follow up that compared two methods of dose escalation (EBRT to 78 Gy compared to 46 Gy with EBRT plus a brachytherapy boost) the biochemical failure was higher for the arm without a brachytherapy boost. There was no difference in severe (grade 3 to 4) late GI toxicity, however there was significantly more severe late GU toxicity in the brachytherapy boost arm (9).

Clinical studies suggest prostate cancer has biologic characteristics that make it more sensitive to larger doses of daily radiotherapy, known as hypofractionation (7). This technique achieves better tumour kill while delivering an isoeffective dose to the prostate (7). Hypofractionation should achieve similar late toxicity compared to CFRT (7, 11). Therefore, hypofractionation is an attractive means of dose escalation without prolonging treatment time (7). Recently, moderate hypofractionation (MH) regimens (2.4-4 Gy per fraction) have been investigated in several international multi-centre Phase III clinical trials comparing MH to CFRT (5, 12-14). These trials have demonstrated that disease-related outcomes are quite similar between approaches, but toxicity outcomes differ among the studies.

There are four superiority trials that together only provide weak evidence that MH can improve efficacy compared to dose escalated EBRT, as only the two smallest studies were positive (15). The HYPRO study is the largest superiority study and randomized 804 patients with intermediate or high risk disease to CFRT or MH (64.6 Gy in 19 fractions; 3.4 Gy per fraction) (14). This trial was negative as there was no statistical difference between the five year relapse free survival between the two arms (MH 77.1% vs. CFRT 80.5%) (14). Combined with the three other smaller superiority trials (Arcangeli et al., Hoffman et al., and Pollack et al.) there was no demonstrated difference in efficacy after five years based on metastasis-free, cancer-specific survival, or overall survival (8). The HYPRO trial found the incidence of acute GI and late GU toxicity was significantly higher in the MH arm (14). However, the other three randomized superiority trials did not demonstrate any significant difference in toxicity between the CFRT and MH arms, and the dose per fraction ranged from 2.4-3.1 Gy (15). Based on the HYPRO trial findings a MH regimen of 3.4 Gy per fraction could not be recommended (15).

The CHHiP trial is the largest non-inferiority RCT and concluded there was no difference in five year biochemical failure-free survival between the MH (60 Gy in 20 fractions, 3 Gy per fraction) and CFRT arms (90.6% and 88.3%) (5). Regarding toxicity the only significant difference was acute GI toxicity, which was 38% in the MH arm and 25% in the CFRT arm (5). The RTOG 0415 non-inferiority study found the MH arm (70 Gy in 28 fractions, 2.5 Gy per fraction) to be non-inferior to CFRT for five year disease free and biochemical recurrence survival (86.3% vs. 85.3%) (13). With regards to toxicity the MH arm had significantly higher late grades 2 to 3 GI (22.4% vs. 14.0%) and GU (29.7% vs. 22.8%) toxicity (13). The PROFIT non-inferiority trial demonstrated no difference in efficacy between the MH

(60 Gy in 20 fractions) and CFRT arms (12). Acute GI grade 2 or greater toxicity was significantly higher in the MH arm (16.7% vs. 10.5%), however conversely late GI grade 2 or greater toxicity was significantly lower in the MH arm (8.9% vs. 13.9%) (12). In summary, there may be up to a 6% to 10% increased risk of moderate acute and late GI and GU toxicity for MH compared to CFRT.

Two review papers recently examined the hypofractionation trials, and concluded MH regimens of 2.5 to 3 Gy per fraction for localized prostate cancer is safe, effective, and non-inferior to CFRT (8, 15). In regimens of 3 Gy per fraction or higher more acute toxicity is observed, though late GI and GU toxicity is similar between MH and CFRT (15). All of the reviewed RCTs compared MH regimens (usually 60 Gy in 20 fractions) to CFRT regimens that varied in each paper but ranged from 73.8 to 80 Gy in 37 to 42 fractions (seven and a half to eight and a half weeks) (15). Overall, the superiority trials included less than 1500 men, while the three large non-inferiority trials enrolled more than 5000 men (15). The review papers concluded that based on these trials MH at around 3 Gy per fraction could be considered standard of care for localized prostate cancer (8).

Stereotactic body radiotherapy (SBRT) is an advanced radiotherapy technique that requires IGRT to ensure the highly conformal dose is delivered accurately to the target volume, and that toxicity to OARs is minimized at each treatment (fraction) by accounting for daily changes in anatomy and positioning (16). When SBRT is used to treat localized prostate cancer IGRT is usually achieved by transperineally inserting fiducial markers into the prostate via ultrasound guidance (16). This is an outpatient procedure done with local anesthetic and the fiducial markers are usually gold and permanently remain in the prostate (16).

SBRT clinical trials have demonstrated excellent long-term results as the eight year biochemical failure rate was only 5.0% in two phase II trials (11). In the largest cohort treated with SBRT the 10 year biochemical disease-free survival was 93.6% for low risk disease, and at 8 years was 94.3% for intermediate and 94.6% for low risk disease (11). Severe GU and GI side effects (RTOG grade 3 or worse) were rare. A review paper that pooled 835 patients with a median follow up of 63 months found severe acute side effects were 0.6% and 0.3% for GU and GI toxicity respectively, and in the late term were 2.6% for GU and 1.0% for GI toxicity (11). A pan-Canadian propensity-based analysis matched and compared the outcomes for 602 low-risk prostate cancer patients treated with SBRT, CFRT, or brachytherapy (11). With an average of over five years of follow up there was no significant difference between those treated with SBRT and brachytherapy, though SBRT patients had significantly lower biochemical failure compared to EBRT patients (11). However, a review paper published in 2017 noted that outcomes were likely worse for intermediate and high risk patients, as pooled data from 1 100 patients treated at eight institutions in two phase II trials found the five year freedom from biochemical failure rate was 95.2% for low risk, 84.1% for intermediate risk, and 81.2% for high risk disease (17).

A recently published phase III RCT compared intermediate and high-risk prostate cancer patients treated with CFRT 78 Gy in 2 Gy per daily fraction over 8 weeks, or ultra-fractionated radiotherapy 42.7 Gy in 6.1 Gy fractions with 3 fractions given weekly over 2.5 weeks (18). The failure free survival rate after five years of follow up was the same in both arms (84%) (18). In the ultra-fractionated arm there was a significant difference of 5% increased acute, and 4% late, GU toxicity that was grade 2 or worse (18).

The 2019 summary guideline published by ASTRO, ASCO, and AUA clearly supported MH as a new standard of care when EBRT is selected to treat localized prostate cancer (19). The guideline was not as clear or supportive for SBRT (19). There is currently one phase III RCT with maturing data to examine SBRT to treat localized prostate cancer as noted in chapter 1, however until the data are published clinicians will continue to have reservations about prescribing SBRT in this patient group off trial. Additionally, even for MH, long-term data to examine late toxicity are not yet available, which may make some clinicians hesitant to prescribe MH. It is important to determine patient preferences and opinions regarding toxicity, and if participants would trade-off potentially increased toxicity for the benefit of reduced treatment time.

Theory of discrete choice experiments

There are two main underlying theories of DCEs. First is Lancaster's theory of value that argues individuals view an alternative as a combination of its separate partial utilities for each attribute (*X*) to make up the total utility (*U*) (20). It is assumed a reasonable individual would choose the alternative that yields their highest individual benefit, known as utility. This theory argues that an individual should choose alternative A over B if:

Equation 1 $$U\,(X_A, Z) > U\,(X_B, Z),$$

where *U* represents the individual's total utility function based on the presented alternatives with certain attributes and levels, X_A represents the attributes and levels in alternative A and X_B the attributes and levels of alternative B, and *Z* represents the characteristics of the individual (e.g. socioeconomic) that influence their decision (20). The distinct choices observed in a DCE are assumed to reveal an underlying (latent) utility function.

The second theory is random utility theory (RUT), where the value individuals place on an attribute is composed of an explainable (systematic) and an unexplainable (random) component (20). The explainable component is the attributes in the DCE, and the unexplainable is unobserved factors unknown to the researcher (20). Therefore, the error term due to unknown factors that influences respondents' decisions needs to be included in the utility function. Thus, a respondent will choose alternative A over B if:

Equation 2 $$V\,(X_A, Z) + \varepsilon_A > V\,(X_B, Z) + \varepsilon_B,$$

where *V* is the measurable or systematic component of utility estimated empirically, and ε reflects the unknown factors.

A statistical model is used to analyze the influence of attribute levels on choice decisions, and then the strength of preferences for the attributes and attribute levels can be estimated (21). The relative significance of attributes can be evaluated by taking a ratio of the coefficients of each possible pair of attributes to provide MRS values (20). The MRS analysis provides important interpretation of the results that are relevant for clinical practice as the parameter estimates are not intrinsically interpretable (22).

Methodology of discrete choice experiments

In this section I review key guidelines for designing DCEs, seminal systematic reviews and papers in DCEs, and crucial DCEs conducted in the field of oncology pertinent to my study. I then use these papers to provide an overview of the steps involved in conducting a DCE.

Guidelines

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has published three guidelines for DCEs. The first paper published in 2013 was a checklist to facilitate and endorse good research practices for the application of conjoint analysis in the health care setting (23). The second guideline published in 2013 summarized and compared several approaches for constructing experimental designs to assist researchers in evaluating the best approach for their research question (24). The third paper was published in 2016 and contrasted and reviewed several methods of statistical analysis (21). The next sections in this chapter outline the key parts of a DCE based on these guidelines. Any issues identified in the literature and how to resolve them are discussed in detail as they were applied to the development of my study's methods and are expanded upon in chapter 3.

Preference-based studies quality assurance

The checklist for good research practice published by ISPOR included 10 broad 'steps' for DCEs: (1) define research question and the study's context; (2) set the attributes and levels; (3) construction of tasks, (4-5) experimental design and preference elicitation occur simultaneously and are inter-related; (6-7) instrument design and data collection; (8-9) statistical analyses, results, and conclusions; and finally (10) study presentation (23). Bridges et al. provide a solid introduction to the overall approach to designing and implementing a DCE, though there was limited discussion about each stage of the DCE process (23).

A 2013 scoping review by Joy et al. developed a quality assessment checklist for preference-based studies (25). This novel checklist, PREFS, has five criteria and stands for; Purpose, Respondents, Explanation, Findings, and Significance (25). Joy et al. reported that there were no universal standards to review preference research. This paper found of the 61 included studies only one met all five of their quality standards, while 12 studies met four of the criteria (25). This may reflect poor quality across this field of study, or unfortunately the authors did not routinely include this checklist information in their write ups. This may have occurred because this information is not a requirement or part of a common guideline, and they did not expect their work to be included in a systematic review (25). Additionally, this method of quality assessment may need to be improved as it has not been validated, and the authors indicated next steps included a task force to validate a suitable instrument for assessing quality and risk of bias in preference studies (25). The contents of these checklists are discussed in future chapters as they were utilized as an initial planning tool to develop a broad framework for my DCE and to support good research practice.

Health care DCE literature reviews

Three systematic reviews have compared the design of published DCEs in different time periods. Ryan et al. discussed the application of DCEs during 1990 to 2000, de Bekker-Grob et al. updated this paper by reviewing the literature between 2001 to 2008 and comparing their results to Ryan et al.'s findings, and Clark et al. reviewed 149 DCEs published between 2009 to 2012 and compared trends to DCEs discussed in the first two reviews (1, 2, 26). The focus of de Bekker-Grob's paper was experimental design, estimation procedures, and validity of responses (1). I discuss these three papers in the context of changes in practice and the trends in health care DCEs for the issues identified in the next sections. After completion of this search a fourth systematic review with identical criteria was published in 2019 that reviewed the literature between 2013 and 2017 (27). This most recent paper which was published after this study was not used to design the DCE and is not discussed further.

Oncology DCEs

In this section I provide an overview of DCEs completed in the field of oncology that are relevant to my study. I first discuss a systematic review and three pertinent examples of non-prostate cancer oncology DCEs. All DCEs involving prostate cancer patients published in English are then critiqued as the included attributes, and magnitude and significance of part-worth estimates, were important to consider in the design phase of my study (Table 1).

A systematic review by Bien et al. identified 28 DCEs conducted to elicit cancer patients' preferences for treatment between January 2010 and April 2016 (3). There were 29 DCEs published in the PubMed database in this area of oncology since May 2016. This higher number in a shorter time frame indicates DCEs are increasing in popularity. Bien et al.'s review found 56% of the included studies received a score lower than 4 on the PREFS 5-point scale (3). The highest proportion (70%) of included attributes were related to outcome (treatment effectiveness and adverse effects) and had the highest likelihood of being significant to patients' choices (81%). Treatment process and cost attributes were included in the studies less often (25% and 5% respectively), but are generally still important to respondents as 73% and 67% respectively of these attributes were significant (3). With regards to my study, this would indicate the three risk attributes for PSA recurrence and toxicity were likely to be more important to participants, though the treatment process attributes (corresponding to length of EBRT and fiducial markers implantation) tended to be significant when they are included in DCEs.

Non-prostate cancer specific oncology DCEs

In 2014 Havrilesky et al. conducted a DCE out of Duke University that found women with advanced or recurrent ovarian cancer were willing to trade-off progression free survival (PFS) for the benefit of decreased treatment related toxicity (28). Interestingly, this study also elucidated patient preferences with ratings and rankings of the attributes (28), which was duplicated in my study to compare the quantitative DCE findings to this ranking information. The efficacy attribute was PFS because the primary endpoint in trials is usually PFS (28), which was interesting as I also based attributes and levels from clinical trial findings.

In 2015 Muhlbacher and Bethge conducted a DCE in Germany to determine patient preferences for treatment of non-small cell lung cancer (NSCLC). This study found respondents had strong preference for PFS and relief from tumor associated symptoms, while mode of administration was less important but still significant (29). This is noteworthy for my study as it demonstrated the treatment process was significant to patient decisions, and the importance of collecting socioeconomic status (SES) information as three groups of preference selection based on patient factors data were identified (29).

In 2015 de Bekker-Grob et al. conducted a DCE in the Netherlands for surgical management of esophageal cancer that demonstrated patients were willing to trade-off some 5-year survival for an improvement in quality of life after surgery (30). This was a well-designed study that provided evidence in oncology that survival was not the only consideration for patients' curative treatment decisions.

Prostate cancer specific DCEs

Table 1 details the five DCEs conducted with prostate cancer patients, the final two rows are qualitative studies completed with this patient population and are discussed in chapter 3.

King et al. found that respondents were willing to trade off several months of survival benefit to avoid GI and GU side effects that could be expected from EBRT (31). The parameter estimates for the 'mild' levels of erectile dysfunction, loss of libido, and fatigue attributes were not significant, and interestingly corresponded to the actual level of toxicity expected from EBRT (31). The specific attributes and parameter estimates are found in Table 1. These findings provided some evidence that fatigue and sexual side effects did not need to be included as attributes in my study. This study also demonstrated that

Table 1: Five DCEs and two qualitative preference studies for prostate cancer treatment.

Authors	Year	Study description	Attributes/ characteristics	Coefficients
King et al.	2012	DCE in Australia with prostate cancer patients who are recurrence free 3 years after treatment and a random sample of age matched controls exploring importance of attributes decision between EBRT, ADT, and surgery. Negative values indicate no side effects were preferred.	- Urinary leakage - Mild[4] - Severe[5] - Urinary blockage - Mild[4] - Severe[5] - Erectile dysfunction - Mild - Severe - Loss of libido - Mild - Severe - Bowel symptoms - Mild[4] - Severe[5] - Fatigue - Mild - Severe	 -0.20 -1.33 -0.23 -1.08 NS -0.30 NS -0.21 NS -1.23 NS -0.62
de Bekker-Grob et al.	2013	DCE in Europe with prostate cancer patients with localized disease and urologists comparing surgery, EBRT, and active surveillance. The patient group's parameter estimates for EBRT are listed. The negative values indicate less risk is preferred and that active surveillance is preferred to EBRT by patients.	- Risk of urinary incontinence[5] - Risk of erectile dysfunction - EBRT alternative specific - Substantial risk of bowel problems and frequent urination[5] - Main aim is cure - Once a year PSA testing with no risk of new biopsies	-0.18 NS -5.33
Howard et al.	2014	DCE in Australia to determine preferences for PSA screening to detect prostate cancer; population was men aged 40-69 (split into three age groups: 40-59; 50-59; and 60-69 years old). Part-worth utilities varied significantly based on age, younger men more strongly preferred less toxicity (parameter estimates larger in magnitude).	- Men who will die from prostate cancer[3] - Men diagnosed with prostate cancer including over-diagnosed - Unnecessary biopsies[2] - On-going impotence - Ongoing urinary incontinence or moderate-severe bowel problems[5]	40-59: 0.63 50-59: 0.40 60-69: 0.08 NS 40-59: NS 50-59: -0.17 60-69: -0.06 NS 40-59: -2.07 50-59: -0.54 60-69: -0.17

			- Approximate out of pocket expense[1] -0.1 (mean score)
Eliasson et al.	2017	DCE in Europe conducted with metastatic prostate cancer patients to determine hypothetical treatment preferences by determining attribute importance and calculating trade-offs participants were willing to make. A larger part-worth magnitude means an attribute was more important to participants' decisions.	1. Bone pain 2.49 2. Fogginess 0.75 3. Time before chemo needed 0.55 4. Possible drug interactions 0.33 5. Fatigue 0.31 6. Food restrictions related to treatment NS
Nakayama et al.	2018	Japanese DCE; patients with advanced and non-advanced prostate cancer that included attributes for a general treatment. Patients ranked the attributes and these results agreed with the quantitative findings (larger magnitude indicates higher importance).	1. Treatment accessibility[1] 0.507 2. Quality of life[4,5] 0.340 3. Side effects[4,5] 0.262 4. Treatment effectiveness[3] 0.182
Feldman-Stewart et al.	2004	Interview, listing, and trade-off exercises conducted in Canada to determine important characteristics to patient's treatment decision for localized prostate cancer; 10 were identified as the most important; median 4 per participant.	1. Effect on bladder functioning[4,5] 2. Chances of PSA rising[3] 3. Procedure[1,2] 4. Effect on sexual functioning 5. Chances of dying of cancer[3] 6. Effect on length of life[3] 7. Effect on quality of life[1,2,4,5] 8. How treatment works[1,2] 9. Doctor's advice 10. Effect on bowel (colostomy)[5]
Helou et al.	2016	Interviews with newly diagnosed localized prostate cancer patients in the United States to identify categories important to participants' treatment decision; listed categories pertinent to EBRT in Canada; most frequently disliked EBRT due to long duration.	- Evidence of treatment for cure[3] - Likelihood of side effects[4,5] - Duration of treatment, loss of income, travel expense[1] - Length of recovery[4] - Recommendations and general preferences

Parameter estimates with coefficients in parentheses for the DCEs; NS: not significant. Superscript numbers refer to related attributes in my DCE: [1]length of radiotherapy, [2]fiducial markers implanted, [3]risk of PSA recurrence within 5 years of treatment, [4]risk of acute moderate GI and GU toxicity, and [5]risk of late moderate GI and GU toxicity.

including men without a diagnosis of prostate cancer in the sample resulted in preference heterogeneity, indicating men without prostate cancer may make different decisions than men diagnosed with prostate cancer. This finding reduced the external validity and ability to implement the results in clinical practice.

de Bekker-Grob et al. demonstrated in their DCE that for the prostate cancer patient group only one attribute - ‘risk of urinary incontinence’ - was significant for the radiotherapy alternative (32). This finding suggested that GU toxicity is significant for radiotherapy patients and should be an included attribute in my study. This DCE demonstrated urologists had different preferences than prostate cancer patients (Table 1). Since the goal of my DCE was to quantify the trade-offs patients are willing to make and are not concerned with physician beliefs about patient preferences, thus only prostate cancer patients were included in my study.

Howard et al. demonstrated the attribute part-worth estimates for unnecessary biopsies and GI and GU side effects varied by age and were larger for younger men (33). This indicated younger respondents had stronger preference to avoid these attributes than older men (33). This finding was important for my study as it indicated respondent age should be tested for any interaction with toxicity and treatment process attributes (Table 1).

Eliasson et al. conducted a large comprehensive DCE that found management of cancer symptoms was important to patients, even compared with potential increased treatment toxicity (34), in the metastatic castrate resistant prostate cancer population regarding their hypothetical treatment preferences (34). A larger value indicated the specified attribute was more important to respondents’ decisions. For example, avoiding bone pain from metastases was overall more important to respondents than the risk of experiencing fogginess as a symptom. This was pertinent to my study as it provided some evidence that prostate cancer patients were willing to accept increased toxicity if they could be guaranteed what they consider a beneficial outcome (Table 1).

Nakayama et al.’s DCE found opposing preferences between the included groups. This indicated only patients with localized prostate cancer and experience with EBRT should be included in my study to improve the interpretation of the results in the correct context. As in Havrilesky et al.'s study the ranking and quantified DCE results agreed with each other and provided a good method to assess the validity (Table 1).

Attribute selection

A range of attributes relevant to the clinical context should be included in a DCE, and ideally these ought to be identified from multiple sources, such as review of the literature and interviews with the target population (23). Researchers should strike a balance between characteristics important to respondents and what is relevant to the particular decision-making environment (23). The Clark et al. systematic review found a trend towards DCEs including more attributes, with the average 4-6 attributes as found in earlier studies but with a higher proportion of studies including 7-9 attributes (26). Next, consultation with experts and qualitative or preliminary studies with patients can be used to determine the final set of key attributes (23). Qualitative methods involving the target population can improve the face validity of the planned DCE (1). Helter et al. indicated there are four steps for attribute selection: (1) raw data collection; (2) data reduction to the upper limit of 6-7 attributes to minimize the burden on respondents; (3) removing inappropriate attributes; and (4) wording (35). Ryan et al. reported appropriate attributes should be salient to patients and/or policy makers, plausible, and capable of being traded (2). In an expert overview Coast et al. added four further criteria to Ryan et al.'s such that attributes: should include all those important for an individual to make their decision; not too close to the latent construct of the DCE; cannot have such a large impact on the decision that based on the level of this attribute the choices become deterministic; and must be experimentally manipulatable (36). Attribute selection is a multi-step process that requires an extensive literature review, discussion with experts and patients, and careful thought about the final language used to describe the attributes.

Attributes were selected based on a literature search in all the selected oncology DCEs (studies highlighted in the previous section), qualitative interviews with participants in four of the eight studies, interviews with care givers in only one, and interviews with clinical experts in seven (28, 30-34, 37).

Experts were asked what would be relevant to patients and the context of previous findings from clinical trials. The number of attributes selected varied from four to eight. All of the selected oncology DCEs first conducted a pilot study in an interview-based setting, size ranging from 5 to 12, to assess interpretation and validity (28-30).

Level selection

The levels for each attribute should be relevant to the target population and can include plausible and/or hypothetical levels, though improbable level ranges could result in respondents' rejection of a choice decision (23, 38). Two of the oncology DCEs set the levels based on data from randomized clinical trials (28, 32). Levels do not need to be clinically possible as participants can be told the levels describe possible future options (38). All studies had two to four levels per attribute, which is consistent with the suggested range (28, 30-34, 37).

Level ranges should be wide enough to induce trade-offs away from strongly preferred attributes such as efficacy (38). If a wider range is used for continuous attributes this can lead to higher importance being placed on that attribute (39). Levels are generally within one standard deviation (SD) of the mean, or include values from within an interquartile range, and evenly spaced attribute levels can be useful for interpreting the estimated effects of numerical attributes (40).

Experimental design

The experimental design is a key component of conducting a DCE. A full factorial design includes all possible combinations of attributes (A) and levels (L) which gives L^A possible alternatives (1). For example, a DCE with four attributes each with three levels results in 81 potential combinations of attributes and levels. A fractional factorial design consists of a smaller sub-set of choice sets are selected using research design principles (e.g. balance and orthogonality). This creates a DCE experimental design that can obtain sufficient information about the trade-offs between each attribute while also minimizing respondent effort (1). D-efficient designs allow researchers to assume *a priori* that parameters are equal to zero, or to use prior assumptions from a literature search or pilot study data (1). When there is uncertainty over the prior assumptions for the parameter estimates a Bayesian efficient design (resulting in a D_b-error) is recommended (41).

Construction of choice sets

Choice sets are the mechanism by which profiles are presented to respondents. Each choice set consists of two alternatives (in a forced choice DCE) or two alternatives and an opt-out option. There are numerous ways to design these choice sets and based on the purpose and target population of the preference study the best method can be deduced by researchers (41). Statistical and response efficiency must be balanced in DCEs. Statistical efficiency is improved by asking many difficult trade-off questions, and response efficiency is improved by asking a smaller number of comparatively easier trade-off questions (24).The optimal amount of choice sets depends on the number and complexity of the attributes, and the target population (23). Doubling the number of choice sets from 10 to 20 per respondent is more effective than increasing the precision of the survey with twice the number of respondents (42). However, this must be balanced with cognitive fatigue as when more than 20 choice sets were included in conjoint analysis surveys the respondents tended to complete the questionnaires much faster and reliability was lost (42). From among the relevant oncology studies Muhlbacher and Bethge reduced the cognitive burden on the participants who had stage IV NSCLC by systematically overlapping the ‘moderate’ attribute levels to ensure the number of differences respondents had to evaluate was reduced in each choice set (29). My study aimed to include less than 20 choice sets per respondent and allowed overlap of levels to reduce the cognitive effort for respondents.

Statistical and response efficiency can be balanced by incorporating blocking into the experimental design (24). Blocking is where each respondent randomly answers only a portion of the entire choice set of questions that the researcher included in the experimental design (24). For example, in de Bekker-Grob et al.'s esophageal cancer and Muhlbacher and Bethge's DCEs they divided 24 choice sets over two versions of the questionnaire (12 choice sets per participant) (29, 30). Interestingly, Havrilesky et al. generated 120 choice sets and blocked them into 10 questionnaire versions (28). Blocking increases response efficiency by reducing the cognitive effort for respondents, however the limitation of this was the desired statistical property of no correlation among attribute levels may not hold for each block due to disruptions to orthogonality and balance (24). Of note, more recent studies tend to present more choices per respondent with the highest proportion of DCEs presenting 9 to 16 choice sets (26). From the relevant oncology DCEs Howard et al.'s study had 15 choices sets per participant (33).

All the selected oncology DCEs reported using the experimental design methods discussed in the ISPOR paper (24). The de Bekker-Grob et al., Muhlbacher and Bethge, and Howard et al. studies used Ngene software to generate designs that were highly efficient (maximized D-efficiency or minimized D-error) (29, 30, 33). Havrilesky et al. used Sawtooth Conjoint Value Analysis designer to generate the paired alternatives for the choice sets using a balanced-overlap experimental design (28). Eliasson et al. and King et al. used orthogonal fractional factorial design with combinations paired with fold-over that resulted in 18 choice sets with two hypothetical scenarios per choice question in the Eliasson et al. study, and 18 choice sets with three versions in the King et al. study (31, 34). These latter methods are now considered outdated and recent DCEs utilize design software such as Ngene, which is further detailed in chapter 3 as part of my methods (26).

Opt-out or status quo alternatives

The inclusion of an opt-out option (neither, status quo, or no treatment) helps prevent a respondent making a forced choice (i.e. one of the alternatives must be chosen), which can falsely increase the strength of preference associated with any of the alternatives (43). Lancsar and Louviere instruct that the types of choice formats must be evaluated carefully and should simulate the actual choice of interest as closely as possible, for example whether a binary option or including an opt-out alternative would be most realistic for the situation at hand (40). Allowing respondents to opt-out and select no alternative, or a status quo option with set levels for each attribute, should be considered. However, if in the real world patients would have to select one of the alternatives in the scenario, then no opt out is required and a forced choice analysis would have more easily interpretable results (40). If the opt-out option is selected by the majority of the participants there is the risk of not obtaining sufficient information on preferences to estimate the statistical model (43). From the relevant oncology DCEs Howard et al.'s study had an opt out (no screening) option and was the only one examined oncology study that did so (33). Since PSA screening is not recommended to the general population, unlike the treatments to the target population of cancer patients in the other DCE studies, an opt-out option was ideal in that scenario.

DCE type: Generic vs. labelled

A generic DCE is one in which the alternatives are unlabelled, such that the choice set offers two scenarios that only differ in terms of the specified attributes and levels (20). For example, each alternative is only labelled by A or B and described by the attribute labels. An alternative-specific DCE compares labelled alternatives that have meaning to the participants (20), such as in de Bekker-Grob et al.'s prostate cancer DCE where one alternative is always surgery and the other is always radiotherapy (32). The benefit of a labelled DCE is the participants are familiar with the context of the scenarios which reduces the cognitive burden, however this may lead to participants ignoring the trade-offs between the specified attributes in the DCE and instead using unknown attributes or personal biases to make their choice (20). For example, a participant believes they would never want surgery they may select the radiotherapy alternative in each choice set. Therefore, the generic version is more common in health services literature and the rest of the selected oncology DCEs were generic (20).

Sample size

The minimum sample size to estimate models based on DCE data has many approaches in the literature. In Orme's seminal paper he noted for investigational work and developing hypotheses between 30 and 60 respondents is a plausible sample size (44). Orme also proposed a rule of thumb for minimum required sample size that used the number of choice sets, number of alternatives per task, and the largest number of levels for any one attribute (44). Muhlbacher and Bethge used this reasoning in their oncology DCE to determine the minimum required sample size (29). Leaders in the field Lancsar and Louviere report in their experience 20 respondents per questionnaire version was able to estimate reliable models (40). However, if there was significant post hoc analysis to identify co-variate effects then a larger sample size would be required (40). Eliasson et al. targeted 300 participants as per the ISPOR paper recommendation, but had difficulty reaching this target and ended up recruiting 285 participants (34). In the prostate DCE de Bekker-Grob et al. reported at least 40 to 100 participants were sufficient for reliable statistical analyses (32). There are disadvantages to using rules of thumb to calculate sample size as they are not always accurate or reliable since they do not consider parameter effect sizes (45).

A minimum sample size can be calculated based on Rose and Bliemer's parametric approach, where one assumes that the estimated coefficient is normally distributed (46). This approach focuses on the minimum sample size required based on the most "critical" parameter, in other words a minimum sample size is provided for each parameter (46). This parametric approach can only be used if prior parameter estimates are available and not equal to zero (46). Rose and Bliemer's approach may be more reliable than rules of thumb, however it is not specific to a certain hypothesis and does not depend on a desired power level (45).

To determine the optimal sample size a systematic review by de Bekker-Grob et al. indicated that there were five required elements: (1) significance level; (2) power level; (3) statistical model used; (4) initial belief about parameter values (requires pilot study); and (5) DCE design (45). The ability to recruit participants tends to be limited in health care settings as there are only a set number of eligible individuals, and helpfully this paper provides a method to calculate minimum required sample size based on the desired power and significance level, as well as the power level based on a certain sample size (24, 45). From the relevant oncology studies Havrilesky et al. reported their sample size of 95 (consented 100 initially) was sufficient to demonstrate statistical significance for six of the seven attributes, though a larger sample size would have provided tighter confidence intervals (CI) and additional power to make comparisons between groups (28). This approach is most statistically sound as an investigator is able to calculate the power and significance of their study.

Analysis approach in DCEs

In DCEs the impact of each attribute is the independent variable and the participants' choices are the dependent variables (34). A regression analysis can be used to linearly relate the probability of choosing one alternative over another based on the attributes and levels presented in the choice task. This model assumes the probability of selecting one alternative is a linear function of the attribute levels:

Equation 3 $$\mathbf{P(choice)} = \boldsymbol{\beta_0 + \Sigma_i \beta_i X_i},$$

where X_i is the level of attribute i, β_0 is the intercept, and β_i is the preference weight for attribute i (21). Based on the relation between choices and attribute levels the preference weights can be estimated through several types of estimators (21).

DCE choices have commonly been analyzed using McFadden's conditional logit model (MNL) (21). This is a limited dependent-variable method and differs from regular logistic regression as the data are grouped and the likelihood is calculated to each group (21). This is unique from a logit model because McFadden applied it to choice behaviour and demonstrated it was consistent with economic theory and RUT, and he derived a regression model that relates choice to the attribute levels present (21). From equation (eq.) (3), the *P(choice)* variable is equal to 1 for the alternative that was chosen, and equal to 0 for the alternative that was not (in a forced choice DCE) (21). As seen in eq. (1) and (2), the utility associated with an alternative is assumed to be a function of known attribute levels and unknown characteristics (random error term), and that an individual will select the alternative that maximizes their utility (21). A statistically significant coefficient indicates the respondents differentiated between attribute levels (30). The utility score has no direct interpretation but can be used to compare to other scenarios within the same DCE (30).

McFadden then applied the assumption of extreme-value distribution of ε_i where the error terms are independent of each other, for example ε_M for M alternative is unrelated to ε_N for another N alternative. This resulted in the logit model:

Equation 4 $$\mathbf{P(choice}=\boldsymbol{i}) = e^{V(\beta, xi)} / \Sigma_j e^{V(\beta, xj)},$$

where $V(\beta, x_i)$ is the known attribute levels for alternative *i*, and *i* is one alternative among a set of *j* alternatives (21). Therefore, the probability of selecting alternative *i* is a function of both the attribute levels of *i* and the attribute levels of all other alternatives in a choice task (21). The probability of choosing the other alternative in a forced choice DCE (total two alternatives) is 1 minus the probability of choosing *i* (21). This results in a model that never has less than 0%, or greater than 100%, chance of selecting a specific alternative (21). Thus, MNL models result in good estimates of patient preference data. However, as discussed in the next section there are key limitations of the MNL model that has led to most researchers analysing the data with more flexible models (38).

MNL models assume the utility is linear and additive in parameters, and the interaction of factors such as age and health history can be tested (38). Two key limitations of MNL are that the model does not account for scale nor preference heterogeneity (21). Since DCE data are commonly panel in nature results can be biased, and in small samples or if substantial preference heterogeneity is present the model may not converge (21).

Assumptions of MNL

MNL has four assumptions: (1) constant error terms (homoscedastic) across observations; (2) respondents have homogeneous preferences (or that preferences depend on observable characteristics);

(3) no panel data (no correlation allowed for within responses); and (4) the IIA property known as independent errors where there is independence of irrelevant alternatives (for a specific respondent the ratio of the choice probability of any two alternatives is unaffected by other alternatives) (21). The issue with the first assumption is that individual preferences not captured by observable characteristics (attributes and levels) will carry that preference across choice occasions. This would not allow for multiple observations from respondents (21). Preference heterogeneity is very unlikely to not exist within a population of individuals, and the issue with MNL is it assumes that all respondents have the same preferences that will yield a single set of preference weights. MNL does not account for systematic variations in preferences across respondents (21). The assumption of independence of irrelevant alternatives (IIA), also known as Sen's property, states: if an alternative A is chosen from a set (A,B), and A is also an element of a subset (A,B,C), then A must be chosen from (A,B,C). That is, eliminating some of the unchosen alternatives shouldn't affect the selection of A as the best option. Therefore, C is irrelevant to the choice between A and B. Some experts believe this is too strict of an axiom, and it has been shown that human behaviour rarely adheres to this axiom which leads to unrealistic predictions with the MNL model (21). These assumptions result in the MNL model being too restrictive in describing human behaviour, which likely restricts the realism of the analysis. There are multiple further models to analyze choice data that is obtained from different assumptions about the distribution of the random terms and is believed to increase the behavioural realism of choice models (21). One key limitation these other models relax are heterogeneity, discussed in more detail in the next two sections.

Scale heterogeneity

Scale heterogeneity is the differences in the error variance of respondents' preferences, and therefore impacts the size of the parameter estimates (47). The difference in variance between individuals is introduced because some participants may be quite certain of their choices or use decision heuristics that reduce variance, while some may not fully understand the task or make mistakes which increases the variance (43). A systematic review of health care DCEs published between 1990 and 2016 by Wright et al. found that scale heterogeneity is generally not addressed even though scale heterogeneity was present in 65% of studies that tested for it (47). Wright et al. recommended that identification and addressing scale heterogeneity should be a fundamental component of DCE analysis, otherwise the results are likely to be biased (47). Methods used to test for scale heterogeneity include coefficient plots, heteroscedastic MNL models, Swait and Louviere tests, generalized multinomial logit models, and scale-adjusted latent class analysis (47).

Preference heterogeneity

Preference heterogeneity is nearly always present as preferences vary across individuals since people have different beliefs and values (47). Several systematic reviews of health care related DCEs found significant heterogeneity in patient responses (47-49). This is important to my study as the objective was to determine if there are certain groups or characteristics of prostate cancer patients with localized disease who were willing (or more willing) to trade-off increased risk of PSA recurrence and radiotherapy toxicity for the benefit of shorter treatment time. There was evident heterogeneity of preferences in the oncology DCE by Havrilesky et al., who found about a quarter of respondents ranked an attribute other than PFS (efficacy) as most important, and 79% of women accepted reductions in PFS for improvements in toxicity or treatment convenience (28). This demonstrated that a large proportion of participants were making choices based on attributes other than efficacy (28). Clark et al.'s systematic review found a trend in more recent DCEs towards flexible and appropriate econometric methods to analyze DCEs to ensure preference heterogeneity is assessed for to inform the results (26).

The SES and health information is generally collected from each respondent to allow testing for systematic differences and heterogeneity in preferences that might be associated with variability in choice behaviour (23, 38). It is pertinent to assess respondents previous history with event (e.g. side effect) severity and frequency (if any) to understand fully how this affects patient choice (50). SES data was collected in all the oncology DCEs, and generally included: age, gender, ethnicity, level of education, and household situation. Most studies also collected clinical data that usually included cancer stage, previous treatment experience, performance status, and some studies collected patient-reported outcome (PRO) data with validated symptom and quality of life surveys. From the relevant oncology DCEs Eliasson et al. and de Bekker-Grob et al.'s esophageal study reported the SES statistics descriptively, and King et al. calculated likelihood ratio tests for the effect of age on the attributes (30, 31, 34). In de Bekker-Grob et al.'s prostate DCE the interaction between treatment choice and participants' SES factors entered the model and were tested for significance (32). Three studies ran separate subgroup analyses to test for differences in SES and/or clinical factors (e.g. non-advanced vs. advanced prostate cancer) (28, 33, 37). One study analysed the latent class responses accounting for SES and clinical factors (29).

Models to address heterogeneity

Several methods of analysis have been used to address the issues with conditional logit (21). Clark et al.'s systematic review found a trend towards use of more flexible econometric methods to analyze DCEs to ensure preference heterogeneity was assessed for (26). One method is the mixed multinomial logit (MXL), also called random parameters model, that is an extension of the conditional model. MXL relaxes restrictive assumptions by allowing for preference heterogeneity of attributes and correlation between the choice tasks completed by each individual (21). MXL models assume there is a continuous distribution of preference weights across the sample reflecting differences in preference, and it models the distribution for each attribute level (43). However, this model is more difficult to use, requires researchers to make assumptions about random parameters and their distributions, and may require larger sample sizes (21).

Another method is latent class (LC) modeling that accounts for heterogeneous effects by determining a number of classes or groups of respondents at the analysis stage (47), and within each class the preference weights are estimated using conditional logit (21). This assumes there is no difference in preference heterogeneity between groups, and results in discrete distributions for the part worth-estimates (47). The latent-class model (LCM) provides a good method for measuring preference heterogeneity and does not require the researcher to make specific assumptions about the distribution of parameters across individuals. However, researchers must select the classes themselves, the model likely requires larger sample sizes due to the higher number of parameters to be estimated, and specialized software is necessary (21). This method works well for DCEs with the objective of estimating how preferences vary by individual respondent characteristics (23), which was one of the objectives of my study.

A paper by Ku et al. proposed a method to detect outliers who have unusually high respondent variability called utility deviation (UD) (51). The identified outliers were those with a large UD and had vague preferences when making choices based on reliability and validity tests (51). This outlier detection could be used to exclude misleading data which may improve the quality of estimation and resultant models (51). Utility outliers and respondents with non-rational responses could have their preference responses removed from the data set. This would improve internal validity but reduce external validity (52, 53). In the literature it is recently being advocated that all data should be included, as removing data can increase sample selection bias and decrease statistical efficiency and power of the models (23, 52).

From the relevant oncology DCEs Eliasson et al. and Nakayama et al. both used the MNL model, and Eliasson et al. reported not accounting for heterogeneity was a limitation of their study (34, 37). Four oncology studies analysed the data with MXL models to take preference heterogeneity and correlation between the choice tasks into account (28, 30, 31, 33), and two studies utilized LCM (29, 32).

The MRS is the rate at which an individual is likely to give up one level of an attribute in exchange for another attribute, and this was calculated by multiple oncology DCE papers (29-31, 33). The esophageal de Bekker-Grob et al. study calculated the willingness to trade-off 5-year survival to achieve an improvement in side effects using the ratios between coefficients with the attribute ‘5-year survival’ as the denominator (30). Havrilesky et al. and Eliasson et al. calculated odds ratios (ORs) controlling for unobserved heterogeneity in preferences, which is similar to MRS as the goal was to estimate the rate participants' were willing to trade off units of one attribute in exchange for those of another attribute, while maintaining the same level of utility (28, 34).

Limitations of discrete choice experiments

In this section several areas of concern that threaten the validity and generalizability of DCEs are discussed. The limitations of the relevant oncology DCEs highlighted previously are reported.

Simplifying heuristics

DCEs are partly so cognitively demanding due to the simultaneous comparisons of multiple attributes and levels of attributes that must be contemplated for each choice an individual makes (54). Simplifying heuristics could be used by participants to reduce the number of attributes they have to weigh in their decision by ignoring certain attributes (39). It should be assumed that participants select their preferred alternative according to the overall value they place on each treatment (39). Some participants may focus on a subset of attributes using simplifying heuristics to streamline their choices, and this method is appropriate if patients have considered all attributes and decided that some are unimportant regardless of level (39). However, if participants do not trade-off due to other reasons like not understanding a certain attribute, or a range of levels does not seem appropriate to them, then this can lead to biased responses (39). One study compared response rates and respondent's self-reported perception of the uncertainty of their answers for DCEs with either 5, 9, or 17 choice sets (55). There was higher response variance when the 17 and 5 choice set DCEs were compared, and the authors believed this indicated higher cognitive burden above a certain threshold (55). Some differences in preferences were found across the three groups, suggesting respondents changed their decision making strategies during the answering process, though the investigators could not determine if this was due to positive learning where respondents become more convinced about their decision rule, or that respondents changed their strategy to a less cognitively demanding one such as adaptation of heuristics and use of non-compensatory decision making (55). Overall, the literature indicates respondents could manage multiple choice sets without problems and there was no serious influence on response behavior with additional choice sets (55). If there is concern about the cognitive burden the number of choice sets can be reduced or investigators could consider face-to-face elicitation (48). There is a gap in the literature identifying how participants decide to make their choices in DCEs, and how uncertain or confused they are about the tasks asked of them.

Omitted variable bias

Researchers may cause omitted variable bias by excluding key attributes, and this risk needs to be weighed against increasing the complexity of a DCE by including too many attributes (40). Not every attribute important to every possible participant needs to be included in a DCE, though attributes important to a majority of participants should be included (40). Omitted variable bias can be avoided at the design stage via a pilot study that includes potential participants from the target population to identify possible attributes, and ensure participants are not considering any other characteristics of the options that would alter their evaluation of the included attributes (3, 22). Two recent systematic reviews of health care related DCEs noted a decrease of qualitative methods to inform attribute selection, which could lead to omitted variable bias if the target populations decision framework is not known prior (26, 56). Over three quarters of the investigators in one of the reviews used interviews as the qualitative method, usually for attribute and/or level selection (66%), and less commonly to pilot the DCE (18%) (56). Unfortunately use of qualitative methods was not consistently reported, and this lack of reporting may lead to the incorrect inference that qualitative methods are not useful in designing DCEs (56).

Harrison et al. published a systematic review comparing DCEs published during 1995-2015 to examine patient and health care provider preferences (57). These types of DCEs were more likely to include attributes related to process (63%) compared to outcomes (29%) than in the DCEs designed for only cancer patients as in Bien et al.'s systematic review (3, 57). Patients tended to place more importance on issues of process attributes (e.g. safety, delivery, and timing) than health care providers (57). The greatest discordance was mortality, as health care providers placed higher weight to this outcome than patients tended to (57). Therefore, the target population needs to be surveyed during the design stage of a DCE to ensure attributes important to them and that may be missed by researchers are included.

Understanding of risk

Special attention is required to properly describe risk attributes as participants may have difficulty interpreting probabilities (40). There is a general lack of understanding about risk across all demographics and an overall low level of numeracy in the general population, and risk has not always been successfully communicated to DCE participants (58, 59). For example, participants may have difficulty evaluating small probabilities, and conversely even have difficulty understanding if the outcome would be a disastrous one (59). A person's familiarity with the risk, and possible heuristics that individuals use to translate population risks into personal risk, need to be considered (59). Individuals make more informed decisions when the information is relayed in an understandable format that reduces the cognitive effort required (60). For example, keeping the denominator constant, incorporating visuals, not using decimals, and displaying absolute risks instead of relative risks should reduce effort and increase comprehension (60). Harrison et al. conducted a systematic review and published key recommendations for explaining risk to respondents and incorporating better communication of risk in the choice sets, such as visual aids (61). The authors noted improvements in reporting and representation of risk throughout the entire design and analysis stages of DCEs were needed (61). Interestingly, a study by Veldwijk et al. found opposing results to Harrison et al.'s review (62). The authors reported that independent of educational level and health literacy that words, compared to graphics, resulted in higher choice consistency and more valid attribute level estimates (62). The authors of this study recognized that these findings were opposed to the results of multiple other studies in this area, and reasoned it may be more difficult for respondents to compare multiple graphics and this may reduce the comparability between studies (62).

A qualitative analysis of how participants understood a DCE found only a small number of participants had difficulty evaluating probabilities, and was generally related to deciding on the importance of rare but severe side effects (63). Most participants in this study were able to accurately quantify the effectiveness attribute (63). It is important to consider potential participants previous exposure to the topic, understanding of risk, and framing, and ensure the background information about each attribute presented to participants is done such that no attribute is biased as more important (61). The framing (positive or negative) used to present risk has an influence on respondents' choices, however there is a lack of evidence supporting if positive or negative framing is best (58). Researchers should also consider showing the probability of both the event and non-event (39). Therefore, close attention to how risk is presented and described to respondents will help the less numerate make informed decisions. This is important for my study as risk was used as levels for the effectiveness and toxicity attributes.

Validity and reliability

Preference studies require a logical soundness of the patient preference information to allow for validity of results (52). Validating stated-preference assessments is difficult due to their hypothetical nature, and whether respondents would make the same choices if they were presented with these options in the real world and had to experience the consequences is unclear (52).

A systematic review by Mandeville et al. collated a list of 13 criteria to assess the validity of DCEs at the choice task design, experimental design, conduct, and analysis stages (43). The researchers drew on the quality checklist published by Lancsar and Louviere and de Bekker-Grob et al.'s systematic review (1, 40). All studies included in this systematic review failed at least some criteria, which underlines that best practices at each stage of design must be considered to ensure validity (43). An overview published by Janssen et al. discussed the tests of validity and reliability that have been applied to DCEs in the literature to determine if choice decisions were reliable and valid (52). There were four intersecting groups; validity, reliability, measurement, and choice (52). Measures of validity capture how accurately an instrument measures the outcome of interest, while measures of reliability capture how consistently an instrument measures the outcome of interest (52). Choice validity and choice reliability assess accuracy and consistency, as well as determine if the results of the study adhere to assumptions specific to choice behaviour (52). The tests for these four overlapping concepts of internal validity are outlined in detail in Appendix B and their implementation in my study are expanded upon in chapter 3.

Response bias

Self-completed DCEs via mail or online are common in the literature, however they tend to have low response rates and may result in lower quality data than in-person interviews (23, 40). Interviews with a person present to answer participants questions when necessary improved face validity (23, 40). Interviewer-led studies may be less prone to measurement error (63). However, in the systematic review by Bien et al. only 18% of surveys were interviewer-led (3). This was likely related to the increased cost and larger time commitment required, and in-person interviews have their own biases due to the recruitment methods and possible incentives used to enhance the response rate (3). However, online interviews may not be representative of the target population, especially if an older target population is needed (40, 64). In de Bekker-Grob et al.'s 2012 systematic review compared to Ryan et al.'s earlier findings there was increased use of interviewer-led DCEs (19% compared to 9%), and the authors indicated that internet-based surveys needed to be further researched (1). A 2013 study comparing in person and online modes of assessment found no significant difference in the choices and preferences between the two samples, though the in person assessments were significantly longer and therefore there was concern the online sample did not fully engage with the task (64). Additionally, the online respondents were found to be more educated which is concerning if a DCE is to be completed with older participants who may also be ill (64), and therefore may have greater difficulty with online applications.

The selected oncology studies varied in how they contacted eligible participants. Some studies contacted eligible patients in clinic, however most mailed out an information letter with the DCE interview if it was on paper, or a link to access an informative video and the computer assisted questionnaire. Eliasson et al. had both online and paper versions based on patient preference (34). The Muhlbacher and Bethge, Eliasson et al., and Howard et al. studies used an external research institute to contact eligible participants (29, 33, 34). For purely mailed letters the response rate varied from 68% in de Bekker-Grob et al.'s esophageal study to 92% in King et al.'s study, and for the clinic and mail approach the percentage of patients that enrolled was only 39% in Havrilesky et al.'s study up to 73% of patients in the prostate de Bekker-Grob DCE (28, 30, 32).

The response rate, and basic sociodemographic data of non-responders, should be recorded (40) as non-response bias may exist if responders are systematically different from non-responders (39). The response rate was generally an issue in the reviewed oncology studies that reported this rate, and de Bekker-Grob et al. noted that selection bias could not be excluded for their study's response rate of 68% (30). Two studies reported the characteristics of non-participants was unknown and therefore selection bias could not be ruled out (29, 34).

Patient preference survey methods

DCEs are an ideal method to determine patient preferences for several reasons. First, compared to other stated preference techniques such as willingness to pay (WTP), visual analogue scale (VAS), standard gamble (SG), time trade-off (TTO), rate or scale alternatives, and other conjoint analysis methods such as best worse scaling (BWS), a DCE presents a relatively straight forward scenario and one that more closely imitates a real world decision (20). Second, DCEs are best suited in assisting health care practitioners to make clinical management decisions using patient preferences since DCEs are the ordinal preference method that can calculate trade-offs between alternatives (20). These quantitative trade-off values can be difficult to elucidate using ranking or cardinal preference methods such as TTO or SG (20).

DCEs are of greater benefit than ranking exercises as individuals can find it more difficult to distinguish between less-preferred alternatives (those at the bottom), thus producing biased rankings (20).

Additionally, this complete ranking can be more cognitively challenging than a DCE (20). However, it has been argued that ranking studies can produce more efficient preference estimation compared to DCEs (20), and DCE tasks can be cognitively demanding because respondents must make simultaneous comparisons of multiple attributes and levels before they make their choice (65).

DCEs are advantageous over TTO and SG studies as they allow for the simultaneous assessment of multiple attributes, rather than a dichotomous choice between one attribute and survival (54). Individuals may value many aspects of quality of life and more treatment attributes than survival (54). Conjoint analysis based experiments that involve subjective judgements by many decision makers are better able to determine individuals' willingness to accept treatment-related risks in exchange for treatment benefits, while direct elicitation methods such as SG and rating exercises determine the relative importance of each of these different types of outcomes (59).

In the last few years maximum difference conjoint analysis, known as BWS, has become increasingly popular (66). This type of conjoint analysis can be relatively easy to determine participants' preferences over a wide range of health care topics (66). One type of BWS, called multi-profile, is most similar to a DCE. The BWS elicits additional information than a DCE on the least preferred alternative as individuals are presented with a minimum of three alternatives and respondents are asked to indicate the best and worst (66). The benefit of this type of study is it requires less participants, and the drawback is it is more cognitively demanding (66). There is greater consistency within DCEs (tested with repeat choice sets) even after controlling for the higher likelihood of selecting the correct repeat choice tasks by chance alone in DCEs (two compared to three or more alternatives in BWS) (67). Potoglou et al. compared DCEs and a simpler type of BWS called profile case where respondents only choose between attributes (not full choice sets) (68). This study found similar preference patterns and no significant difference between the part-worth estimates derived from the analyses (68). Though both methods are valid there is greater trading between attributes and levels in DCEs as more participants exhibited dominant preferences for a single attribute in profile case BWS tasks (67). Therefore, DCEs provide better trade-off information for researchers and may have greater internal validity.

Suggestions for Future Research

There is growing sentiment that benefit-risk evaluations should take patient perspectives into consideration, and this requires that patients' values be quantified (59). Physicians tend to underestimate the side effects and the impact of the characteristics of a treatment on patients' decision making, and overestimate the effectiveness of treatments (69). In many studies it was reported that patients focused on the quality of the process (69). Therefore, preferences of different stakeholders are not always congruent, and patients' preferences should be sought (69). There is evidence in the literature that a preference for convenience exists among respondents' independent of health outcome (70). Convenience related processes include distance to travel, duration of intervention, frequency of intervention, and waiting time (70). Multiple studies have demonstrated that treatment related and convenience attributes can be just as important, and sometimes more important, to patients than efficacy attributes (48, 70, 71). This is supported by the selected oncology DCE studies that overall found a modest but positive value placed by patients on convenience attributes. Therefore, it is feasible to design a DCE with the objective of quantifying how the length of radiotherapy impacts prostate cancer patients' EBRT decision to treat localized disease.

Suitable prostate cancer patients with localized disease can be treated with CFRT, MH, and sometimes SBRT as well. Therefore, a DCE is an ideal experimental design to determine patients' preferences for hypofractionation and their willingness to trade off the possible increased risk of toxicity and extra intervention with fiducial marker implantation, for the convenience of shorter treatment time. No patient preference data currently exists exploring preferences of patients with localized prostate cancer for EBRT. My study provides great value to the literature as it reveals what aspects of EBRT prostate cancer patients specifically value and how this varies based on age and other patient factors. A DCE conducted with this patient population is ideal as multiple attributes and trade-offs can be quantified and applied to the clinical trial findings, and patient factors can be tested to determine if they significantly impact decisions. A DCE is the most realistic and similar to a real-world scenario of the available patient preference survey methods, while still being at an understandable level to this population that includes many older men. The results of this study can help clinicians in their decisions on how best to treat their specific patient with localized prostate cancer by considering all pertinent patient factors and the multiple methods available to treat these patients.

Few published studies have compared risk tolerance estimates to actual clinical data, a necessary condition if the results of these studies are to be used to inform regulatory benefit-risk decisions (59). The design of this DCE incorporated clinical trial efficacy and toxicity outcomes into the attribute levels, which was done in some of the selected oncology DCEs but is new to EBRT preference research with localized prostate cancer patients.

This DCE was designed by following the best practices obtained from the most recent literature on DCEs that were discussed throughout this chapter. The pros and cons of the selected oncology DCEs were carefully considered during the design for this study. Therefore, based on the best practices the ensuing steps and processes were followed in this study: pilot study; all included attributes were based upon a thorough literature review; level selection was only based on RCTs; efficient choice set construction; best methods to calculate minimum sample size; SES information and possible pertinent factors were collected from responders and non-responders; face to face elicitation instead of computer-assisted; and data were analyzed with extended models after initial analysis with conditional logit. These steps are further elucidated in chapter 3.

References

1. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012 Feb;21(2):145-72.
2. Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. Appl Health Econ Health Pol. 2003 Feb;2(1):55-64.
3. Bien DR, Danner M, Vennedey V, Civello D, Evers SM, Hiligsmann M. Patients' preferences for outcome, process and cost attributes in cancer treatment: a systematic review of discrete choice experiments. Patient. 2017 Oct;10(5):553-65.
4. 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. 2018 Mar;199(3):683-90.
5. Dearnaley D, Syndikus I, Mossop H, Khoo V, Birtle A, Bloomfield D, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. Lancet Oncol. 2016 Aug;17(8):1047-60.
6. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys. 1995 Mar;31(5):1341-6.
7. Moon DH, Efstathiou JA, Chen RC. What is the best way to radiate the prostate in 2016? Urol Oncol. 2017 Feb;35(2):59-68.
8. Benjamin LC, Tree AC, Dearnaley DP. The role of hypofractionated radiotherapy in prostate cancer. Curr Oncol Rep. 2017 Apr;19(4):30.
9. Rodda S, Tyldesley S, Morris WJ, Keyes M, Halperin R, Pai H, et al. ASCENDE-RT: an analysis of treatment-related morbidity for a randomized trial comparing a low-dose-rate brachytherapy boost with a dose-escalated external beam boost for high- and intermediate-risk prostate cancer. Int J Radiat Oncol Biol Phys. 2017 Jun;98(2):286-95.
10. Chin J, Rumble RB, Kollmeier M, Heath E, Efstathiou J, Dorff T, et al. Brachytherapy for patients with prostate cancer: American Society of Clinical Oncology/Cancer Care Ontario joint guideline update. J Clin Oncol. 2017 May;35(15):1737-43.
11. Loblaw A, Liu S, Cheung P. Stereotactic ablative body radiotherapy in patients with prostate cancer. Transl Androl Urol. 2018 Jun;7(3):330-40.
12. Catton CN, Lukka H, Gu CS, Martin JM, Supiot S, Chung PWM, et al. Randomized trial of a hypofractionated radiation regimen for the treatment of localized prostate cancer. J Clin Oncol. 2017 Jun;35(17):1884-90.
13. Lee WR, Dignam JJ, Amin MB, Bruner DW, Low D, Swanson GP, et al. Randomized phase III noninferiority study comparing two radiotherapy fractionation schedules in patients with low-risk prostate cancer. J Clin Oncol. 2016 Jul;34(20):2325-32.
14. Incrocci L, Wortel RC, Alemayehu WG, Aluwini S, Schimmel E, Krol S, et al. Hypofractionated versus conventionally fractionated radiotherapy for patients with localised prostate cancer (HYPRO): final efficacy results from a randomised, multicentre, open-label, phase 3 trial. Lancet Oncol. 2016 Aug;17(8):1061-9.
15. Lee WR, Koontz BF. Moderate hypofractionation for prostate cancer. Transl Androl Urol. 2018 Jun;7(3):321-9.
16. Dang A, Kupelian PA, Cao M, Agazaryan N, Kishan AU. Image-guided radiotherapy for prostate cancer. Transl Androl Urol. 2018 Jun;7(3):308-20.
17. Haque W, Butler EB, Teh BS. Stereotactic body radiation therapy for prostate cancer-a review. Chin Clin Oncol. 2017 Sep;6(Suppl 2):S10.
18. Widmark A, Gunnlaugsson A, Beckman L, Thellenberg-Karlsson C, Hoyer M, Lagerlund M, et al. Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year

outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial. Lancet. 2019 Aug;394(10196):385-95.
19. Morgan SC, Hoffman K, Loblaw DA, Buyyounouski MK, Patton C, Barocas D, et al. Hypofractionated radiation therapy for localized prostate cancer: executive summary of an ASTRO, ASCO and AUA evidence-based guideline. J Urol. 2019 Mar;201(3):528-34.
20. Ali S, Ronaldson S. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. Br Med Bull. 2012 Sep;103(1):21-44.
21. Hauber AB, Gonzalez JM, Groothuis-Oudshoorn CG, Prior T, Marshall DA, Cunningham C, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR conjoint analysis good research practices task force. Value Health. 2016 Jun;19(4):300-15.
22. Lancsar E, Fiebig DG, Hole AR. Discrete choice experiments: a guide to model specification, estimation and software. Pharmacoeconomics. 2017 Jul;35(7):697-716.
23. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health--a checklist: a report of the ISPOR good research practices for conjoint analysis task force. Value Health. 2011 Jun;14(4):403-13.
24. Johnson RF, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, Regier DA, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health. 2013 Jan;16(1):3-13.
25. Joy SM, Little E, Maruthur NM, Purnell TS, Bridges JF. Patient preferences for the treatment of type 2 diabetes: a scoping review. Pharmacoeconomics. 2013 Oct;31(10):877-92.
26. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014 Sep;32(9):883-902.
27. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics. 2019 Feb;37(2):201-26.
28. Havrilesky LJ, Alvarez Secord A, Ehrisman JA, Berchuck A, Valea FA, Lee PS, et al. Patient preferences in advanced or recurrent ovarian cancer. Cancer. 2014 Dec;120(23):3651-9.
29. Muhlbacher AC, Bethge S. Patients' preferences: a discrete-choice experiment for treatment of non-small-cell lung cancer. Eur J Health Econ. 2015 Jul;16(6):657-70.
30. de Bekker-Grob EW, Niers EJ, van Lanschot JJ, Steyerberg EW, Wijnhoven BP. Patients' preferences for surgical management of esophageal cancer: a discrete choice experiment. World J Surg. 2015 Feb;39(10):2492-9.
31. King MT, Viney R, Smith DP, Hossain I, Street D, Savage E, et al. Survival gains needed to offset persistent adverse treatment effects in localised prostate cancer. Br J Cancer. 2012 Feb;106(4):638-45.
32. de Bekker-Grob EW, Bliemer MC, Donkers B, Essink-Bot ML, Korfage IJ, Roobol MJ, et al. Patients' and urologists' preferences for prostate cancer treatment: a discrete choice experiment. Br J Cancer. 2013 Aug;109(3):633-40.
33. Howard K, Salkeld GP, Patel MI, Mann GJ, Pignone MP. Men's preferences and trade-offs for prostate cancer screening: a discrete choice experiment. Health Expect. 2015 Dec;18(6):3123-35.
34. Eliasson L, de Freitas HM, Dearden L, Calimlim B, Lloyd AJ. Patients' preferences for the treatment of metastatic castrate-resistant prostate cancer: a discrete choice experiment. Clin Ther. 2017 Apr;39(4):723-37.
35. Helter TM, Boehler CE. Developing attributes for discrete choice experiments in health: a systematic literature review and case study of alcohol misuse interventions. Journal Subst Use. 2016 Nov;21(6):662-8.
36. Coast J, Al-Janabi H, Sutton EJ, Horrocks SA, Vosper A, Swancutt DR, et al. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. Health Econ. 2012 Jun;21(6):730-41.

37. Nakayama M, Kobayashi H, Okazaki M, Imanaka K, Yoshizawa K, Mahlich J. Patient preferences and urologist judgments on prostate cancer therapy in Japan. Am J Mens Health. 2018 Jul;12(4):1094-101.
38. Muhlbacher A, Johnson FR. Choice experiments to quantify preferences for health and healthcare: state of the practice. Appl Health Econ Health Pol. 2016 Jun;14(3):253-66.
39. Hazlewood GS. Measuring patient preferences: an overview of methods with a focus on discrete choice experiments. Rheum Dis Clin N Am. 2018 May;44(2):337-47.
40. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics. 2008 Feb;26(8):661-77.
41. Bliemer MCJ, Rose JM. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. Transport Res B-Meth. 2010 Jul;44(6):720-34.
42. Huertas-Garcia R, Gazquez-Abad JC, Martinez-Lopez FJ, Esteban-Millat I. Using response surface methodology to optimise factors in conjoint experiments. Int J Mark Res. 2013 Mar;55(2):267-88.
43. Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. BMC Health Serv Res. 2014 Sep;14:367.
44. Orme B. Getting started with conjoint analysis: strategies for product design and pricing research. 2nd ed. Madison, WI: Research Publishers LLC.; 2010.
45. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient. 2015 Oct;8(5):373-84.
46. Rose JM, Bliemer MCJ. Sample size requirements for stated choice experiments. Transportation. 2013 Sep;40(5):1021-41.
47. Wright SJ, Vass CM, Sim G, Burton M, Fiebig DG, Payne K. Accounting for scale heterogeneity in healthcare-related discrete choice experiments when comparing stated preferences: a systematic review. Patient. 2018 Oct;11(5):475-88.
48. Marsh K, Caro JJ, Hamed A, Zaiser E. Amplifying each patient's voice: a systematic review of multi-criteria decision analyses involving patients. Appl Health Econ Health Pol. 2017 Apr;15(2):155-62.
49. Hiligsmann M, Bours SP, Boonen A. A review of patient preferences for osteoporosis drug treatment. Curr Rheumatol Rep. 2015 Sep;17(9):61.
50. von Arx LB, Kjeer T. The patient perspective of diabetes care: a systematic review of stated preference research. Patient. 2014 Sep;7(3):283-300.
51. Ku Y-C, Chiang T-F, Chang S-M. Is what you choose what you want?-Outlier detection in choice-based conjoint analysis. Mark Lett. 2017 Mar;28(1):29-42.
52. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Rev Pharmacoecon Outcomes Res. 2017 Dec;17(6):531-42.
53. Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006 Aug;15(8):797-811.
54. Blinman P, King M, Norman R, Viney R, Stockler MR. Preferences for cancer treatments: an overview of methods and applications in oncology. Ann Oncol. 2012 May;23(5):1104-10.
55. Bech M, Kjaer T, Lauridsen J. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. Health Econ. 2011 Mar;20(3):273-86.
56. Vass C, Rigby D, Payne K. The role of qualitative research methods in discrete choice experiments: a systematic review and survey of authors. Med Decis Making. 2017 Apr;37(3):298-313.
57. Harrison M, Milbers K, Hudson M, Bansback N. Do patients and health care providers have discordant preferences about which aspects of treatments matter most? Evidence from a systematic review of discrete choice experiments. BMJ Open. 2017 May;7(5):e014719.
58. Vass CM, Payne K. Using discrete choice experiments to inform the benefit-risk assessment of medicines: are we ready yet? Pharmacoeconomics. 2017 Sep;35(9):859-66.

59. Hauber BA, Fairchild AO, Reed Johnson F. Quantifying benefit-risk preferences for medical interventions: an overview of a growing empirical literature. Appl Health Econ Health Pol. 2013 Aug;11(4):319-29.
60. Peters E. Numeracy and the perception and communication of risk. Ann N Y Acad Sci. 2008 Apr;1128:1-7.
61. Harrison M, Rigby D, Vass C, Flynn T, Louviere J, Payne K. Risk as an attribute in discrete choice experiments: a systematic review of the literature. Patient. 2014 Jun;7(2):151-70.
62. Veldwijk J, Lambooij MS, van Til JA, Groothuis-Oudshoorn CG, Smit HA, de Wit G. Words or graphics to present a discrete choice experiment: does it matter? Patient Educ Couns. 2015 Nov;98(11):1376-84.
63. Kenny P, Hall J, Viney R, Haas M. Do participants understand a stated preference health survey? a qualitative approach to assessing validity. Int J Technol Assess Health Care. 2003 Dec;19(4):664-81.
64. Mulhern B, Longworth L, Brazier J, Rowen D, Bansback N, Devlin N, et al. Binary choice health state valuation and mode of administration: head-to-head comparison of online and CAPI. Value Health. 2013 Jan;16(1):104-13.
65. Street DJ, Burgess L, Louviere JJ. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. Int J Res Mark. 2005 Dec;22(4):459-70.
66. Cheung KL, Wijnen BF, Hollin IL, Janssen EM, Bridges JF, Evers SM, et al. Using best-worst scaling to investigate preferences in health care. Pharmacoeconomics. 2016 Dec;34(12):1195-209.
67. Whitty JA, Oliveira Goncalves AS. A systematic review comparing the acceptability, validity and concordance of discrete choice experiments and best-worst scaling for eliciting preferences in healthcare. Patient. 2018 Jun;11(3):301-17.
68. Potoglou D, Burge P, Flynn T, Netten A, Malley J, Forder J, et al. Best-worst scaling vs. discrete choice experiments: An empirical comparison using social care data. Social Science & Medicine. 2011;72(10):1717-27.
69. Muhlbacher AC, Juhnke C. Patient preferences versus physicians' judgement: does it make a difference in healthcare decision making? Appl Health Econ Health Pol. 2013 Jun;11(3):163-80.
70. Higgins A, Barnett J, Meads C, Singh J, Longworth L. Does convenience matter in health care delivery? A systematic review of convenience-based aspects of process utility. Value Health. 2014 Dec;17(8):877-87.
71. Ryden A, Chen S, Flood E, Romero B, Grandy S. Discrete choice experiment attribute selection using a multinational interview study: treatment features important to patients with type 2 diabetes mellitus. Patient. 2017 Aug;10(4):475-87.

Chapter 3

Methods

Introduction

This DCE was designed by following the evidence-based best practices discussed extensively in chapter 2. The pros and cons of the selected oncology DCEs were carefully considered during the design for this study. Therefore, based on the best practices the following steps and processes were followed and are expanded upon in the subsequent sections: completing a pilot study; included attributes were based upon a literature review, insight of an expert (staff Radiation Oncologist), and patient input; level selection was based on RCTs data and expert opinion; efficient choice set construction; best methods to calculate minimum sample size; collecting sociodemographic and disease specific information on possible pertinent factors from responders and non-responders; in-person interviews instead of computer-assisted; and analyzing the data with extended models after initial analysis with conditional logit.

Questionnaire design

Attribute selection

In this study the attributes were set based on the clinical trial literature that included five DCEs and two qualitative studies undertaken with prostate cancer patients to determine what was important to their treatment decision making process (1-5). The decision was also guided by the overarching goal to determine individuals' willingness to trade off the risk of increased toxicity for the convenience of shorter treatment time within the framework of the clinical trial findings. Five attributes were included; 'length of radiotherapy', 'fiducial markers implanted', 'risk of PSA rising within 5-years', 'risk of acute moderate GI and GU toxicity', and 'risk of late moderate GI and GU toxicity' (Table 2).

Table 2: Attributes and levels in questionnaire.

Treatment attributes	Levels
Length of radiotherapy (Length)	2 weeks (5 treatments) 4 weeks (20 treatments) 8 weeks (40 treatments)
Fiducial marker implantation (Markers)	Yes No
Risk of PSA recurrence within 5 years of treatment (PSA)	6% 12% 18%
Risk of acute toxicity (Acute)	20% 35% 50%
Risk of late toxicity (Late)	10% 15% 20%

Attribute short form notation in parentheses. PSA – Prostate specific antigen.

As discussed in chapter 2, to increase measurement validity there was no conceptual overlap between the selected attributes and each was uni-dimensional (6). The ‘fiducial markers implanted’ (Markers) attribute’s effect was studied independently even though in clinical practice it is required only for SBRT. The two toxicity attributes (Acute and Late) only incorporated GI and GU toxicity since sexual side effect attributes tended to be either non-significant or of minimal influence in prostate cancer DCEs (Table 1). There is a higher risk of erectile dysfunction after prostatectomy, thus I decided to exclude sexual side effects since this attribute would likely not be of great importance to patients’ preferences for EBRT scenarios as the associated risk levels would be low. Fatigue was only included in two DCEs due to it being a main side effect of androgen deprivation therapy (ADT). Toxicity was defined as moderate (RTOG 2) since mild side effects were non-significant in previous DCEs, and severe toxicity is rare in EBRT. This ensured that a wide range of plausible levels were presented to participants.

The five attributes were assessed based on previous qualitative work undertaken in a similar patient population. Qualitative data was included to increase measurement validity, determine if patients reported these attributes as important to their decision making, and inspect if any key attributes were missing. A qualitative study undertaken at the Cancer Centre of Southeastern Ontario in 2004 interviewed

60 patients with early-stage prostate cancer to determine what treatment they preferred as they could be treated with EBRT, surgery, or active surveillance (see Table 1) (7). These three treatments were the standard options with no changes in practice until recently with the addition of brachytherapy and introduction of hypofractionation as outlined in chapter 2, thus the findings of this 2004 study are likely still relevant to current patient preferences for EBRT. However, there has been no comparable qualitative study over the past 15 years and perceptions of cancer and survival expectations may have shifted.

This qualitative study identified a median of four important attributes per patient (7). Aspects from eight of the 10 overall most important attributes were incorporated into the five attributes in my study (listed in Table 2). The 'length of radiotherapy' (Length) attribute incorporated 'quality of life' and 'how treatment works'; Markers incorporated 'procedure' and 'how treatment works', 'risk of PSA rising within 5-years' (PSA) incorporated 'chances of PSA rising', 'chances of dying of cancer', and 'effect on length of life', and both of the toxicity attributes incorporated 'effect on bladder functioning', 'effect on quality of life', and 'effect on bowel functioning' (7). The two characteristics not incorporated were either not pertinent or previously found not to be significant in the prostate cancer DCEs.

The second qualitative study reported in Table 1 was completed in the United States in 2016 and corroborated the preferences identified in the earlier study in this patient population (8). In Table 1 I highlight pertinent categories identified as important by newly diagnosed prostate cancer patients to their treatment decision. I omitted characteristics not pertinent to EBRT or the Canadian health care system, for example the category 'insurance coverage of treatment' is not relevant since all costs of cancer treatment are covered by the provincial health care system for Canadian citizens and residents. The category 'recommendations and general preferences' was not directly related to one of the attributes in my DCE. However, in my study participants were informed that in the hypothetical scenario their Radiation Oncologist has recommended both the alternatives they were deciding between, and it was up to their personal preference to select their treatment.

In summary, the five included attributes were demonstrated in DCEs and qualitative studies to be pertinent to this patient population's decision making, and each attribute was considered independent of each other.

Level selection

The three levels for the four quantitative attributes (Length, PSA, Acute, and Late) were selected based on results from clinical trials in the localized prostate cancer population to ensure realism (Table 2). For Length the plausible lengths of 2, 4, and 8 weeks were chosen (Table 2). Radiotherapy is delivered Monday to Friday, and the 2 weeks level corresponded to SBRT (5 fractions every other day for a total of 10 days), 4 weeks to MH (20 fractions daily), and 8 weeks to CFRT. The Markers attribute was dummy coded with two levels (yes and no).

For the outcome attributes the phase III RCT data comparing MH and CFRT, which included studies that enrolled prostate cancer patients with localized disease from all three risk groups, were used as a baseline (9, 10). No similar data for SBRT was published at the beginning of this study, though the phase II SBRT clinical trials suggested similar outcomes to CFRT and MH (11). The superiority HYPRO trial data were excluded since the toxicity was significantly higher for the MH arm, possibly explained by the higher dose per fraction of 3.4 Gy. The six other trials MH arm doses range from 2.4-3.1 Gy per fraction (9). Efficacy and toxicity data were used from both the CFRT and MH arms in the six trials. In the trials the 5-year biochemical relapse free results ranged from 79% to 96%, with a mean of 87.4% and SD of 5.8% (9). The PSA attribute levels were framed negatively, which provided a range of 4% to 21% and a mean risk of recurrence of 12.6%. The middle risk level was set to 12%. The highest risk level was determined by adding one SD, rounded to 6%, to the middle level which equaled 18%, and the minimum level by subtracting one SD from the middle level to equal 6% (Table 2). In the trials the acute moderate GI and GU toxicity risk ranged from 10.3% to 49% (9). This is a large range with a mean of 34.5% and SD of 12.9%. The mean was rounded to 35% and set as the middle risk level. The SD was rounded to

15%, and the highest risk level was set to 50% (35% + 15%) and the lowest risk level was set to 20% (35% - 15%) (Table 2). In the trials the risk of late moderate GI and GU toxicity ranged from 5.1% to 29.7% (9). The mean is 15.2% with a SD of 5.9%. The SD was rounded down to 5%, and the highest risk level was set to 20% (15% + 5%) and the lowest risk level was set to 10% (15% - 5%) (Table 2). For both toxicity attributes the levels were rounded to whole numbers that ensured equal spacing that should improve participants' comprehension (the closest number ending in 0 or 5).

Efficient design

My study included 12 choice sets in the experimental design. This number was determined based on literature that suggested, in order to ensure statistical sufficiency, the number of choice sets included in a DCE should not be smaller than:

Equation 5 $$K/(J\text{-}1),$$

where *K* is the number of parameters to be estimated, and *J* is the number of alternatives (12). In this study *K* = 5 if the quantitative attributes were coded as continuous variables. Sufficiency was obtained by varying the attribute levels within and across choice sets to ensure enough combinations of attribute levels were presented to participants that would require them to weigh the trade-offs of each attribute (12). If these four attributes were coded as categorical variables *K* = 9 since there was three levels for four attributes (one reference level per attribute), plus one parameter for the dummy coded attribute Markers. As in all forced choice DCEs *J* = 2. Thus, the two minimum numbers were 5 and 9, so in my study greater than nine choice sets were included in the experimental design. The appropriate number of choice sets is context specific and a pilot test was conducted to ensure the number was feasible for this target population (13).

Risk and framing

As discussed in chapter 2, choice set presentation was considered carefully to ensure the best comprehension possible by participants. Whole numbers were used with a consistent denominator (e.g. compared to 100 men) and diagrams were included for the risk attributes (Figure 4). Efficacy and toxicity attributes were all framed negatively in the diagrams to provide consistency and to increase comprehension since toxicity is generally described by health care practitioners negatively (e.g. one in five patients may experience this side effect). The positive framing was included in words underneath the diagram (Figure 4).

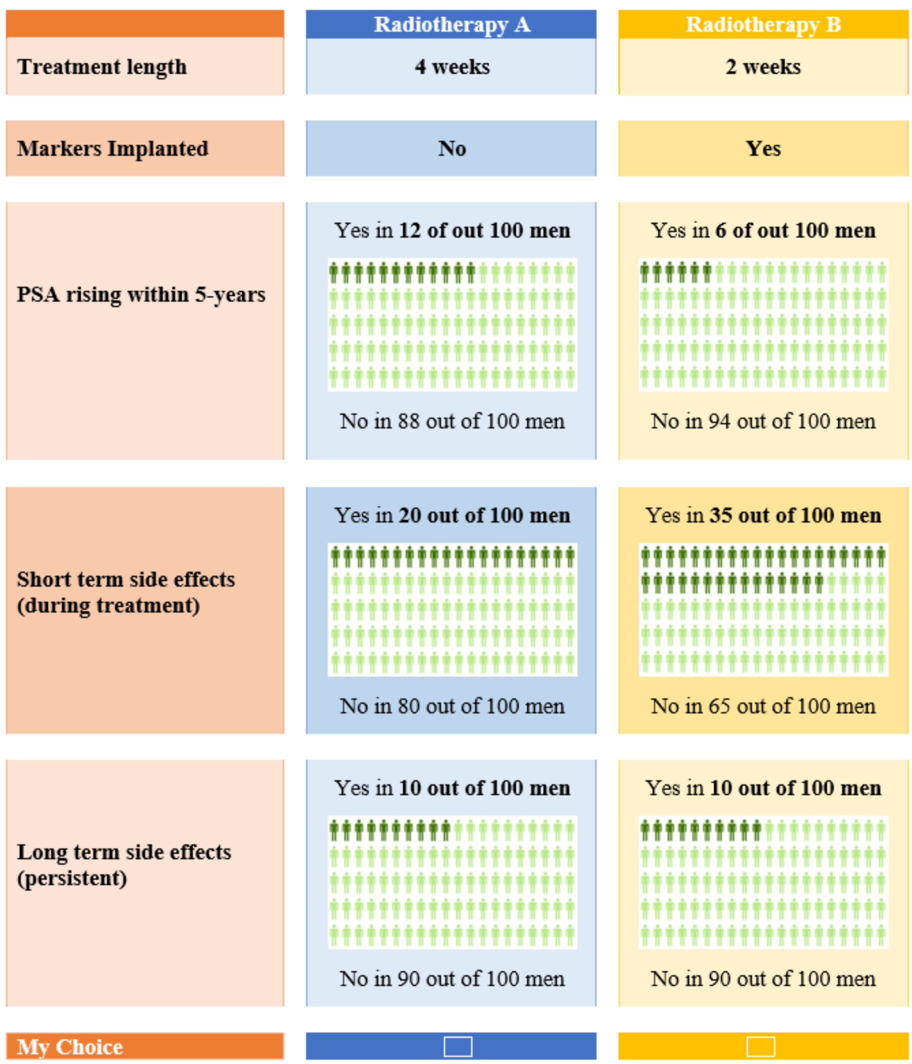


Figure 4: Example of choice set that participants used from experimental design.

Pilot study

A pilot test of the DCE choice sets was required to assess the feasibility of the DCE in this patient population and calculate appropriate part-worth estimates for the design of the full DCE. A total of six pilot participants were involved. The pilot parameter estimates were set for each attribute based on the characteristics and the associated parameter estimates from the prostate cancer DCEs listed in Table 1. The Length attribute (referred to as 1 in Table 1) was hypothesized to be negative in magnitude as participants should prefer shorter treatment time, and for the pilot study was given a value of -0.4. The Markers attribute (referred to as 2 in Table 1) had the least amount of previous evidence since this specific procedure had not been included in a DCE prior to my study. This dummy coded attribute was hypothesized to be negative in magnitude with the reference level 'no'. For the pilot study it was given a value of -0.2 since it was hypothesized to have similar importance to participants as acute toxicity. The PSA attribute (referred to as 3 in Table 1) represented treatment efficacy and was hypothesized to be negative in magnitude as participants should prefer less risk of their PSA rising. For the pilot study this was given a value of -1.3, larger in magnitude than previous studies' findings but this was expected to be the attribute with the highest importance. The two toxicity attributes (Acute and Late referred to as 4 and 5 respectively in Table 1) encompassed numerous urinary and bowel side effects noted in the previous published DCEs. The Acute attribute was given a value of -0.2 which was at the lower range of parameter values found for side effect attributes, and the late toxicity attribute a value of -1.2 from the higher range. These continuous attributes were hypothesized to both be negative in magnitude as participants were expected to prefer less risk of side effects.

The experimental design was coded in the software program Ngene with the attributes Length and Late set as random parameters with a normal distribution and a SD of 0.1 and 0.2 respectively. These attributes were set as random due to the higher uncertainty about these estimates and belief they would be calculated as Bayesian in the full DCE. The most efficient design estimated had a D_b-error of 0.009762 for random parameter analysis and a fixed D_b-error of 0.008177. There was a total of 24 choice sets divided into two blocks, and three participants answered each version of 12 choice sets from this initial experimental design. Six additional choice sets were included to test internal validity (repeatability, monotonicity, transitivity, and version consistency) that resulted in a total of 18 choice tasks per pilot participant.

A semi-structured interview with each pilot study participant was completed. It was discovered that all participants understood the attributes and levels, did not have comprehension difficulties, and could not identify any additional attributes important to their decision making. None of the pilot participants reported ignoring any attributes (e.g. attribute non-attendance). The five attributes were labeled as ‘treatment length’, ‘seeds inserted’, ‘risk of PSA rising within 5-years’, ‘short term side effects’, and ‘long term side effects’ with the goal of simplifying language and increasing participants’ comprehension. All six participants reported at least considering and understanding each attribute after a thorough explanation by the interviewer.

The six pilot study participants ranged in age from 68 to 80 years old, and at the time three were undergoing EBRT and the other three had completed EBRT. They had diverse educational backgrounds, three having obtained university/college degrees and the remaining three had completed high school. The time it took the participants to drive to the cancer centre varied from less than 30 minutes to over 2 hours. Through the semi-structured interviews all participants stated that all attributes that were of importance to them had been included in the DCE. Participants generally found the concept confusing at the beginning of the study, as one respondent noted “due to the similarity between the questions was hard to determine what I had to decide”. Interestingly, three of the participants indicated that after the first few choice sets were answered they did not find any remaining tasks confusing, suggesting a learning effect. Two participants provided feedback that they thought the attribute ‘seeds inserted’ sounded like brachytherapy since the term seeds were used with that procedure. Two participants noted that they found the diagrams helpful. Five of the respondents reported comparing the absolute differences between the levels, and only one noted he didn’t compare the absolute numbers and only determined if the risk level was higher or lower than in the other alternative. All participants reported that even though there was the smallest absolute difference in the risk for 'long term side effects' that this attribute was considered carefully and was very important to them. In fact, this attribute was ranked as most important by four participants and second most important by two participants. The efficacy attribute 'PSA rising within 5-years' was ranked as most important by two participants and second most important by three participants. The attribute 'treatment length' was deemed most important by only one participant, and it was generally in the middle of the priorities list as only one participant ranked it as least important. When asked why a certain attribute was most important to them, one respondent who indicated 'treatment length' as most important noted "driving far" was his highest concern, and another participant who indicated 'long term side effects' was most important to him reported "chronic side effects would bother me more than anything." One respondent who reported the attribute 'PSA rising within 5-years' as most important to their decision making stated that "getting rid of the cancer" was most important to them.

Two dominant choice sets were placed at the beginning of the DCE (outside of the two versions of 12 choice sets each) to test comprehension. As noted in chapter 2, a dominant choice task contains one alternative where all of the attribute levels are as good as or better than the other alternative (for example alternative A has a shorter length of EBRT, no fiducial markers implanted, and less risk of PSA recurrence and toxicity and therefore should be preferred over alternative B by all participants). All six participants answered these correctly. The last choice set, number 18, was a repeat of the third choice set and all six participants answered these the same. If this was due to chance alone it would be 50% for each participant. All six participants' responses to the choice sets resulted in the principle of transitivity being upheld.

The results of the pilot study were analyzed in two ways, as originally coded (MNL continuous in Table 3), and as all categorical (MNL categorical in Table 3). When the results were initially analysed, with only Markers as dummy coded, the parameter estimates were close to the expected values (Table 3). The Markers parameter estimate was highest in magnitude due to its categorical coding. All five parameter estimates were negative as participants preferred the lower value levels of these attributes (e.g. shorter treatment time, no markers implanted, less risk of PSA rising, less risk of toxicity). Each attribute was then dummy coded with the lowest value set as the reference level for the categorical model. This produced nine parameter estimates, as there were two for the four attributes with three levels, and still only one parameter estimate for the Markers attribute (Table 3). When linearity was not assumed the parameter estimates fit the respondents' interview responses well, and the model fit was slightly improved based on the log likelihood statistic (Table 3). In this model it was easier to compare the relative importance of the attributes. The PSA attribute had the largest magnitude for the highest risk level (18%) and the Late attribute had the second largest magnitude for the highest risk level (20%), where respectively the 6% and 10% levels were the reference categories with a value of zero. This was consistent with the interview results as all participants rated these two attributes as first or second most important to their decision making. In keeping with the interview results the Markers attribute had a lower value, indicating less importance to participants' choices, than the highest risk levels and 8-week EBRT level.

Table 3: Pilot test results for continuous and categorical variables.

Attributes	**MNL continuous**	**MNL categorical**
Length β_1 (ref. 2 weeks)	-0.1278	
4 weeks β_{1a}		-0.2805
8 weeks β_{1b}		-0.9110
Markers β_2 (ref. no markers)	-0.6211	-0.6079
PSA (ref. 6% β_3)	-0.2490	
12% β_{3a}		-1.3691
18% β_{3b}		-3.0745
Acute (ref 20% β_4)	-0.0429	
35% β_{4a}		-0.6833
50% β_{4b}		-1.2889
Late (ref. 10% β_5)	-0.1837	
15% β_{5a}		-1.1195
20% β_{5b}		-1.8951
Goodness of fit		
Log likelihood	-45.49	-45.32

An interaction term between Length and Acute was included in the design of the pilot study. It was thought participants may associate the acute treatment side effects lasting over the course of the radiotherapy, though they were instructed that acute toxicity takes weeks to 3 months to resolve after the completion of EBRT. The pilot participants reported all five attributes were considered independently in the decision-making process. Additionally, the sample size required to take interaction into account with adequate statistical significance and power was not realistic for this study. Therefore, it was decided going forward this interaction term would not be accounted for at the design stage.

In summary, participants were able to tolerate 18 choice sets without undergoing comprehension fatigue. Furthermore, no important attributes to respondents' decision making were missing from the DCE. The DCE passed the reliability and validity tests at this point, though there was a small sample size. These choice sets were included for initial testing of comprehension and to ensure participants could tolerate 18 choice sets. The task complexity was deemed appropriate and the length of the interview and the attributes and levels remained unedited except for the description of the Markers attribute to exclude the term 'seeds' when described to participants.

Experimental design

Based on the pilot test results the DCE was planned in the experimental design software Ngene with all attributes categorically coded. The nine parameter estimates (MNL categorical in Table 3) and SDs were used to assume a fixed variable as the prior estimate for the Markers attribute, while Bayesian priors with normal distributions were assumed for Length, PSA, Acute, and Late attributes. This was to ensure the most efficient design as possible since the parameter values were obtained from the pilot study which had a small sample size and large uncertainty. The experimental design was specified to have 24 total choice sets in two blocks with dominated alternatives excluded. This Bayesian approach estimated a D_b-error of 0.02468 for conditional logit analysis and a D_b-error of 0.0715 for random parameter analysis.

Validity

Internal validity

Six choice sets were included in the DCE interview to test for internal validity: two dominant choice sets to test for monotonicity and comprehension; a repeat of a profile from the experimental design; two transitive choice sets based on a profile from the experimental design within Ngene to complete the set of three; and a holdout profile from the other version the respondent had not seen before. Therefore, each respondent saw 18 profiles. The first dominant choice set in the questionnaire was the first profile, which also served to ease respondents gradually into the survey. The second dominant choice set was the 15th choice set in each version (the other dominant choice set was first followed by the 12 choice sets from the experimental design). Repeatability was assessed by repeating the third choice set from within the experimental design at the 14th position, and intra-rater agreement was tested with Cohen's kappa (κ) statistic. Transitivity was tested by adding two choice sets outside the experimental design to determine if the preference relation was consistent between three choice profiles within each respondents' choices. By necessity these choice sets differed between the blocks, however they were similar in principle and were not more difficult in either version and placed at the 16th and 17th positions. Version consistency was tested via holdout prediction and was the 18th choice set presented to respondents. Since there were two versions of the DCE respondents all answered one choice set from the opposite block they had not seen previously. The purpose was to test if the model could predict choices outside the choice model. The utility for each alternative was calculated using the mixed multinomial logit (MXL) model. This model only incorporated choice data from the choice sets within the experimental design. For example, the holdout choice set for block 1 (in the 18th position) was part of the validity check and these choices did not enter the choice model, while this same choice set in block 2 was part of the experimental design and model estimation. Block 1 participants' choices for this holdout choice set were only used to compare the choices to what would be 'expected' using the κ statistic. The alternative with the less negative utility was more likely to be chosen.

Task or attribute non-attendance was assessed via three methods: (1) it was considered if respondents appeared to have a pattern to their choices (e.g. selected all A); (2) participants were asked if they understood the concept of making choices between the two alternatives in the DCE; and (3) each participant was asked how many attributes they considered when making their choices, and the interviewer also reported how many attributes they believed the participant considered based on the participants comments during the DCE and the debriefing survey. Each participants' answers in the DCE were examined for non-compensatory decision making by reviewing if anyone always selected the alternative with the best level for their preferred attribute (e.g. if a participant most preferred to reduce risk of recurrence and always chose the alternative with the least risk of PSA recurrence).

External validity

The sample population consisted of prostate cancer patients who were undergoing or had completed EBRT targeted to the prostate at the time of recruitment. This was similar to the larger population of interest; prostate cancer patients with localized disease considering different fractionation regimens of EBRT (14). The more the respondents approximated the target population, the better the external validity of the results (14). It was important that the situation presented to participants was familiar to them, otherwise responses may have been artificial (14). Thus, only prostate cancer patients who had at least started EBRT were eligible for this study. The use of concrete values (attributes and levels based on RCTs) presented more lifelike scenarios to respondents, and therefore increased the generalizability of the results (14). The realism of this study was improved by completing the DCE in one session and in a natural setting, in this case a clinic room in the cancer centre (14). To increase the likelihood that the DCE was capturing the latent beliefs and choices of respondents, the respondents were asked to deliberate carefully as their responses had the possibility to help determine future research directions at the institution. Thus, they understood there were real effects from the outcome of the survey even though it would not affect them on an individual level.

Ranking, rating, and qualitative information

After the pilot study was completed the planned debriefing questions all participants in the full DCE answered were adjusted (Figure 5). All the respondents to the full DCE were asked to assess their self-reported overall confusion on a five-point Likert scale where the end points of 1 corresponded to ‘not confused’, and a score of 5 to ‘very confused’ (Figure 5 question 1). To help assess attribute non-attendance participants were asked in multiple choice format how they made their decision between the alternatives, and they selected if they were unsure or considered one, some, most, or all of the attributes (Figure 5 question 3) (14). There was a short qualitative question where respondents were asked why one, or a subset, of attributes were most important to them and their answers were written down verbatim by the interviewer with their answer repeated back to the respondent to ensure accuracy (Figure 5 question 6). These answers were examined with content analysis (16). To slightly mimic the trade offs participants made between the attributes in the DCE participants were asked to allocate points to various attributes based on their importance (Figure 5 question 7) (17). Specifically, participants were told to; “Assume you have a total of 100 importance points to assign to the five radiotherapy characteristics. Please show how important each item is to you by distributing your 100 points among the five items; to show that one item is more important to you than another you should give more points to it. You do not need to give points to all the items, but remember the total needs to equal 100” (17). The internal consistency of the implicit ranking of the attributes by the DCE were compared to the explicit ranking and rating exercises completed by all participants. The explicit ranking was determined by first calculating the sample-wide (overall mean) for each attribute for the rating and point assignments questions in Figure 5 (questions 5 and 7). Comparison was made on a class basis (from the results of the LCM) using one-way analysis of variance (ANOVA) comparison of means between groups to determine if there was a significant difference. Second, these two results were assessed to determine if the ratings and points were consistent with the DCE model estimates (17).

1. How confused are you about the choices you made, on a scale of 1-5 where 1 is not confused at all and 5 is very confused?

   1 **2** **3** **4** **5**

2. On average how easy or difficult was it to make a choice between the 2 radiation treatment options (A or B)?

1	2	3	4	5
Very Easy	**Easy**	**Neutral**	**Difficult**	**Very Difficult**

3. How did you make your choices when deciding which treatment (A or B) was best for you?
   a. I only focused on the **1** characteristic most important to me
   b. I focused on **some** of the characteristics most important to me
   c. I considered **most** of the characteristics important to me
   d. I considered **all** of the characteristics
   e. I'm not sure or couldn't say

4. Did you understand the concept of making choices between the 2 options? **Yes** / **No**

5. Rate the importance of each characteristic to your decision making on a scale of 1-10, where 10 is very important and 1 is not important.

Length of radiotherapy:	1	2	3	4	5	6	7	8	9	10
Markers implanted:	1	2	3	4	5	6	7	8	9	10
Risk of PSA rising in 5 years:	1	2	3	4	5	6	7	8	9	10
Risk short-term side effects:	1	2	3	4	5	6	7	8	9	10
Risk of long-term side effects:	1	2	3	4	5	6	7	8	9	10

6. Why is one or more of the characteristics most important to you?
   ____

7. Assume you have a **total of 100 importance points** to assign to the five characteristics. Please show how important each item is to you by distributing your 100 points among the five items; to show that one item is more important to you than another you should give more points to it. You do not need to give points to all the items, but remember the **total needs to equal 100**.

   Length of radiotherapy: ______
   Markers implanted: ______
   Risk of PSA rising within 5 years: ______
   Risk of short-term side effects: ______
   Risk of long-term side effects: ______
   Total = **100**

Figure 5: Sample questionnaire participants completed after DCE.

The interviewer answered similar questions about how participants responded to the DCE to determine inter-rater agreement using the κ statistic and several other factors about participants (Figure 6). The interviewer reported categorical scores were tested for association with χ^2 Test of Independence.

1. How confusing do you think the participant found the choices tasks? 1 2 3 4 5
2. Do you think the participant understood the concept of making choices between the 2 treatment options (**yes, no uncertain**)?
3. Do you think the participant considered all of the attributes when making their choices?
    a. One
    b. Some
    c. Most
    d. All
    e. Unsure
4. Do you think the participant found it difficult to decide between the different treatments in each question?
    1. Very easy
    2. Easy
    3. Neutral
    4. Difficult
    5. Very difficult
5. Which attributes/levels do you think drove the participants choices?
6. Do you think the participant was actively trading between attributes (**yes, no uncertain**)?
7. Did the participant engage in the task and make an effort in choosing (**yes, no, uncertain**)?
8. Did the participant find it difficult to choose between the 2 alternatives (**yes, no, uncertain**)?
9. Did the participant appear to **know** their preferences, or did they **develop** as they went through the choice tasks (**uncertain**)?
10. Did the participant find the task easier as they went through (**yes, no, uncertain**)?
    a. If yes did this appear to be due to **learning** effects or adopting a **simplifying** strategy?
11. How long did the participant take to complete the DCE survey (**quick, average, slow**)?

Bolded: options interviewer could select.

Figure 6: Sample interviewer form interviewer.

Participant accrual

Recruitment

Potential participants were patients of the Cancer Centre of Southeastern Ontario who had been diagnosed with localized prostate cancer and had completed or were receiving curative intent EBRT to their prostate at the time of recruitment. This target population was appropriate for these research objectives since this was a forced choice DCE between two EBRT options. The sampling frame was representative of the target population as all patients receiving EBRT to their prostate between February and April 2019 were approached and invited to participate in the study by the attending Radiation

Oncologist and researcher. Additionally, patients from three Radiation Oncologists' follow up clinics were also approached to participate in the study. If possible, all eligible patients in a clinic were asked to participate; some patients were missed if their appointment was completed while the interviewer was finishing the DCE with a different patient. Basic information about non-responders (patients who declined to take part in the study) was collected, including age and distance from the cancer centre.

Background Information

A general information hand out was provided to participants. This information was repeated at the time of the DCE (Figure 7). Participants were also informed about: a planning CT scan occurred prior to radiotherapy in all cases; placement of fiducial markers for SBRT; the meaning of moderate GI and GU side effects; and all other side effects (e.g. fatigue, erectile dysfunction) should be considered to be the same between the two alternatives in the DCE. The explanation of toxicity seen in Figure 7 was taken from the descriptions of grade 2 (moderate) GU and GI toxicity in the RTOG schema (18).

Sample size

One rule of thumb to calculate required sample size was proposed by Orme:

Equation 6 $$[nta]/c \geq 500,$$

where n is the number of respondents, t is the number of tasks (choice sets in the experimental design), a is the number of alternatives per task (not including the opt-out), and c is the number of analysis cells (equal to the largest number of levels for any one attribute when using main effects, and if considering two way interactions then equal to the largest product of levels of any two attributes) (19). In this study $t = 12$, $a = 2$, and $c = 3$, and therefore the suggested minimum sample size based on this method was 63. Interaction between two attributes was only investigated if the two attributes were related, and in my study no attributes were expected to interact (12).

Treatment length

In the scenarios, for the 8- and 4-week treatment lengths you would be coming to the cancer centre everyday Monday to Friday until the treatment is completed. The 2-week treatment is only every other day for a total of 5 treatments. For example, if you start on a Monday, you would come back on Wednesday, then the Friday, then Monday and Wednesday of the next week and then be done. In all the scenarios you have the weekends off.

Markers

Having fiducial markers implanted is a day procedure done at the hospital and you can leave the same day. It feels like a prostate biopsy. You will feel discomfort and/or pressure during the procedure, and some feel mild pain afterwards. The procedure should only take up to 15 minutes. This is one of the newer parts of radiotherapy, so you did not experience this during your treatment. This is done so the prostate can be better seen on CT imaging to target the radiation dose to the prostate. In the scenarios in this study you will be told if this procedure will or will not happen. In all scenarios, you still require a planning CT scan 1-2 weeks before the start of your radiotherapy.

PSA

For most men, PSA will never rise again after treatment. This means the cancer has been eliminated and the radiotherapy was successful. Some men who have treatment will eventually have their PSA rise again, which can occur months or years after treatment. The PSA rising does not mean you will experience symptoms from your cancer or die from it since there are many more treatments available to you. Most men do not experience any further symptoms and eventually die from other causes as prostate cancer grows slowly. In the scenarios presented to you in this study there will be listed chances of your PSA rising in the next 5 years. These are within reasonable ranges but are not your specific risk. For these scenarios, assume further treatments are available to if your PSA rises (e.g. hormones).

Short term side effects

Most men need to empty their bladder more frequently while undergoing radiotherapy. Some men have a bit more serious (moderate) side effects, such as discomfort or difficulty emptying their bladder. Radiotherapy can cause some men to move their bowels more often, and they may even notice some discomfort like urgency or mucous with their bowel movements. These risks are going to vary in the scenarios presented to you based on ranges we see in patients, but are not your specific risk in the real world. These side effects get better within a few months after treatment ends.

Long term side effects

Several months or years after treatment some men will lose some control of their bladder and have dribbling, may have to empty their bladder up to twice as many times as prior to treatment (but less often then every hour), and/or experience occasional bleeding seen as small amounts of blood in the urine. Sometimes scar tissue can develop in the bowel causing increased urgency, cramping, small amounts of blood with bowel movements, and/or diarrhea (more than 5 bowel movements a day). These problems can usually be treated with medication. These risks are going to vary in the scenarios presented to you based on ranges we see in patients, but are not your specific risk in the real world.

Consider fatigue and sexual side effects to be the same

Tiredness and sexual side effects can happen, for this study you are to assume they are the same for each scenario. Patients often take 2-3 weeks to feel like they can return to normal activities after radiotherapy ends. The most common side effect is tiredness during treatment, and there would be no difference in the severity of the tiredness you would experience between the different scenarios presented to you in this study. Some men have reduced ability to get an erection, this may be partial or complete.

Figure 7: Explanation of attributes and side effects the interviewer provided to each participant.

In the parametric approach (parameter estimates assumed to be normally distributed) the minimum sample size required is based on the most "critical" parameter. This parameter was the Markers attribute based on the Ngene calculation with a minimum required sample size of 42 for random parameter analysis (20). This approach was likely more reliable than rules of thumb since it required prior parameter estimates and utilized a 95% significance level. However, this sample size method was not specific to a certain hypothesis and did not depend on a desired power level (21).

The final formula used in this study to determine the optimal sample size for conditional logit analysis took into account: (1) a two-tailed significance level of 5% (to rule out findings due to chance alone at 2.5% on either side of the mean); (2) an 80% power level, (3) the number of attributes, (4) levels, and (5) number of choice sets in the DCE (21). The prior estimates from the pilot study were also included (21). The paper by de Bekker-Grob et al. provided an R code freely available online for this application (21). This was the most accurate sample size estimation that determined the critical parameter was the Late attribute and at least 28 respondents per version (block), a total of 56 across both versions, were required to significantly estimate the Late toxicity parameter estimate. Therefore, since this sample size calculation tested for the significance of the null hypotheses and had an associated power level this was the minimum sample size aimed for prior to start of recruitment.

Participant sociodemographic information

Sociodemographic data and prostate cancer specific information was collected from all participants prior to the DCE (Figure 8).

Analysis

The DCE was analysed by taking each choice among the two alternatives as an observation. No data were planned to be excluded regardless of participants' responses to the internal validity choice sets. The pilot study data were not included in the analysis, though five of the six pilot participants were later recruited to take part in the study from follow up clinics. As for all of the reviewed oncology DCEs, a statistically significant attribute ($p < 0.05$) was interpreted to indicate that respondents considered the attribute important and differentiated between one attribute (or attribute level) and another when they made their choices in the DCE. *A priori* all attributes were expected to be important, and all attribute levels (except the reference level)would have a negative parameter estimate based on the framing of the DCE since respondents should find it preferable to have shorter treatment, no markers, and less risk of PSA recurrence and GI and GU toxicity. The variables were initially all categorically coded.

Please tell us about yourself. Age: _________

Relationship status:

☐ Married/ committed partner ☐ Single ☐ Other ______________

Who do you live with?

☐ Spouse/ partner ☐ Someone other than spouse/partner e.g. parent, child, friend ☐ I live alone

How far away do you live from the cancer centre in Kingston?

☐ Less than a 30-minute drive ☐ Less than 1 hour but more than a 30-minute drive ☐ Over 1 hour drive

Highest level of education:

☐ High school ☐ Vocational/ technical certificate ☐ College/ University

Employment status:

☐ Employed ☐ Retired ☐ Not currently working

Time of prostate cancer diagnosis:

☐ Less than 1 year ago ☐ More than 1 year ago

Current treatment status:

☐ No treatment, regular checkups ☐ Radiotherapy
☐ Drug therapy (hormones, chemotherapy) ☐ Other ______________

Previous treatment you have received besides radiotherapy:

☐ Surgery to remove prostate ☐ Brachytherapy
☐ Drug therapy (hormones, chemotherapy) ☐ None

When you were undergoing radiotherapy, how severe were your side effects? For example, increased frequency of urination, pain while urinating, diarrhea, and/or abdominal cramping.

☐ Not at all ☐ Mild ☐ Moderate ☐ Severe ☐ Not sure

Do you think you have any lasting side effects after radiotherapy? For example, increased frequency of urination, blood in your urine, diarrhea, abdominal cramping, and/or mucous or blood in your stool. If so, please indicate how severe they are.

☐ Not at all ☐ Mild ☐ Moderate ☐ Severe ☐ N/A (not completed radiotherapy)

Thank you for participating in this research study and taking the time to answer these questions about yourself!

Figure 8: Form for sociodemographic and prostate cancer specific information.

Conditional logit models

The observations were analyzed using a conditional, also known as multinomial, logit regression model (MNL) that assumed all attributes had an independent influence on a participant's preference:

Equation 7 $$V = \beta_0 + \beta_{1a}LENGTH4 + \beta_{1b}LENGTH8 + \beta_2MARKERS + \beta_{3a}PSA12\% + \beta_{3b}PSA18\% + \beta_{4a}ACUTE35\% + \beta_{4b}ACUTE50\% + \beta_{5a}LATE15\% + \beta_{5b}LATE20\%,$$

where V represents the utility derived from EBRT for localized prostate cancer and $\beta_{1\text{-}5}$ are coefficients that indicate the relative importance of each attribute (Table 4). β_0 was a dummy coded alternative specific constant (ASC) where the base level 0 was set to alternative A, and alternative B coded with a 1. This was included initially to test for ordering effect; if respondents were more likely to choose the option that was presented first or second. If the coefficient for this constant was not significantly different from 0 it would be excluded from further models. The F test for linear correlation was used to verify the Length attribute was sufficiently linear ($p < 0.05$), and then the Length attribute would enter the model as a continuously coded parameter. The MRS estimation requires the parameter on the denominator (the comparison attribute) to be continuous. If the Length attribute was significantly linearly related and β_0 was not significant equation 7 would simplify to:

Equation 8 $$V = \beta_1LENGTH + \beta_2MARKERS + \beta_{3a}PSA12\% + \beta_{3b}PSA18\% + \beta_{4a}ACUTE35\% + \beta_{4b}ACUTE50\% + \beta_{5a}LATE15\% + \beta_{5b}LATE20\%,$$

where the $\beta_{1\text{-}5}$ variables are described in Table 4.

As discussed in chapter 2, there are four main limitations of MNL. The assumption of independence of irrelevant alternatives (IIA) did not apply in this study since it was a forced choice DCE and there were only two alternatives per choice set. However, the other three assumptions of MNL limit estimation of true human behaviour, especially the assumption of preference homogeneity, which likely restricted the realism of the analysis.

Table 4: Description of coefficients and the corresponding attributes.

Coefficients	Coding	The coefficient represents the attribute…
$\beta_1 LENGTH$	Continuous	'Length of radiotherapy'.
$\beta_{1a} LENGTH4$ $\beta_{1b} LENGTH8$	Dummy	'Length of radiotherapy' with '2 weeks' as the reference level.
$\beta_2 MARKERS$	Dummy	'Fiducial marker implantation' with 'no marker' as the reference level.
$\beta_{3a} PSA12\%$ $\beta_{3b} PSA18\%$	Dummy	'Risk of PSA recurrence' with '6% by 5 years' as the reference level.
$\beta_{4a} ACUTE35\%$ $\beta_{4b} ACUTE50\%$	Dummy	'Risk of acute moderate GI and GU toxicity' with '20%' as the reference level.
$\beta_{5a} LATE15\%$ $\beta_{5b} LATE20\%$	Dummy	'Risk of late moderate GI and GU toxicity' with '10%' as the reference level.
β_0	Dummy	Alternative specific constant (ASC) with reference level set to alternative A to test for ordering effect.

Goodness of fit

Conditional logit models are not able to measure absolute goodness of fit. An indicator used to determine the relative explanatory power of a model was the log likelihood value, which was developed to mimic the R^2 calculation. Higher (less negative) log likelihood values are associated with greater ability of a model to explain the pattern of choices in the data. These values cannot be used alone since they are a function of sample size. The likelihood ratio χ^2 test provided a method to determine whether including attribute-level variables significantly improved the model fit compared to a model without any attribute-level variables, and whether one or more of the preference weights were expected to be different from 0. The critical value for this test can be calculated as:

Equation 9 $$\textit{Likelihood ratio } \chi^2 = -2\ (\textit{LL model without predictors- LL of model}),$$

where *LL* is the log-likelihood value (22). The likelihood ratio has a χ^2 distribution where the degrees of freedom (df) are equal to the number of preference weights estimated in the model minus the number of alternative-specific constants (22). The model with the higher ratio can be assumed to fit the data better if each model has the same degrees of freedom. McFadden's pseudo R^2 value improves with the number of explanatory variables. In general, a measure between 0.2 to 0.4 indicates a good model fit. between 0.2 to 0.4 can be considered a good model fit.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are commonly used for model selection. AIC minimizes the mean square error of predictions. BIC is more consistent than AIC and tends to select the 'truer' model as sample size increases. When heterogeneity is small AIC is more likely to perform well, but since heterogeneity is expected BIC likely outperformed AIC in my DCE due to the stronger penalty afforded. A third criterion called Consistent Akaike Information Criterion (CAIC) has a stronger penalty for overparametrized models, and thus was used for comparison of latent classes due to the small sample size per group.

Subgroup analysis

Objective 1 included the hypothesis that preferences would vary based on respondent factors. It was expected that older respondents and those who lived further from the cancer centre would value length of EBRT more strongly than younger participants and those who lived closer to the cancer centre. The MNL model was estimated for two subgroups: (1) respondents younger than 70 years old and those 70 years and older; and (2) respondents who lived less than a 60-minute drive to the cancer centre and those whose drive was 60-minutes and longer. The MNL model was used for this objective because it was likely to be more accurate with the small sample sizes in each subgroup. More flexible analyses, such as MXL and LC, require researchers to make assumptions about random parameters and their distributions and likely need larger sample sizes (22).

Mixed multinomial logit model

An MXL model was initially tested with the dummy alternative ASC as the only randomly distributed variable, if this was not significant it was removed from the model. Next, each of the five attributes were added to the model as a random normal distribution variable, either as continuous (Length), dummy (Markers), or categorical (PSA, Acute, Late) based on the MNL model. After the addition of each randomly distributed variable the resulting model was compared with the previous model using a likelihood (LL) ratio test (Eq. 9). If the SDs for the mean values were significant and the model fit improved significantly the specified attribute level remained in the model as a random parameter. However, if these two conditions were not met then the attribute level was entered as a fixed parameter. Random parameter variables (β_i) were drawn from normal distributions (N):

Equation 10 $$\beta_i \sim N(\mu_i, \sigma_i),$$

where μ_i is the mean coefficient and σ_i is the SD of the original distribution. The distribution of the predicted individual β coefficients for each of the random parameters were plotted to examine if they followed the assumed posterior normal distribution. Individual preference weights are not directly interpretable and it is often difficult to determine the distributional characteristics of preferences (22). The proportion of respondents with positive parameter coefficients (opposite of what was expected) for each attribute estimated with the MXL model were examined, and if present was a good indication preference heterogeneity was present. This would be evidence that LC analysis could be useful to characterize which respondents had different preferences.

Latent class analysis

An alternative model called the LC model was used to express the potential preference heterogeneity with the best possible model fit for a set number of classes. For each class, different coefficients (or discrete random effects) were estimated for each attribute. The optimal number of classes was determined in an iterative procedure by comparing models' CAIC, AIC, and BIC values that varied by number of classes (23). LCM assumes individuals are implicitly sorted into a set of Q classes, but which class contains any specific respondent is unknown to the researcher (24). The model for a discrete choice among J_i alternatives:

Equation 11 $$\text{Prob(choice } j \text{ by individual } i \text{ in choice situation } t \,|\text{class } q| = e^{(x'it;j\,\beta q)} / \Sigma_{j=1}\, e^{(x'it,j\,\beta q)}$$
$$= F\,(i, t, j\,|q),$$

by individual i observed in T_i choice situations (24). This logit model posits the behaviour of individuals depends on observable attributes and on latent heterogeneity that varies with factors unknown to the investigator (24). Thus, the probability of selecting a certain alternative in the choice set depended on the class the participant was sorted into. With the estimated parameter coefficients the unconditioned prior probability for each individual to be in class q was modeled based on observable characteristics or risk factors (24). Using these probabilities and respondents' estimated choice probabilities based on Bayes theorem, the conditioned posterior estimates of the latent class probabilities were calculated (24).

The association of class membership with sociodemographic and prostate cancer factors were explored in the 2- and 3- class models by adding each as a covariate for predicting class membership. All covariates found to significantly impact class membership status were together estimated in the final LC model. The significant sociodemographic and prostate cancer factors identified within the LCM were tested for association with each class using Fisher's exact test.

Comparing models

This study used several different estimation models to quantify the preferences of prostate cancer patients for aspects of EBRT to treat localized disease. Choosing the optimal model to describe the respondents' choices was challenging given the many domains of contrast (24). The MXL model incorporated preference heterogeneity, and the LCM was able to identify different groups of participants based on their heterogeneity of preferences. Since the MXL and LC models are not nested, comparison of the likelihood ratio tests is not appropriate (24), and instead comparison of AIC and BIC was employed.

The willingness to trade off EBRT length to achieve less risk or recurrence, less risk of toxicity, or no marker implantation was estimated using marginal rate of substitution (MRS). The MRS was calculated for each attribute level by dividing the specified level coefficient by the coefficient for the continuous attribute Length (β_1). These results and comparisons are discussed in the next two chapters. Preference scores (also known as relative utility) for EBRT were generated based on the MXL and LC models. This resulted in relative utility scores for hypothetical EBRT scenarios. These scenarios were created to illustrate possible differences in patient preference for EBRT that would vary by model type. The higher (less negative) relative utility score indicated participants in that model (or class) would prefer that particular scenario over another scenario with a lower utility. The absolute values of utility (V) had no direct interpretation and could not be compared between models. However, the ranking of the hypothetical scenarios within each model could be compared. The preference score was first calculated for a 'base case' scenario using eq. (8) where the attribute levels were set as the middle levels (4 weeks, 12% risk of PSA recurrence, 35% risk of acute toxicity, and 15% risk of late toxicity) and no markers implanted. One or more attribute levels was then changed to gain insight about how each attribute systematically affected the utility score and rank relative to the base case. This is explored further in chapter 4.

References

1. de Bekker-Grob EW, Bliemer MC, Donkers B, Essink-Bot ML, Korfage IJ, Roobol MJ, et al. Patients' and urologists' preferences for prostate cancer treatment: a discrete choice experiment. B J Cancer. 2013 Aug;109(3):633-40.
2. Eliasson L, de Freitas HM, Dearden L, Calimlim B, Lloyd AJ. Patients' preferences for the treatment of metastatic castrate-resistant prostate cancer: a discrete choice experiment. Clin Ther. 2017 Apr;39(4):723-37.
3. King MT, Viney R, Smith DP, Hossain I, Street D, Savage E, et al. Survival gains needed to offset persistent adverse treatment effects in localised prostate cancer. Br J Cancer. 2012 Feb;106(4):638-45.
4. Howard K, Salkeld GP, Patel MI, Mann GJ, Pignone MP. Men's preferences and trade-offs for prostate cancer screening: a discrete choice experiment. Health Expect. 2015 Dec;18(6):3123-35.
5. Nakayama M, Kobayashi H, Okazaki M, Imanaka K, Yoshizawa K, Mahlich J. Patient preferences and urologist judgments on prostate cancer therapy in Japan. Am J Mens Health. 2018 Jul;12(4):1094-101.
6. Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. BMC Health Services Research. 2014 Sep;14:367.
7. Feldman-Stewart D, Brundage MD, Van Manen L, Svenson O. Patient-focussed decision-making in early-stage prostate cancer: insights from a cognitively based decision aid. Health Expect. 2004 Jun;7(2):126-41.
8. Holmboe ES, Concato J. Treatment decisions for localized prostate cancer: asking men what's important. J Gen Intern Med. 2000 Oct;15(10):694-701.
9. Benjamin LC, Tree AC, Dearnaley DP. The role of hypofractionated radiotherapy in prostate cancer. Curr Oncol Rep. 2017 Apr;19(4):30.
10. Haque W, Butler EB, Teh BS. Stereotactic body radiation therapy for prostate cancer-a review. Chin Clin Oncol. 2017 Sep;6(Suppl 2):S10.
11. Loblaw A, Liu S, Cheung P. Stereotactic ablative body radiotherapy in patients with prostate cancer. Transl Androl Urol. 2018 Jun;7(3):330-40.
12. Johnson RF, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, Regier DA, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. Value Health. 2013 Jan;16(1):3-13.
13. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012 Feb;21(2):145-72.
14. Aguinis H, Bradley KJ. Best practice recommendations for designing and implementing experimental vignette methodology studies. Organ Res Methods. 2014 Oct;17(4):351-71.
15. Bech M, Kjaer T, Lauridsen J. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. Health Econ. 2011 Mar;20(3):273-86.
16. Neuendorf KA. The content analysis guidebook. 2nd ed. Thousand Oaks, CA: SAGE Publications, Inc.; 2017.
17. Christie M, Azevedo CD. Testing the consistency between standard contingent valuation, repeated contingent valuation and choice experiments. J Agric Econ. 2009 Jan;60(1)154-70.
18. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys. 1995 Mar;31(5):1341-6.
19. Orme B. Getting started with conjoint analysis: strategies for product design and pricing research. 2nd ed. Madison, WI: Research Publishers LLC.; 2010.
20. Rose JM, Bliemer MCJ. Sample size requirements for stated choice experiments. Transportation. 2013 Sep;40(5):1021-41.

21. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. Patient. 2015 Oct;8(5):373-84.
22. Hauber AB, Gonzalez JM, Groothuis-Oudshoorn CG, Prior T, Marshall DA, Cunningham C, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR conjoint analysis good research practices task force. Value Health. 2016 Jun;19(4):300-15.
23. Pacifico D, Yoo HI. lclogit: a Stata module for estimating latent class conditional logit models via the expectation-maximization algorithm [Internet]. School of Economics, The University of New South Wales; 2012. Available from: https://ssrn.com/abstract=2174146
24. Greene WH, Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed logit. Transport Res B-Meth. 2003 Sep;37(1):681-98.

Chapter 4

Results

Sample population

Respondents

Fifty-eight participants were recruited for this study, 29 completed each version of the questionnaire. Respondent characteristics are detailed in Table 5. More than 70% of participants were between the ages of 65 to 80, with a mean age of 72 (Table 5). Almost 80% of participants described no late toxicity, which included the almost half of participants who were undergoing EBRT when they completed the questionnaire and the time for this side effect to develop had not passed yet. Of the remaining 33 participants who had previously completed EBRT five reported mild, six moderate, and one severe late toxicity.

Non-responders

Twenty-four potential participants declined to be part of the study, giving a response rate of 70.7% (58/82). The mean age of non-responders was the same as responders, 72 years old, and similar to the overall sample about a third of participants lived less than 30 minutes from the cancer centre. Response rates were similar between younger patients and those 70 years or older (38/52, 73.1%; and 20/30, 66.7% respectively). Two of the non-responders initially agreed to take part in the study but after an initial explanation of what a DCE was they felt it was too confusing and declined to take part. One participant informed the researcher the study's purpose did not make sense to him, and the other indicated their physician should make this choice so he would not be able to make choices in the DCE.

Table 5: Characteristics of the respondents.

Characteristics	Value n=58 (%)
Age	
<65	7 (12.1)
65-69	13 (22.4)
70-74	12 (20.7)
75-79	16 (27.6)
≥80	10 (17.2)
Household	
Single	7 (12.1)
With partner	51 (87.9)
Driving distance from local cancer centre	
< 30 minutes	18 (31.0)
30-60 minutes	16 (27.6)
> 60 minutes	24 (41.4)
Education	
University, college, or technical certificate	37 (63.8)
High school or less	21 (36.2)
Employment status	
Currently working part or full time	9 (15.5)
Retired or on disability	49 (84.5)
Currently receiving radiotherapy	
Yes	25 (43.1)
No	33 (56.9)
Received systemic therapy for prostate cancer	
Yes	27 (46.6)
No	31 (53.4)
Undergone surgery for prostate cancer	
Yes	8 (13.8)
No	50 (86.2)
Undergone brachytherapy for prostate cancer	
Yes	6 (10.3)
No	52 (89.7)
Self-reported severity of acute toxicity	
None	14 (24.1)
Mild	20 (34.5)
Moderate	19 (32.8)
Severe	5 (8.6)
Self-reported severity of late toxicity	
None or not applicable	46 (79.3)
Mild	5 (8.6)
Moderate	6 (10.4)
Severe	1 (1.7)

Models

A total of 1392 observations (12 choice sets thus 24 observations per respondent) were used for each model. No data were excluded and there were no missing values. The DCE responses were analysed with three model types discussed in detail in chapter 3. The multinomial logit (MNL), mixed multinomial logit (MXL), and latent class model (LCM) analyses, and goodness of fit are examined in this section. As indicated in chapter 3, attributes that were statistically significant ($p < 0.05$) were interpreted to be important to how participants made their choices in the DCE. This indicated respondents differentiated between one attribute (or attribute level) and another in making their choices in the DCE. Non-significant results were of unknown significance as they could not be determined to be not important to participants' choices, especially in the subgroup and latent class analyses since they were underpowered due to the small sample size. *A priori* all attributes were expected to have a negative effect based on the framing of the attributes. As explained in chapter 3 the Length attribute parameters were tested for a linear relationship, and the Markers, PSA, Acute, and Late attributes were dummy coded.

Objective 1: Parameter estimates

MNL

The first objective of my thesis was to quantify the coefficient values, also known as parameter estimates, for each attribute. The null hypotheses were that the estimates for each attribute would not be significantly different from zero. The sign of a coefficient indicated whether the attribute had a positive or negative impact on the utility score (32). The alternative specific term was not significantly different from 0 ($p= 0.291$), thus there was no ordering effect present in the model that would have represented an underlying preference for radiotherapy alternative A or B. Consequently, none of the models presented in this chapter include an ASC. The F-test of linearity for the attribute length of radiotherapy was significant ($p= 0.004$); therefore, this attribute was fitted as a continuous attribute in all analyses going forward (Table 6). In the MNL model all attributes were statistically significant; thus, it was interpreted that all attributes were important to participants' choices (Table 7). The *a priori* hypothesis that the signs for all the parameter estimates would be negative was correct. This indicated participants preferred a shorter radiotherapy treatment without fiducial markers and less risk of PSA recurrence and toxicity (Table 7).

Table 6: Test of linearity for Length attribute.

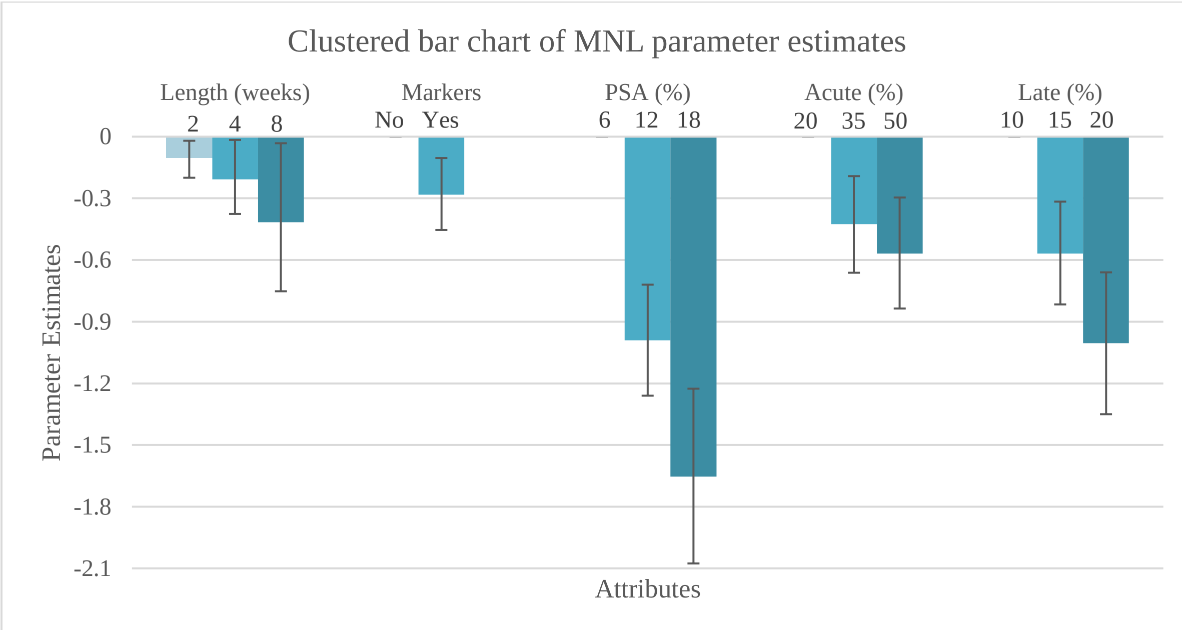
Parameter description	Reference (2 weeks)	β_{1a} (4 weeks)	β_{1b} (8 weeks)
Coefficients	0	-0.137 (.295)	-0.420 (.003)
F- test	R^2 = 1.000 (.004)		

p-values in parentheses.

Table 7: Comparison of MNL and MXL models.

	MNL		MXL		
Characteristic	*β*	*p-value*	*β*	*p-value*	*SD (p-value)*
Length β_1	-0.052*	.021	-0.091*	.017	0.176 (<.001)
Markers β_2 (ref. no markers)	-0.282*	.002	-0.563*	.004	1.043 (<.001)
PSA (ref. 6%)					
12% β_{3a}	-0.990*	<.001	-1.456*	<.001	-
18% β_{3b}	-1.653*	<.001	-2.581*	<.001	1.772 (<.001)
Acute (ref. 20%)					
35% β_{4a}	-0.426*	<.001	-0.724*	<.001	-
50% β_{4b}	-0.568*	<.001	-1.005*	<.001	0.675 (.008)
Late (ref. 10%)					
15% β_{5a}	-0.568*	<.001	-0.969*	<.001	-
20% β_{5b}	-1.005*	<.001	-1.854*	<.001	1.431 (.001)
Goodness of fit					
Log likelihood	-430.33		-395.26		
Degrees of freedom (df)	8		8		
Likelihood ratio χ^2	4.16		70.14		
Prob > χ^2 than model with df-1	.041		.000		
Pseudo R^2	0.1080		-		
AIC	876.68		816.53		
BIC	918.58		884.63		

*indicates parameter estimate is significant at 0.05 level; **-** indicates not applicable.



Reference levels set at 0: No, 6%, 20%, 10%.

Figure 9: Clustered bar chart of MNL parameter estimates.

The largest parameter estimate was the decrease in risk of PSA recurrence from 18% to 6% (-1.653). The second largest was the parameter estimate for a decreased risk of late toxicity from 20% to 10% (-1.055). The smallest coefficient for a non-continuous variable was the Markers attribute (-0.282). Overall the smallest coefficient was the continuous Length variable (-0.052). The magnitude of the attribute coefficients corresponded with the relative importance of the attributes, for example on average in the MNL model risk of late toxicity was more important to participants than risk of acute toxicity even though the 95% CI overlapped (Figure 9). For correct interpretation of the comparison of the parameter estimates within one model (e.g. only considering MNL) the units of measurement must be considered carefully. The continuous attribute Length had a coefficient of -0.052, which indicated for a 1 week increase in treatment length the average utility decreased by this number. For example, the difference in utility between the levels 2 weeks and 8 weeks was -0.312 (-0.052 x 6) (visualized in Figure 9). The coefficient in the MNL model for the PSA attribute at the 12% level was -0.990, which corresponded to the average decrease in utility of a change from 12% to 6% risk. With higher risk there was a larger associated difference in utility between the 18% and 6% levels (-1.653). The Acute attribute was dummy coded, and respondents had greater preference to decrease risk of acute toxicity from 35% to 20% (-0.426) than 50% to 35% (-0.142 = -0.568 – (-0.426); Figure 9). This larger difference in utility between the middle and lowest risk levels was demonstrated by all three risk attributes (Figure 9), though the difference between the Late attribute parameter estimates resulted in relative marginal utility differences that were more proportional. When the risk of late toxicity was reduced from 15% to 10% the utility changed by -0.587, while 20% to 15% was similar with a parameter estimate of -0.474 (-1.061 – (-0.587)). Therefore, on average (though some of the 95% CI overlapped) the ranking of the attributes from most to least important to the participants was: (1) risk of PSA recurrence; (2) risk of late toxicity; (3) risk of acute toxicity; (4) length of EBRT; (5) and fiducial marker implantation (Figure 9).

Subgroup analysis

As part of my first objective I hypothesized older respondents and those who lived further from the cancer centre would value length of EBRT more strongly than younger participants and those who lived closer to the cancer centre. Two subgroup analyses were conducted using the MNL model: (1) respondents under and over 70 years old (Table 8); and (2) those who lived less than 60 minutes, compared to a 60 minute or longer drive to the cancer centre (Table 9). The under 70 years old subgroup model fit the data well even though there were only 480 observations (20 respondents in this subgroup). The pseudo R^2 value was within the good fit range of 0.2 to 0.4 with a value of 0.243, and the AIC and BIC values were low. The over 70 years old subgroup did not have as good a fit leading to the conclusion older participants likely had more heterogeneous preferences, though all parameter estimates were significantly different from zero (Table 8). Similarly, the model fit for respondents who lived over 60 minutes from the cancer centre fit the data better than the model for those who lived closer, even though this latter group had a larger sample size (Table 9). The pseudo R^2 value for the subgroup who lived further away was 0.1559, closer to the good fit range than the overall MNL model, even though it only included 24 participants.

As expected, the parameter estimates for the subgroups varied from those in the main MNL model. These results suggest that age and distance acted as effect modifiers on the relationship between preferences and the choices participants made in the DCE. In the subgroups some attributes were not statistically significant, thus they were of unknown significance to participants' decisions. All attributes were statistically significantly important to older participants' (70 years or older) choices, which was similar to the main MNL model. For participants who lived closer (less than an hour drive to cancer centre), length of EBRT was of unknown significance to their choices. Those who lived further away statistically significantly considered length important. These findings proved my *a priori* hypothesis that age and distance would influence respondents' choices for length of EBRT was correct.

Table 8: MNL subgroup analysis by age.

Age subgroup	**<70 years old**			**≥70 years old**		
Characteristic	β	*p-value*	*95% CI*	β	*p-value*	*95% CI*
Length β_1	-0.04	.551	(-0.11, 0.04)	-0.06*	.024	(-0.12, -0.01)
Markers β_2	-0.16	.346	(-0.57, 0.12)	-0.31*	.006	(-0.52, -0.09)
PSA						
12% β_{3a}	-1.60*	<.001	(-2.15, -1.06)	-0.77*	<.001	(-1.11, -0.44)
18% β_{3b}	-3.00*	<.001	(-3.86, -2.15)	-1.11*	<.001	(-1.62, -0.59)
Acute						
35% β_{4a}	-0.75*	.001	(-1.19, -0.27)	-0.37*	.011	(-0.66, -0.09)
50% β_{4b}	-1.29*	<.001	(-1.88, -0.71)	-0.39*	.017	(-0.71, -0.07)
Late						
15% β_{5a}	-0.65*	.005	(-0.98, -0.09)	-0.42*	.009	(-0.73, -0.11)
20% β_{5b}	-1.06*	.003	(-1.89, -0.54)	-0.98*	<.001	(-1.41, -0.54)
Goodness of fit						
Log likelihood	-257.14			-168.50		
LL ratio χ^2	51.34			62.25		
Prob > χ^2	.000			.000		
Pseudo R^2	0.0908			0.1559		
# observations	816			576		
AIC	530.27			353.00		
BIC	567.91			387.85		

*indicates significance at 0.05 level; obs: observations.

Table 9: MNL subgroup analysis by driving distance from cancer centre.

Distance subgroup	**< 60 min**			**≥ 60 min**		
Characteristic	*β*	*p-value*	*95% CI*	*β*	*p-value*	*95% CI*
Length β_1	-0.026	.377	(-0.09, 0.02)	-0.083	.023	(-0.14, -0.02)
Markers β_2	-0.393	.001	(-0.54, -0.11)	-0.167	.245	(-0.53, 0.09)
PSA						
12% β_{3a}	-0.887	<.001	(-1.16, -0.49)	-1.259	<.001	(-1.85, -0.74)
18% β_{3b}	-1.466	<.001	(-1.82, -0.79)	-2.104	<.001	(-2.78, -1.51)
Acute						
35% β_{4a}	-0.353	.023	(-0.70, -0.05)	-0.586	.003	(-0.96, -0.11)
50% β_{4b}	-0.553	.002	(-0.99, -0.10)	-0.741	.002	(-1.21, -0.29)
Late						
15% β_{5a}	-0.716	<.001	(-0.97, -0.08)	-0.351	.069	(-0.71, 0.01)
20% β_{5b}	-1.083	<.001	(-1.87, -0.56)	-0.914	.001	(-1.38, -0.57)
Goodness of fit						
Log likelihood	-257.14			-168.50		
LL ratio χ^2	51.34			62.25		
Prob > χ^2	.000			.000		
Pseudo R^2	0.0908			0.1559		
# observations	816			576		
AIC	530.27			353.00		
BIC	567.91			387.85		

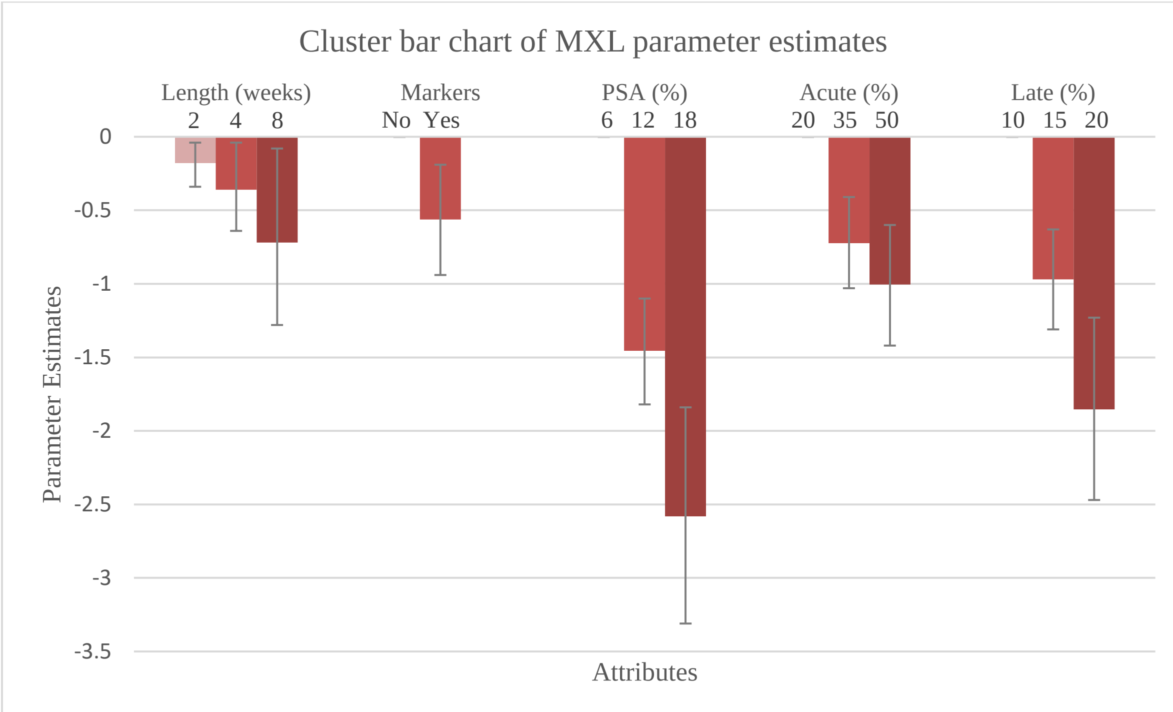
*indicates significance at 0.05 level; obs: observations.

Thus, it was determined heterogeneity was present in participants' choices based on some known factors that the MNL model was unable to account for. This was evidenced in the improved, but not ideal pseudo R^2 values in the subgroup analyses compared to the main MNL model. The next section explores the MXL analysis that considered unknown factors influencing preference heterogeneity.

MXL

An MXL analysis accounts for the relative weight each respondent placed on a certain attribute, known as preference heterogeneity, with randomly distributed parameters. As in the MNL model all attributes in the MXL model were statistically different from zero and therefore the null hypothesis was rejected for each attribute (Table 7). The negative signs of the coefficients indicated participants preferred a shorter radiotherapy treatment without fiducial markers and less risk of PSA recurrence and toxicity. Five of the eight attributes significantly improved the fit when they entered the model as random parameters (Length, Markers, PSA at the 18% level, Acute at the 50% level, and Late at the 20% level). These five parameters all had significant SDs which indicated participants had heterogenous preferences for all five attributes of EBRT management for localized prostate cancer and they should be modeled as randomly distributed parameters (Table 7). The SDs for the middle levels of the risk attributes (PSA at the 12% level, Acute at the 35% level, and Late at the 15% level) were not statistically significant, and therefore these three attribute levels entered the model as fixed parameters.

The ranking of the mean magnitude of the coefficients were the same as for the MNL model; the largest coefficient was the PSA attribute to decrease risk from 18% to 6% (-2.581), the second largest was the Late attribute to decrease risk from 20% to 10% (-1.854), and the smallest coefficient for a non-continuous variable was the Markers attribute (-0.563) (Figure 10). Overall the smallest coefficient was for the continuously related Length variable (-0.091). As reported for the MNL model, the interpretation of the magnitude of the coefficients were carefully considered (visualized in Figure 10). For example, the random parameter coefficient in the MXL model for the Acute attribute at the 50% level was -1.005 and implied the average decrease in utility of a change from 50% to 20% risk was this value (-1.005 x 1) (Table 7). The fixed parameter coefficient in the MXL model for this same attribute but at the 35% level was -0.724 (Table 7). The continuous Length coefficient was -0.091 and indicated for each 1 week increase in length the average utility decreased by this number. For example, the difference in utility between 8 and 2 weeks was 0.546 (-0.091 x 6) (visualized in Figure 10), thus the Length attribute levels had the smallest changes in utility (Figure 10). In other words, worse utility is associated with the 8-week level overall, however the difference to shorten treatment to 2 weeks resulted in a smaller utility change than avoiding Markers (level change from yes to no), so participants in general prefer to avoid fiducial marker implantation even at the expense of an 8 week EBRT regimen. Therefore, the ranking of the attributes in this model from most to least important (using the largest expected difference in utility



Reference levels set at 0: No, 6%, 20%, 10%.

Figure 10: Clustered bar chart MXL parameter estimates.

between the best and worst levels) was; risk of PSA recurrence, risk of late toxicity, risk of acute toxicity, fiducial marker implantation, and length of EBRT. For correct comparison between models only the ranking of the coefficients could be assessed and not the actual estimate due to scale differences between types of analyses. The ranking of the attributes for MNL and MXL models was the same for the risk attributes. In the MNL analysis length was more important than avoiding fiducial marker implantation, while in the MXL analysis that accounted for preference heterogeneity avoiding fiducial marker implantation was more important to participants than shortening EBRT.

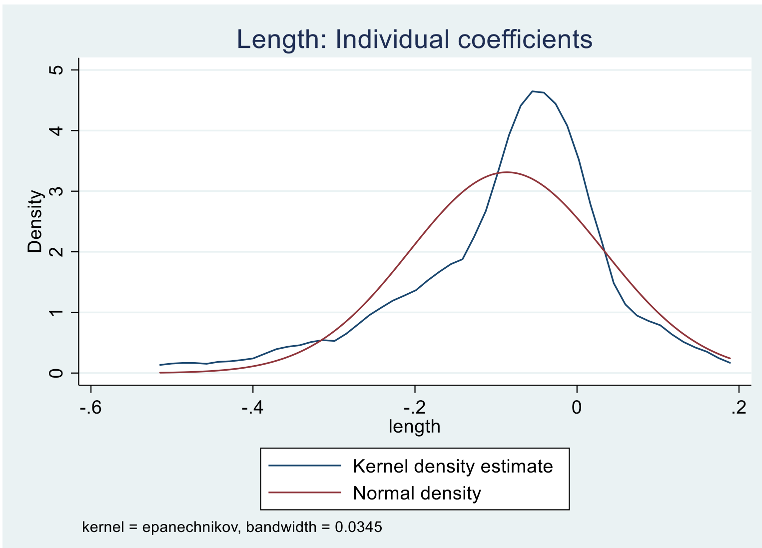
To assess goodness of fit the log likelihood values were directly compared between the models since they had the same number of degrees of freedom. Since the MXL model had the lower log likelihood, AIC, and BIC values, and a higher likelihood χ^2 ratio, the data was concluded to fit the MXL model better than the MNL model (Table 7). As seen in Table 10 a proportion of participants had opposite preferences than expected for each attribute. The risk attributes had the least amount of preference heterogeneity (range 5.9% to 9.8%), while the Markers and Length attributes had the most (29.5% to 30.3%). This is interesting since both the MNL and MXL models found the risk attributes were most important to participants, and the Length and Markers attributes were least important. The fit of the MXL model was further evaluated by inspecting whether the predicted individual coefficients for each of the five random parameters followed the assumed posterior normal distribution (Figure 11 A-E). All five attributes' individual coefficients approximated the normal distribution and therefore the MXL model was accepted as fitting the data better than the MNL model. This evidence provided a good picture of the heterogeneity of preferences, and together with the significant SD for each attribute this indicated there was preference heterogeneity that needed to be further explored for relationships and categories within these preferences. This was studied with latent class modeling and discussed in detail in the next section.

LCM

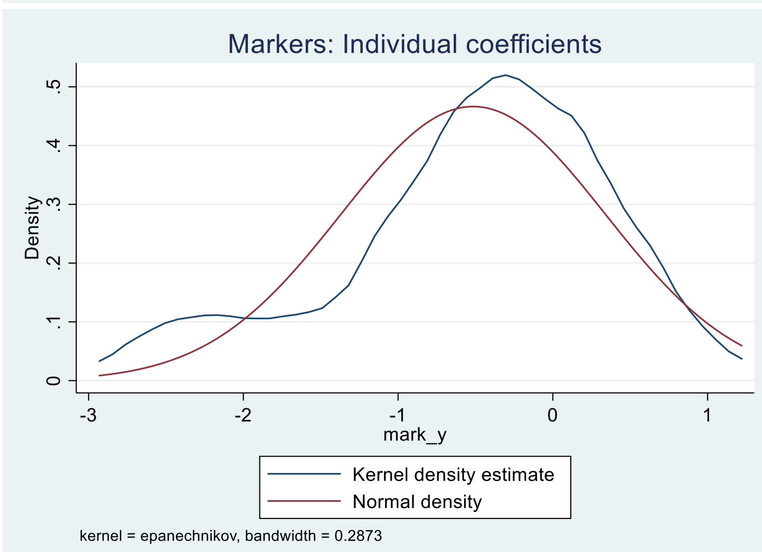
The LCM model with the optimal number of classes generally has the lowest CAIC value (90). The CAIC values for 2- and 3-classes were the lowest, and by a small margin was lowest for 2-classes (Figure 12 plots CAIC values for models with 1 to 10 classes). Commonly, the curve is 'U' shaped and the CAIC values progressively increase for higher classes after the optimal number of classes is found (90). This was not seen in Figure 12, likely due to the small sample size as at least 20 to 30 participants are required per class to sufficiently model preferences (25). Thus, class sizes larger than 3 could not be estimated with the sample size of 58 in this study. The 3-class model parameter estimates are displayed in Table 11. The 3-class analysis is reported here as this model fit the data better than the 2-class LCM based on goodness of fit, conditioned and unconditioned probability, and the posterior probability (available in Appendix C Table 22). Each parameter significantly improved the model fit as calculated with eq. (9). In the LCM three preference types were identified and each class incorporated about a third of the participants. In the two class LCM classes 3.1 and 3.3 from the three class LCM (Table 11) were

Table 10: Proportion of respondents with positive individual coefficients for each randomly distributed parameter in the MXL model (visualized in Figure 8).

Attribute (Figure 8)	Z value	Proportion with positive individual coefficients
Length (A)	-0.5171	0.3026
Markers (B)	-0.5398	0.2947
PSA at 18% level (C)	-1.4566	0.0726
Acute at 50% level (D)	-1.5630	0.0590
Late at 20% level (E)	-1.2956	0.0976

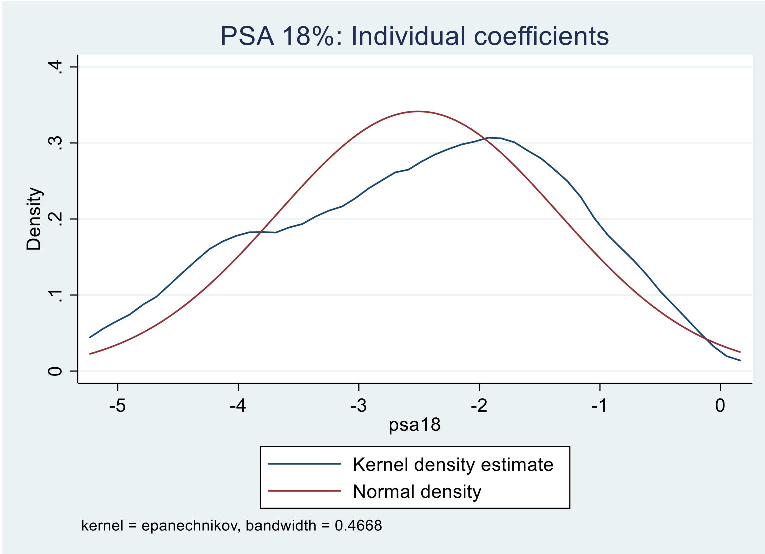


A.

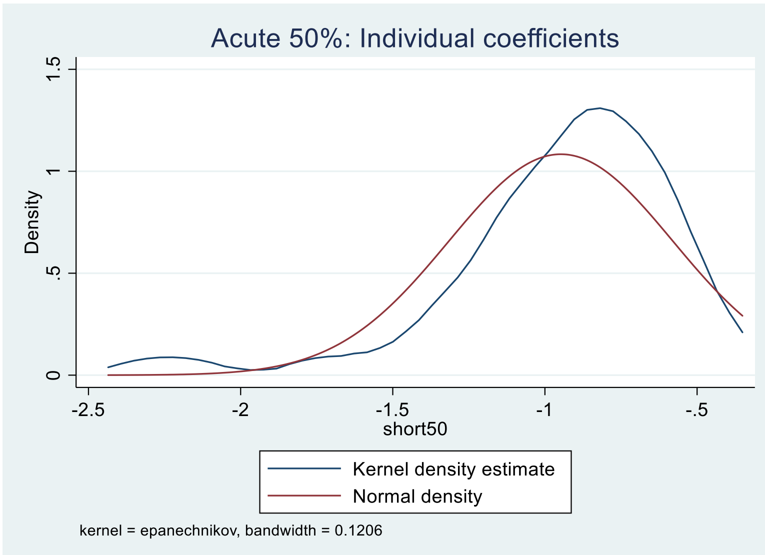


B.

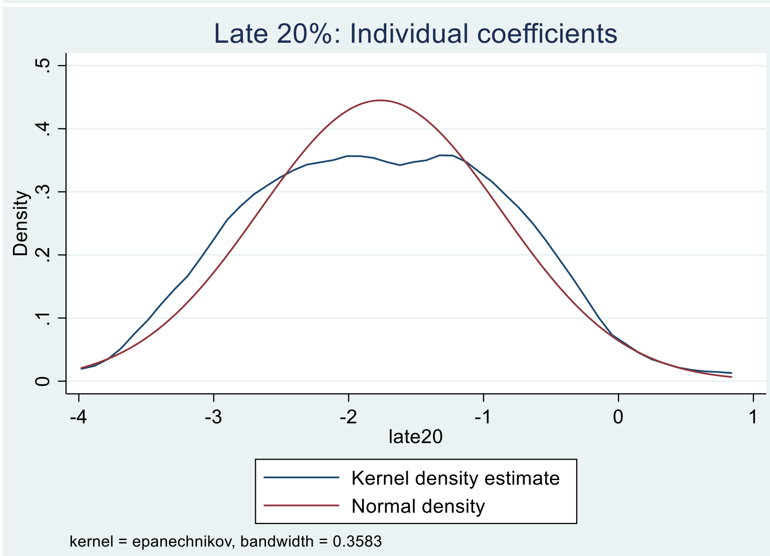
Figure 11: Kernel density charts for individual's β coefficient for the five randomly distributed parameters in the MXL model (A-B; C-E next page).


PSA 18%: Individual coefficients
Density
psa18
Kernel density estimate
Normal density
kernel = epanechnikov, bandwidth = 0.4668

C.


Acute 50%: Individual coefficients
Density
short50
Kernel density estimate
Normal density
kernel = epanechnikov, bandwidth = 0.1206

D.


Late 20%: Individual coefficients
Density
late20
Kernel density estimate
Normal density
kernel = epanechnikov, bandwidth = 0.3583

E.

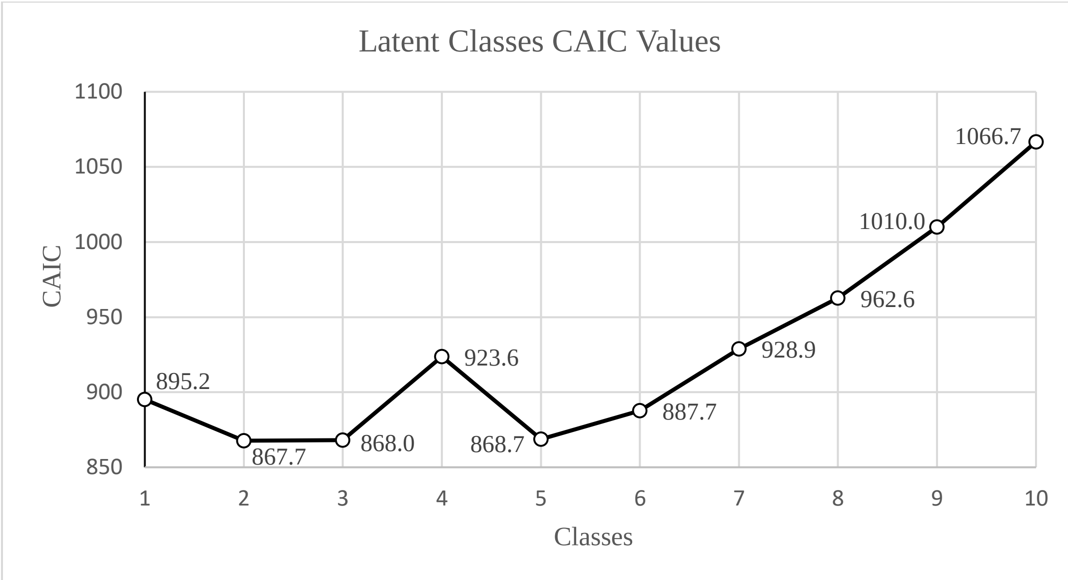


Figure 12: Plot of CAIC values for LC models.

combined into class 2.2 (Appendix C Table 22), and based on the preference patterns discussed in the next paragraphs three classes fit the data better.

Class 3.1 participants, the “toxicity group”, preferred to avoid marker implantation and toxicity, and avoid risk of PSA recurrence at the 12% level, but not the higher 18% level (Table 11). Compared to class 3.3, class 3.1 more strongly preferred to avoid the highest (20%) risk of late toxicity (95% CI did not overlap). Class 3.2, the “convenience group”, was the only group who preferred shorter treatment length. This class also preferred to avoid fiducial marker implantation and risk of PSA recurrence at the 12% level. Class 3.3, the “efficacy group”, most strongly made their choices based on avoiding risk of PSA recurrence and was the only class where both levels of the PSA attribute significantly impacted respondents’ choices.

The highest posterior probability of class membership was 0.993, indicating the model did very well in distinguishing among different underlying preferences for the observed choice behaviour (90). Thus, there was little uncertainty regarding the classes that respondents belonged to. The model’s ability to make in-sample predictions of the actual choice outcomes was examined. The average unconditional

Table 11: 3-class LCM.

Characteristics	Class 3.1 "Toxicity group"		Class 3.2 "Convenience group"		Class 3.3 "Efficacy group"	
	β	*p-value*	*β*	*p-value*	*β*	*p-value*
Length β_1	0.041	.429	-0.243*	<.001	-0.028	.580
Markers β_2 (ref. no markers)	-0.545*	.038	-1.379*	<.001	0.024	.909
PSA (ref. 6%)						
12% β_{3a}	-1.277*	.001	-1.014*	.001	-2.735*	<.001
18% β_{3b}	-1.198	.065	-0.537	.238	-5.766*	<.001
Acute (ref 20%)						
35% β_{4a}	-1.077*	<.001	-0.470	.058	-0.789*	.011
50% β_{4b}	-2.334*	<.001	0.293	.289	-1.222*	.001
Late (ref. 10%)						
15% β_{5a}	-2.235*	<.001	-0.465	.099	-0.964*	.002
20% β_{5b}	-3.602**	<.001	-0.103	.758	-1.475**	.001
Class share	0.311		0.328		0.360	
≥70 years old	+		+		-	
Working	-		-		+	
Prior systemic therapy	+		+		-	
Posterior probability	0.993					
Unconditional probability	0.570		0.550		0.651	
Conditional probability	0.738		0.705		0.757	
Observations	216		228		252	
Goodness of fit						
Log likelihood	-339.50					
AIC	795.29					
BIC	861.22					
CAIC	893.22					

* : parameter estimate significantly different from 0
** : coefficient significantly different from other classes estimates (based on 95% CI exclusion).
+ : higher probability characteristic present in class members.

choice probability was higher than 0.5, which is the value a naïve model would predict given there were 2 alternatives per choice set (90). The average conditional probability was higher than this number in all three classes and was over 75% in class 3.3. This confirmed the model described the choice behaviour well. As seen in Table 12 the fit of the model significantly improved when membership for age (dummy coded for 70 years and older as 1), working status (full time or part time work dummy coded as 1), and previous systemic therapy (dummy coded 1 for ADT or chemotherapy) were included. When class membership was tested with Fisher's exact test classes 3.1 and 3.2 tended to have older members ($p= 0.094$), class 3.3 did not have any members who had been treated with systemic therapy ($p= 0.053$), and there was a significant association between working at the time of the DCE and being a member of class 3.3 ($p= 0.029$). The association of these three factors by class are delineated in Table 13 and positive association was indicated in Table 11 with '+' when participants in that class were more likely to be associated with that characteristic. Overall the model fit was significantly improved by including all three characteristics, however the only significant association between a class and characteristic was work status with class 3.3. Of note there was no relationship between class and distance from the cancer centre.

The LCM identified three groups of participants' preferences: (1) a group most concerned about toxicity and almost all choices were made based on the alternative with the lowest risk of toxicity; (2) a group who considered convenience as most important and participants tended to select alternatives with shorter treatment lengths and no marker implantation; and (3) a group that considered efficacy as most important and almost all choices were based on the alternative with lowest risk of PSA recurrence.

Table 12: Covariates relationship with class membership in LCM.

Covariate	**df**	**Log Likelihood**	**Likelihood Ratio χ^2**	**Prob> χ^2**
Base	-	-347.70	-	-
Age (<70 or ≥70)	1	-344.97	5.46	.02*
Partner status	1	-347.58	0.24	.62
Distance	1	-347.25	0.45	.50
Education level	1	-347.55	0.30	.58
Work status	1	-345.29	4.82	.03*
Time since diagnosis	1	-347.67	0.08	.78
Received EBRT at time of DCE	1	-347.37	0.66	.42
Prior systemic therapy (ADT or chemotherapy)	1	-344.63	6.14	.01*
Received brachytherapy	1	-346.09	3.22	.07
Self-reported: acute toxicity	1	-346.97	1.46	.23
Self-reported: late toxicity	1	-347.68	0.04	.84
Age ≥70, work, prior systemic therapy	**3**	**-339.50**	**16.60**	**.009***

* : statistically significant.

Table 13: Significant membership covariates and listed attributes important to participants decisions compared by class.

Characteristic	Number of participants (%)
Over 70 years old	Total: 38 (65.52%)
Class 3.1	12 (66.67%)
Class 3.2	15 (83.33%)
Class 3.3	11 (50.00%)
Work	Total: 9 (15.52%)
Class 3.1	1 (5.56%)
Class 3.2	1 (5.56%)
Class 3.3	7 (31.82%)
Previous systemic therapy	Total: 7 (12.07%)
Class 3.1	4 (22.22%)
Class 3.2	3 (16.67%)
Class 3.3	0
Age	**Mean:** 72.88
Class 3.1	72.17
Class 3.2	75.56
Class 3.3	71.27

Objective 2: Trade-offs

The second objective of my thesis was to quantify how systematically varying the levels assigned to each attribute impacted the respondents' willingness to trade-off each attribute for the ease of shorter treatment time. To compare the model results I had to calculate the trade-offs participants were prepared to make between the attributes as the absolute coefficient values could not be compared directly (91). Two methods were explored to answer this objective: (1) marginal rates of substitution (MRS) discussed in chapter 3; and (2) calculating the utility for theoretical EBRT scenarios and comparing the ranking of these cases for different models.

Contrasts in MRS are informative, unlike a direct comparison of absolute parameter estimates due to scale differences (91). For example, the MRS to avoid a 5% increased risk of late toxicity (difference between 15% and 10%) was estimated as $\beta_{5a}/\text{-}\beta_1$, where β_1 represented the importance of a one week change in length of EBRT. The trade-offs between the units/levels of each attribute and length of treatment was calculated using the MXL and LC models. The MRS values for several models discussed in this chapter are reported in Table 14. Respondents expressed they were prepared to lengthen EBRT to avoid increased risk of PSA recurrence, as well as acute and late toxicity, in the MXL model as the 95% CIs excluded zero. The participants in the MXL model were not willing to lengthen EBRT to avoid fiducial marker implantation since the associated MRS 95% CI included zero (Table 14). The MXL model found avoiding the risk of PSA recurrence was most important to participants as they were willing to lengthen EBRT by 18 weeks to reduce risk of PSA recurrence from 12% to 6%, and 30 weeks to reduce risk from 18% to 6%. However, for these estimates the 95% CI were wide and indicated a range of 1 to 35 weeks, and 3 to 57 weeks, respectively (Table 14). To avoid risk of acute and late moderate GI and GU toxicity participants were willing to significantly lengthen EBRT by many weeks, though there was a wide 95% CI (Table 14). In the clinical context these trade-offs indicated the relative importance of these attributes, though the MRS estimates do not correspond to EBRT regimens as the longest treatment is CFRT at 8 weeks. Additionally, the estimates were not precise as demonstrated by the wide 95% CI.

The LC model could not be used to assess MRS values between the classes because Length was only significantly different from zero in class 3.2 ("convenience group"). Class 3.2 members were the only group of participants who would trade off length of EBRT to avoid fiducial markers, they were the least willing to lengthen EBRT to reduce risk of PSA recurrence from 12% to 6%, and the only group who would not increase length of EBRT to reduce risk of PSA recurrence from 18% to 6% (Table 13). Considering the subgroup analyses, older participants (70 years or older) were less willing to trade-off reduced risk of PSA recurrence for longer EBRT and were only willing to lengthen EBRT to reduce risk of late toxicity from 20% to 10%. This subgroup has the closest MRS values to class 3.2, which is logical since this class was significantly associated with age over 70 years old. Participants who lived further from the cancer centre were more willing to trade-off longer EBRT to reduce higher level risks, but the model was unable to determine if these respondents would reduce risk from middle to lowest levels (e.g. to reduce risk of acute toxicity from 35% to 20%, and late toxicity from 15% to 10%) since the associated MRS 95% CI included zero (Table 14). The subgroup analyses identified different groups of participants with unique preferences compared to the LC analysis.

Theoretical EBRT scenarios

Hypothetical alternatives not in the DCE that mimicked clinically realistic scenarios based on the attribute levels were created. Eq. 7 in chapter 3 was used to calculate utility scores estimated from the MXL and LC models (Tables 15 and 16). Reasonable respondents were expected to select the option associated with the highest (least negative) utility. The preference score, or utility, for a base case scenario where the attribute levels were set as 4 weeks, no markers implanted, 12% risk of PSA recurrence, 35% risk of acute toxicity, and 15% risk of late toxicity was calculated (Table 14). The relative ranking of these clinically realistic scenarios was compared (Tables 15 and 16) where a rank of 1 would be most preferred by respondents.

Table 14: MRS estimates for Length compared to other four attributes.

	MNL	**MNL Age: ≥70 years old**	**MNL Distance: ≥60 minutes**	**MXL**	**LC 3.2 "Convenience group"**	**Interpretation** Willing to lengthen EBRT by…
Markers β_2 (ref. no markers)	-5.39 (-11.00, 0.22)	-4.84 (-10.30, 0.61)	-2.02 (-5.87, 1.83)	-7.09 (-14.98, 0.80)	-5.66* (-7.97, -3.35)	-weeks to avoid markers implanted
PSA (ref. 6%)						
12% β_{3a}	-18.94* (-35.97, -1.90)	-12.29* (-23.91, -0.66)	-15.18* (-29.95, -0.42)	-18.16* (-35.27, -1.02)	-4.14* (-6.77, -1.51)	-weeks to decrease risk from 12% to 6%
18% β_{3b}	-31.64* (-57.63, -5.65)	-17.57* (-32.78, -2.36)	-25.38* (-46.73, -4.03)	-30.16* (-57.26, -3.06)	-2.21 (-4.39, 0.30)	-weeks to decrease risk from 18% to 6%
Acute (ref. 20%)						
35% β_{4a}	-8.14* (-16.03, -0.26)	-5.93 (-12.26, 0.40)	-7.07 (-14.63, 0.50)	-8.23* (-15.82, -0.64)	-1.93 (-4.78, 0.98)	-weeks to decrease risk from 35% to 20%
50% β_{4b}	-10.87* (-20.35, -1.39)	-6.16 (-12.62, 0.30)	-8.94* (-17.38, -0.50)	-11.90* (-23.41, -0.39)	1.21 (-1.31, 3.90)	-weeks to decrease risk from 50% to 20%
Late (ref. 10%)						
15% β_{5a}	-10.88* (-20.57, -1.18)	-6.60 (-13.54, 0.34)	-4.23 (-9.55, 1.10)	-11.73* (-22.83, -0.64)	1.91 (-0.68, 4.77)	-weeks to decrease risk from 15% to 10%
20% β_{5b}	-19.24* (-34.37, -4.11)	-15.49* (-27.96, -3.03)	-11.02* (-20.29, -1.76)	-21.13* (-38.59, -3.67)	-0.42 (-3.41, 2.40)	-weeks to decrease risk from 20% to 10%

*: significant, 95% CI in parentheses

Table 15: Participants' ranks and utility scores of EBRT scenarios estimated with MXL model.

Case	Length (weeks)	Markers (no=0)	PSA (%)	Acute (%)	Late (%)	MXL Utility	Difference	Rank
Base	4	0	12	35	15	-6.966	0	11
One level improvement								
1)	**2**	0	12	35	15	-6.774	0.192	10
2)	4	0	**6**	35	15	-5.484	1.482	4
3)	4	0	12	**20**	15	-6.168	0.798	7
4)	4	0	12	35	**10**	-6.026	0.940	6
One level worsening								
5)	**8**	0	12	35	15	-7.350	-0.384	14
6)	4	**1**	12	35	15	-7.525	-0.559	15
7)	4	0	**18**	35	15	-8.448	-1.482	18
8)	4	0	12	**50**	15	-7.206	-0.240	13
9)	4	0	12	35	**20**	-7.906	-0.940	17
Mixed level change								
10) CFRT best	**8**	0	**6**	**20**	**10**	-4.130	2.836	3
11) CFRT worst	**8**	0	**18**	**50**	**20**	-10.012	-3.046	21
12) MH best	4	0	**6**	**20**	**10**	-3.746	3.22	1
13) MH worst	4	0	**18**	**50**	**20**	-9.628	-2.662	19
14) SBRT best	**2**	**1**	**6**	**20**	**10**	-4.113	2.853	2
15) SBRT worst	**2**	**1**	**18**	**50**	**20**	-9.995	-3.029	20
If higher risk acute toxicity with hypofractionation								
16) CFRT	**8**	0	12	**20**	15	-6.552	0.414	8
17) SBRT	**2**	**1**	12	**50**	15	-7.573	-0.607	16
If higher risk acute and late toxicity with hypofractionation but improved PSA								
18) CFRT	**8**	0	12	**20**	**10**	-5.612	1.354	5
19) HF	4	0	**6**	**50**	**20**	-6.664	0.302	9
20) SBRT	**2**	**1**	**6**	**50**	**20**	-7.031	-0.065	12

Using the MXL model, the base case was ranked 11 out of 21 hypothetical EBRT scenarios to treat localized prostate cancer (Table 15). When only one attribute level was improved or worsened and all other attribute levels held constant at the base case level, the PSA attribute caused the largest change in rank (cases 2 and 7). Change in the risk of late toxicity when all other levels were held constant caused the second largest change in rank (cases 4 and 9). This was in keeping with what was expected as these two attributes had the largest parameter estimates in the MXL model. When two attributes' levels were altered, such as Length and Acute, the utility difference and ranks of scenario 5 (an EBRT option that differed from the base case because it was 8 weeks instead of 4 weeks) and scenario 8 (risk of acute toxicity was 50% instead of 35% in base case) were compared. As can be seen in Table 15, the utility of scenario 5 was -7.350, and for scenario 8 was -7.206. Thus, to maximize utility most respondents would select scenario 8 over scenario 5. Therefore, in general based on the MXL model respondents would prefer to avoid an 8-week long EBRT associated with a 35% risk of acute toxicity and would prefer a shorter 4-week EBRT with a higher 50% risk of acute toxicity. This is also demonstrated by comparing the rank of the scenarios; case 8 was ranked 13th and case 5 was ranked 14th out of 21 scenarios.

These sample scenarios were used to construct theoretical clinical EBRT options. Stereotactic body radiotherapy (SBRT) is 2 weeks long and requires markers to be implanted, while moderate hypofractionation (MH; 4 weeks) and conventionally fractionated radiotherapy (CFRT; 8 weeks) regimens do not require markers. I found that the top ranked scenario was MH with the least amount possible of associated risk (case 12), second ranked was SBRT also with minimal risk (case 14), and third CFRT again with minimal toxicity (case 10). A similar order held when the three cases with the worse possible associated toxicity (highest attribute risk levels) were compared: (1) MH ranked 19th; (2) SBRT ranked 20th; and (3) CFRT ranked 21st. Thus, when toxicity was held constant respondents preferred MH over SBRT and CFRT, since MH does not involve marker implantation and is shorter than CFRT. If I postulate hypofractionation results in a higher risk of acute toxicity than CFRT, for example 35% vs. 20%, my model demonstrates respondents would prefer CFRT. Case 16 corresponds to this CFRT scenario with 20% risk of acute toxicity and was ranked 8 out of 21, which was better than the corresponding 11th out of 21 cases for the MH scenario with 35% risk of acute toxicity (base case). If late toxicity risk was higher for hypofractionated regimens this would clearly result in preference for CFRT since the Late attribute was more strongly avoided by respondents and impacted utility score more than the Acute attribute. I lowered the level for the PSA attribute and increased the level for the Late attribute in the MH and SBRT regimens (cases 18, 19, and 20 in Table 15) leading to cases with hypofractionation scenarios with lower chance of PSA recurrence and higher risk of late toxicity. This resulted in CFRT being preferred (rank 5 out of 21) above the MH (ranked 9th) and SBRT (ranked 12th) regimens.

Table 16: LCM ranking of hypothetical scenarios.

Case	LC 3.1: Toxicity		LC 3.2: Convenience		LC 3.3: Outcome	
	Difference	Rank	Difference	Rank	Difference	Rank
Best and worse scenarios						
10) CFRT best	4.589	1.5	0.042	6.5	4.488	2
11) CFRT worst	-1.347	7	0.042	6.5	-3.975	9.5
12) MH best	4.589	1.5	1.014	2	4.488	2
13) MH worst	-1.347	7	1.104	2	-3.975	9.5
14) SBRT best	4.044	3	0.121	4.5	4.488	2
15) SBRT worst	-3.169	11	-0.893	8.5	-6.710	11
If higher risk acute toxicity with hypofractionation						
16) CFRT	1.077	5	-0.972	10.5	0.789	7
17) SBRT	-1.802	9	-0.893	8.5	-0.433	8
If higher risk acute and late toxicity, and lower risk of PSA recurrence, with hypofractionation						
18) CFRT	3.312	4	-0.972	10.5	1.753	6
19) MH	-1.347	7	1.014	2	1.791	4.5
20) SBRT	-1.892	10	0.121	4.5	1.791	4.5

Attribute level details in Table 15.

The LCM was used to compare the three preference groups' ranking of the hypothetical CFRT, MH, and SBRT scenarios (Table 16). The Length and highest level of PSA risk (18%) parameter estimates for class 3.1 ("toxicity group") were not significantly different from zero, thus these terms were removed when eq. 8 was used to estimate utility for this class. The utility equation for class 3.2 ("convenience group") was only influenced by the parameter estimates for Length, Markers, and the middle level risk of the PSA attribute (12%). The parameter estimates for the risk attributes in class 3.3 ("efficacy group") were all significantly different from zero and thus included in the utility equation (Table 11). The toxicity group's most preferred scenarios were a tie between the MH and CFRT scenarios with the least associated risk. This was similar for the efficacy group except the first-place tie was between all three regimens since this class did not consider fiducial marker implantation in their utility calculation (Table 16). Interestingly, my model estimated the convenience group would most prefer MH no matter the associated risk levels (Table 16). Case 12 had the lowest levels for the three risk attributes, while case 13 had the highest level for all three. Since the convenience group significantly preferred to avoid Markers the next most preferred scenarios were SBRT without increased risk of acute toxicity (cases 14 and 20 in Table 16), then CFRT regimens (cases 10 and 11), followed by SBRT with increased risk of toxicity (cases 15 and 17), and lastly CFRT with increased risk of toxicity (cases 16 and 18).

Objective 3: Internal validity

All participants completed a debriefing questionnaire after the DCE that was used to assess internal validity (refer to Figure 5). The interviewer present for every DCE (myself) assessed how the participant completed the DCE (refer to Figure 6). These findings are reviewed and contrasted in the following sections outlined in greater detail in Appendix B; measurement validity, measurement reliability, choice validity, and choice reliability. As mentioned in chapter 3, the experimental DCE design included 24 choice sets divided into two versions so each participant answered 12 of these. Additionally, I generated six choice sets designed to test four aspects of internal validity and thus each participant answered 18 choice sets.

Measurement validity

Face validity was acceptable since the magnitude and direction of the parameter estimates were in keeping with what was expected *a priori*. Construct validity was measured with rating questions posed to the participants following the DCE to compare to the MNL, MXL, and LC model results. Overall the respondents rated the attributes PSA and Late highest on a 10-point scale of importance, followed by Length, Markers, and Acute (Table 17). The willingness to trade off was assessed by asking each respondent to distribute 100 importance points among the five attributes, resulting in the same rank order as the rating out of ten (Table 17). This ranking agreed with the MNL and MXL model findings for the two most important attributes (PSA and Late), however the rating exercises found Length was third most important (Table 17). Participants were asked what attributes were most important to their choices in the DCE and were able to list up to three each (Table 17). The largest proportion of participants reported the PSA and Late attributes were important to their decision-making (Table 17), which was in keeping with

Table 17: Mean ratings and points compared with ANOVA; attributes listed as important per class.

Attribute	Mean rating out of 10 (SD)	Mean importance points out of 100 (SD)	Attributes listed as most important to decision (%)
Length			
Overall	5.29 (3.17)	18.12 (21.22)	18/58 (31.03%)
Class 3.1	5.22 (3.04)	12.22 (13.20)	4/18 (22.22%)
Class 3.2	6.83 (3.07)	27.22 (23.72)	10/18 (55.56%)
Class 3.3	4.09 (2.93)	15.50 (22.75)	4/22 (18.18%)
One-way ANOVA	F= 4.124* p= 0.022	F= 2.667 p= 0.078	
Markers			
Overall	4.36 (3.29)	11.43 (18.51)	10/58 (17.24%)
Class 3.1	3.44 (2.53)	3.61 (4.47)	0/18
Class 3.2	6.50 (3.54)	26.11 (26.71)	9/18 (50.00%)
Class 3.3	3.36 (2.89)	5.82 (8.13)	1/22 (4.55%)
One-way ANOVA	F= 6.617* p= 0.003	F= 11.258* p< 0.001	
PSA			
Overall	7.70 (2.63)	36.64 (27.98)	36/58 (62.07%)
Class 3.1	6.94 (3.12)	30.28 (24.64)	11/18 (61.11%)
Class 3.2	6.72 (2.29)	18.06 (16.55)	5/18 (27.78%)
Class 3.3	8.86 (2.05)	57.05 (25.39)	20/22 (90.91%)
One-way ANOVA	F= 3.871* p= 0.027	F= 15.533* p< 0.001	
Acute			
Overall	4.29 (2.51)	7.81 (7.59)	8/58 (13.79%)
Class 3.1	4.94 (2.80)	11.22 (9.87)	4/18 (22.22%)
Class 3.2	4.11 (2.54)	7.22 (5.75)	1/18 (5.56%)
Class 3.3	3.91 (2.24)	5.50 (5.92)	3/22 (13.64%)
One-way ANOVA	F= 0.906 p= 0.410	F= 3.102 p= 0.053	
Late			
Overall	7.12 (2.57)	25.52 (22.78)	32/58 (55.17%)
Class 3.1	8.78 (1.77)	40.83 (26.64)	15/18 (83.33%)
Class 3.2	6.72 (2.78)	21.67 (20.15)	7/18 (38.89%)
Class 3.3	6.09 (2.35)	16.14 (14.14)	10/22 (45.45%)
One-way ANOVA	F= 6.919* p= 0.002	F= 7.630* p= 0.001	

*: significant at 0.05 level

the models. However, Length was listed as third most important overall which was consistent with the rating and allocation of points responses (Table 17). This demonstrated preference heterogeneity was present and best modeled with LC analysis since the class apportionment corresponded closely to the rating, allocated points, and listed important attributes (Table 17). For example, class 3.3's (efficacy group) mean rating and allocated points to PSA was significantly higher than the overall mean, while the scores for the other four attributes were lower than the overall mean (Table 17). For all three classes the findings of the debriefing survey were in keeping with the preferences estimated by the LCM.

The responses to the qualitative question ('why is one or more of the characteristics most important to you?' in Figure 5) were examined with content analysis to elicit categories of participants' preferences (88). Each category was either positive (preferred longer EBRT), neutral (did not strongly prefer the attribute), or negative (preferred shorter EBRT, no markers, and less risk of PSA recurrence and toxicity) (Table 18). A negative category indicated the preference would influence the associated attribute's parameter estimate negatively, for example preferring less risk of late toxicity (Table 18). No respondents preferred higher risk of PSA recurrence or toxicity. Neutral categories emerged for all three risk categories indicating no preference for that attribute (Table 18). Interestingly, for the Length attribute there were positive, negative, and neutral categories. Some respondents were not concerned about the length of EBRT (neutral), and as expected several negative categories emerged (Table 18). One respondent believed longer EBRT would impact his health negatively by increasing what he termed his 'radiation exposure' despite the interviewer's efforts to explain all three EBRT regimens resulted in similar biologically effective doses (Table 18). Further, there were three respondents who were members of the positive category 'shorter EBRT not ideal treatment' where two respondents did not believe SBRT of 2 weeks would be sufficient treatment, and one respondent who had heard of MH before and wanted to avoid both the MH and SBRT regimens (Table 18).

Face and construct validity were satisfactory as the overall and by class rating and ranking of the attributes, the association of patient factors and preferences, and participants' reasoning were consistent with prior expectations. While similar results from the DCE and rating exercises were obtained, the DCE allowed a more in-depth analysis of how participants traded off between the attributes and levels.

Table 18: Identified categories with representative quotes to illustrate the respondents' reasons for preference.

Attribute	Categories	Quotes (participant ID)
Length	**Positive (preferred longer EBRT):**	
	Shorter EBRT not ideal treatment	"don't trust the 2 weeks, how could the treatments be equivalent" (1032) "I think 8 weeks is less radiation" (1015) "due to my knowledge of radiation I want to avoid higher dose per treatment" (1008)
	Neutral:	
	Do not mind longer EBRT	"time frame not a negative experience, not a hassle to come here as I live 15 minutes away" (1024) "already got through 7 ½ weeks of treatment" (1026)
	Negative (preferred shorter EBRT):	
	Long EBRT is inconvenient	"drive in myself from out of town so want to reduce that time" (1020) "8 weeks is a nuisance" (1028) "stayed at hotel during treatment, disruptive to be away from home" (1040) "get back to a normal life" (1048) "live far away so it's a long drive" (1054)
	Long EBRT results in worse quality of life	"really don't like 8 weeks, if I went that long I would be too tired" (1014) "8 weeks is long, any stress or psychological discomfort will be better when treatment is over" (1027)
	Radiation exposure	"2-week option best, otherwise risk of radiation exposure over a longer time" (1043)
Markers	**Neutral:**	
	Do not mind markers	"markers is quick in a day" (1021)
	Negative (avoid markers):	
	Avoid invasive procedure	"Not big on invasive procedure" (1011) "avoid procedure as I had an infection after the biopsy that was terrible" (1014) "most important to me is least invasive as possible" (1037) "most important to avoid complications" (1034) "don't like the idea of implants, more invasive" (1055)
PSA	**Neutral:**	
	Further treatment is available	"if PSA goes high there is another treatment" (1017) "PSA is just a number, doesn't mean you will die tomorrow" (1005)

		"PSA can be dealt with" (1047)
	Negative (preferred less risk): Indicator of oncological outcome	"if PSA doesn't rise it means I am doing well" (1014) "if numbers are down it means keeping the cancer away- what treatment is all about" (1018) "PSA going down after treatment felt like an accomplishment to me" (1025) "it tells you if the cancer is still active" (1029) "PSA is signal prostate cancer may be back" (1041) "it is my ailment why else am I getting treatment" (1042)
	Do not want further treatment	"hormones are awful" (1013)
Acute	**Neutral:** Toxicity will resolve	"short term side effects will go away" (1010) "short term side effects just an inconvenience" (1013) "can tolerate in the short term" (1039)
	Negative (preferred less risk): High overall risk	"minimize side effects and short-term side effects since higher risk numbers" (1033)
Late	**Neutral:** Low overall risk	"15% and 20% not a big difference" (1028) "all within that 15% range so not as important" (1052) "bigger difference between 50% and 20% for short term than 15% and 20% in long term" (1056)
	Side effects tolerable	"at my age long term side effects not important to me, if I was younger may be different" (1040) "PSA worries me the most, I can deal with the side effects" (1045) "I can live with the side effects" (1058)
	Negative (preferred less risk): Quality of life	"Quality of life long term most important" (1009) "Long term could be more of a problem" (1011) "don't want nothing down the road" (1019) "don't want to live with it for the rest of my life" (1022) "want to be back to normal activities" (1005) "living with side effects long term will bother me the most" (1035) "long term most important to me, I don't want to spoil my retirement" (1037) "quality of life most important to me" (1038)

Measurement reliability

To determine if this DCE was able to capture participants' preferences reliably tests for internal consistency were analysed. The participants all repeated the third choice set from the experimental design and the observed overall agreement was 84.48%. There were nine participants who answered the repeat question differently, for example selected alternative A first and on repeat chose alternative B. The expected agreement was at least 50% due to chance alone. The κ statistic was calculated as 0.69 which indicated substantial agreement. The participants were not informed that a repeat choice set was present, and none voiced the concern that there was a repeat. Based on the placement of these two choice sets (fourth overall and 14th) I do not believe any participant was cognizant of the repeat choice set. For the nine participants who failed the test-retest the mean self-reported confusion score of 2.33 (SD 0.71) and the interviewer-reported mean of 2.22 (SD 1.39) was higher than the overall average for participants. The interviewer tended to agree with the participant reported scores and were within one point on the 5-point scale of each other except for two participants (ID: 1013 and 1044). Participant 1013 reported a confusion score of 3 (corresponding to 'somewhat confused') while the interviewer assessed them as 'not confused' (score of 1), and participant 1044 reported a score of 3 while the interviewer scored them a 5 (corresponding to 'very confused').

To test for version consistency a holdout choice set was included as part of the extra six validity check choice sets. In block 1 the holdout was the 8th choice set from the experimental design of block 2. In block 2, the holdout was the 9th choice set from the experimental design of block 1. This ensured every respondent answered two of the same choice sets no matter which version they were assigned to complete. For the block 1 holdout alternative A had a utility of -2.763 and alternative B had a utility of -4.779 based on the MXL model. Block 1 participants' selections for this choice set were not part of the model estimation, and instead their choices were examined to determine if the alternative with the less negative utility was chosen more often (alternative A should have been selected by more participants). In block 1 25 participants chose alternative A, compared to 19 in block 2, out of a total of 29 participants in each block. As would be expected, participants tended to prefer alternative A and the observed inter-rater agreement was 79.31% between the two versions, while the probability of expected random agreement for alternative A was 56.48%. The version consistency was determined with the κ statistic as discussed in chapter 3 and found moderate inter-rater agreement of 0.52. The block 2 holdout alternative A had a utility of -3.508, and alternative B had a utility of -3.331. Fifteen participants in block 1 selected A, and in block 2 nine participants selected A. The observed inter-rater agreement was again 79.31% between the two blocks, while the probability of expected random agreement for alternative A was only 16.05%. The version consistency was substantial with a κ agreement of 0.75. Therefore, there was moderate to substantial agreement of the holdout choice sets, and substantial agreement demonstrating reasonable measurement reliability was present in this DCE.

Choice: Internal validity

DCE choices would not be valid or reliable unless respondents comprehended what was being asked of them. The participants' and interviewer-reported scores for comprehension and difficulty completing the DCE were tested for inter-rater agreement. Self-reported mean confusion was lower (1.93, SD 1.10) than interviewer assessed participant confusion (mean 2.12, SD 1.36) on a 5-point Likert-scale where 1 corresponded to 'not confused', 2 to 'mildly confused', and 5 to 'very confused' (Table 19). The fair overall agreement (κ= 0.322) is visualized in Figure 13 as interviewer assessment was usually within one or two points on the 5-point scale of the self-reported confusion score. The 11 (18.97%) participants assessed by the interviewer to be moderately or very confused tended to make choices in the DCE in keeping with the self-reported attributes most important to them. One respondent (ID: 1012) rated Length, PSA, and Acute as most important to him (all 10 on a 10-point scale) which spans the three preference types identified by the LCM (toxicity, convenience, and efficacy). This participant was classified as class3.2 (the "convenience group"). The other 10 respondents reported attributes most important to their

Table 19: Tabulation of participant- and interviewer- reported confusion on a 5-point scale.

	Interviewer					
Participant	1	2	3	4	5	Total
1	19	1	2	1	0	23
2	9	6	5	1	1	22
3	1	2	2	3	1	9
4	0	0	0	1	1	2
5	0	0	0	0	2	2
Total	29	9	9	6	5	58
Mean (SD)	Participant-reported: 1.93 (1.01)			Interviewer-reported: 2.12 (1.36)		
Inter-rater agreement	κ: 0.322 (p< .001)					

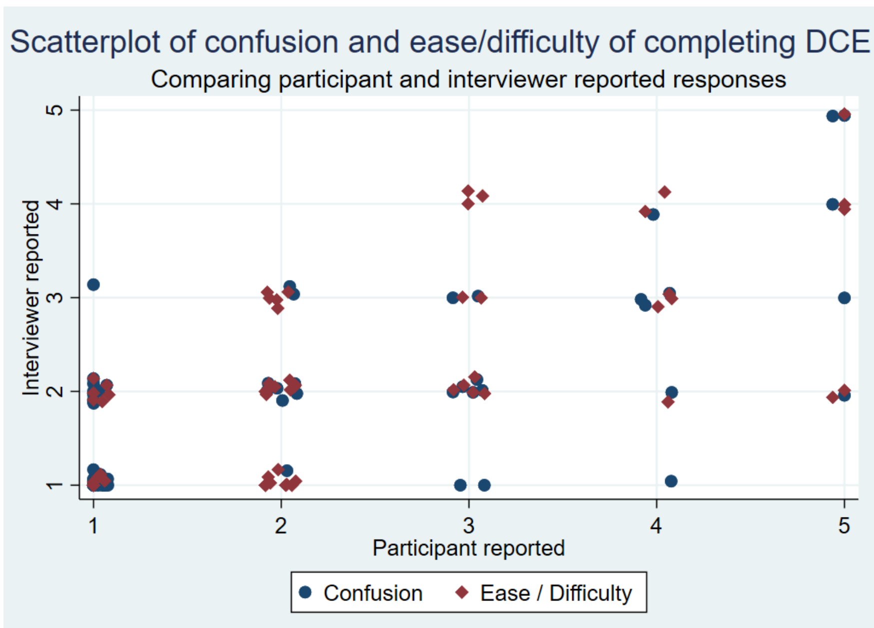


Figure 13: Participant- and interviewer-reported confusion and difficulty completing DCE.

choices were consistent with their membership class. This suggests in general the participants understood the task as their vocalized preference matched with the LC they were assessed to belong by the model.

Respondents were asked how easy or difficult it was to choose between the alternatives on a 5-point scale (where 1 corresponded to 'very easy', 3 to 'neutral', and 5 to 'very difficult'), and the interviewer assessed the respondents with the same metric. The mean self-reported score was 2.17 (SD 1.05), score of 2 corresponded to 'easy', and the interviewer agreed with this finding with a mean score of

Table 20: Tabulation of participant- and interviewer-reported ease or difficulty completing DCE.

	Interviewer-reported ease or difficulty					
Participant-reported	1	2	3	4	5	Total
1	8	9	0	0	0	17
2	6	9	5	1	2	23
3	0	5	2	3	0	10
4	0	0	3	2	2	7
5	0	0	0	0	1	1
Total	14	23	10	6	5	58
Mean (SD)	Participant-reported: 2.17 (1.05)			Interviewer-reported: 2.40 (1.21)		
Inter-rater agreement	κ: 0.148 (p= .023)					

2.40 (SD 1.21) (Table 20, Figure 13). The interviewer reported 47 of the 58 (81.03%) respondents seemed to know their preferences from the beginning of the DCE and did not change their decision-making strategy throughout the DCE. The interviewer reported 16 (27.59%) respondents found the DCE easier to complete as they moved through the choice sets, always due to learning and not a simplifying strategy. Overall there was slight agreement (κ= 0.148) between the participant- and interviewer- reported scores (Table 20, Figure 13).

The mean time it took for participants to complete the DCE was 2.40, scored by the interviewer on a 3-point scale where 1 was ‘quicker than average’, 2 ‘average’, and 3 corresponded to ‘slower than average’. The χ^2 Test for Independence values were insignificant ($p > 0.05$) indicating no association between time and participant- or interviewer-reported confusion. The interviewer considered this to be due to two reasons: (1) confused participants slowed down to try to increase their understanding; or (2) confused participants quickly went through the choice sets to finish sooner. Four of the five respondents assessed by the interviewer to have ‘great difficulty’ (score of 5 on a 5-point scale) making choices were ‘slower than average’, while no respondents who found the DCE ‘very easy’ or ‘easy’ were. This association was significant ($p < 0.001$) thus participants who had greater difficulty with the tasks took longer to complete the DCE.

Choice validity

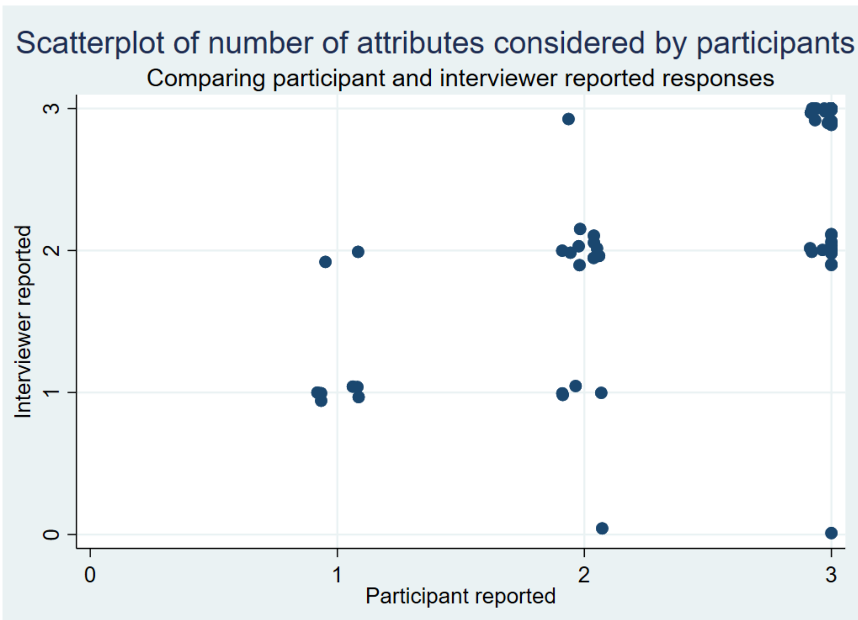
The concept of choice validity was explored by assessing rational decision-making with dominant choice sets and exploring if task-nonattendance was present in the participants responses. Two dominant choice tasks (all attribute levels for alternative A were better than alternative B) were assessed to determine if participants were making choices “as expected”. No respondents selected the non-dominant alternative, therefore 100% of respondents passed this basic rationality test of monotonicity. There were three checks for task-nonattendance: (1) pattern between the selected alternatives; (2) understanding; and (3) number of attributes considered by respondents. There was no discernable pattern of responses to the DCE, for example no respondents always chose A over B. The majority (57 out of 58; 98.28%) of participants reported they understood the concept of choosing between the alternatives. The one participant (ID: 1016) who reported not understanding listed PSA as the most important attribute to him, and he was classified as most likely to belong to class 3.3 (the “efficacy group”). By this measure his choices in the DCE led him to be correctly categorized by the LCM, and likely his choices were valid. The interviewer reported 2 out of 58 (3.45%) participants (ID: 1012 and 1015) had not understood the concept of making choices in the DCE. Both participants reported being ‘very confused’ (corresponding to a score of 5 on a 5-point scale) and the interviewer agreed with this assessment. Both participants were classified as belonging to class 3.2 (“convenience group”), and both participants listed Length as one of the most important attributes.

There was moderate inter-rater agreement ($\kappa= 0.434$) for assessment of number of attributes the participants considered to make their choices (Table 21). As seen in Figure 14 the largest discrepancy was for two participants who were unsure of how many attributes they considered, though the interviewer found they considered more than one attribute. The interviewer reported 24 respondents considered less than five attributes (one to four), and thus utilized attribute non-attendance as a simplifying heuristic to make their choices in the DCE. Eight out of 58 (13.79%) participants considered one attribute when making their choices. These eight, and two other participants assessed by the interviewer, did not trade

Table 21: Tabulation of participant- and interviewer-reported number of attributes considered.

	Interviewer-reported number of attributes considered			
Participant-reported	1	2	3	Total
0	0	1	1	2
1	6	4	0	10
2	2	10	12	24
3	0	1	21	22
Total	8	16	34	58
Inter-rater agreement	κ: 0.434 (p< .001)			

0: unsure; 1: Focused on 1 attribute; 2: considered 2-4 attributes; 3: considered all 5 attributes.



0: unsure; 1: Focused on 1 attribute; 2: considered 2-4 attributes; 3: considered all 5 attributes.

Figure 14: Comparison of number of attributes considered.

between attributes and thus in total 10 (17.24%) participants did not use compensatory preferences to make choices. The mean self-reported confusion score (1.70; SD 0.82) for these 10 participants was lower than the sample wide mean, and the interviewer reported a similar mean to the sample (mean 2.10, SD 1.20). In summary, some task non-attendance was present though the responses were likely valid as these participants' quantified preferences in the LC analysis agreed with the groups they would have belonged to based on their listed most important attributes, and non-compensatory decision makers did not have worse comprehension than other responders.

Choice reliability

Transitivity was assessed via three related choice sets to determine the preference relation between three choice profiles. Among the participants 55 out of 58 (94.83%) answered all three choice sets upholding the transitive principle of decision theory. The three participants (ID: 1015, 1020, and 1055) who did not uphold this principle had varying levels of confusion and difficulty making their choices in the DCE. All three participants listed attributes most important to their choices accurately corresponded to the class they belonged to estimated by the LC model. For example, participant 1020 identified Length as the attribute most important to him and was classified in the convenience group (class 3.2). Thus, the principle of transitivity was upheld by most participants and though three participants violated this principle their choices still appeared to be valid since their listed attributes important to their decisions corresponded to their class membership in the LC model.

The attributes participants listed as most important to their choices was assessed for consistency with the class the participant was allocated to based on the LC analysis (Table 15 final column). This finding generally held for participants assessed by themselves or the interviewer to have not understood the task or did not trade between attributes. For example, within class 3.2 three participants (ID: 1010, 1012, 1033) did not list Length or Markers as important to their decision making. Two of these participants (ID: 1010 and 1033) rated the Length attribute an 8 on a 10-point scale and the interviewer assessed that these two patients engaged in the task and traded between attributes. Participant 1012 was assessed by himself and the interviewer to be ‘very confused’ (score of 5 on a 5-point scale) and found the DCE ‘difficult’ (score of 4 on a 5-point scale). This participant rated the Length, PSA, and Late attributes 10 on a 10-point scale. Likely these three participants favoured shorter length during the DCE but failed to mention this attribute when asked to list attributes important to their decision making. This pattern carried through to the other two classes; five participants in class 3.1 did not list the Acute or Late attributes but rated one or both highly out of 10. In summary, most participants' choices in the DCE corresponded to the attributes they listed as important to their decision making in the debriefing survey, and for the few participants whose choices did not when examined more closely most of these participants rated these attributes highly on a 10-point scale of importance.

Next steps

Chapter 5 starts with a summary of the findings discussed in this chapter, followed by the main limitations of this study, and finally implications for future research and my conclusion.

References

1. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health--a checklist: a report of the ISPOR good research practices for conjoint analysis task force. Value Health. 2011 Jun;14(4):403-13.
2. de Bekker-Grob EW, Niers EJ, van Lanschot JJ, Steyerberg EW, Wijnhoven BP. Patients' preferences for surgical management of esophageal cancer: a discrete choice experiment. World J Surg. 2015 Oct;39(10):2492-9.
3. Pacifico D, Yoo HI. lclogit: a Stata module for estimating latent class conditional logit models via the expectation-maximization algorithm [Internet]. School of Economics, The University of New South Wales; 2012. Available from: https://ssrn.com/abstract=2174146
4. Greene WH, Hensher DA. A latent class model for discrete choice analysis: contrasts with mixed logit. Transport Res B-Meth. 2003 Sep;37(1):681-98.

Chapter 5

General Discussion and Conclusion

Summary of findings

Objective 1

In the overall MNL and MXL models each attribute's parameter estimate was significantly different from zero, thus the null hypotheses were rejected (Table 7). As hypothesized *a priori*, the part-worth estimates for each attribute were negative because respondents preferred treatment alternatives that had a lower risk of not being effective, a lower risk of side effects, and preferred not to have a procedure to implant fiducial markers, along with a shorter treatment time (Table 7). *A priori*, it was expected that preference heterogeneity would exist within the sample, which was proven by demonstrating that a portion of respondents had preferences opposite to what was expected (Table 11). The potential association of preference for certain attributes and respondent characteristics was explored via several analysis methods. For example, the subgroup of respondents 70 years and older significantly preferred shorter EBRT in the MNL analysis, while this attribute was of unknown significance in the younger subgroup (Table 8). Additionally, the LC analysis identified three classes of respondents, one of which significantly preferred shorter length of EBRT (Table 11). There was a trend for this class to be 70 years or older (Table 13). The class of respondents who most strongly preferred to reduce risk of PSA recurrence was significantly more likely to be working at the time of the DCE than the other two classes identified in the model (Tables 11 and 13). The attribute levels' parameter estimates that excluded zero impacted participants' choices in the DCE, but to determine the clinical significance of these attributes of EBRT for prostate cancer the trade-offs participants are willing to make should be estimated.

Objective 2

MRS estimates were used to quantify the trade-offs respondents were willing to make for shorter EBRT, for example to accept fiducial marker implantation, or to accept increased levels of the risk attributes (Table 14). Most of the MRS values were negative, demonstrating that individuals were willing to increase length of EBRT to avoid fiducial markers and risk of worse efficacy or toxicity. Insignificant values, some of which were positive, suggested that those respondents were not willing to increase length of EBRT to avoid markers or reduce risk. However, the 95% CI were wide, especially for the latent class 3.2 (the "convenience group) and the subgroup analyses, likely since these MRS estimates were underpowered. A willingness to shorten EBRT by 4 weeks corresponded to a change from CFRT (8 weeks) to MH (4 weeks) and SBRT (2 weeks), and a willingness to shorten EBRT by 6 weeks corresponded to a willingness to accept SBRT (2 weeks) over CFRT. Any MRS values larger than 6 weeks were not clinically meaningful as the model only accommodated for up to a 6-week difference, and values larger than 8 weeks are not clinically possible. About a third of respondents belonged to class 3.2 ("convenience group") and were willing to increase EBRT by almost 6 weeks to avoid markers (Table 14). This corresponded to preferring CFRT (8 weeks) to SBRT (2 weeks) or MH (4 weeks). This class, like all the other models, were willing to trade-off longer EBRT to reduce risk of PSA recurrence from 12% to 6%, however there may be an issue with this model since the MRS value to reduce risk of PSA recurrence from 18% to 6% was not significant (Table 14). This class was not willing to lengthen EBRT to reduce risk of toxicity at any level within the model (Table 14). The MRS values for the other models were either larger than feasible for the model, indicating respondents were willing to give up many more weeks than even clinically possible to reduce risk, or were not significant. For example, the MRS Markers value for the MXL model was not significant, thus conclusions about participants' willingness to lengthen EBRT to avoid fiducial marker implantation cannot be made based on this model, possibly due to the lack of power.

The preferred rank order of theoretical EBRT scenarios based on the MXL and LC models was assessed. The MXL analysis found when toxicity was held constant respondents preferred MH over both SBRT and CFRT since this regimen maximized utility by being the middle level for Length and not involving marker implantation (Table 15). Class 3.1 (the "toxicity group") preferred CFRT and MH, since SBRT requires markers, with toxicity set at the lowest Acute and Late attribute levels (Table 15). Class 3.2 (the "convenience group") members preferred MH scenarios regardless of the level of toxicity (Table 16). Class 3.3 (the "efficacy group") most strongly preferred scenarios that minimized risk of PSA recurrence, and the Length attribute was of unknown significance, thus the top three scenarios were the ones with the least risk of PSA recurrence and toxicity (Table 16).

As discussed in chapter 2, the RCTs found the efficacy of MH and CFRT to be equivalent with the possibility of up to 6% to 10% increased risk of acute and late GI and GU toxicity (1-3). Further SBRT RCT data are required, though one ultra-fractionated RCT found efficacy was similar between SBRT and CFRT with an increased risk of 4% to 5% toxicity (4). Thus, based on the MRS values and scenarios I conclude that prostate cancer patients with localized disease who have similar characteristics to class 3.2 (trend for this class to be 70 years and older) prefer MH over CFRT.

Objective 3

The results of the DCE are valid since most respondents passed the internal validity tests. All respondents passed the test of rationality, there was moderate to substantial agreement for internal and version consistency, and 55 of the 58 respondents upheld the principle of transitivity. A post-DCE questionnaire was completed by participants and the interviewer to examine construct validity. Self- and interviewer-reported scores for comprehension, ease or difficulty completing the DCE, and how participants made their choices in the DCE had variable inter-rater agreeability (Tables 18, 19, and 20). The respondents who the interviewer deemed moderately or very confused (11 of 58; 18.97%), or not to understand the attributes and levels (n=2, 3.45%), made choices in the DCE that were consistent with the preferences they expressed in the debriefing survey. Thus, it can be concluded that the model findings and estimates are accurate representations of participants' stated preferences. This was confirmed by comparing the respondents' explicit ratings and allocation of importance points to the attributes by class membership assigned by the LCM (Table 17). These ratings were not used to estimate the LCM, and therefore consistent findings between the rating scale and LCM helped validate the LCM findings. As expected, class 3.2 ("convenience group") members rated the Length and Markers attributes significantly higher on average than the other two classes, significantly allocated more points to the Markers attribute, and there was a trend to allocate more points to the Length attribute (Table 17). This was significantly demonstrated for class 3.1 ("toxicity group") for the Late attribute, and class 3.3 ("efficacy group") for the PSA attribute (Table 17).

Face validity was assessed by identifying that the magnitude and direction of the parameter estimates for the attributes were consistent with my hypotheses, and through qualitative methods that can be compared to a researcher's intuition of whether the preference data made sense (5, 6). Overall the qualitative content analysis provided reasonable explanations for why the 'best' levels of the attributes were preferred, for example to improve long term quality of life by minimizing risk of late toxicity (Table 18). This analysis also offered intriguing reasoning for participants' preferences. Interestingly there was only a 'positive' category for the Late attribute, indicating that some participants reported a specific reason to prefer increased length of EBRT (Table 18). These respondents' reasoning was either that they did not think two weeks was enough treatment, or, for one respondent, knowing hypofractionation resulted in higher doses of radiotherapy per fraction led to a preference to avoid this (Table 18). There were neutral attributes for all five attributes indicating that some respondents had no preference for the specified attribute (Table 18). Therefore, face validity was adequate as the rating and explanation of preference for the attributes was largely consistent with *a priori* expectations.

Study strengths and limitations

The strengths of this study are first summarized, and then the potential limitations of this study specific to DCEs discussed in chapter 2 are elucidated in this section.

Strengths

Findings from this research add new knowledge, understanding, and perspectives about the preferences for radiotherapy to treat localized prostate cancer. The research context aligns with the emerging priority for patient-oriented research, and patient-centred care where input about clinically meaningful topics is sought from patients and included in treatment decisions. The DCE included multiple tests of internal validity, and results from the DCE were confirmed in the subsequent rating and qualitative questions. The understanding by participants and validity of the results were enhanced by utilizing in person rather than the more common computer assisted surveys. This is the first DCE to include fiducial markers as an attribute, and the first radiotherapy alternative specific DCE for any cancer site. This work may be meaningful to Radiation Oncologists to facilitate their discussions with patients regarding concerns patients have about radiotherapy, and how CFRT, MH, or SBRT would best fit an individual patient's preferences with the counseling of their physician that standard of care options include both MH and CFRT.

Limitations

Interaction effects

This study utilized a main effects design that assumed all attributes were value-independent of each other (all interactions between attributes were zero). This was reasonable since main effects typically account for 70-90% of the explained variance in DCEs, though specific combinations of attribute levels may have specific effects that remained unidentified (7). An interaction term between the attributes Length and Acute was included in the design of the pilot study, and as reported in chapter 4, the interaction term was not significant, and participants stated they did not associate the acute toxicity varying with length of EBRT. To ensure this continued into the main study, all participants were instructed that acute toxicity takes weeks to 3 months to resolve after the completion of EBRT. Additionally, if interaction terms are estimated more degrees of freedom are required resulting in larger designs and sample sizes for adequate statistical significance and power, which was beyond the scope of this thesis. Therefore, interaction terms were not accounted for at the design stage and since the study was not powered to detect these interaction effects, they were not tested for at the analysis stage.

Level recoding

It is critical for researchers to know how participants interpret the attributes and levels and ultimately make their decision (8). Respondents may have thought of the presented risk numbers as ‘low’, ‘medium’, and ‘high’ categories, known as level recoding (5). I did not assess the extent respondents processed the actual risk numbers, though this would only be important for continuous risk attributes since the parameter estimates were uniquely estimated for each risk level for the PSA, Late, and Acute attributes. The Length attribute levels were not assessed for level recoding and this is not a limitation of my study for two reasons; (1) the levels were absolute numbers (length in weeks) and not risk which is more difficult for individuals to understand; and (2) the attribute levels were found to be significantly linearly related.

Cognitive discordance

Patients, irrespective of their type of experience treatment, may ‘defend’ their own treatment (the MH 4 week and the 8-week CFRT regimens were both offered at the institution in this study). These phenomena are known as status quo bias or cognitive discordance and can bias results (9). To reduce this bias, preferences would have to be elicited from prostate cancer patients with localized disease who had underwent EBRT and those treated differently, for example surgery or brachytherapy. Due to the complexity of EBRT and the scope of this study, I decided to only recruit prostate cancer patients experienced with EBRT, with the acceptance that newly diagnosed prostate cancer patients who had not yet been treated may have different preferences than my study sample.

External validity

Multiple tests of internal validity were undertaken and reported in chapters 4 and 5, however for DCEs the best test of validity is one that demonstrates external validity; whether behaviour stated in a hypothetical context is revealed (carried through) in choices made in the real world. External validity is the strongest form of evidence for validity in DCEs, and yet it is the most difficult to demonstrate. Respondents with non-rational or non-compensatory (did not trade off between the attribute levels) choices could have their preference responses removed from the data set. This would improve internal validity but reduce external validity (6, 10). Therefore, in this study no responses were removed.

The hypothetical bias occurs when researchers assume participants' choices reflect their true underlying preferences if faced with an actual decision (11). To increase the likelihood that a DCE is capturing participants' latent preferences, participants should believe the surveys have meaning (12). Furthermore, it was important to match the study sample to the larger population of interest as I did in this DCE since the likelihood of genuine responses increases when participants are familiar with the presented scenarios (13). In this study participants were asked to deliberate carefully and participants were assured no one else besides the interviewer would know their answers (anonymized data) in an attempt to avoid participants from selecting answers they may have thought they were 'supposed' to choose (11).

External validity should be tested against a gold standard, and in the case of stated preferences research, a revealed preferences study would be needed. However, this is difficult to investigate in Canada's publicly funded health care system where patients have limited choice and do not pay for the services. In non-health care related DCEs there is evidence of external validity (14). Imaginative solutions to this problem are framed field (in this case would take place in the clinical setting with subjects knowing they are part of an experiment) or natural field (without subjects knowing they are part of the experiment) experiments (15). The latter is the stronger external validity test because in a framed field experiment an individual's desire to make a choice they believe the researcher wants may influence their behaviour (16). There are ethical concerns about a natural field experiment for this study, and a framed field experiment could not be undertaken since all participants had already completed EBRT. The choices the respondents in this DCE had to make were realistic, yet real life decisions had already been made. It is possible that preferences presented here differ from those of patients who are EBRT naïve. The current study does not allow stated preferences to be validated by revealed preferences from actual behaviour. This interesting comparison could be an area for further research, where this DCE is repeated prospectively in treatment naïve prostate cancer patients with localized disease who had agreed to be treated with EBRT but had not yet started treatment, and then followed by asking the participants to select a treatment to receive in the real world, with the participants knowing it was an experiment. A prospective design study like this would also prevent previous experience from interfering with patients' choices.

Non-compensatory decision making

Compensatory decision making is required in DCEs to measure trade-offs, and occurs when respondents trade between the attribute levels and do not always select the alternative with the 'best' level for their most preferred attribute (10). The opposite of this concept, called non-compensatory decision making, occurred in this DCE for each attribute except Acute, as 10 of 58 (17.24%) respondents always selected the 'best' level of the attribute they dominantly preferred. When respondents did not trade between attributes the MRS could not be estimated from their data (10). However, these 10 respondents could have used simplifying heuristics to make their choices. This is a valid method in DCEs if the participants understood the attributes and levels, and considered all attributes and decided that some were unimportant regardless of level (11). Several studies have reported separate analyses to investigate the implication of retaining these participants with dominant preferences, and found the findings were not significantly different (10). Furthermore, deleting these respondents' choices may result in the removal of valid preferences, lead to selection bias, and reduce the statistical efficiency and power of the estimated choice models. Therefore, it is recommended no data be removed from the statistical analysis and my study followed this recommendation (10).

Selection bias

In the literature there is indication that DCEs probably cause some extra selection bias compared with general questionnaire surveys due to the complexity of a DCE task (8). Given the complexity of a DCE task, selection of respondents by education level or language skill is likely. Four of the 24 non-responders stopped the DCE at the first choice set after hearing the explanation and background of the study (and signing consent) and indicated they did not understand, found it too complicated, and no longer wanted to take part in the study. This is a small proportion (16.7%), however some of the other non-responders may have declined due to not wanting to take on the complexity of a preference survey after the initial basic explanation, though education level of non-responders was not collected. This type of selection bias was limited in this study by keeping the DCE survey simple with clear language, low number of overall attributes and levels, and taking numeracy issues into account by providing diagrams for the risk values. Additionally, this design was piloted with six individuals who found the DCE not confusing or only mildly confusing, and all passed the rationality (dominance) tests.

A strength of my study was the good overall response rate of 70.7% (58 of 82 individuals). In the reviewed oncology DCEs in chapter 2, of the three studies that enrolled patients in the clinic and followed up by mail, the response rate varied from 39% to 73% (17-19). The two subgroups', age and distance from the cancer centre, response rate varied to some extent. Individuals younger than 70 years old had a response rate of 66.7%, and those 70 years or older had a response rate of 73.1%. The response rate in individuals living closer than a 60-minute drive to the cancer centre was 68.0%, while it was 75.0% for individuals living further than a 60-minute drive. My results suggest participants who are older and/or live a 60-minute drive or longer from the cancer centre prefer shorter length of EBRT, while this attribute was not significant for younger participants. This may have led to selection bias and therefore a more positive attitude towards preferring reduced length of EBRT in the large group models (for example MXL risk categorical). I stratified the data based on both covariates which should eliminate this bias in the subgroup analyses, plus the difference in response rate between these subgroups was small. The response and non-response rates by age and distance were similar, however a limitation of my study is there may have been a non-response bias based on systematic differences between responders and non-responders in a sociodemographic or prostate cancer specific factor that was not collected.

Framing

The direction of framing of the risk levels may have influenced the results. The framing used to present risk has an influence on respondents' choices, though there is a lack of evidence supporting if positive or negative framing is best (20). The risk levels were framed positively to ensure similarity to participants' prior experience since clinicians usually explain risk to patients as the chance the event would happen to them, and not the risk it would not. This bias was limited by presenting the probability of the non-event underneath the diagram describing the risk level (see Figure 4).

Confounding

There is the possibility of an unobserved confounder that correlates with a covariate and impacts choice. If present, this would result in making incorrect inferences about participants' choices that really is not about the covariate at all but is related to this unobserved confounder. Unobserved confounders cause problems in modeling and a model's predictions depend on including all relevant variables. For example, in the LC modeling no relationship between the distance respondents lived from the cancer centre and class membership was found. This may be due to an unobserved confounder, for example that participants who lived a certain distance from the cancer centre qualified for subsidized stays in a hotel near the cancer centre on weekdays. This could compensate for living further away, and therefore the distance respondents live from the cancer centre would not impact the relationship to preference for length of EBRT, rather the ability to access a hotel does. Though this is one explanation, this does not explain the difference in preference found by stratifying the two subgroups of those who lived less than and greater than a 60 minute drive from the cancer centre (Table 9), and living in a hotel while undergoing cancer treatment is inconvenient and one would assume participants would want to minimize length of EBRT.

Another concern with confounding, is in the regression models the preference weights were confounded with error variance, which would vary in scale between models (21). Therefore, to compare the different models only the rank of the attributes or the ratio between two attributes (MRS) could be likened, and this was carefully carried out in chapter 4.

Model assumptions

There are three main assumptions of the MNL model that cause limitations for my DCE: (1) homoscedastic error terms not accounting for unobserved variations between respondents; (2) homogeneous preferences though this is unlikely within any sample and heterogenous preferences was demonstrated in my study; and (3) the panel data of the data is not accounted for (21). MXL is a flexible model that relaxes these assumptions of MNL and is considered an improved fit for human choice behaviour. However, models that account for scale heterogeneity are preferred to MXL as the normal distribution may be mis-specified (22). If scale heterogeneity was present in my study then much of the heterogeneity in attribute weights would be accounted for by a scale effect that varied across individuals due to differing strength of preferences and certainty of choices. This more advanced model could have been applied to my study, however since the LCM was planned to determine classes of participants based on preference heterogeneity, I believed this scale-MNL model was unnecessary. Furthermore, I proved that the individual preference estimates for each attribute were normally distributed (see Figure 11 A-E).

Implications

In this section the implications of this study to treating prostate cancer patients with localized disease, patient preference research, and future research directions is discussed.

EBRT for localized prostate cancer

Several phase III RCTs of CFRT and MH regimens were discussed in chapter 2. Based on 5 year follow up data only two trials found increased risk of acute toxicity (about 10% to 15%) in the MH arms (1, 2). Two of the smaller superiority trials found MH can improve efficacy, however the larger superiority trials and the non-inferiority trials concluded the efficacy was not significantly different between the two EBRT regimens (23). At the recommended dose of 3.0 Gy per fraction for MH only two trials found a significant difference in risk of toxicity between the regimens, one an increased risk of late toxicity for the MH arm while the other found the opposite (2, 3). Longer follow up data are required to determine conclusively that there is no difference of late toxicity and biochemical recurrence between these two regimens. One phase III RCT investigating ultra-hypofractionated EBRT found significantly worse acute and late toxicity in the investigative arm, though the difference was only 5% for acute and 4% for late toxicity. More phase III trial data that includes SBRT regimens specifically are required.

These issues are pertinent to my findings as some prostate cancer patients with localized disease, especially older patients and those who lived further from the cancer centre, preferred a shorter EBRT regimen with the trade-off of increased risk of acute toxicity within the risk levels found in RCTs. For patients who preferred not to risk increased acute toxicity, a candid discussion is required with their clinician as not all trials demonstrated this increased risk. In addition, a smaller group (LCM analysis 3.2 'convenience group') were willing to risk increased late toxicity for the benefit of shorter EBRT. As hypothesized *a priori*, no models found participants would trade increased risk of PSA recurrence for shorter treatment time. Therefore, if the long-term data demonstrate increased late toxicity is associated with MH this must be disclosed to patients so they can make informed decisions about their treatment.

Patient preferences

The impetus toward patient-centred care can be traced back to the Institute of Medicine (IOM) as it first conceptualized patient centredness as “providing care that is respectful of and responsive to individual patient preferences, needs and values” (24). The IOM provided this benchmark that prompted efforts globally to integrate a patient-centred approach in healthcare across jurisdictions. In Ontario, the formal commitment to a patient-centred approach came with the passing of the Excellent Care for All Act (ECFAS) in 2010 and was further solidified with the Patient’s First Act (2016) that required hospitals to have patient engagement indicators as part of their Quality Improvement Plans (25, 26). This trend to place increasing importance on patients’ perspectives and preferences for treatments can better inform both clinical practice and resource allocation decisions about patient management by respecting patients’ expectations and priorities. My study is in keeping with this trend as the findings quantify patient concerns and values. This DCE identified groups of prostate cancer patients with localized disease who were more amenable to MH, in this case older patients and those who lived further from a cancer centre.

Similar to the literature discussed in chapter 2 (27-29), the results of my study demonstrated patients have preferences and concerns about the EBRT regimen selected for them by their Radiation Oncologists. Therefore, as MH becomes the standard of care for prostate cancer patients with localized disease who are treated with EBRT, it is important for clinicians to remember to discuss the risks and benefits with each patient and include them in the decision-making process. Since some participants in this study were willing to trade-off increased risk of acute and late toxicity for the benefit of shorter treatment length, as per discussion in studies by Sharieff et al. and Muralidhar et al., there is potential for savings to the health care system since hypofractionation has been proven to be more cost effective than conventional radiotherapy, likely due to the decreased incremental cost-effectiveness ratio (ICER) of hypofractionation compared to CFRT (30, 31). Also, as discussed by Brazier et al. and Bien et al., and demonstrated with the results from this DCE, the cost effectiveness of therapies can be increased by aligning care with patient preferences as this should lead to improved compliance (21, 32). Questions still remain on how to balance patient and public preferences with experts' judgement in healthcare, and when physicians make decisions for their patients they need to respect patients' expectations and priorities (27). This study also demonstrated that when given the different choices, along with clear explanations of the trade-offs, patients were very interested and benefitted from being part of the decision-making process since preference heterogeneity was present in the estimated models as hypothesized *a priori*. Patients are both the potential beneficiaries of increasingly sophisticated EBRT methods, but also the ones who accept the risks associated with these treatments and their preferences warrant attention.

As reported in chapter 2, there is a gap in the DCE literature on perceiving how participants make their choices, and how uncertain or confused they are about the tasks asked of them. My study explored these issues in depth and found, while most respondents do not consider all attributes, their responses were likely to be valid since they were on average only 'mildly confused' by the DCE (similar to the sample wide mean). Simplifying heuristics were utilized to improve respondents' understanding and nonetheless resulted in the LC analysis class characterization matching with respondents' rating of the attributes.

Further research

The trend towards patient focused care has been increasing, though a recent international survey found almost half (47%) of cancer patients believe they were not sufficiently involved in treatment decisions (33). My study demonstrates prostate cancer patients have preferences regarding the type of EBRT offered to them, and further studies to determine if these results are reproducible in prostate cancer patients who have not yet undergone EBRT are warranted. The Canadian Institutes of Health Research (CIHR) has invested in a Strategy for Patient- Oriented Research (SPOR) with the goal of actively involving patients in health research and health care delivery (34). Thus, there is the possibility of interest and funding for further investigative studies in this field. An important potential extension of this work could use the patient preference results to enhance current prostate cancer patient decision aids to assist their understanding of radiotherapy choices available to them, as well as consider their own values and how these might influence their treatment preferences. Furthermore, my study could be adapted to other cancer sites where hypofractionation is being investigated to expand the sample size as the trade-offs the participants in my study were willing to make may not be generalizable to other populations. My study was conducted in Kingston, Ontario where a large proportion of the participants live in small towns or rural areas and must drive to the regional cancer centre for EBRT which is not in their hometown. It would be interesting to reproduce this study at a cancer centre located in a city with a larger population in which many of the participants would live in the same city as the cancer centre. This would also increase the diversity of the sample as it would likely introduce a wider range of socioeconomic and demographic factors, as most participants in this study were Caucasian. This study could also be scaled up to recruit prostate cancer patients with localized disease across Canada, though this would require computer assisted DCE administration.

Conclusion

At the beginning of this chapter the findings of this thesis were summarized and discussed. This DCE provided insight that all five of the selected attributes of radiotherapy are significant to prostate cancer patients' preferences, and whether these attributes and attribute levels are significant to several subgroups of participants. Additionally, the relative importance of the attributes, the rate at which participants were willing to trade between attributes, and the overall ranking of several theoretical EBRT scenarios was estimated. However, the main limitation of this study is the external validity, which was discussed at length in the limitations section of this chapter with suggestions for future research to clarify if these preferences are reproducible with actual choices and generalizable to other populations. Therefore, the main outcome of this DCE is to estimate the relative importance of these attributes of EBRT for prostate cancer patients with localized disease and not to predict real choice.

This study demonstrated a high response rate is achievable in this patient population. Since the sample population were all familiar with the context of the treatment, no participants failed the rationality test and there were no major concerns with internal validity. Therefore, patient preference data can be quantified with conjoint analysis methods and it is possible to attempt to reproduce this study in similar populations.

The DCE conducted in this thesis provides valuable insights into patients' reasoning and preferences, and these results should be considered for future clinical decision-making processes and best practice decisions regarding EBRT.

References

1. Dearnaley D, Syndikus I, Mossop H, Khoo V, Birtle A, Bloomfield D, et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. Lancet Oncol. 2016 Aug;17(8):1047-60.
2. Catton CN, Lukka H, Gu CS, Martin JM, Supiot S, Chung PWM, et al. Randomized trial of a hypofractionated radiation regimen for the treatment of localized prostate cancer. J Clin Oncol. 2017 Jun;35(17):1884-90.
3. Lee WR, Dignam JJ, Amin MB, Bruner DW, Low D, Swanson GP, et al. Randomized phase III noninferiority study comparing two radiotherapy fractionation schedules in patients with low-risk prostate cancer. J Clin Oncol. 2016 Jul;34(20):2325-32.
4. Widmark A, Gunnlaugsson A, Beckman L, Thellenberg-Karlsson C, Hoyer M, Lagerlund M, et al. Ultra-hypofractionated versus conventionally fractionated radiotherapy for prostate cancer: 5-year outcomes of the HYPO-RT-PC randomised, non-inferiority, phase 3 trial. Lancet. 2019 Aug;394(10196):385-95.
5. Mandeville KL, Lagarde M, Hanson K. The use of discrete choice experiments to inform health workforce policy: a systematic review. BMC Health Serv Res. 2014 Sep;14:367.
6. Janssen EM, Marshall DA, Hauber AB, Bridges JFP. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Rev Pharmacoecon Outcomes Res. 2017 Dec;17(6):531-42.
7. Louviere J, Hensher D, Swait J. Stated choice methods: analysis and application. Cambridge: Cambridge University Press; 2000.
8. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ. 2012 Feb;21(2):145-72.
9. Festinger L. A theory of cognitive dissonance. Stanford, CA: Stanford University Press; 1957.
10. Lancsar E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ. 2006 Aug;15(8):797-811.
11. Hazlewood GS. Measuring patient preferences: an overview of methods with a focus on discrete choice experiments. Rheum Dis Clin N Am. 2018 May;44(2):337-47.
12. Czajkowski M, Vossler CA, Budzinski W, Wisniewska A, Zawojska E. Addressing empirical challenges related to the incentive compatibility of stated preferences methods. J Econ Behav Organ. 2017 Oct;142:47-63.
13. Aguinis H, Bradley KJ. Best practice recommendations for designing and implementing experimental vignette methodology studies. Organ Res Methods. 2014 Oct;17(4):351-71.
14. Carlsson F, Martinsson P. Do hypothetical and actual marginal willingness to pay differ in choice experiments?: application to the valuation of the environment. J Environ Econ Manag. 2001 Mar;41(2):179-92.
15. List JA, Shogren JF. Calibration of the difference between actual and hypothetical valuations in a field experiment. J Econ Behav Organ. 1998 Oct;37(2):193-205.
16. Levitt SD, List JA. What do laboratory experiments measuring social preferences reveal about the real world? Journal Econ Perspect. 2007 May;21(2):153-74.
17. Havrilesky LJ, Alvarez Secord A, Ehrisman JA, Berchuck A, Valea FA, Lee PS, et al. Patient preferences in advanced or recurrent ovarian cancer. Cancer. 2014 Dec;120(23):3651-9.
18. de Bekker-Grob EW, Niers EJ, van Lanschot JJ, Steyerberg EW, Wijnhoven BP. Patients' preferences for surgical management of esophageal cancer: a discrete choice experiment. World J Surg. 2015 Oct;39(10):2492-9.
19. de Bekker-Grob EW, Bliemer MC, Donkers B, Essink-Bot ML, Korfage IJ, Roobol MJ, et al. Patients' and urologists' preferences for prostate cancer treatment: a discrete choice experiment. Br J Cancer. 2013 Aug;109(3):633-40.

20. Vass CM, Payne K. Using discrete choice experiments to inform the benefit-risk assessment of medicines: are we ready yet? Pharmacoeconomics. 2017 Sep;35(9):859-66.
21. Brazier JE, Dixon S, Ratcliffe J. The role of patient preferences in cost-effectiveness analysis: a conflict of values? Pharmacoeconomics. 2009 Sep;27(9):705-12.
22. Lancsar E, Fiebig DG, Hole AR. Discrete choice experiments: a guide to model specification, estimation and software. Pharmacoeconomics. 2017 Jul;35(7):697-716.
23. Lee WR, Koontz BF. Moderate hypofractionation for prostate cancer. Transl Androl and Urol. 2018 Jun;7(3):321-9.
24. Institute of Medicine Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academies Press.; 2001.
25. Bill 41, Patients First Act, 2016 [Internet]. Legislative Assembly of Ontario. Available from: https://www.ontario.ca/laws/statute/S16030
26. Bill 46, Excellent Care For All Act, 2010 [Internet]. Legislative Assembly of Ontario. Available from: https://www.ola.org/en/legislative-business/bills/parliament-39/session-2/bill-46
27. Muhlbacher AC, Juhnke C. Patient preferences versus physicians' judgement: does it make a difference in healthcare decision making? Appl Health Econ Health Pol. 2013 Jun;11(3):163-80.
28. Blinman P, King M, Norman R, Viney R, Stockler MR. Preferences for cancer treatments: an overview of methods and applications in oncology. Ann Oncol. 2012 May;23(5):1104-10.
29. Nakayama M, Kobayashi H, Okazaki M, Imanaka K, Yoshizawa K, Mahlich J. Patient preferences and urologist judgments on prostate cancer therapy in Japan. Am J Mens Health. 2018 Jul;12(4):1094-101.
30. Sharieff W, Greenspoon JN, Dayes I, Chow T, Wright J, Lukka H. The technique, resources and costs of stereotactic body radiotherapy of prostate cancer: a comparison of dose regimens and delivery systems. Technol Cancer Res T. 2016 Feb;15(1):171-8.
31. Muralidhar V, Nguyen PL. Maximizing resources in the local treatment of prostate cancer: a summary of cost-effectiveness studies. Urol Oncol. 2017 Feb;35(2):76-85.
32. Bien DR, Danner M, Vennedey V, Civello D, Evers SM, Hiligsmann M. Patients' preferences for outcome, process and cost attributes in cancer treatment: a systematic review of discrete choice experiments. Patient. 2017 Oct;10(5):553-65.
33. All.Can International. Patient insights on cancer care: opportunities for improving efficiency [Internet]. 2019. Available from: https://www.all-can.org/what-we-do/research/patient-survey/
34. Canadian Institutes of Health Research. Strategy for Patient-Oriented Research [Internet]. Government of Canada; 2019. Available from: http://www.cihr-irsc.gc.ca/e/41204.html

Appendix A

Research Ethics Approval

QUEEN'S UNIVERSITY HEALTH SCIENCES & AFFILIATED TEACHING HOSPITALS RESEARCH ETHICS BOARD (HSREB)

HSREB Initial Ethics Clearance

October 31, 2018

Dr. Sigurdson
Department of Oncology
Queen's University

ROMEO/TRAQ #: 6025029
Department Code: ONGY-545-18
Study Title: "Prostate cancer patients' preferences for hypofractionated radiotherapy"
Supervisor: Dr. Michael Brundage
Review Type: Delegated
Date Ethics Clearance Issued: October 31, 2018
Ethics Clearance Expiry Date: October 31, 2019

Dear Dr. Sigurdson,

The Queen's University Health Sciences & Affiliated Teaching Hospitals Research Ethics Board (HSREB) has reviewed the application and granted ethics clearance for the documents listed below. Ethics clearance is granted until the expiration date noted above.

- Letter of Information/Consent Form v.2018OCT29
- SES Health Data Survey v.2018OCT16
- Sample Survey Question v.2018OCT16

Documents Acknowledged:

- CORE training certificate S. Sigurdson 2015NOV09

Amendments: No deviations from, or changes to the protocol should be initiated without prior written clearance of an appropriate amendment from the HSREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involve(s) only administrative or logistical aspects of the trial.

Renewals: Prior to the expiration of your ethics clearance you will be reminded to submit your renewal report through ROMEO. Any lapses in ethical clearance will be documented on the renewal form.

Completion/Termination: The HSREB must be notified of the completion or termination of this study through the completion of a renewal report in ROMEO.

Reporting of Serious Adverse Events: Any unexpected serious adverse event occurring locally must be reported within 2 working days or earlier if required by the study sponsor. All other serious adverse events must be reported within 15 days after study team members have become aware of the information.

Reporting of Complaints: Any complaints made by participants or persons acting on behalf of participants must be reported to the Research Ethics Board within 7 days of study team members becoming aware of the complaint.

<u>Note: All documents supplied to participants must include the contact information for the Research Ethics Board. Investigators: please note that if your trial is registered by the sponsor, you must take responsibility to ensure that the registration information is accurate and complete.</u>

Yours sincerely,

Albert F. Clark.

Albert F. Clark, PhD

Chair, Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board

The HSREB operates in compliance with, and is constituted in accordance with, the requirements of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations, and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is qualified through the CTO REB Qualification Program and is registered with the U.S. Department of Health and Human Services (DHHS) Office for Human Research Protection (OHRP). Federalwide Assurance Number: FWA#:00004184, IRB#:00001173

HSREB members involved in the research project do not participate in the review, discussion or decision

Appendix B
Validity and Reliability

Measurement validity

Measurement validity determines whether DCE results meet content and face validity by identifying how accurately the DCE measures preference (53, 62). To ensure content validity the choice of attributes and levels should be grounded in qualitative work, as discussed in this chapter in the Attribute selection section (62). Qualitative research methods have commonly been used at the DCE design phase, however they have rarely been reported as an approach to evaluating questionnaire validity (72). To draw valid inferences from a survey, most members of the target population should interpret the survey as the researchers intended (72). Qualitative debriefing questions and respondents' rated comprehension can help test construct validity of the DCE (73). Kenny et al. examined the validity of a DCE via a quantitative analysis and found the DCE conclusions were consistent with the qualitative results (72). This provided assurance there were no major inconsistences between participants' views and inferences drawn from the quantitative analysis (72). Among the selected oncology DCEs Havrilesky et al., Eliasson et al., and Nakayama et al. asked participants to rank the attributes from most to least important (23). Additionally, Havrilesky et al. had participants use the Likert scale (5-point where 1 is not concerned and 5 is deeply concerned) to rate their concern for each of the attributes (23, 46, 49). The rating and ranking results were tested with paired t-tests, and in Havrilesky et al.'s paper was consistent with the quantitative DCE results (23). Construct validity can partly be tested by determining if the expected associations between patient variables and patterns of risk preferences are present (50). Additionally, the reasoning for an opt-out or status quo alternative, or justification of a forced choice, should be noted (53).

Face validity, also referred to as theoretical validity, is the degree to which the results of a DCE are consistent with *a priori* expectations (62). Face validity of the results can be assessed through qualitative methods and compared to *a priori* hypotheses about magnitude and direction of part worth utilities for the attributes (53, 62). Additionally, the results can also be compared with a researcher's intuition of whether the preference data makes sense (62). Assessing face validity with these two tests is fairly simple and about 60% of DCEs published between 2009-2012 completed this test (62).

Measurement validity also measures the external validity of the results by assessing how generalizable the choice preferences are in other circumstances (53, 62). To examine external validity researchers technically must compare the DCE results to a gold standard, which are choices people make in the 'real world' known as revealed preferences data (50). Lancsar et al. published an interesting article that discussed seven points of good survey design for DCE-based stated preference models and reported the required knowledge about how decisions were made, and what information was required to make them by participants (74). In health care relatively few studies have examined external validity of DCEs since opportunities to observe preferences through real-life choices are limited (53). Additionally, actual treatment choices are often driven by physicians' preferences and to determine 'real world' choices would require a prospective study (50). From the selected oncology DCEs, Muhlbacher and Bethge report a limitation of their study was subjects may make different decisions after discussion with their physician and family members as hypothetical choices do not have clinical, financial, and emotional consequences (27, 42).

Hypothetical bias occurs when investigators assume participants' choices reflect their true underlying preferences if faced with an actual decision, when they may not (50). To increase the likelihood that a DCE is capturing participants' latent preferences all respondents should believe their responses could lead to real consequences (75). This can be achieved by asking participants to deliberate carefully, and if possible the surveys should be anonymous to avoid participants selecting answers they think they are 'supposed' to (50). There was a hypothetical bias of the electronic nature in Havrilesky et al.'s oncology study as respondents may not be as emotionally connected to each choice, especially since the effectiveness of an educational video instead of an interactive approach is unknown (23).

It is important to match the study sample to the larger population of interest and the presented scenarios are familiar to participants, as otherwise responses may be artificial (76). Realism can be improved by completing the DCE in one session and in the setting that is natural to participants (76). From the selected oncology DCEs Havrilesky et al. noted the small sample size of 95 and the heterogenous study population limit the generalizability of their study (23). Muhlbacher and Bethge and de Bekker-Grob et al. reported in their oncology DCEs the external validity could have been improved if patients from different hospitals were included as patient preferences can depend on cultural background and the existing healthcare system (32, 42). Since participants were only be recruited from one institution this was important for the implications of my study.

Measurement reliability

Measurement reliability captures whether a DCE produces similar results under consistent conditions (53). This will never be perfect due to simple variation in responses, though there are several tests for preference reliability such as test-retest, version consistency, and holdout prediction (62). In the health care setting older patients will be asked to complete DCEs. A study by Milte et al. investigated the effect of cognitive functioning on the consistency of older participants responses to a DCE (77). The minimum age was 65 and cognitive impairment was assessed with the Mini-Mental State Examination (77). The study found there was no significant difference of consistency of responses between participants with mild or no cognitive impairment, and those with mild cognitive impairment were able to fully complete the DCE and provide meaningful responses (77).

Choice validity

Choice validity examines whether participants engage with the choice task “as expected” based on assumptions about consumer behaviour, and whether choices comply with assumptions about preferences, trading, and responsiveness to the attributes (53). There are three main tests for choice validity discussed in the literature; monotonicity, compensatory preferences, and task nonattendance (53). Monotonicity (non-satiation) means participants will not prefer 'worse' levels of an attribute to 'better' levels assessed. This is assessed with a 'test of rationality' via a dominant choice task such that all of the attribute levels for one profile are as good as or better than the other profile (53, 62). Howard et al. included a monotonicity or consistency check (where one alternative is set to be dominant across all attributes) and only two respondents failed this question, indicating the vast majority of men understood what was being asked of them (45). Compensatory preferences refer to the extent that participants trade between all attributes of the choice profiles. DCEs assume participants consider all attributes, however non-compensatory preferences can be indicated by either attribute dominance or attribute nonattendance (53). Dominance is when all but the dominant attribute is ignored, and the opposite is attribute nonattendance when a respondent ignores a specific attribute (53). The order the attributes are presented to participants in the choice tasks should be considered carefully as the attributes at the top can be prescribed more importance by participants (50). Randomizing the attribute order can minimize this (50), however Norman et al. proved that order of presentation of attributes had no effect on part-worth estimates (78). The group conducted a large DCE with over 2000 individuals who were randomized between 60 versions of 16 choice sets, where the order of the attributes was randomized within each version (78). It can be difficult to distinguish between true attribute non-attendance (e.g. the attribute is important, but the individual is ignoring it to simplify the task) and genuinely non-important attributes to the respondent. Similar to this concept is task nonattendance, which is the extent respondents pay attention to the choice tasks (21). If the task is too cognitively burdensome, the participant is fatigued, or the scenarios are not realistic then individuals may not carefully consider their decisions which can lead to invalid conclusions (62). One easy assessment for task nonattendance in a non-labelled DCE is to check if any respondents selected all the alternatives on the right or on the left (e.g. only selected 'A' or 'B') (62).

Since the purpose of a DCE is to mimic real-world decisions it is valid for participants to use any reasoning they prefer to make choices within the DCE. It is critical for researchers to know how participants interpret the attributes and levels and ultimately make their decision (21). If individuals do not understand all the information, they will likely not understand the tasks being asked of them and therefore would not make reliable and stable choices that reflect their true preferences. It is important to comprehensively explain the motivation behind a DCE and describe the tasks that will be asked of respondents, including the required medical information in plain language, to improve understanding (31). If sufficient material is provided participants can understand the choice tasks, attributes, and levels and are successful in using complex decision strategies (22). Choices which appear 'irrational' may be due to poor study design or be due to factors unknown to the researcher but are valid to the individual (63). For example, a pilot study should identify if attribute and task nonattendance were present.

Choice reliability

Choice reliability examines whether participants make consistent choices in accordance with assumptions about consumer choices (53). There are three main concepts of choice reliability; transitivity, consistency, and level recoding (62). Transitivity and consistency can be tested by adding specific choice sets outside the experimental design to determine the preference relation between three or more choice profiles (62). Level recoding refers to the extent that participants process the actual numbers presented in continuous attributes, as they may instead think of the presented risk numbers as 'low', 'medium', and 'high' categories (53). From the oncology DCEs Muhlbacher and Bethge reported they did not include a test for test for recoding and this was a limitation as such a test would have improved the internal validity (42). It is critical for researchers to know how participants interpret the attributes and levels and ultimately make their decision (21). As discussed in the choice validity section, if individuals do not understand all the information, they will likely not understand the tasks being asked of them, and therefore would not make reliable and stable choices that reflect their true preferences.

Appendix C

Supplementary Table

Table 22: 2-class LCM.

Characteristics	**Class 2.1 "Convenience group"**		**Class 2.2 "Toxicity and efficacy group"**	
	β	p-value	β	p-value
Length β_1	-0.256	<.001	0.012	.687
Markers β_2 (ref. no markers)	-1.466	<.001	0.090	.468
PSA (ref. 6%)				
12% β_{3a}	-1.136	.003	-1.316	<.001
18% β_{3b}	-0.568	.257	-2.642	.002
Acute (ref 20%)				
35% β_{4a}	-0.482	.063	-0.494	.002
50% β_{4b}	0.345	.304	-1.142	<.001
Late (ref. 10%)				
15% β_{5a}	0.469	.115	-1.154	<.001
20% β_{5b}	-0.131	.716	-1.696	<.001
Class share	0.318		0.682	
≥70 years old	+		-	
Working	-		-	
Had prior systemic therapy	-		-	
Posterior probability	0.979			
Unconditional probability	0.549		0.588	
Conditional probability	0.719		0.641	
Observations	216		480	
Goodness of fit				
Log likelihood	-380.20			
AIC	817.38			
BIC	854.47			
CAIC	872.47			