DESIGN AND ANALYSIS ISSUES OF NON-INFERIORITY CLINICAL TRIALS

STATISTICAL AND METHODOLOGICAL ISSUES IN THE DESIGN AND ANALYSIS OF NON-INFERIORITY CLINICAL TRIALS OF RADIOTHERAPY IN WOMEN WITH EARLY STAGE BREAST CANCER

By

SAMEER PARPIA, B.Sc. (Honours), M.Sc. (Biostatistics)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2014) Hamilton, Ontario

(Health Research Methodology – Biostatistics Specialization)

TITLE: Statistical and Methodological Issues in the Design and Analysis of Non-inferiority Clinical Trials of Radiotherapy in Women with Early Stage Breast Cancer

AUTHOR: Sameer Parpia

B.Sc. (McMaster University)

M.Sc. (University of Western Ontario)

SUPERVISOR: Professor Lehana Thabane

NUMBER OF PAGES: x, 103

ABSTRACT

Background and Objectives

We investigate three statistical and methodological issues within the context of non-inferiority randomized controlled trials (RCTs), specifically those of radiotherapy regimens for the prevention of local recurrence in patients with early stage breast cancer who have undergone breast conserving surgery. These issues are: (1) the analysis of multiple time-to-event outcomes in non-inferiority RCTs; (2) the interim analysis of a binary outcome that is repeatedly assessed at pre-specified times; and (3) determining the optimal analysis population for dealing with crossovers in non-inferiority RCTs.

Methods

Issue 1: We investigated and compared the properties of four statistical models (proportional hazards model, competing risk model, marginal model and frailty model) for analyzing radiotherapy non-inferiority RCTs of patients with early stage breast cancer who are at risk for and may experience multiple failure types. We applied the four methods to data from an existing trial in which subjects with breast cancer could experience local recurrence (the primary outcome), distant recurrence, death, or a combination of these events. In addition, we compared these models using simulated examples of similar non-inferiority trials with varying hazards of each failure type.

Issue 2: We investigated and compared the properties of three methods for estimating the event proportions for an interim analysis in RCTs with a binary outcome that is

repeatedly assessed at pre-specified times. Generally, interim analyses are performed after half or more of the subjects have completed full follow-up. However, depending on the duration of accrual relative to the length of follow-up, this may be inefficient, since there is a possibility that the trial will have completed accrual prior to the interim analysis. We focussed our simulations on situations where delaying the interim analysis until half or more of subjects have completed full follow-up is an inefficient approach. The methods include: 1) estimation of the event proportion based on subjects who have been followed for a pre-specified time (less than the full follow-up duration) or who experienced the outcome; 2) estimation of the event proportion based on all available data from subjects randomized by the time of the interim analysis; and 3) the Kaplan-Meier approach to estimate the event proportion. We varied the risk of the outcome, the treatment effect and the probability of an event occurring at each pre-specified time. We compared the three methods in terms of overall type I and II errors, as well as the probability of stopping early for benefit.

Issue 3: We explored the effect of subject crossover from the experimental to the standard radiotherapy arm prior to treatment initiation on the intention-to-treat, per-protocol, as-treated and combined intention-to-treat and per-protocol analysis in non-inferiority RCTs of radiotherapy for the prevention of local recurrence in patients with early stage breast cancer. We varied the non-inferiority margin, the percent of subjects who cross over and evaluated random and non-random crossover. The main comparison of the methods was done using overall type I error. In addition, we compared the methods based on estimate bias and standard error of the estimate.

Results and Conclusions

Issue 1: All four models produced similar results for the existing trial (i.e. non-inferiority was observed regardless of the method used). Simulations showed that the event-specific methods yielded contrasting results when the distribution of distant recurrence or death differed between treatment groups. We conclude that multiple models should be used as part of a comprehensive analysis.

Issue 2: We showed that conducting an interim analysis when a considerable number of subjects have completed a portion of their full follow-up duration is an efficient approach under certain scenarios where event distribution probabilities are similar between treatment groups. Under these specific scenarios, all three methods preserved the type I and II errors. In these cases, we recommend using the Kaplan-Meier method because it incorporates all the available data and has greater probability of early stopping.

Issue 3: The as-treated analysis had the best performance in terms of type I error rate. However, it can be recommended only in scenarios where crossover is random. It performed poorly in scenarios with greater than 2% non-random crossover. The intention-to-treat and per-protocol analysis performed poorly under both random and non-random crossover scenarios.

ACKNOWLEDGEMENTS

This achievement would not have been possible without the encouragement, support, mentorship and guidance from many individuals.

First, I would like to thank my supervisor, Lehana Thabane. I am extremely appreciative of your faith in me, your patience, endless encouragement and your guidance throughout the completion of this degree.

Second, I would like to express my gratitude to my mentor and committee member, Jim Julian. You took me "under your wing" many years ago and have guided me through this journey step by step. Your mentorship and friendship is unparalleled. Thanks to you, I am not only a better statistician but also a better person.

I would also like to thank my committee member, Mark Levine, for his continuous support and mentorship, as well as for the opportunity to participate in several cancer and thrombosis research studies, the results of which have directly improved patient care.

I owe gratitude to the Ontario Clinical Oncology Group. The knowledge and skills that I have gained working with the group cannot be taught in any classroom. I would specifically like to thank my colleagues Chushu Gu, Denise Julian, Timothy Whelan and Kathryn Cline.

I also owe special thanks to my fellow HRM students, Ilia Ferrusi, Morgan Lim, Natalia Diaz-Granados and Nathan Souza. I could not have asked for a better group of scholars

with whom to share this experience, and I hope to continue to work with you in the future.

To my dearest family and friends – at some point over the last few years, you have heard about the trials and tribulations of my PhD program, and have provided me with tremendous support and encouragement. I am eternally grateful to you.

To my beloved wife, Arti Sharma Parpia, you provided me with strength, perspective and humour when I had lost mine. Thank you for keeping me sane and always believing in me.

Finally, I would like to thank my parents, Nizar and Chandni Parpia. Without them and their sacrifices, none of this would have been possible.

Sameer Parpia

Hamilton, Ontario
June 2014

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

DECLARATION OF ACADEMIC ACHEIVEMENT

This thesis is a “sandwich thesis”, in that it is composed of three individual projects prepared for publication in peer-reviewed journals. The following are the contributions of Sameer Parpia in all of the papers included in the dissertation: developing the research ideas and research questions; developing and designing the studies; developing the analysis and simulation plans; conducting all statistical analyses and simulations; interpreting the results; and writing all drafts of the manuscripts. My co-authors contributed to the development and design of the studies, providing clinical and methodological expertise and critical revision of the manuscripts. The work of this thesis was conducted between September 2007 and June 2014. The first two papers have been published and the third paper has been submitted for publication to a peer-reviewed journal.

CHAPTER 1

INTRODUCTION

Breast cancer is one of the most common cancers in women and is the second leading cause of cancer deaths among females [1]. It is estimated that approximately 23,800 Canadian women were diagnosed with breast cancer in 2013, with 9300 of these occurring in Ontario [2]. Approximately 5000 women per year die from breast cancer in Canada [2].

Breast cancers present as palpable or non-palpable tumours. Non-palpable tumours detected by screening tools such as mammography tend to be smaller in size compared with palpable tumours. The extent of breast cancer is generally classified into four stages (I to IV) and this correlates with prognosis, where a higher stage indicates a worse prognosis. Tumour size, presence of tumour in the axillary nodes and presence of metastases are components that define the stage of breast cancer. This is commonly done using the International Union Against Cancer TNM classification where T refers to the tumour size (T1: $\leq$ 2cm, T2: > 2 to $\leq$ 5cm, T3: > 5cm), N refers the presence of cancer in the lymph nodes (N0: 0 nodes, N1: 1-3 nodes, N2: 4-9 nodes, N3: $\geq$ 10 nodes) and M refers to the presence of metastases (M0: no metastases, M1: metastases present) [3].

Randomized clinical trials (RCTs) are considered the gold standard for evaluating therapeutic interventions for various diseases, including breast cancer. Mortality and recurrence rates from early stage breast cancer have been declining over the past few

decades due to advances in treatments, and the rigorous evaluation of these treatments through RCTs.

Breast cancer is commonly treated using each or a combination of surgery, radiotherapy, chemotherapy and endocrine therapy. The course of treatment is determined by stage as well as by other clinical factors such as age and presence of biomarkers. For patients with early stage breast cancer, surgery is usually the first step of treatment. Surgical treatment can be either a mastectomy, which consists of removal of the whole breast, or a lumpectomy, which involves removal of the portion of the breast containing the tumour, thus aiming to preserve the appearance and sensation of the breast. The 20-year follow-up results of the National Surgical Adjuvant Breast and Bowel Project trial comparing mastectomy with lumpectomy showed that the overall survival was similar between the two procedures [4].

The second phase of treatment is generally radiotherapy. Breast conserving treatment (BCT), which consists of lumpectomy or breast conserving surgery (BCS) followed by radiotherapy, has been shown to be superior to BCS alone in the prevention of ipsilateral cancer recurrence in the breast in patients with early stage breast cancer, and is the preferred option for treating this population [4-9]. Furthermore, BCT has been shown to be superior to BCS alone with respect to breast cancer death [9].

In addition to surgery and radiotherapy, some patients also receive adjuvant chemotherapy and endocrine therapy. The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) review of RCTs investigating the effect of chemotherapy and

endocrine therapy concluded that "*some of the widely practicable adjuvant drug treatments that were being tested in the 1980s, which substantially reduced 5-year recurrence rates, also substantially reduce 15-year mortality rates*" [10].

The advancement of statistical and research methodology over the past few decades has played a significant role in establishing improvement in breast cancer outcomes and in population health in general [11]. Many methodological and statistical challenges that were inherent to breast cancer clinical trials have been studied, and solutions to overcome these challenges have been proposed [11-17].

A major advancement was the development of the non-inferiority trial design [18, 19]. The non-inferiority design aims to show that an experimental treatment is no worse than the active control by a pre-specified tolerance margin. This design is generally considered when the experimental treatment is less toxic, or less invasive, or more convenient or cheaper, compared with the active control. There are a number of weaknesses in this design that do not exist in the traditional superiority RCTs. These include the choice of non-inferiority margin, the issue of assay sensitivity, which is the ability of a trial to distinguish an effective therapy from one that is less or not effective, and the notion of bio-creep which refers to the phenomenon that a slightly inferior treatment becomes the active control for future non-inferiority trials and after a series of trials, the active control becomes no better than placebo. These issues, among others in the design, conduct and analysis of non-inferiority trials, have been discussed by several authors [20-24].

Nonetheless, the non-inferiority design has played an important role in evaluating radiotherapy treatment of women with early stage breast cancer. For example, hypofractionation radiotherapy, which delivers a higher dose of radiation over a fewer number of treatments, has been shown to be non-inferior to standard whole breast irradiation for the prevention of local recurrence [25-27]. Furthermore, techniques such as accelerated partial breast irradiation, which delivers multiple doses of radiotherapy per day to only part of the breast tissue, are currently being compared with whole breast irradiation in several on-going non-inferiority trials [28, 29].

However, various unresolved challenges in the design and analysis of non-inferiority RCTs still require statistical and methodological investigation, and current methods for designing and analyzing non-inferiority radiotherapy trials in breast cancer can be improved upon. The objective of this dissertation is to address some of the recent challenges surrounding non-inferiority radiotherapy breast cancer clinical trials through simulations and re-analyses of previous trials, and to provide direction for future research. We investigated three specific methodological issues: (1) the analysis of multiple time-to-event outcomes in non-inferiority RCTs, (2) the interim analysis of a binary outcome that is assessed repeatedly at pre-specified times, and (3) determining the optimal analysis population for dealing with cross overs in non-inferiority RCTs.

Issue 1: The Analysis of Multiple Time-to-event Outcomes in Non-inferiority RCTs

Women with early stage breast cancer who have been treated with radiotherapy for their disease are at risk of their disease recurring. The recurrence of disease can be local (i.e.

within the treated breast), contralateral or untreated breast, regional (i.e. in the axillary, supraclavicular, infraclavicular, or the internal mammary lymph nodes of the treated breast) or distant (metastases in the liver, bone, lung, brain etc.). In addition, women are also at risk of a new primary cancer occurring at another site, and at risk of death. These events can be related to each other, and patients can experience any number of them over a specified period of time.

Several methods exist for analyzing data consisting of multiple time-to-event outcomes. These include the proportional hazards model [30], the competing risk model [31], the marginal model [32] and the frailty model [33]. While some methods completely ignore the plausible relationships between the events, others account for a correlation using various modelling techniques. The use of these methods to analyze such data arising from non-inferiority trials has not been investigated thoroughly. The effect of the plausible relationships between events on the performance of the models within the non-inferiority framework has not yet been demonstrated in the published literature.

Issue 2: Interim Analysis of a Binary outcome that is Assessed Repeatedly at Pre-specified Times

Analyses that evaluate the treatment effect during the conduct of a trial are referred to as interim analyses. These analyses are generally included in the design of RCTs for ethical and economic reasons as they provide guidance for early stopping of the trial for extremely positive or harmful results, or for futility. For example, if the results of a mid-trial interim analysis demonstrate that the experimental therapy is unsafe, then the

principal investigators should consider stopping the trial. If, on the other hand, the results show that the experimental therapy is much better than the standard therapy, then early termination should also be considered since it would be unethical to continue the trial.

A major concern when delivering radiotherapy to the breast is the effect of radiation on the cosmetic appearance of the breast. Therefore, many RCTs evaluating new high-dose radiotherapies include adverse breast cosmesis as an important safety outcome [28, 29]. In some cases, although the primary objective of the trial is the prevention of local recurrence, in the non-inferiority setting, it is not uncommon to perform an interim analysis of safety based on adverse cosmesis using the superiority framework [29].

Breast cosmesis of a subject is assessed typically at pre-specified follow-up visits (e.g. 1, 2, 3 and 5 years post-randomization) by a nurse using a questionnaire such as the European Organisation for Research and Treatment of Cancer (EORTC) Cosmetic Rating System for Breast Cancer in which the treated breast is compared with the untreated breast in terms of size, shape, location of the areola and nipple, appearance of the surgical scar, presence of telangiectasia, and an overall global cosmetic score.

The global cosmetic score is generally the focal point of the interim analysis, and is often analyzed as a dichotomy, either adverse or normal cosmesis. A failure in cosmetic outcome can be identified in two ways: a) whether a subject has been assessed as having adverse cosmesis at a pre-specified follow-up visit (e.g. the 3-year assessment) or b) whether a subject has been assessed as having adverse cosmesis at any visit prior to the pre-specified visit. In our research, we focus on the latter. Similar situations are seen in

venous thrombosis trials investigating the effect of an intervention for preventing post-thrombotic syndrome, where subjects are assessed every 6 months for up to 24 months using a disease-specific questionnaire [34, 35]. A failure is defined if the score exceeds a threshold at any visit.

Interim analyses in RCTs with binary outcomes are performed typically after half or more of the subjects have completed follow-up. However, in some cases, depending on the duration of accrual relative to the length of follow-up, this may not be an efficient undertaking because it may be possible that the trial will have completed accrual and all subjects will have been treated prior to the interim analysis. An alternative is to plan the interim analysis after subjects have completed follow-up to a specified time that is less than the fixed full follow-up duration. However, this option has never been investigated in the clinical trial methodology literature, and little is known about the effect of conducting an interim analysis using data on subjects who have completed follow-up to a time that is less than the fixed full follow-up duration, especially in RCTs where the outcome is assessed at pre-specified visits.

Issue 3: The Optimal Analysis Population for Dealing with Crossovers in Non-inferiority RCTs

In most RCTs, it is likely that some subjects will cross over from one treatment arm to another. This may occur prior to initiating any study treatment, midway through the trial or after a study treatment has been completed. Subjects crossing over from the experimental intervention to the standard therapy are most common, and they can occur

for a number of reasons [36]. There may be complications in delivering the allocated treatment. A physician may conclude that a subject is not responding to the allocated treatment and therefore recommends switching treatments. Moreover, a subject may choose to switch due to side-effects related to their current treatment.

In non-inferiority RCTs of radiotherapy in women with early stage breast cancer, it is inevitable that some subjects will cross over from the experimental arm to the standard arm prior to initiation of any treatment due to complexities that occur in delivering the experimental radiotherapy, physician preference, or subject preference. Generally, crossovers from the standard arm to the experimental arm are not permitted, especially if the experimental radiotherapy is not offered outside of the trial.

During the analysis of a trial, crossovers pose a challenge for methodologists and statisticians since they are a potential cause for bias [36]. For superiority trials, it is well known that the *intention-to-treat* (ITT) analysis which analyzes subjects according to their randomized group regardless of what they received preserves randomization and should be used for the primary analysis of the trial [37]. Since crossovers are a source of contamination, the ITT analysis tends to dilute the treatment effect and therefore provides a conservative estimate of the treatment effect [38]. However, some argue that the opposite is true for non-inferiority trials. This argument suggests that since non-inferiority trials aim to show that the experimental therapy is not inferior to the standard therapy (i.e. that they are similar), the ITT analysis in the presence of crossovers is anti-conservative [21, 38, 39]. Alternative strategies are to conduct a *per-protocol* (PP) analysis which

includes only subjects who have fully complied with their allocated treatment, an *as-treated* (AT) analysis which analyzes subjects according to the treatment they actually received, or a combined ITT and PP (ITT+PP) analysis which requires that both the ITT and PP analyses reject the null hypothesis of inferiority in order to conclude non-inferiority of the experimental therapy [21, 39].

The effect of crossovers from the experimental to standard therapy prior to initiation of any treatment on the ITT, PP, AT and combined ITT and PP analyses using time-to-event outcomes has not been studied extensively. The current literature provides little guidance on the most appropriate approach for conducting analyses of non-inferiority RCTs in the presence of crossovers.

Summary of Chapters

This is a "sandwich thesis" of three manuscripts, each relating to one of the issues described above. The papers are separated into three chapters beginning with Chapter 2.

Chapter 2 deals with the analysis of multiple correlated time-to-event outcomes in a non-inferiority trial. We compared four statistical methods for analyzing multiple time-to-event outcomes using a previously reported non-inferiority trial of hypofractionated radiotherapy versus standard therapy for the prevention of local recurrence in women with early stage breast cancer [27]. In addition, we compared the methods using simulations of similar non-inferiority trials using a latent failure time approach. The methods under investigation include the proportional hazards model, the competing risk

model, the marginal model and the frailty model. We considered two scenarios in which we varied the hazards for experiencing each of the three events: local recurrence, distant recurrence and death. The methods were evaluated in terms of whether they concluded non-inferiority of the experimental arm.

Chapter 3 discusses the assessment of methods of estimating the event proportions when conducting an interim analysis in RCTs evaluating a binary outcome that is assessed repeatedly at pre-specified times. We restricted our attention to the situation where it was impractical to conduct an interim analysis on only those subjects who have completed their follow-up. In our simulations, we varied the "true" event proportion in the control group, the treatment effect, and the probability of experiencing the event at each of the pre-specified times. We evaluated three approaches to estimate the event proportions: 1) based only on subjects who have been followed for a pre-specified time (less than the full follow-up duration) or who experienced the outcome; 2) based on all available data from subjects randomized by the time of the interim analysis; and 3) using the Kaplan-Meier approach. Assessment of the three approaches was done using the overall type I and II errors and the probability of early stopping for benefit of the experimental treatment.

Chapter 4 focuses on the effect of crossovers from the experimental to the standard therapy on the ITT, PP, AT and ITT+PP analyses of non-inferiority RCTs. We used simulations to investigate the effect of random and non-random crossovers on each of the analysis approaches. Inputs for the simulation were also based on previously reported non-inferiority RCTs of hypofractionated radiotherapy versus standard therapy for the

prevention of local recurrence in women with early stage breast cancer [26, 27]. In our simulations, we varied the non-inferiority margin of the RCT, the percentage of subjects who crossed over, and the degree of non-random crossover based on a prognostic variable. We evaluated each approach in terms of type I error. In addition, we compared bias and standard errors for the ITT, PP and AT approaches.

Lastly, Chapter 5 discusses the key findings, limitations and implications of the thesis. The overall objective of the thesis is to advance our understanding on some of the issues related to non-inferiority RCTs of radiotherapy in subjects with early stage breast cancer. Results reported in the three papers will enhance the literature on the design and analysis of non-inferiority RCTs in breast cancer.

References

1. Jemal A, Siegel R, Xu J, Ward E: Cancer statistics, 2010. *CA Cancer J Clin* 2010, 60:277-300.

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics: Canadian Cancer Statistics 2013. *Canadian Cancer Society* 2013

3. Sobin LH, Gospodarowicz MK, Wittekind C: *TNM Classification of Malignant Tumours*, 7th Edition. London, UK: Wiley; 2009

4. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, Jeong JH, Wolmark N: Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *N Engl J Med* 2002, 347:1233-1241.

5. Veronesi U, Luini A, Del Vecchio M, Greco M, Galimberti V, Merson M, Rilke F, Sacchini V, Saccozzi R, Savio T: Radiotherapy after breast-preserving surgery in women with localized cancer of the breast. *N Engl J Med* 1993, 328:1587-1591.

6. Forrest AP, Stewart HJ, Everington D, Prescott RJ, McArdle CS, Harnett AN, Smith DC, George WD: Randomised controlled trial of conservation therapy for breast cancer: 6-year analysis of the Scottish trial. Scottish Cancer Trials Breast Group. *Lancet* 1996, 348:708-713.

7. Clark RM, Whelan T, Levine M, Roberts R, Willan A, McCulloch P, Lipa M, Wilkinson RH, Mahoney LJ: Randomized clinical trial of breast irradiation following lumpectomy and axillary dissection for node-negative breast cancer: an update. Ontario Clinical Oncology Group. *J Natl Cancer Inst* 1996, 88:1659-1664.

8. Clarke M, Collins R, Darby S, Davies C, Elphinstone P, Evans E, Godwin J, Gray R, Hicks C, James S, MacKinnon E, McGale P, McHugh T, Peto R, Taylor C, Wang Y, Early Breast Cancer Trialists' Collaborative Group (EBCTCG): Effects of radiotherapy

and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lanc*et 2005, 366:2087-2106.

9. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S, McGale P, Correa C, Taylor C, Arriagada R, Clarke M, Cutter D, Davies C, Ewertz M, Godwin J, Gray R, Pierce L, Whelan T, Wang Y, Peto R: Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. *Lancet* 2011, 378:1707-1716.

10. Early Breast Cancer Trialists' Collaborative Group (EBCTCG): Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* 2005, 365:1687-1717.

11. Sylvester R, Van Glabbeke M, Collette L, Suciu S, Baron B, Legrand C, Gorlia T, Collins G, Coens C, Declerck L, Therasse P: Statistical methodology of phase III cancer clinical trials: advances and future perspectives. *Eur J Cancer* 2002, 38 Suppl 4:S162-8.

12. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976, 34:585-612.

13. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer* 1977, 35:1-39.

14. Gelman R: Statistical methods for early breast cancer trials. *Cancer Treat Res* 1992, 60:27-53.

15. Altman DG, Machin D: Current statistical issues in clinical cancer research. *Br J Cancer* 1993, 68:455-456.

16. Machin D: On the evolution of statistical methods as applied to clinical trials. *J Intern Med* 2004, 255:521-528.

17. Ryan RP, Woodall WH: The most-cited statistical papers. *J of Applied Stat* 2005, 32:461-474.

18. Dunnett CW, Gent M: Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics* 1977, 33:593-602.

19. Blackwelder WC: "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982, 3:345-353.

20. James Hung HM, Wang SJ, Tsong Y, Lawrence J, O'Neil RT: Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003, 22:213-225.

21. D'Agostino RB S, Massaro JM, Sullivan LM: Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med* 2003, 22:169-186.

22. Fleming TR: Current issues in non-inferiority trials. *Stat Med* 2008, 27:317-332.

23. Fleming TR, Odem-Davis K, Rothmann MD, Li Shen Y: Some essential considerations in the design and conduct of non-inferiority trials. *Clin Trials* 2011, 8:432-439.

24. DeMets DL, Friedman L: Some Thoughts on Challenges for Non-inferiority Study Designs. *Drug Info J* 2012, 12:420-427.

25. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P,

Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial A of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet Oncol* 2008, 9:331-341.

26. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bentzen SM, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P, Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial B of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet* 2008, 371:1098-1107.

27. Whelan TJ, Pignol JP, Levine MN, Julian JA, MacKenzie R, Parpia S, Shelley W, Grimard L, Bowen J, Lukka H, Perera F, Fyles A, Schneider K, Gulavita S, Freeman C: Long-term results of hypofractionated radiation therapy for breast cancer. *N Engl J Med* 2010, 362:513-520.

28. Vicini FA, Chen P, Wallace M, Mitchell C, Hasan Y, Grills I, Kestin L, Schell S, Goldstein NS, Kunzman J, Gilbert S, Martinez A: Interim cosmetic results and toxicity using 3D conformal external beam radiotherapy to deliver accelerated partial breast irradiation in patients with early-stage breast cancer treated with breast-conserving therapy. *Int J Radiat Oncol Biol Phys* 2007, 69:1124-1130.

29. Olivotto IA, Whelan TJ, Parpia S, Kim DH, Berrang T, Truong PT, Kong I, Cochrane B, Nichol A, Roy I, Germain I, Akra M, Reed M, Fyles A, Trotter T, Perera F, Beckham W, Levine MN, Julian JA: Interim cosmetic and toxicity results from RAPID: a randomized trial of accelerated partial breast irradiation using three-dimensional conformal external beam radiation therapy. *J Clin Oncol* 2013, 31:4038-4045.

30. Cox DR: Regression models and life-tables. *J Royal Stat Soc Series B (Methodol)* 1972, 43:187-220.

31. Fine J, Gray RJ: A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 1999, 94:496-509.

32. Wei LJ, Glidden DV: An overview of statistical methods for multiple failure time data in clinical trials. *Stat Med* 1997, 16:833-9.

33. Wienke A: *Frailty Models in Survival Analysis.* Boca Raton, FL: Chapman & Hall/CRC; 2010

34. Enden T, Haig Y, Klow NE, Slagsvold CE, Sandvik L, Ghanima W, Hafsahl G, Holme PA, Holmen LO, Njaastad AM, Sandbaek G, Sandset PM, CaVenT Study Group: Long-term outcome after additional catheter-directed thrombolysis versus standard treatment for acute iliofemoral deep vein thrombosis (the CaVenT study): a randomised controlled trial. *Lancet* 2012, 379:31-38.

35. Vedantham S, Goldhaber SZ, Kahn SR, Julian J, Magnuson E, Jaff MR, Murphy TP, Cohen DJ, Comerota AJ, Gornik HL, Razavi MK, Lewis L, Kearon C: Rationale and design of the ATTRACT Study: a multicenter randomized trial to evaluate pharmacomechanical catheter-directed thrombolysis for the prevention of postthrombotic syndrome in patients with proximal deep vein thrombosis. *Am Heart J* 2013, 165:523-530.e3.

36. Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ: Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Med Res Methodol* 2011, 11:4-2288-11-4.

37. Montori VM, Guyatt GH: Intention-to-treat principle. *CMAJ* 2001, 165:1339-1341.

38. Matsuyama Y: A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. *Stat Med* 2010, 29:2107-2116.

39. Matilde Sanchez M, Chen X: Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Stat Med* 2006, 25:1169-1181.

CHAPTER 2

EMPIRICAL COMPARISON OF METHODS FOR ANALYZING MULTIPLE TIME-TO-EVENT OUTCOMES IN A NON-INFERIORITY TRIAL: A BREAST CANCER STUDY

S. Parpia[a], L. Thabane[bc], J. A. Julian[a], T.J. Whelan[ad] and M. N. Levine[ad]

a Ontario Clinical Oncology Group, Department of Oncology, McMaster University, 711 Concession Street – G (60) Wing 1st Floor, Hamilton, ON, Canada. L8V 1C3

b Centre of Evaluation of Medicines, St Joseph's Healthcare - Hamilton, 50 Charlton Avenue East , Hamilton, ON, Canada. L8N 4A6

c Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON Canada

d Juravinski Cancer Centre, 699 Concession Street, Hamilton, ON, Canada. L8V 5C2

Abstract

Background: Subjects with breast cancer enrolled in trials may experience multiple events such as local recurrence, distant recurrence or death. These events are not independent; the occurrence of one may increase the risk of another, or prevent another from occurring. The most commonly used Cox proportional hazards (Cox-PH) model ignores the relationships between events, resulting in a potential impact on the treatment effect and conclusions. The use of statistical methods to analyze multiple time-to-event events has mainly been focused on superiority trials. However, their application to non-inferiority trials is limited. We evaluate four statistical methods for multiple time-to-event endpoints in the context of a non-inferiority trial.

Methods: Three methods for analyzing multiple events data, namely, i) the competing risks (CR) model, ii) the marginal model, and iii) the frailty model were compared with the Cox-PH model using data from a previously-reported non-inferiority trial comparing hypofractionated radiotherapy with conventional radiotherapy for the prevention of local recurrence in patients with early stage breast cancer who had undergone breast conserving surgery. These methods were also compared using two simulated examples, scenario A where the hazards for distant recurrence and death were higher in the control group, and scenario B, where the hazards of distant recurrence and death were higher in the experimental group. Both scenarios were designed to have a non-inferiority margin of 1.50.

Results: In the breast cancer trial, the methods produced primary outcome results similar to those using the Cox-PH model: namely, a local recurrence hazard ratio (HR) of 0.95 and a 95% confidence interval (CI) of 0.62 to 1.46. In Scenario A, non-inferiority was observed with the Cox-PH model (HR=1.04; CI of 0.80 to 1.35), but not with the CR model (HR=1.37; CI of 1.06 to 1.79), and the average marginal and frailty model showed a positive effect of the experimental treatment. The results in Scenario A contrasted with Scenario B with non-inferiority being observed with the CR model (HR=1.10; CI of 0.87 to 1.39), but not with the Cox-PH model (HR=1.46; CI of 1.15 to 1.85), and the marginal and frailty model showed a negative effect of the experimental treatment.

Conclusion: When subjects are at risk for multiple events in non-inferiority trials, researchers need to consider using the CR, marginal and frailty models in addition to the Cox-PH model in order to provide additional information in describing the disease process and to assess the robustness of the results. In the presence of competing risks, the Cox-PH model is appropriate for investigating the biologic effect of treatment, whereas the CR models yields the actual effect of treatment in the study.

Keywords: non-inferiority, Cox model, correlation, marginal model, frailty model, competing risks

1. Introduction

Randomized controlled trials are considered to be the gold standard for evaluating therapeutic interventions in many different diseases including those in oncology. Unlike studies in other diseases, cancer trials typically follow subjects beyond the planned intervention, often for many years. During this time, subjects may be at risk for several events. For example, subjects in breast cancer trials can experience local recurrence in the treated breast, distant recurrence, death or a combination of these. In most trials, only one of these events is considered the primary outcome and the others are secondary outcomes. The occurrence of multiple events per subject over a period of time is sometimes referred to as event history data [1].

One of the most commonly used statistical approaches for analyzing such data is the Cox proportional hazards (Cox-PH) model, which models the time from randomization to a specific event [2]. However, analyzing each outcome separately using the Cox-PH model does not make use of all the available information because it fails to account for the plausible relationships or correlations between events. For instance, it is possible that experiencing one event increases the risk of experiencing another event. Conversely, it is also possible that the occurrence of one event may even prevent others from occurring, a situation known as competing risks [3]. Standard survival analysis techniques have been shown to bias results in such circumstances [4–6]. The effect of treatment may differ depending on whether or not intermediary events are incorporated into the analysis.

Several statistical methods exist to analyze event history data. These include: the Cox-PH models [2], the competing risk (CR) model [7], the marginal model [8], and the frailty model [9]. The majority of research has focused on using these methods in the analysis of superiority trials where the intervention is expected to be superior to the standard treatment, but their application to non-inferiority trials is lacking. Marginal and frailty models are efficient methods of estimating treatment effect in studies where patients have multiple events of the same type, such as recurrence of asthma attacks. In addition, they are used in studies where treatment can have an effect on multiple events using the same biological pathway. Research on the CR models in superiority trials has shown the Kaplan-Meier approach over-estimates the event rate in the presence of competing risks. However, the relative treatment effect from the CR model remains unchanged compared to the Cox-PH model unless treatment affects the competing event [10].

Non-inferiority randomized trials generally compare the standard treatment with a new treatment that is expected to be less toxic or less expensive or less invasive but "no worse" within a tolerance margin than the standard treatment in terms of clinical outcome.

The purpose of this manuscript is to compare empirically these different approaches in the analysis of a non-inferiority trial in which a subject can experience more than one type of outcome event. In addition, we compare these methods using simulated examples of trials. We first provide a brief overview of the methods, and then apply them to a previously-reported randomized trial of hypofractionated radiotherapy in patients with

breast cancer [11,12], and to the simulated trial examples. For the purposes of this study, we will consider the Cox-PH model for each type of event as the primary analysis.

2. Methods

2.1 Cox Proportional Hazards (PH) Model

The instantaneous rate of failure known as the hazard rate is defined as the probability of failing in the next small time interval, given that one has already survived until the beginning of the interval [13]. The standard Cox-PH model has become the most frequently used method for modeling hazards and covariates. The model does not make a distributional assumption about the baseline hazard, and assumes that the covariate acts multiplicatively on the hazard independent of time. The model is given by the following:

$$\lambda_i(t \mid X) = \lambda_0(t)\exp(\beta X)$$

where $\lambda_i(t \mid X)$ is the hazard of subject *i* conditional on covariate X at time t, $\lambda_0(t)$ is the baseline hazard at time t, X is the covariate (e.g. 1 = experimental group, 0 = control group), and β is the coefficient representing the effect of treatment independent of time. In cancer trials, the constant treatment effect is represented by the ratio of hazards for the experimental group relative to the control group, or hazard ratio (HR), given by $\exp(\beta)$.

2.2 Competing Risks (CR) Model

Kalbfleisch and Prentice developed the cumulative incidence function (CIF) to analyze competing risk data [14]. The CIF estimates the hazard of an event of interest in the presence of other competing events, known as hazard of the sub-distribution. The estimation of the CIF is similar to that used by Kaplan-Meier ; however, the CIF does not censor subjects who experience a competing event and therefore does not require the assumption of independence between the event of interest *j* and the competing events [4,15]. Fine and Gray [7] proposed a proportional hazards model that models the effects of covariates on the hazards of the CIF by distinguishing between competing events and truly censored subjects [15]. Similar to the Cox-PH model, the CR model is given by:

$$\lambda_{ij}^{sub}\left(t \mid X\right) = \lambda_{0j}^{sub}\left(t\right)\exp\left(\beta_j^{sub} X\right)$$

where $\lambda_{ij}^{sub}\left(t \mid X\right)$ is the hazard of the sub-distribution for cause *j*; $\lambda_{0j}^{sub}\left(t\right)$ is the baseline hazard of the sub-distribution; and β_j^{sub} is the treatment effect of the sub-distribution. This model reduces to the standard Cox-PH model when competing risks are absent.

2.3 Marginal Model

Wei, Lin and Weissfeld [8] proposed a marginal model (WLW model) where a subject is assumed to be simultaneously at risk for all events, and is at risk for each event until this event occurs [16]. The WLW model estimates the treatment effect using independent Cox-PH models for each event, and, therefore, the relationship structure between event

times does not need to be known [17,18]. For each event *j* for subject *i*, the model is given by

$$\lambda_{ij}(t \mid X) = \lambda_{0j}(t)\exp(\beta_j X).$$

Stratification by event *j* allows for varying underlying baseline hazards λ_{0j} for each event. In addition, treatment by event interactions allows for estimation of event-specific treatment effects [18,19]. The WLW model also estimates the 'average effect' of treatment $\overline{\beta}$ on all events using a weighted average of $\hat{\beta}_j$, which we will call the average WLW model. Dependencies between observed event times are adjusted for by the use of a robust sandwich estimate of the variance. In the presence of competing risks, the WLW model models both the marginal hazard for death and the cause-specific hazard for recurrences [19].

2.4 Frailty Model

Frailty models are survival random effects models in which a parameter for heterogeneity is incorporated into the model. The model is given by:

$$\lambda_i(t \mid X) = \lambda_0(t)\exp(\beta X + \gamma_i)$$

where γ_i is the frailty parameter that can also be used to model associations between event times [20]. A large parameter value corresponds to a large correlation between event times for a subject, and also describes the *frailty* or excess risk within a subject

[9,21,22]. This model assumes that event times within a subject are independent given the frailty parameter [20]. Similar to other random effects model, this one also yields effects specific to the subjects in the trial. Several published books provide excellent reviews on frailty models [9,21,23].

2.5 The Hypofractionation Trial

Between April 1993 and September 1996, 1234 patients with early stage breast cancer who had undergone breast conserving surgery were randomly allocated to receive either 42.5 Gray of radiotherapy in 16 fractions (the experimental arm) or 50 Gray in 25 fractions (the standard arm) to the breast for the prevention of local breast recurrence; details and long-term results are described elsewhere [11,12]. The primary outcome of local recurrence was compared was using a point-in-time comparison of local recurrence failure probabilities at five and 10 years [11,12].

For the purpose of this paper, HRs rather than point-in-time failure probabilities will be used. The hypofractionation trial was designed with a control arm local recurrence rate of 7% at 5 years. The non-inferiority margin was set at 5% to tolerate an increase in local recurrence to 12% in the experimental arm. This translates into a HR = $ln(0.88)/ln(0.93)$ = 1.76. Additional events of interest were distant recurrence, new primary cancer and death. Because of the difficulty in differentiating new primaries from distant recurrences, these will be combined in the distant recurrence category. In addition, we consider only the first occurrence of each type of event.

2.6 Simulated Examples

Suppose that a randomized non-inferiority trial similar to the Hypofractionation Trial were designed to demonstrate that an experimental therapy *E* is as good as a control therapy *C* for the prevention of local recurrence in a subset of breast cancer patients. Assuming that the rate of local recurrence at five years in the control arm is 10.0%, and that the maximum tolerable rate of local recurrence at five years in the experimental arm is 14.6% (HR=1.50), then 1000 patients per treatment arm would be required, giving 90% power and a one-sided alpha of 0.025.

As with the Hypofractionation trial, we assume that these patients will also be at risk for distant recurrence and death. We simulated two possible outcome scenarios (A and B) for this trial using a latent failure time approach. For each treatment group, data were generated using two independent bivariate exponential models based on the hazards in **Table 1** (24); one model for local recurrence (l) and death (d_1) with correlation of 0.2, and the other for distant recurrence (m) and death (d_2) with correlation of 0.6. Time of death (d) is given by:

$$d = \begin{cases} d_1 & \text{if } \min(l, d_1, m, d_2) = d_1 \\ d_2 & \text{if } \min(l, d_1, m, d_2) = d_2 \\ d_1 & \text{if } \min(l, d_1, m, d_2) = l \\ d_2 & \text{if } \min(l, d_1, m, d_2) = m \end{cases}$$

which essentially is the time of death if death is the first event, or the time of death that is linked to the first recurrence (local or distant). Independent censoring was generated so

that approximately 40% of the observations were censored. Survival times for event and censored observations were calculated for each subject by combining event times and censoring times. Recurrences could occur only prior do death (i.e. recurrence times were less than the death time). Similarly, local recurrence could not occur after distant recurrence. Censoring could occur prior to any events occurring, or after recurrences have occurred. Based on the standard error estimated from the Cox-PH model using hypofractionation trial data, 1000 simulations would produce an estimate to within at least 1.5 percent of the true coefficient.

2.7 Analysis

For the Cox-PH model, we structured the data in a "wide" format (i.e. one record per subject). We fit Cox-PH models for each event separately. For the local recurrence model, death and distant recurrences are censored, and for the distant recurrence model, death is treated as a censored observation and local recurrence is ignored. Any recurrence is ignored for the death model. Similarly, for the CR approach, we fit Fine and Gray's model (7) for each event. Death and distant recurrences are treated as competing events for the local recurrence model, and death is treated similarly for the distant recurrence model. The analysis for death is equivalent to the standard Cox-PH model because death is always observable.

Data for the WLW model is set up in a "long format" where every subject has three records, one for each event, whether censored or observed. The events are treated as independent strata in the model, and time is expressed from randomization to each event.

Table 2 shows the time and censoring mechanism for each event given a subject's event experience.

The frailty model is fit using an extension of the Cox-PH model that includes the frailty parameter that assumes a gamma distribution because the events are assumed to be positively correlated (25,26). For this analysis, every subject has at least one record representing vital status (i.e. alive or dead) at the end of the study, and each recurrence is represented by an additional record.

For the simulation, the HRs and the standard errors of the HRs were averaged on the log scale (1000 replications). All analyses were performed using SAS 9.2 (SAS Institute, Cary, NC) and R 2.13 (www.r-project.org).

3. Results

3.1 The Hypofractionation Trial

Figure 1 shows the results of the treatment effect using each of the methods. The Cox-PH, CR and WLW models all yield almost identical estimates for each of the events of interest. The shorter experimental treatment does not affect the occurrence of local recurrence, distant recurrence or death with HRs (95% CI) of 0.95 (0.62, 1.46), 1.12 (0.86, 1.43) and 0.97 (0.75, 1.24) respectively. Since the upper 95% CI of the HR for local recurrence is less than 1.76, non-inferiority can be concluded. Moreover, the frailty and WLW models also show that treatment does not affect the risk of failure from all events combined.

3.2 Simulated Examples

Results of scenario A (**Figure 2**), the Cox-PH and WLW model yield an upper 95% CI of 1.35 for the HR for local recurrence, thus suggesting that the experimental therapy is non-inferior to the control. These models also suggest a protective effect of experimental therapy on distant recurrence and death. In contrast, the CR model shows that the experimental therapy is not non-inferior to the control with the upper confidence limit of 1.79 crossing the 1.50 margin. In addition, the CR model yields a treatment effect of 0.96 (0.75, 1.21) for distant recurrence. The frailty and average WLW model show that the experimental arm is significantly protective for all events combined.

The results of Scenario B (**Figure 3**) are opposite to that of scenario A. In this case, the Cox-PH and WLW models yield an upper 95% CI for the HR for local recurrence that is greater than the 1.50 margin. On the other hand, the CR model shows that the experimental therapy is non-inferior to the control with respect to local recurrence. Moreover, the CR model shows that treatment has no effect on distant recurrence whereas the Cox-PH and WLW models suggest a detrimental effect of the experimental treatment on distant recurrence. The frailty and average WLW model suggest that the experimental treatment is harmful when considering all events together.

4. Discussion

In non-inferiority clinical trials of patients with breast cancer, patients may be at risk of and may experience multiple failure types. The occurrence of one of these events may

alter the probability of occurrence of other events. Moreover, the influence of treatment may differ depending on whether another event has occurred, thus affecting the conclusions of the trial. This paper discusses, and applies four approaches of analyzing non-inferiority trials with multiple events, by using data from an existing trial in which subjects with breast cancer could experience local recurrence, distant recurrence, death, or a combination of these events. In addition, we compared the methods using simulated examples of non-inferiority trials.

The analysis of the Hypofractionation Trial showed that treatment was not associated with increased risk of any of the events of interest either individually or in combination. The results for each event using the Cox-PH model and the CR model are similar, suggesting that the impact of competing risks in this data set is minimal. The treatment estimates for each event from the WLW model are identical to those of the standard Cox-PH model since the estimates of the regression coefficients are calculated using equivalent methods. However, the adjustment of correlation in the variance estimate of the WLW model leads to slightly different confidence intervals when compared with the Cox-PH model. The WLW model is also susceptible to the competing risk problem since subjects are at risk for events until they occur, but the model yields unbiased estimates when treatment does not influence the competing events [27].

Scenarios A and B provide evidence that the presence of multiple events could alter the conclusions of the trial depending on the method of final analysis. The Cox-PH and WLW local recurrence models ignore the hazards for distant recurrence and death, thus

resulting in different conclusions for local recurrence when compared to with CR model. Similarly, the Cox-PH and WLW distant recurrence models ignore the hazard for death. By ignoring the competing risks, the Cox-PH and WLW methods model the cause-specific hazard or the marginal failure times, and the effect of treatment can be interpreted as the "pure effect" or the biologic effect of treatment on the event of interest [28]. This is the effect of treatment under the assumption that the competing risk had not occurred, which can be of interest to investigators.

Unlike the Cox-PH and WLW models, the CR model does not censor patients who have had a distant recurrence or death, but rather assumes that these patients will have a zero risk of local recurrence once distant recurrence or death is observed. Censoring assumes that the patient is still at risk for local recurrence. Therefore, in the CR model, the treatment group with higher relative hazards of distant recurrence and death will have a relatively lower hazard of local recurrence, and the HR for local recurrence will favor this treatment group. This approach models the hazard of the sub-distribution, and the effect of treatment can be described as the "real effect" or the actual effect seen in the data [28,29].

The CR model does provide additional information about the treatment when competing events are present. The Cox-PH model declares non-inferiority of local recurrence, but the CR model shows that the absolute effect of treatment is inferior in the study because the control group has a higher hazard of competing events (scenario A). However, in some situations (scenario B), the results from the CR model should be interpreted with

caution since the CR model may show that the experimental group is non-inferior to the control for local recurrence, but at the expense of increased distant recurrence or death, which are clinically worse outcomes. If this is a concern, one may opt to design the trial using an outcome such as disease-free survival which encompasses local and distant recurrence. In addition, CR models have less power than the Cox-PH models to rule out the same non-inferiority margin [30].

The average WLW and frailty models are useful in investigating the overall effect of treatment for any event accounting for the correlation between event times in their respective ways. The main advantage of these approaches is that they are efficient in their estimation of regression coefficients due to their ability to use all the data and to adjust for the association between event times, thus increasing statistical power. However, their use is limited when dealing with dissimilar types of events with different clinical etiology such as local and distant recurrence, because the approach does not provide HRs of treatment and other factors in relation to specific events but rather a combination of all events. Moreover, these models do not correspond to the design of the trial which is evaluating a local treatment and based on rate of local recurrence.

The methods behave similarly in non-inferiority trials as compared with superiority trials. As in superiority trials, competing risks is an issue when treatment affects the competing event. When the distribution of competing events are similar in both treatment groups, the CR model and the Cox-PH model yield similar results, and therefore, the biologic effect and actual effect of treatment in the study are similar. However, similar to superiority

trials, when treatment has a differential effect on competing events, the results of the biologic and actual effect of treatment can contradict each other.

A limitation of this study is that we compared analytic techniques using a single non-inferiority trial. To overcome this, we simulated examples to illustrate that the choice of method may influence the conclusions. However, we simulated only two scenarios using the latent failure time approach, thus limiting the generalizability of the results. Secondly, we generated the data using a latent failure time approach which is not without controversy [31]. However, we did not use the model or its assumptions in any of our analyses, and do not recommend it for use for analysis. Lastly, we considered only the most commonly used methods of analysis which are readily available in current statistical software. Alternative options include jointly modeling all types of events using a joint frailty model where each event has one hazard function [32], or using a multivariate competing risk frailty model [33]. However, such undertakings would be computationally intensive and complex.

5. Conclusions

Our results show that the choice of event-specific models did not affect the non-inferiority conclusion of the Hypofractionation Trial. However, our examples showed that the CR method did yield contrasting conclusions to the Cox-PH and WLW models when competing events were present. In general, the method of analysis should be determined by the research question. The Cox-PH or the WLW model can be used for analysis of non-inferiority trials when the question relates to the biologic effect of treatment. The CR

model should also be used when competing risks are present as it provides valuable information on the actual effect of treatment in the study, especially when treatment has an effect on the competing event. Both models should be part of a comprehensive analysis. The frailty and average WLW provide similar results of the overall effect of treatment on all the events. When subjects are at risk for multiple events in non-inferiority trials, researchers should consider the use of the CR, WLW and frailty models concurrent with the standard Cox-PH model in order to provide additional information in describing the disease process.

Competing Interests

None

Author's Contributions

SP, JAJ, LT and MNL conceived the study. SP conducted literature review, designed and implemented the simulation, preformed data analysis and wrote the initial draft of the manuscript. TJW, JAJ and MNL participated in the design and implementation of the Hypofractionation Trial. All authors reviewed and revised the draft version of the manuscript. All authors read and approved the final version of the manuscript.

References

1. Andersen PK, Borgan O, Gill RD, Keiding N: *Statistical models based on counting processes*. New York: Springer; 1993.

2. Cox D: Regrssion models and life-tables. *J Royal Stat Soc B Methodol* 1972, 43:187-220.

3. Gooley TA, Leisenring W, Crowley J, Storer BE: Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med* 1999, 18:695-706.

4. Kim HT: Cumulative incidence in competing risks data and competing risks regression analysis. *Clin Cancer Res* 2007, 13:559-565.

5. Williamson PR, Kolamunnage-Dona R, Tudur Smith C: The influence of competing-risks setting on the choice of hypothesis test for treatment effect. *Biostat* 2007, 8:689-694.

6. Tai B-C, Grundy RG, Machin D: On the importance of accounting for competing risks in paediatric cancer trials designed to delay or avoid radiotherapy: I. Basic concepts and first analyses. *Int J Radiat Oncol Biol Phys* 2010, 76:1493-1499.

7. Fine J, Gray R: A proportional hazards model for the sub-distribution of a competing risk. *J Am Stat Assoc* 1999, 94:496-509.

8. Wei LJ, Glidden DV: An overview of statistical methods for multiple failure time data in clinical trials. *Stat Med* 1997, 16:833-839.

9. Hougaard P: *Analysis of Multivariate Survival Data.* New York, NY: Springer; 2000.

10. Bakoyannis G, Touloumi G: Practical methods for competing risks data: A review. *Stat Meth Med Res* 2011, 3:257-272

11. Whelan T, MacKenzie R, Julian J, Levine M, Shelley W, Grimard L, Lada B, Lukka H, Perera F, Fyles A, Laukkanen E, Gulavita S, Benk V, Szechtman B: Randomized trial of breast irradiation schedules after lumpectomy for women with lymph node-negative breast cancer. *J Natl Cancer Inst* 2002, 94:1143-50.

12. Whelan TJ, Pignol J-P, Levine MN, Julian JA, MacKenzie R, Parpia S, Shelley W, Grimard L, Bowen J, Lukka H, Perera F, Fyles A, Schneider K, Gulavita S, Freeman C: Long-term results of hypofractionated radiation therapy for breast cancer. *N Engl J Med* 2010, 362:513-520.

13. Parmar M, Machin D: *Survival Analysis: A Practical Approach.* Chichester, UK: John Wiley and Sons; 1995.

14. Kalbfleisch J, Prentice R: *The Statistical Analysis of Failure Time Data.* New York, USA: John Wiley and Sons; 1980.

15. Parpia S, Julian JA, Thabane L, Lee AYY, Rickles FR, Levine MN: Competing events in patients with malignant disease who are at risk for recurrent venous thromboembolism. *Contemp Clin Trials 2011*, 32:829-833.

16. Ghosh D: Methods for analysis of multiple events in the presence of death. *Control Clin Trials* 2000, 21:115-126.

17. Lim HJ, Liu J, Melzer-Lange M: Comparison of methods for analyzing recurrent events data: application to the Emergency Department Visits of Pediatric Firearm Victims. *Accident Anal Prev* 2007, 39:290-299.

18. Metcalfe C, Thompson SG: Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Stat Meth Med Res* 2007, 16:103-122.

19. Li QH, Lagakos SW: Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event. *Stat Med* 1997, 16:925-940.

20. Wienke A: *Frailty Models in Survival Analysis.* Boca Raton, FL: Chapman & Hall/CRC; 2010.

21. Therneau T, Grambsch P: *Modeling Survival Data*. New York, NY: Springer; 2000.

22. Duchateau L, Janssen P: Evolution of recurrent asthma event rate over time in frailty models. *J Royal Stat Soc C App Stat* 2003, 52:355-363.

23. Duchateau L, Janssen P: *The Frailty Model*. New York, NY: Springer; 2010.

24. Freidlin B. Korn EL: Testing treatment effects in the presence of competing risks. *Stat Med* 2005, 24:1703-1712

25. Clayton D: A model for association in bivariate life tables and its application in epidemiological of familial studies tendency in chronic disease incidence. *Biometrika* 1978, 65:141-151.

26. Hougaard P: A Class of Multivariate Failure Time Distributions. *Biometrika* 1986, 73:671-678.

27. Tai B-Choo, Stavola BLD, Gruttola VD, Gebski V, Machin D: First-event or marginal estimation of cause-specific hazards for analysing correlated multivariate failure-time data ? *Stat Med* 2008, 27:922-936.

28. Pintilie M. *Competing Risks: A Practical Perspective.* Chichester, UK: John Wiley and Sons; 2006.

29. Koller MT, Raatz H, Steyerberg EW, Wolbers M: Competing risks and the clinical community: irrelevance or ignorance? *Stat Med* 2011,31:1089-1097.

30. Tai B-C, Wee J, Machin D: Analysis and design of randomised clinical trials involving competing risks endpoints. *Trials* 2011, 12:127.

31. Allignol A, Schumacher M, Wanner C, Drechsler C, Beyersmann J. Understanding competing risks: a simulation point of view. *BMC Med Res Methodol* 2011,11:86.

32. Liu L, Huang X: The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat Med* 2008, 27:2665-2683.

33. Dixon SN, Darlington GA, Desmond AF: A competing risks model for correlated data based on the subdistribution hazard. *Lifetime Data Anal* 2011, 17:473-495.

Table 1. Hazards for simulated scenarios of non-inferiority trials

Scenario	Outcome	Hazard Rate		Marginal Hazard Ratio
		Experimental	Control	
A	LR	0.02	0.02	1.00
	MR	0.02	0.03	0.67
	DT	0.02	0.04	0.50
B	LR	0.03	0.02	1.50
	MR	0.03	0.02	1.50
	DT	0.04	0.02	2.00

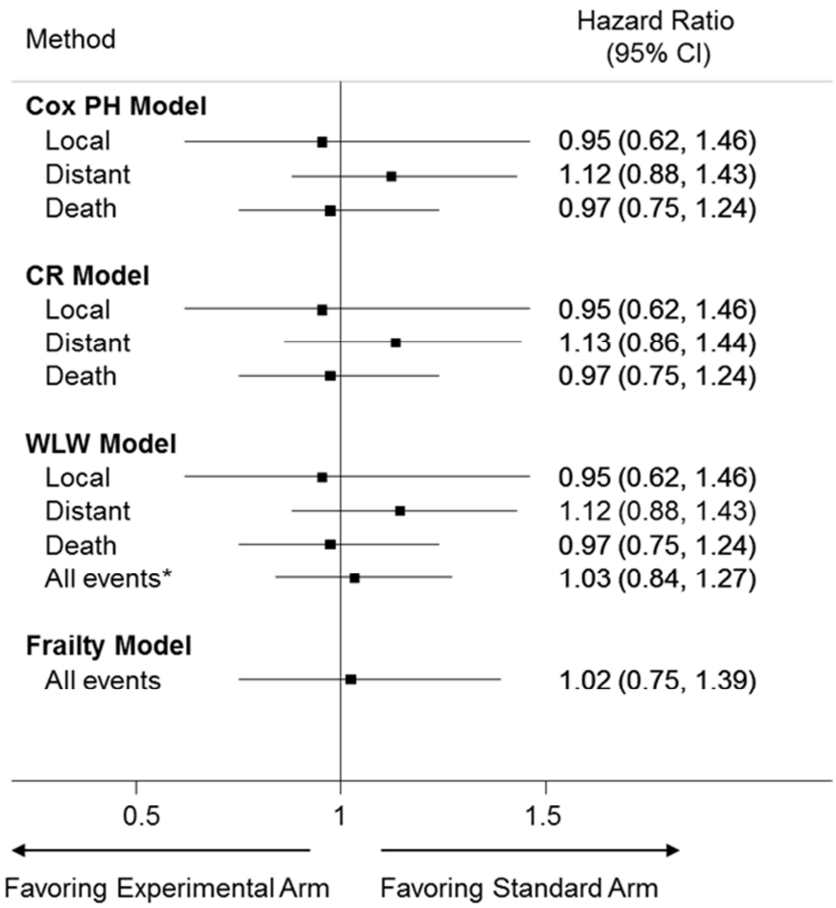
LR = local recurrence, MR = distant recurrence, DT = Death

Table 2: Data structure for the WLW model for all possible combinations of events

	Event Stratum		
Events	Local	Distant	Death
L, M, D	L	M	D
L, M	L	M	E^+
L, D	L	D^+	D
M, D	M^+	M	D
L	L	E^+	E^+
M	M^+	M	E^+
D	D^+	D^+	D

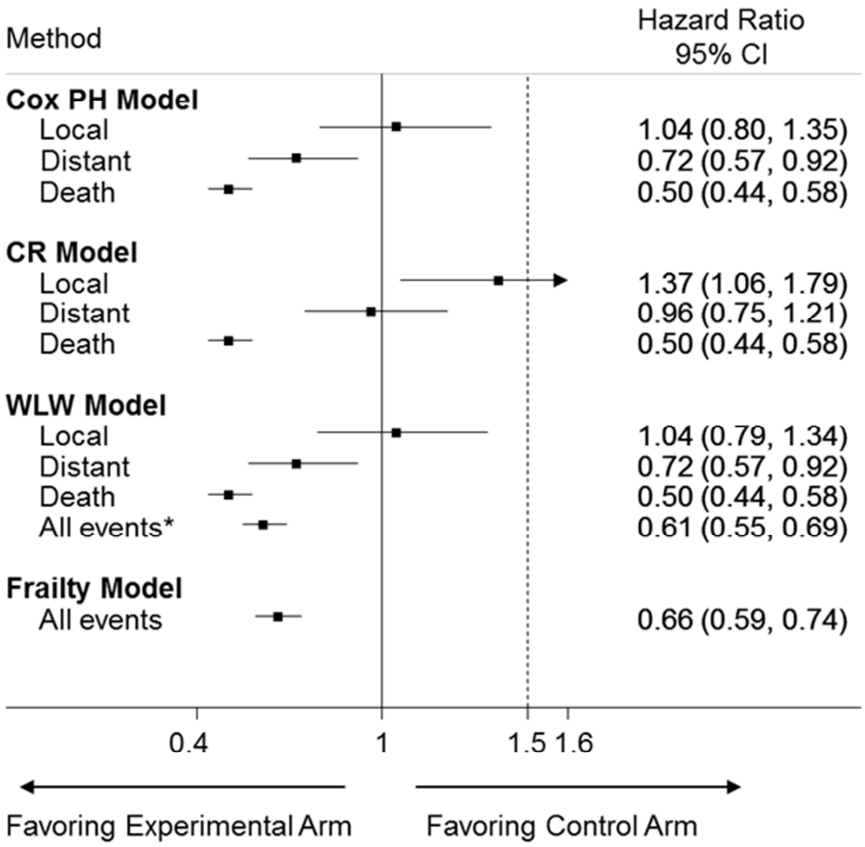
L= time to local recurrence, M= time to distant recurrence, D= time to death, E= time at end of follow-up, + = censoring indicator

Figure 1. Forest plot showing the treatment effect in the Hypofractionation Trial using each of the analysis methods



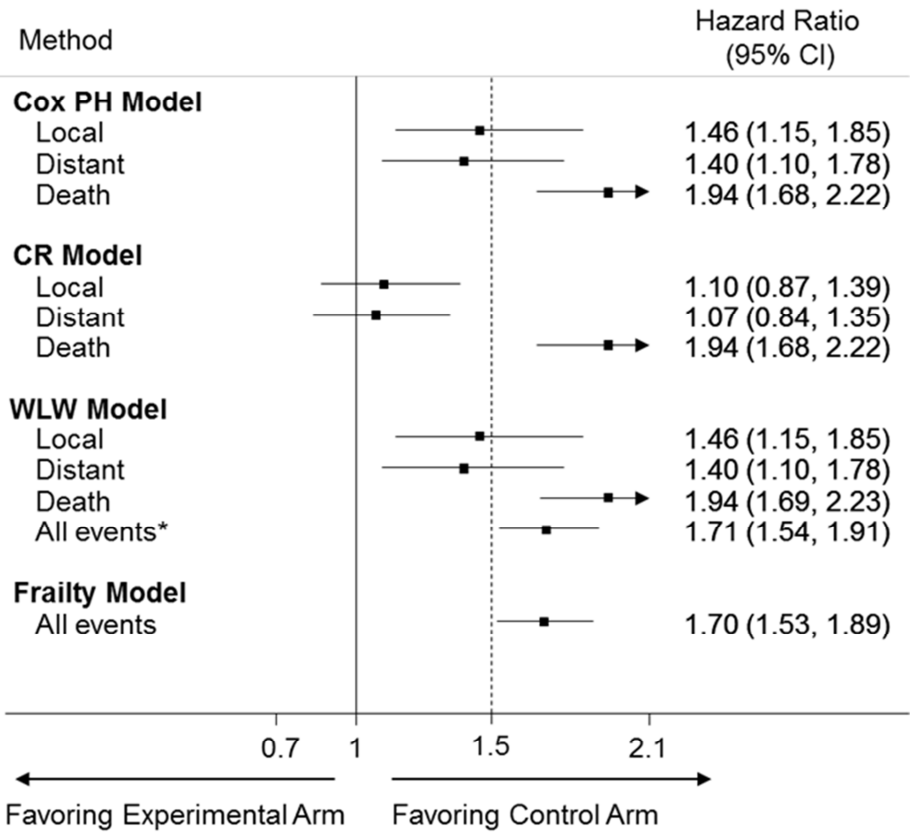
PH = proportional hazards, CR = competing risks, WLW = Wei, Lin and Weissfeld, CI = confidence interval, * average WLW model.

Figure 2. Forest plot showing the treatment effect in Scenario A using each of the analysis methods



PH = proportional hazards, CR = competing risks, WLW = Wei, Lin and Weissfeld, CI = confidence interval, * average WLW model.

Figure 3. Forest plot showing the treatment effect in Scenario B using each of the analysis methods



PH = proportional hazards, CR = competing risks, WLW = Wei, Lin and Weissfeld, CI = confidence interval, * average WLW model.

CHAPTER 3

INTERIM ANALYSIS OF BINARY OUTCOME TRIALS WITH A LONG FIXED FOLLOW-UP TIME AND REPEATED OUTCOME ASSESSMENTS AT PRE-SPECIFIED TIMES

S. Parpia[a], J. A. Julian[a], C. Gu[a], L. Thabane[bc] and M. N. Levine[ad]

a Ontario Clinical Oncology Group, Department of Oncology, McMaster University, 711 Concession Street – G (60) Wing 1st Floor, Hamilton, ON, Canada. L8V 1C3

b Biostatistics Unit - FSORC, St Joseph's Healthcare - Hamilton, 50 Charlton Avenue East , Hamilton, ON, Canada. L8N 4A6

c Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON Canada

d Juravinski Cancer Centre, 699 Concession Street, Hamilton, ON, Canada. L8V 5C2

Abstract

In trials with binary outcomes, assessed repeatedly at pre-specified times and where the subject is considered to have experienced a failure at the first occurrence of the outcome, interim analyses are performed, generally, after half or more of the subjects have completed follow-up. Depending on the duration of accrual relative to the length of follow-up, this may be inefficient, since there is a possibility that the trial will have completed accrual prior to the interim analysis. An alternative is to plan the interim analysis after subjects have completed follow-up to a time that is less than the fixed full follow-up duration. Using simulations, we evaluated three methods to estimate the event proportion for the interim analysis in terms of type I and II errors and the probability of early stopping. We considered: 1) estimation of the event proportion based on subjects who have been followed for a pre-specified time (less than the full follow-up duration) or who experienced the outcome; 2) estimation of the event proportion based on data from all subjects that have been randomized by the time of the interim analysis; and 3) the Kaplan-Meier approach to estimate the event proportion at the time of the interim analysis. Our results show that all methods preserve and have comparable type I and II errors in certain scenarios. In these cases, we recommend using the Kaplan-Meier method because it incorporates all the available data and has greater probability of early stopping when the treatment effect exists.

Key words: interim analysis, binary outcome, power, type I error

1. Introduction

Interim analyses that permit early stopping of a randomized controlled trial (RCT) for extremely positive results or for futility are included in the design for ethical and economic reasons. Strategies have been developed for interim analyses such that the overall type I error of the entire trial is preserved at a fixed level [1-4].

Often, the primary outcome is whether or not a subject experienced an event over a fixed period of time *T*. In some trials, the outcome is assessed repeatedly at pre-specified times during follow-up, and the subject is considered a failure if the event occurs at any time. For example, in a cardiovascular RCT investigating the effect of an intervention for preventing post-thrombotic syndrome, subjects can be assessed every 6 months for up to 24 months using a disease-specific questionnaire [5, 6]. A failure has occurred if the questionnaire score exceeds a pre-specified threshold. Another example would be a breast cancer radiotherapy RCT where adverse cosmesis (i.e. a dichotomy), assessed at 1, 3 and 5 years post-randomization, would be the primary safety outcome and the focus of the interim analysis.

Interim analyses are generally performed after half or more of the subjects have completed follow-up [7]. Depending on the duration of accrual relative to the length of follow-up, this strategy may be inefficient because it is possible that accrual will have been completed and patients will have finished treatment prior to the interim analysis. If, however, the interim analysis was done earlier and a statistically significant effect was

found, the trial may be stopped, and all future subjects would receive the experimental therapy.

In this situation, one alternative is to plan an interim analysis after a smaller percentage of subjects have completed full follow-up. However, there is a low probability of terminating the trial early when the interim analysis is based on so little information, and, therefore, such an analysis would unnecessarily spend alpha [8]. A second alternative is to plan the interim analysis after half or more of the subjects have completed a specified portion of the follow-up R, where $R<T$, and T is the fixed full follow-up duration for each subject.

Several researchers have studied methods that combine data from subjects who have completed full follow-up with those who have been followed for duration R in situations where the outcome is reversible [9-11]. In our research, however, the situation is different in that the outcome can be ascertained at any of the pre-specified visits during follow-up and is irreversible.

In this paper, we consider 3 methods of estimating the interim event proportion (risk) for each treatment group in an RCT for an interim analysis: 1) estimated event proportion based only on subjects who have been followed for at least duration R or who had an outcome event; 2) the event proportion based on data from subjects that have been randomized by the time of the interim analysis, and 3) the Kaplan-Meier approach to estimate the event proportion. We investigate the effect of each method on the type I and

II errors and the probability of early stopping through computer simulation of various trial scenarios.

2. Materials and Methods

Consider a trial designed to detect an absolute risk reduction (ARR) between the standard group (π_0) and the experimental group (π_1) over the time period 0 to T using a normal approximation Z-test with

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0}}}$$

where $\hat{\pi}_0$ and $\hat{\pi}_1$ are the observed proportions, n_0 and n_1 are the group sample sizes, and we are testing the one-sided hypotheses H_0: $\pi_1 \geq \pi_0$ versus H_1: $\pi_1 < \pi_0$. Furthermore, we assume 90% power, an alpha of 0.025 and a 1:1 randomization. Since the normal distribution is symmetric, the p-value for a one-sided test is equivalent to half of the two-sided p-value.

Suppose the trial requires 4 years for enrolment, each subject is followed for 2 years (i.e. T=24 months), and failures are ascertained at any of the four 6-monthly pre-specified visits post-randomization. Let the start of the trial (calendar time) be denoted by τ_0. Following the notation in **Table 1,** let t_j be the pre-specified visit times in the trial where $t_j \leq T$ and j is the visit number where j = 0, 1, 2… J, and J denotes the number of visits (e.g. J = 4 and t_0 = 0, t_1 = 6, t_2 = 12, t_3 = 18, t_4 = 24 months). Suppose an interim analysis is scheduled to occur when 50% of the subjects have completed R=12 months (t_2=R) of

follow-up which, assuming a uniform recruitment pattern, corresponds to approximately 36 months after the start of the trial, denoted by τ_1 (**Figure 1**). At the interim analysis, the proportion of subjects who fail in each group could be estimated using any of the following approaches.

2.1 Method 1: Event proportion based on subjects followed for at least duration R or who had an event

In RCTs where the length of enrolment relative to follow-up is not an issue, subjects included in the interim analysis are those who have completed their full follow-up *T* or who have had an event prior to completion [7]. A similar approach is used here whereby we include only subjects who have completed at least duration *R* (where $t_r = R$, *r* refers to the visit at which follow-up time equals *R*) of their full follow-up *T*, or have had an event prior to this point. Since the interim analysis occurs after 50% of the subjects have completed at least follow-up of *R*, this approach includes the first 50% of enrolled subjects plus those subjects that have experienced an event but have not completed follow-up of *R*. For each treatment group *i* (0=standard, 1=experimental) at visit time t_j, let m_{ij} be the number of subjects at risk (i.e. have completed visit at t_j without having an event), and let e_{ij} be the number of new events diagnosed. Then the event proportion in treatment group *i* at the time of interim analysis τ_1 is given by:

$$\hat{\pi}_i(\tau_1) = \frac{\sum_{k=1}^{J} e_{ik}}{m_{ir} + \sum_{k=1}^{J} e_{ik}}$$

The individuals who have experienced an event but have not completed duration *R* of follow-up are included in the numerator and the denominator.

2.2 Method 2: Event proportion based on data from subjects that have been randomized by the time of the interim analysis

This simple approach uses data from the subjects randomized by the time of the interim analysis τ_1 (i.e. once 50% of the subjects have been followed for time *R*). Let n_i be the number of subjects who have been randomized to treatment group *i*. Then the event proportion for each group at the time of interim analysis τ_1 is given by

$$\hat{\pi}_i\left(\tau_1\right) = \frac{\sum_{k=1}^{J} e_{ik}}{n_i}$$

which is simply the total number of observed events divided by the number of subjects randomized by τ_1.

2.3 Method 3: Kaplan-Meier Approach

This approach also uses all the data available at the time of the interim analysis τ_1 (i.e. once 50% of the subjects have been followed for time *R*). For individuals who have not completed follow-up time *T* (i.e. the full fixed follow-up duration) and have not had the event, they are simply right-censored at the latest time that they were observed. Then the Kaplan-Meier (KM) estimates can be calculated using all randomized subjects and the event proportion in treatment group *i* at the time of interim analysis τ_1 is given by

$$\hat{\pi}_i\left(\tau_1\right) = 1 - S_i\left(T\right)$$

where $S_i(T)$ is the KM survivor function estimate. Following the notation in **Table 1**, this is equivalent to

$$\hat{\pi}_i(\tau_1) = 1 - \prod_{k=1}^{J}(1 - d_{ik}).$$

We evaluated these methods in terms of overall type I and II errors and the probability of early stopping of the trial for a positive result at the interim. The interim analysis was performed using the Haybittle-Peto [1] and O'Brien-Fleming [4] monitoring boundaries for extreme positive results. These boundaries are conservative and require small p-values for early stopping of the trial. Other less conservative boundaries such as the Pocock approach were not evaluated [12, 13].

2.4 Simulation

We considered six RCTs similar to the trial described in the methods section (see **Table 2**). Data for the binary endpoint were generated using the binomial distribution under the null and alternative hypotheses.

For each subject with an event, the time at which the event occurred was randomly assigned to reflect five clinically-plausible scenarios (**Table 3**), using the following: 1) events were distributed equally across the four time-points with probabilities (0.25, 0.25, 0.25, 0.25) for both groups; 2) the majority of the events occurred in the first two time-points with probabilities (0.35, 0.30, 0.20, 0.15) for both groups; 3) the majority of the events occurred in the last two time-points with probabilities (0.15, 0.20, 0.30, 0.35) for both groups; 4) the standard group follows distribution (3) and the experimental group

follows distribution (2); and 5) the reverse of scenario (4). Entry times for subjects over 48 months were randomly generated from a uniform distribution, and the interim analysis was carried out after 50% of the subjects completed R=12 months of follow-up. We carried out 10,000 replications for each trial. Given that $Z(x)$ and $Z(y)$ are the interim and final test statistics, respectively, the type I error rate,

$P_{H_o}\left(Z(x) > g \text{ or } \left[Z(x) \le g \text{ and } Z(y) > f\right]\right)$, and the type II error,

$P_{H_1}\left(Z(x) \le g \text{ and } Z(y) \le f\right)$, were obtained from data generated under the null and alternative hypotheses, respectively, where g and f are the interim and final critical values of the O'Brien-Fleming (g = 2.797, f = 1.977) and Haybittle-Peto (g = 3.0, f = 1.967) monitoring boundaries. The probability of early stopping, $P_{H_1}\left(Z(x) > g\right)$, was obtained under the alternative hypotheses. All analysis was performed in R 2.15 (www.r-project.org).

3. Results

The results of the type I error rates for the three methods are shown graphically in **Figure 2**. The three methods have comparable type I error rates across each of the trials and event distribution scenarios. The methods in general have nominal or close-to-nominal type I error rates when the event distribution probabilities are equivalent between treatment groups or when the experimental treatment group events occurred earlier in the trial compared with the standard group. However, under these same scenarios, slightly greater-than-nominal type I error rates are seen in the trials where (π_0, π_1) = (0.30, 0.10) and (π_0, π_1) = (0.50, 0.45), where the type I error rates are approximately 0.03. For the

scenario where the experimental group events occurred later in the trial compared with the standard group, the type I error was generally inflated for all methods.

The three methods also have comparable type II error rates (**Figure 3**). In general, under all event distribution scenarios and trials, the type II error rates are comparable to the nominal value of 0.10 regardless of the interim analysis method or stopping boundary rule. Moreover, in the scenario where the experimental group events occurred later in the trial compared with the standard group, the type II errors rates are much lower than the nominal value for the trials with ARRs of 0.05 and 0.10.

Under the alternative hypothesis, methods 1 and 3 have comparable probabilities for early stopping in scenarios where the treatment groups have equivalent event distributions probabilities over time, specifically in the trials where $\pi_0 = 0.30$ (**Figure 4**). Method 3 has a slightly greater probability of early stopping than method 1 in the trials where $\pi_0 = 0.50$. Moreover, method 2 has the smallest probability of early stopping in scenarios where the treatment groups had equivalent event distributions probabilities over time. On the other hand, all methods have comparable probabilities of early stopping in the scenarios where the treatment groups had contrasting event distributions over time. The highest probabilities for early stopping are seen in the trials where the experimental group had a smaller proportion of events occur earlier in the trial compared with the standard group, and the lowest probabilities of early stopping are seen in the opposite scenario. In general, the probability for early stopping is greater using the O'Brien-Fleming boundaries compared with the Haybittle-Peto monitoring boundaries.

4. Discussion

In RCTs with binary endpoints, interim analyses are generally conducted after a considerable percentage of subjects have completed follow-up. However, under certain situations this approach is not optimal since the trial may have completed accrual and all the subjects will have been treated by that time. We evaluated three approaches for an interim analysis when a considerable percentage of subjects complete a follow-up time that is less than the planned trial follow-up.

We observed that the type I error rates were comparable for all three methods. For most trials simulated, under the scenarios where the event distributions were equivalent between treatment groups or the experimental group had events occur earlier than the standard group, the type I error rates were close to the nominal value. These results concur with those of Pedley [7], who showed that conducting the interim analysis after a considerable percentage of subjects had completed full follow-up (using method 2) produced nominal type 1 error rates, albeit in the situation where events could be measured at any time during follow-up and not just at specific time points. However, we also observed that the type I error rate increased with increasing absolute risk reduction for trials with a standard group event proportion of 0.3, thus resulting in slightly higher type I error rates for the trial with ARR to 0.20. In addition, similar slightly higher type I error rates were seen in the trial with a standard group event proportion of 0.5 and the ARR=0.05. This is perhaps due to a combination of less variability and a small sample

size for the former, and a large sample size and small ARR for the latter. Therefore, trialists should be cautious of using either of these methods under these situations.

While there were situations in which the type I errors were slightly inflated with all methods, the methods performed much better with regard to the type II errors under all scenarios, suggesting that these methods will not have a negative effect on the power to detect the hypothesized difference between treatment groups provided the difference exists. Under the scenarios where the experimental group had events occur later compared with the standard group, the methods showed increased overall power because the probability of early stopping was greater in these scenarios. However, under these scenarios, the type I error rates are inflated.

The methods differed on the probability of early stopping under the alternative hypothesis with method 2 having the lowest probability. This is because this approach includes data from all subjects that have been randomized subjects by the time of the interim analysis in the denominator of the estimation of the event proportion even though a subgroup of these patients would not have had any assessment of the outcome since they would not have reached their first time point for outcome assessment. The consequence is the dilution of the interim treatment effect leading to lower interim power. Method 3 also uses all available data from randomized subjects at the time of the interim analysis. However, it employs a conditional probability approach which differentiates between those subjects who have not yet had an assessment visit (i.e. censored) and who are at risk at each assessment visit, thus yielding a greater probability of early stopping. Similarly,

since method 1 uses only a subset of randomized subjects at the time of the interim analysis, the estimated interim treatment effect is less diluted and, therefore, has greater probability for early stopping than method 2. Conversely, since it uses a smaller number of subjects compared with method 3, the probability for early stopping is slightly lower than method 3 in trials where the standard group event proportion is 0.5, because the variability is greater for proportions closer to 0.5. Furthermore, we observed that the probabilities for early stopping are greater using the O'Brien-Fleming boundary compared with the Haybittle-Peto boundary since it is less conservative.

Although the largest probabilities of early stopping under the alternative hypothesis and the smallest type II errors were seen under the scenario where the experimental group had events occurring later compared with the standard group, the type I errors is greatly inflated and, therefore, none of the methods can be recommended in this situation. Since there is a delay in occurrence of the event in the experimental group, this may be perceived as an effect of treatment. However, in situations where investigators are interested in the occurrence of an event over a fixed time period, this scenario, although rare, would still be considered under the null hypothesis.

Our study had some limitations. The generalizability of our findings may be limited since we evaluated six trial scenarios with particular event distributions over time. In diseases where the event distributions over time differ from the ones evaluated in this research, further simulations would be required to evaluate these methods. Secondly, we evaluated trials with one interim analysis after 50% of the subjects completed 12 months of follow-

up using the O'Brien-Fleming or Haybittle-Peto approach. These findings may not be applicable to trials in which interim analyses are required at multiple times or when using the alpha spending function approach to monitor the trial. Finally, the biases of the interim event proportions and treatment effects were not evaluated primarily because it is well known that estimators at the interim are biased, especially for estimators that allow for early stopping for positive results. However, further investigation on the estimators is needed.

5. Conclusion

Nonetheless, we have shown that under certain scenarios, conducting an interim analysis when a considerable number of subjects have some follow-up data, using any of the methods, preserves the type I and II errors. Although all three methods preserve type I and II errors under these scenarios, we recommend using the Kaplan-Meier method because it incorporates all the available data and has greater probability of early stopping when the treatment effect exists. We have also shown that under certain scenarios, none of these methods is suitable for an interim analysis, and trialists should be cautious when using them. Finally, when possible, an interim analysis should be undertaken when data from a considerable number of subjects who have completed full follow-up are available. However, if waiting for a considerable number of subjects to complete full follow-up is not an efficient approach, such as in the examples described, the methods outlines in this paper should be considered and evaluated to fit the specific needs of the trial.

References

1. Haybittle JL: Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971, 44:793-797.

2. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976, 34:585-612.

3. Pocock S: Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977, 64:191-199.

4. O'Brien P, Fleming T: A multiple testing procedure for clinical trials. *Biometrics* 1979, 35:549-556.

5. Enden T, Haig Y, Klow NE, Slagsvold CE, Sandvik L, Ghanima W, Hafsahl G, Holme PA, Holmen LO, Njaastad AM, Sandbaek G, Sandset PM, CaVenT Study Group: Long-term outcome after additional catheter-directed thrombolysis versus standard treatment for acute iliofemoral deep vein thrombosis (the CaVenT study): a randomised controlled trial. *Lancet* 2012, 379:31-38.

6. Vedantham S, Goldhaber SZ, Kahn SR, Julian J, Magnuson E, Jaff MR, Murphy TP, Cohen DJ, Comerota AJ, Gornik HL, Razavi MK, Lewis L, Kearon C: Rationale and design of the ATTRACT Study: a multicenter randomized trial to evaluate pharmacomechanical catheter-directed thrombolysis for the prevention of postthrombotic syndrome in patients with proximal deep vein thrombosis. *Am Heart J* 2013, 165:523-530

7. Pedley A: Applying survival analysis techniques to interim analysis and sample size reassessment of clinical trials with dichotomous endpoint. *Boston University*, 2011

8. Togo K, Iwasaki M: Optimal timing for interim analyses in clinical trials. *J Biopharm Stat* 2013, 23:1067-1080.

9. Marschner IC, Becker SL: Interim monitoring of clinical trials based on long-term binary endpoints. *Stat Med* 2001, 20:177-192.

10. Sooriyarachchi MR, Whitehead J, Whitehead A, Bolland K: The sequential analysis of repeated binary responses: a score test for the case of three time points. *Stat Med* 2006, 25:2196-2214.

11. Whitehead A, Sooriyarachchi MR, Whitehead J, Bolland K: Incorporating intermediate binary responses into interim analyses of clinical trials: a comparison of four methods. *Stat Med* 2008, 27:1646-1666.

12. Pocock SJ: When (not) to stop a clinical trial for benefit. *JAMA* 2005, 294:2228-2230.

13. Freidlin B, Korn EL: Stopping clinical trials early for benefit: impact on estimation. *Clin Trials* 2009, 6:119-125.

Table 1. Notation table for estimation of event proportions

Visit Number j	Visit Time t_j	Subjects at Risk m_j	New Events e_j	Incidence at Visit j d_j
0	t_0 (<6m)	m_0	$e_0 = 0$	$d_0 = 0$
1	t_1 (6m)	m_1	e_1	$d_1 = e_1 / m_1$
2	t_2 (12m)	m_2	e_2	$d_2 = e_2 / m_2$
3	t_3 (18m)	m_3	e_3	$d_3 = e_3 / m_3$
4	t_4 (24m)	m_4	e_4	$d_4 = e_4 / m_4$

Table 2. Summary of six trials considered for simulation with $\beta = 0.10$ and a one-sided $\alpha = 0.025$

Standard Group Event Proportion (π_0)	Experimental Group Event Proportion (π_1)	Absolute Risk Reduction $(\pi_0 - \pi_1)$	N
0.30	0.25	0.05	3342
0.30	0.20	0.10	796
0.30	0.10	0.20	160
0.50	0.45	0.05	4182
0.50	0.40	0.10	1030
0.50	0.30	0.20	242

Table 3. Summary of the event distribution probabilities for the simulated scenarios

Scenario	Event Distribution Probabilities by Visit Time t_1, t_2, t_3, t_4	
	Standard Group	Experimental Group
1	0.25, 0.25, 0.25. 0.25	*same as standard*
2	0.35, 0.30, 0.20, 0.15	*same as standard*
3	0.15, 0.20, 0.30, 0.35	*same as standard*
4	0.15, 0.20, 0.30, 0.35	0.35, 0.30, 0.20, 0.15
5	0.35, 0.30, 0.20, 0.15	0.15, 0.20, 0.30, 0.35

Figure 1. Plot showing the follow-up time in months for 10 subjects and the proposed time for the interim analysis after 5 (50%) subjects have completed 12 months of follow-up

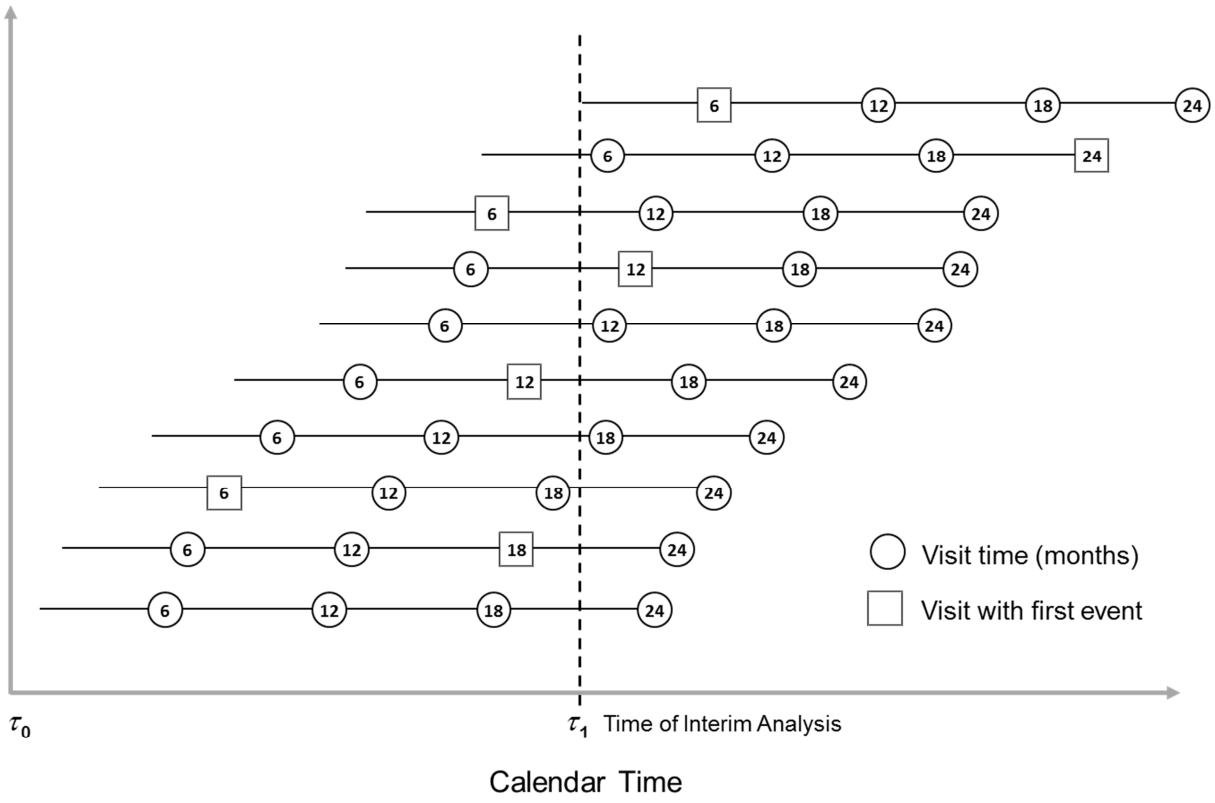


Figure 2. Overall type I error rates for each trial by event distribution scenario

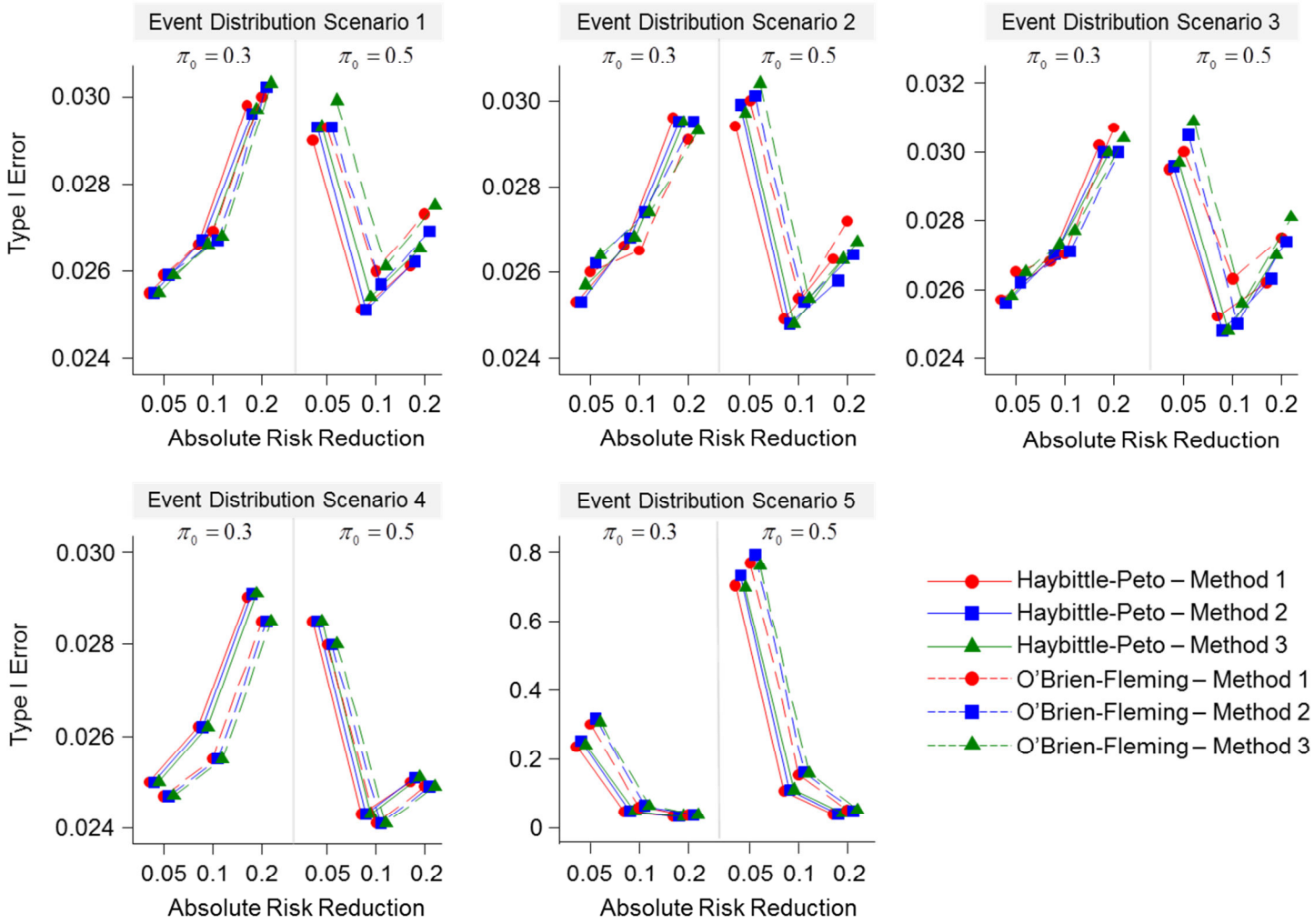


Figure 3. Overall type II error rates for each trial by event distribution scenario

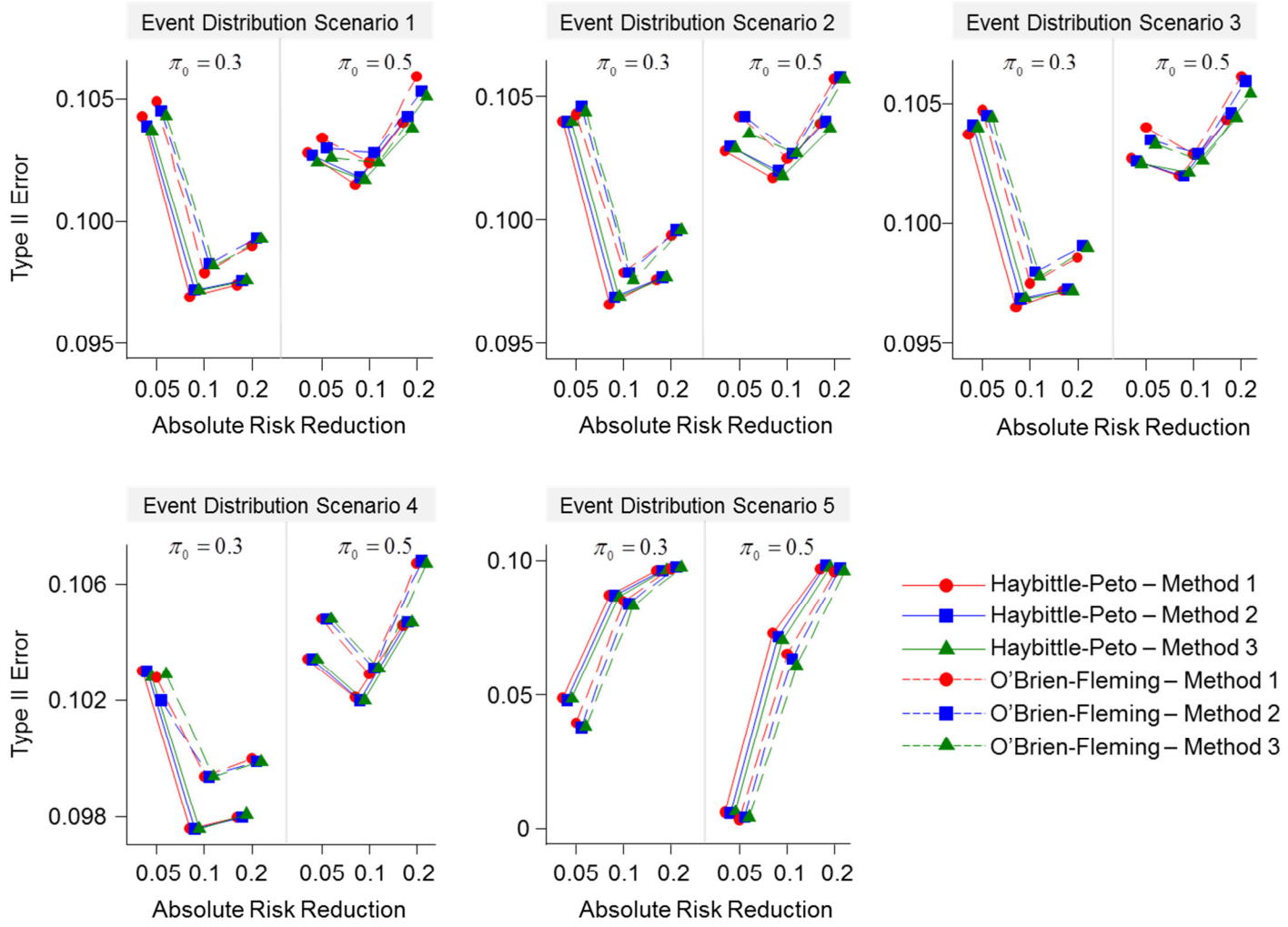
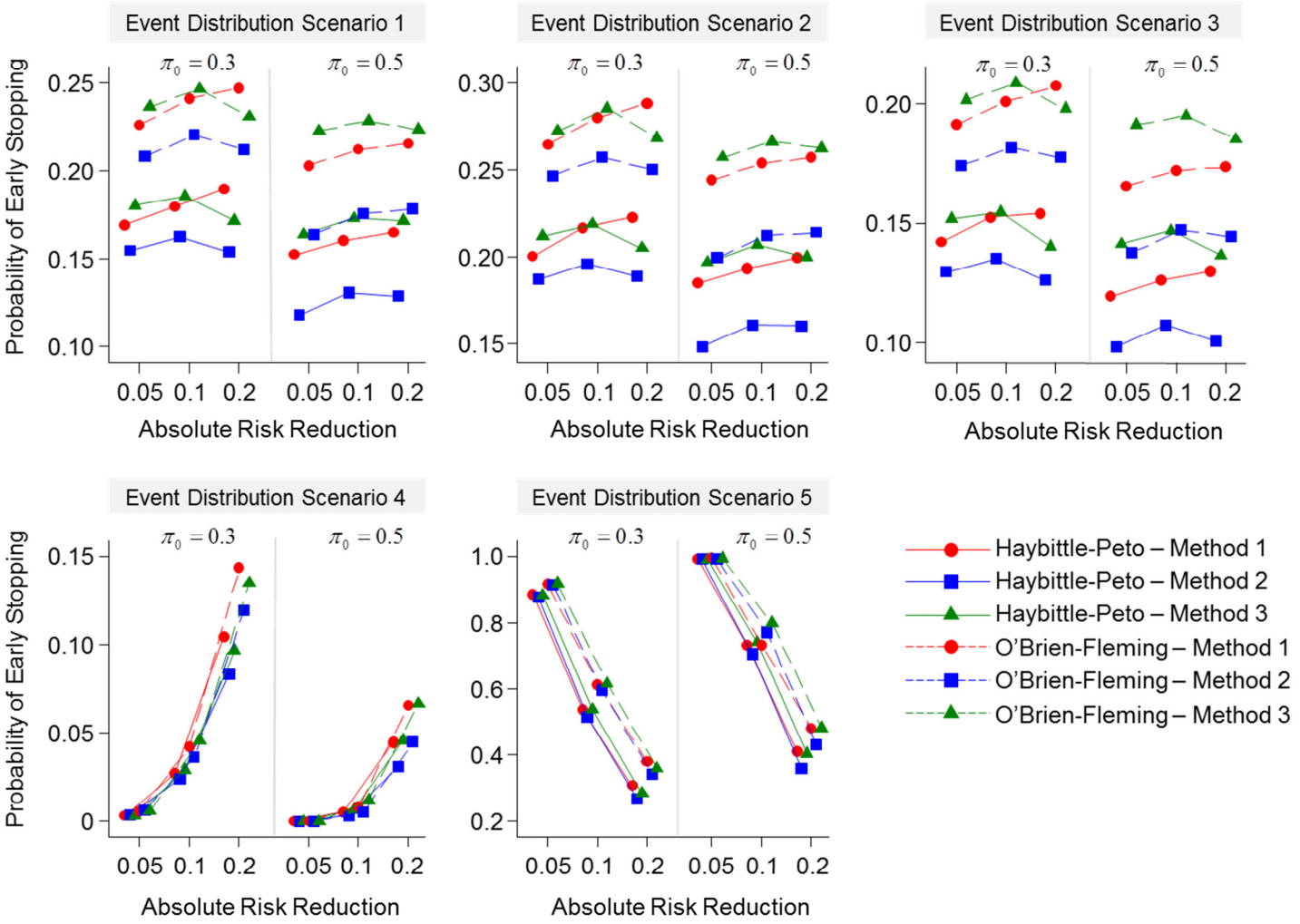


Figure 4. Probabilities for early stopping under the alternative hypothesis for each trial by event distribution scenario



CHAPTER 4

TREATMENT CROSSOVERS IN TIME-TO-EVENT NON-INFERIORITY RANDOMIZED TRIALS OF RADIOTHERAPY IN SUBJECTS WITH BREAST CANCER

S. Parpia[a], J. A. Julian[a], L. Thabane[bc], C. Gu[a], T.J. Whelan[ad] and M. N. Levine[ad]

a Ontario Clinical Oncology Group, Department of Oncology, McMaster University, 711 Concession Street – G (60) Wing 1st Floor, Hamilton, ON, Canada. L8V 1C3

b Centre of Evaluation of Medicines, St Joseph's Healthcare - Hamilton, 50 Charlton Avenue East , Hamilton, ON, Canada. L8N 4A6

c Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON Canada

d Juravinski Cancer Centre, 699 Concession Street, Hamilton, ON, Canada. L8V 5C2

Abstract

Background: In non-inferiority trials of radiotherapy in subjects with early stage breast cancer, it is inevitable that some subjects will cross over from the experimental to standard arm prior to initiation of any treatment due to complexities in treatment planning or subject preference. Although the intention-to-treat (ITT) analysis is the preferred approach for superiority trials, its role in non-inferiority trials is still under debate. This has led to the use of alternative approaches such as the per-protocol (PP) analysis or the as-treated (AT) analysis, despite the inherent biases of such approaches.

Methods: Using simulations, we investigate the effect of 2%, 5% and 10% random and non-random crossovers prior to radiotherapy initiation on the ITT, PP, AT, and the combination of ITT and PP analyses with respect to type I error in trials with time-to-event outcomes. We also evaluate bias and standard error of the estimates from the ITT, PP and AT approaches.

Results: The AT approach had the best performance in terms of type I error, but was anti-conservative as non-random crossover increased. The ITT and PP approaches were anti-conservative under all percentages of random and non-random crossover. Similarly, lowest bias was seen with the AT approach; however, bias increased as the percentage of non-random crossover increased. The ITT and PP had poor performance in terms of bias as crossovers increased.

Conclusions: If minimal crossovers were to occur, we have shown that the AT approach has the lowest type I error rates and smallest opportunity for bias. Results of trials with

high number of crossover should be interpreted with caution especially when crossover is non-random. Attempts to prevent crossovers should be maximized.

1. Introduction

The non-inferiority randomized trial design is frequently used to compare novel experimental breast radiation regimens with standard breast irradiation for the prevention of local recurrence in patients with breast cancer who have undergone breast conserving surgery. For example, hypofractionated radiotherapy that delivers a high dose of radiation per fraction and, therefore, requires a shorter duration of treatment resulting in greater convenience for the patient, has been compared with standard radiotherapy using a non-inferiority design [1-4]. The challenges in the design, conduct and analysis of such trials have been discussed by several authors [5-9]. These include the determination of the non-inferiority margin, and issues related to assay sensitivity, biocreep and the choice of the analysis population [5, 7, 10-14].

Typically, in breast cancer radiotherapy trials, prior to beginning treatment, the patient undergoes a planning process to establish the treatment fields to target the tumour and avoid radiating normal tissue. Such planning generally occurs after randomization. Sometimes planning may reveal that it is not possible to deliver the experimental regimen and therefore the patient is treated with standard therapy. In some cases, after being randomized to experimental radiotherapy, the patient decides to be treated with standard radiotherapy instead. In a trial of 1234 women comparing hypofractionated radiotherapy to standard radiotherapy for the prevention of local recurrence, the crossover percentage was 1.2% [3]. Generally, in trials evaluating new experimental radiotherapy techniques

(that are not currently available as part of usual care), patients are not permitted to cross over from standard therapy to experimental therapy [15].

In randomized superiority trials, it is well established that the analysis should be performed based on the *intention-to-treat* (ITT) principle—which states that subjects are analyzed according to the group they were randomized to regardless of the treatment they received. An ITT analysis tends to produce diluted treatment effect estimates and therefore is considered a conservative approach in analysis of superiority trials, but is anti-conservative in demonstrating non-inferiority [16]. This has led to the use of a *per-protocol* (PP) analysis where the analysis set consists only of subjects who fully comply with their assigned treatment [11], or an *as-treated* (AT) analysis which groups subjects according to the treatment they actually received [12], despite the inherent bias of such analyses [17, 18].

Several researchers have investigated the effect of non-compliance on the ITT and PP analyses in non-inferiority trials with binary or continuous outcomes [11, 19-22]. This research has focussed mainly on issues such as drop outs, missing data and treatment discontinuations. Literature on the effect of crossover bias in non-inferiority trials is limited. Sheng and Kim showed that both the ITT and PP can be biased in trials with binary outcomes [23]. Matsuyama studied the effect of crossovers to the other treatment after initiating the assigned treatment in the time-to-event situation, and suggested that the PP analysis should not be used [16]. Similarly, others have studied the effect of switching treatments mid trial (i.e. after receiving at least some of their original allocated treatment)

in the context of superiority trials [24-29]. However, the effect of crossovers prior to treatment initiation such as in radiotherapy trials is unknown.

Some authors have suggested performing the analysis using the ITT and PP populations, and non-inferiority should only be concluded if the null hypothesis is rejected using both analyses [5, 22, 30]. The Committee of Proprietary Medicinal Products Points-to-Consider states that *'... similar conclusions from both the ITT and PP are required in a non-inferiority trial*' [31]. However, this has not been investigated comprehensively for trials with time-to-event outcomes with crossovers.

In this paper, we focus on non-inferiority trials of radiotherapy with a time-to-event outcome in subjects with early stage breast cancer. Using simulation, we investigate the effect of subject crossover from the experimental to standard therapy (prior to initiation assigned therapy) on the analysis using the ITT, PP, AT and the combination of ITT and PP analysis sets with respect to type I error, bias and standard errors (SE).

2. Methods

2.1 Intention-to-treat (ITT)

The ITT approach uses all subjects that were randomized to the study and analyzes them according to their assigned treatment group regardless of whether they actually received or complied with the treatment. This is advantageous because it preserves the integrity of randomization and therefore ensures that, on average, the treatment groups are comparable. In addition, since it includes all subjects, it helps prevent bias which may

occur when excluding subjects [12]. Furthermore, the ITT approach will produce results that are likely to be observed in the clinic [10]. However, the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use also states: '*in an equivalence or non-inferiority trial use of the full analysis set (ITT) is generally not conservative and its role should be considered very carefully*' [10].

2.2 Per-Protocol (PP)

The PP approach excludes subjects who have not completed their assigned treatment based on the study protocol. This approach aims to measure the 'pure' treatment effect by including only subjects who comply with the protocol and excludes those that have crossed over [32]. The use of the PP analysis in non-inferiority trials has increased because of the apparent anti-conservative nature of the ITT in this setting [11].

2.3 As-Treated (AT)

This approach analyses subjects according to the treatment they actually received, and not the treatment assigned. Therefore, crossover patients are included in the analysis and are grouped with the treatment arm to which they crossed over.

2.4 Combination of ITT and PP Analyses

This combined approach requires that both the ITT and PP analyses are performed and that the null hypothesis is rejected (i.e. declaring non-inferiority) only if both analyses reject the null hypothesis.

2.5 Hypotheses and Assessing Non-inferiority

Following D'Agostino et al [5], let λ_E and λ_S represent the constant hazard rates for the experimental and standard therapy respectively, and $\theta = \lambda_E / \lambda_S$ is the hazard ratio (HR). Let M be the non-inferiority margin, that is, the maximum tolerable amount by which λ_E can be worse than λ_S ($M > 1$). Then the null and alternative hypotheses are:

$$H_0 : \theta = \lambda_E / \lambda_S \geq M$$
$$H_1 : \theta = \lambda_E / \lambda_S < M$$

To test the hypothesis of non-inferiority, we compute the $100(1-2\alpha)\%$ confidence interval (CI) for $\hat{\theta}$. If the upper bound of the CI is less than M, then we can conclude that the experimental therapy is no worse than the standard therapy by a maximum of M, and hence is non-inferior to the standard therapy at a significance level of α.

2.6 Simulation

The 5-year local recurrence rates following radiotherapy in women with early stage breast cancer who have undergone breast conserving surgery is approximately 5%, or $\lambda_S = -\log(0.95)/5$ [3]. In recent trials of radiotherapy in women with breast cancer, non-inferiority margins of 1.5 and 1.7 have been used [2, 3]. Also, it is recommended that a one-sided $\alpha = 0.025$ be used for non-inferiority studies [10, 31]. We considered two non-inferiority trials with a 5-year local recurrence rate of 5%, a one-sided $\alpha = 0.025$, 90% power, 4 years of accrual and an additional 3 years of follow-up. Based on these parameters, we calculated total sample sizes of 5134 and 3004 for trials with non-inferiority margins of 1.5 and 1.7 respectively assuming a 1:1 allocation ratio [33].

Subjects undergoing breast conserving therapy have varying risk of recurrence, and therefore, we considered two risk subgroups (high, low) and assumed that the HR of high vs. low risk to be 1.4 similar to that of a grade III vs. grade I/II tumours [34]. In addition, we assumed that 20% of the subjects were high risk and 80% were low risk.

To evaluate the type I error rates, data were simulated under the null hypothesis where E is inferior to S with a true HR of θ equal to the non-inferiority margin ($\theta = M$). For each trial, we simulated data for two randomly generated treatment groups of equal size. For each subject, the baseline covariate of risk (high vs. low) was generated using the binomial distribution with probability 0.2. Survival times were generated using the formula [35]:

$$T = -\frac{\log(U)}{\lambda_S \exp(\beta' x)}$$

where U is a random variable following a uniform distribution on the interval from 0 to 1, $\lambda_S = -\log(0.95)/5$ is the baseline hazard function, β is the vector of the regression coefficients and x is the vector of covariates. The regression coefficient for treatment (E vs. S) was log(M) and log(1.4) for the risk (high vs. low) covariate. Enrolment times were generated using the uniform distribution from 0 to 4 corresponding to 4 years of accrual. Subjects were censored at the end of the trial if they remained event-free and that time.

We evaluated scenarios where the percentage of subjects that crossed over from experimental to standard therapy were 2%, 5% and 10%. For each of these, we considered the following situations: (a) the crossover of subjects was random, and (b) the

high risk subjects were more likely to crossover (i.e. non-random). This was simulated assuming 50% of the crossover subjects were high risk patients.

For each approach, computation of the 100(1-2α) CI of the estimated HR, $\hat{\theta}$, was done using the Cox proportional hazards (Cox-PH) model with α=0.025. We carried out 10,000 replications for each trial giving a standard error of the estimate of type I error of 0.15%. Type I error was calculated as the proportion of trials that had the null hypothesis of inferiority rejected i.e. the proportion of trials in which the upper CI was less than *M*. Bias for each of the ITT, PP and AT analyses was calculated as the percentage difference between $\hat{\theta}$ and θ, and averaged over the number of simulations. The SE was also averaged over the total number of simulation. All analysis was performed in R 3.0 (www.r-project.org).

3. Results

3.1 Impact on Type I Error

The results of the type I errors for the four approaches are shown in **Table 1**, and graphically in **Figure 1**. The results showed that the AT approach had the best performance with Type I errors closer to nominal for 2% and 5% crossovers, 0.028 and 0.027 respectively (**Figure 1a**). We observed that the combined ITT+PP approach performed better than the separate ITT and PP analyses, and that the ITT and PP approaches had comparable overall type I errors. However, these approaches had type I

errors greater the nominal value regardless of the crossover percentage. In general, overall type I errors increased as the crossover percentage increased for all approaches.

For scenarios with random crossover (**Figure 1b**), the AT approach had nominal or close to nominal type I errors for all crossover percentages. The ITT+PP approach had close to nominal type I error when random crossover was 2%, but performed poorly as the random crossover percentage increased. The individual ITT and PP approaches had greater than nominal type I errors under all scenarios of random crossover.

Under non-random crossover scenarios (**Figure 1c**), all approaches performed poorly irrespective of the crossover percentage with the exception of the AT analyses when the true HR was 1.5 and crossover was 2% (**Table 1**). In general, the PP approach had the worst performance under scenarios of non-random crossover.

3.2 Impact of Bias

The AT approach also had the best performance in terms of overall bias of the HR estimates, whereas the ITT and PP approaches perform similarly (**Figure 2a**). As the percentage of crossover subjects increased, the overall percent bias also increased for all approaches. When the crossovers were random (**Figure 2b**), the AT approach had comparable bias across all levels of crossover percentages, whereas the ITT and PP approaches had greater bias as the crossover percentage increased. The ITT and PP approaches behaved similarly under the random and non-random scenarios, but their bias was larger under the non-random crossover scenario with the PP approach having larger

bias compared with the ITT (**Figure 2c**). Similar to the ITT and PP approaches, the AT approach also showed increased bias as non-random crossover increased, albeit with smaller bias. For each approach, bias was greater when the true HR was 1.7 compared with 1.5.

3.3 Impact on Standard Error

The three approaches had comparable overall SEs across all scenarios (**Table 1**). However, a slight trend was observed where the AT approach had the smallest SEs, followed by the ITT, and then the PP approach. Furthermore, we observed that for each approach, the SEs were comparable under scenarios of random and non-random crossovers. SEs were greater for the trials where the true HR was 1.7 compared with 1.5.

4. Discussion

In randomized non-inferiority trials of radiotherapy regimens in women with early stage breast cancer, it is inevitable that some subjects will cross-over from the experimental arm to the standard arm prior to treatment initiation due to complications in experimental radiotherapy planning or, subject or physician preference, or. In such situations, the ideal population for the final analysis is unclear. We evaluated the performance of the ITT, PP, AT and combined ITT+PP approach under various crossover scenarios.

The AT approach had the best performance under all scenarios in terms of type I error rate. However, it can only be recommended for situations where the crossover is random. Subjects that crossed over had their hazard of outcome determined by that of the standard

group, and were analyzed accordingly by the AT approach. Considering this and the fact that crossover was random, it is not surprising that the AT approach had near nominal type I error rates under these situations.

Moreover, the combined ITT+PP approach performed better than the ITT and PP approaches separately. This is due to the fact that the ITT+PP approach requires both analyses to reject the null hypothesis prior to non-inferiority being concluded, hence adding an extra level of testing compared to individual ITT and PP approaches, and therefore making it *harder* to conclude non-inferiority. Interestingly, neither the ITT nor the PP approach can be recommended under simulated scenarios, adding to the literature that both approaches could provide increased erroneous results [16, 23].

We also observed that the AT approach had the lowest bias of the HR estimate across all crossover percentages. Moreover, the biases of the ITT and PP approaches were comparable across all scenarios. For all three approaches, the bias is in the negative direction, and generally increases as the crossover percentage increases, except for the AT approach under the random crossover scenarios where it is not affected by the percentage crossover. Reasons for this observation are similar to that of its performance in terms of type I error under the same scenarios.

The biases for all methods are larger in scenarios where the true HR is larger because this reflects a greater hazard of event in the experimental arm. Therefore, the crossover subjects have a greater impact on the estimated HR, driving it closer to the null than in

situations where the true HR is smaller.

Since the assessment of non-inferiority is based on the CI approach, a combination of greater bias in the negative direction and smaller SEs would yield a lower upper limit of the 95% CI which is more likely to fall within the non-inferiority margin. Therefore, it is no coincidence that in general the scenarios with the greater bias and smaller SEs corresponded to the scenarios with larger type I error rates. We observed that within each approach, the SEs were comparable for random and non-random crossover, but the bias was larger for non-random crossover suggesting the bias had a greater influence on the type I error rate when comparing non-random versus random crossover within each method.

Our study had some limitations. The generalizability of our findings may be limited since we studied trials with event rates that are pertinent to radiotherapy trials in subjects with early stage breast cancer. However, our methodology and results can be applied to other clinical settings where cross overs occur prior to initiation of treatment. In diseases where the event rates differ from the ones evaluated in this research, further simulations would be required to evaluate these approaches. Secondly, for simplicity we assumed that non-random crossover was based on a single covariate. However, non-random crossover can occur for several reasons and, depending on the reason for crossover, the hazards may also differ. We did not consider adjusting for baseline covariates in the analysis which may improve the estimation of the treatment effect. However, this is less likely in large RCTs. Finally, we did not evaluate the causal proportional hazards estimator [36] because

it is not readily available in standard statistical software.

The choice of analysis population for non-inferiority trials is a difficult issue. We have shown that the AT approach preserves type I error under scenarios of random crossover. However, it is difficult to prove that crossover is random, and therefore assuming random crossover may not be appropriate leading to concerns about the validity of the inference test. Moreover, the PP approach which excludes patients is likely to disturb the prognostic balance achieved by randomization which can also cause erroneous trial results. The advantage of the ITT approach is that it preserves the advantages of randomization, and mirrors what will happen in practice and therefore is pragmatic. On the other hand, it can be anti-conservative in situations where crossover is high. In our experience, the crossover percentage in radiotherapy trials in subjects with early stage breast cancer is less than 2%, and we have shown that the AT and combined ITT+PP approaches are better at handling crossovers than the ITT and PP approaches.

5. Conclusion

The design, conduct and analysis of non-inferiority trials should be performed with extra rigour and to the highest standards. Every effort should be made to minimize the number of crossovers. If minimal percentage of crossovers were to occur, we have shown that the AT approach had the lowest type I error rates and smallest bias. A sensitivity analysis using the combined ITT+PP approach may also be warranted. In addition, both the ITT and PP results should be reported with details of the subjects who crossed over.

References

1. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P, Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial A of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet Oncol* 2008, 9:331-341.

2. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bentzen SM, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P, Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial B of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet* 2008, 371:1098-1107.

3. Whelan TJ, Pignol JP, Levine MN, Julian JA, MacKenzie R, Parpia S, Shelley W, Grimard L, Bowen J, Lukka H, Perera F, Fyles A, Schneider K, Gulavita S, Freeman C: Long-term results of hypofractionated radiation therapy for breast cancer. *N Engl J Med* 2010, 362:513-520.

4. Olivotto IA, Whelan TJ, Parpia S, Kim DH, Berrang T, Truong PT, Kong I, Cochrane B, Nichol A, Roy I, Germain I, Akra M, Reed M, Fyles A, Trotter T, Perera F, Beckham W, Levine MN, Julian JA: Interim cosmetic and toxicity results from RAPID: a randomized trial of accelerated partial breast irradiation using three-dimensional conformal external beam radiation therapy. *J Clin Oncol* 2013, 31:4038-4045.

5. D'Agostino RB S, Massaro JM, Sullivan LM: Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med* 2003, 22:169-186.

6. James Hung HM, Wang SJ, Tsong Y, Lawrence J, O'Neil RT: Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003, 22:213-225.

7. Fleming TR: Current issues in non-inferiority trials. *Stat Med* 2008, 27:317-332.

8. Fleming TR, Odem-Davis K, Rothmann MD, Li Shen Y: Some essential considerations in the design and conduct of non-inferiority trials. *Clin Trials* 2011, 8:432-439.

9. DeMets DL, Friedman L: Some Thoughts on Challenges for Noninferiority Study Designs. *Drug Info J* 2012, 12:420-427.

10. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials - E9. 1998

11. Garrett AD: Therapeutic equivalence: fallacies and falsification. *Stat Med* 2003, 22:741-762.

12. Wiens BL, Zhao W: The role of intention to treat in analysis of noninferiority studies. *Clin Trials* 2007, 4:286-291.

13. Brown D: Noninferiority Trials in Regulatory Guidance and Marketing Authorization Applications: Huge Advances Over the Last 20 Years but Problems Still to be Solved. *Stat Biopharm Res* 2013, 5:223-228.

14. James Hung HM, Wang SJ: Statistical Considerations for Noninferiority Trial Designs Without Placebo. *Stat Biopharm Res* 2013, 5:239-247.

15. Morden JP, Lambert PC, Latimer N, Abrams KR, Wailoo AJ: Assessing methods for dealing with treatment switching in randomised controlled trials: a simulation study. *BMC Med Res Methodol* 2011, 11:4-2288-11-4.

16. Matsuyama Y: A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial. *Stat Med* 2010, 29:2107-2116.

17. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB: Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* 1991, 10:1595-1605.

18. McNamee R: Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. Stat Med 2009, 28(21):2639-2652.

19. Rohmel J: Therapeutic equivalence investigations: statistical considerations. *Stat Med* 1998, 17:1703-1714.

20. Robins JM: Correction for non-compliance in equivalence trials. *Stat Med* 1998, 17:269-302.

21. Hauck W,W., Anderson S: Some Issues in the Design and Analysis of Equivalence Trials. *Drug Info J* 1999, 33:109-118.

22. Matilde Sanchez M, Chen X: Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Stat Med* 2006, 25:1169-1181.

23. Sheng D, Kim MY: The effects of non-compliance on intent-to-treat analysis of equivalence trials. *Stat Med* 2006, 25:1183-1199.

24. Robins JM, Tsiatis A: Correcting for non-compliance in randomized trials using rank-preserving structural failure time models. *Comm Stat - Theory and Methods* 1991, 20:2609-2631.

25. Law MG, Kaldor JM: Survival analyses of randomized clinical trials adjusted for patients who switch treatments. *Stat Med* 1996, 15:2069-2076.

26. Branson M, Whitehead J: Estimating a treatment effect in survival studies in which patients switch treatment. *Stat Med* 2002, 21:2449-2463.

27. Shao J, Chang M, Chow SC: Statistical inference for cancer trials with treatment switching. *Stat Med* 2005, 24:1783-1790.

28. Odondi L, McNamee R: Performance of statistical methods for analysing survival data in the presence of non-random compliance. *Stat Med* 2010, 29:2994-3003.

29. White IR: Uses and limitations of randomization-based efficacy estimators. *Stat Methods Med Res* 2005, 14:327-347.

30. Gomberg-Maitland M, Frison L, Halperin JL: Active-control clinical trials to establish equivalence or noninferiority: methodological and statistical concepts linked to quality. *Am Heart J* 2003, 146:398-403.

31. Committee on Proprietary Medical Products Point-to-Consider: Points to Consider on Switching Between Superiority and Non-inferiority, 2000.

32. Brittain E, Lin D: A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Stat Med* 2005, 24:1-10.

33. Jung SH, Kang SJ, McCall LM, Blumenstein B: Sample size computation for two-sample noninferiority log-rank test. *J Biopharm Stat* 2005, 15:969-979.

34. Voduc KD, Cheang MC, Tyldesley S, Gelmon K, Nielsen TO, Kennecke H: Breast cancer subtypes and the risk of local and regional relapse. *J Clin Oncol* 2010, 28:1684-1691.

35. Bender R, Augustin T, Blettner M: Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005, 24:1713-1723.

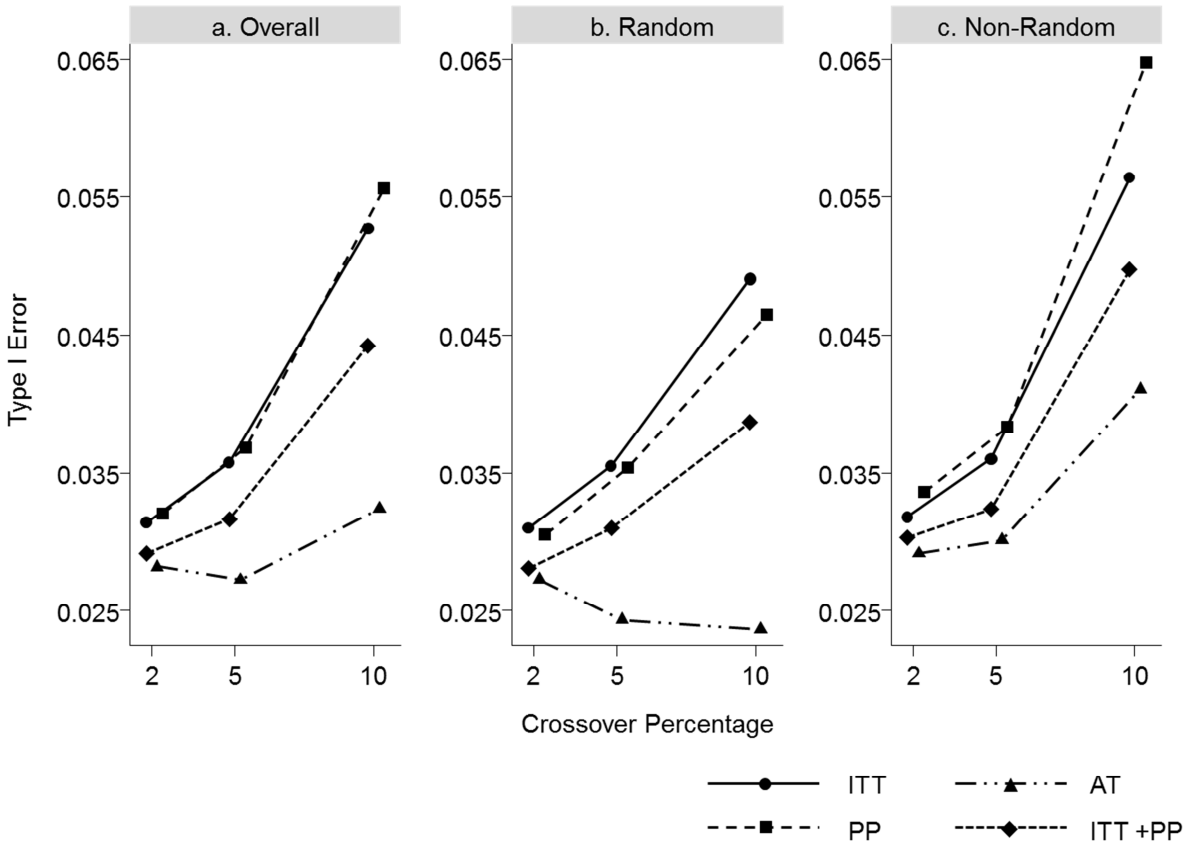
36. Loeys T, Goetghebeur E: A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics* 2003, 59:100-105.

Table 1. Results of type I error, bias and standard error for each approach by non-inferiority margin, crossover percentage and crossover type

True Hazard Ratio (NI Margin)	Cross-over (%)	Crossover Type	Type I Error				Bias (−%)			Standard Error		
			ITT	PP	AT	ITT+ PP	ITT	PP	AT	ITT	PP	AT
1.5	2	Random	0.0317	0.0312	0.0283	0.0288	1.1868	1.1935	0.5350	0.1106	0.1111	0.1103
		Non-random	0.0330	0.0354	0.0313	0.0316	1.1835	1.4029	0.8987	0.1107	0.1112	0.1103
	5	Random	0.0366	0.0367	0.0251	0.0319	2.3329	2.3343	0.6509	0.1108	0.1120	0.1101
		Non-random	0.0348	0.0367	0.0291	0.0313	2.1389	2.7165	1.4218	0.1109	0.1122	0.1100
	10	Random	0.0507	0.0464	0.0227	0.0394	4.1908	4.2561	0.6910	0.1112	0.1137	0.1100
		Non-random	0.0547	0.0653	0.0436	0.0483	4.5385	5.6278	3.0124	0.1112	0.1140	0.1099
1.7	2	Random	0.0302	0.0298	0.0262	0.0273	1.8692	1.8839	0.9979	0.1416	0.1421	0.1410
		Non-random	0.0306	0.0318	0.0270	0.0290	1.6345	1.8810	1.1791	0.1416	0.1422	0.1410
	5	Random	0.0345	0.0341	0.0236	0.0300	2.7744	2.8141	0.6719	0.1420	0.1434	0.1407
		Non-random	0.0373	0.0400	0.0311	0.0334	3.2014	3.7458	2.0289	0.1420	0.1436	0.1407
	10	Random	0.0474	0.0465	0.0246	0.0381	5.3500	5.4620	1.0507	0.1424	0.1454	0.1402
		Non-random	0.0581	0.0643	0.0386	0.0512	5.8370	6.8523	3.2413	0.1425	0.1458	0.1402

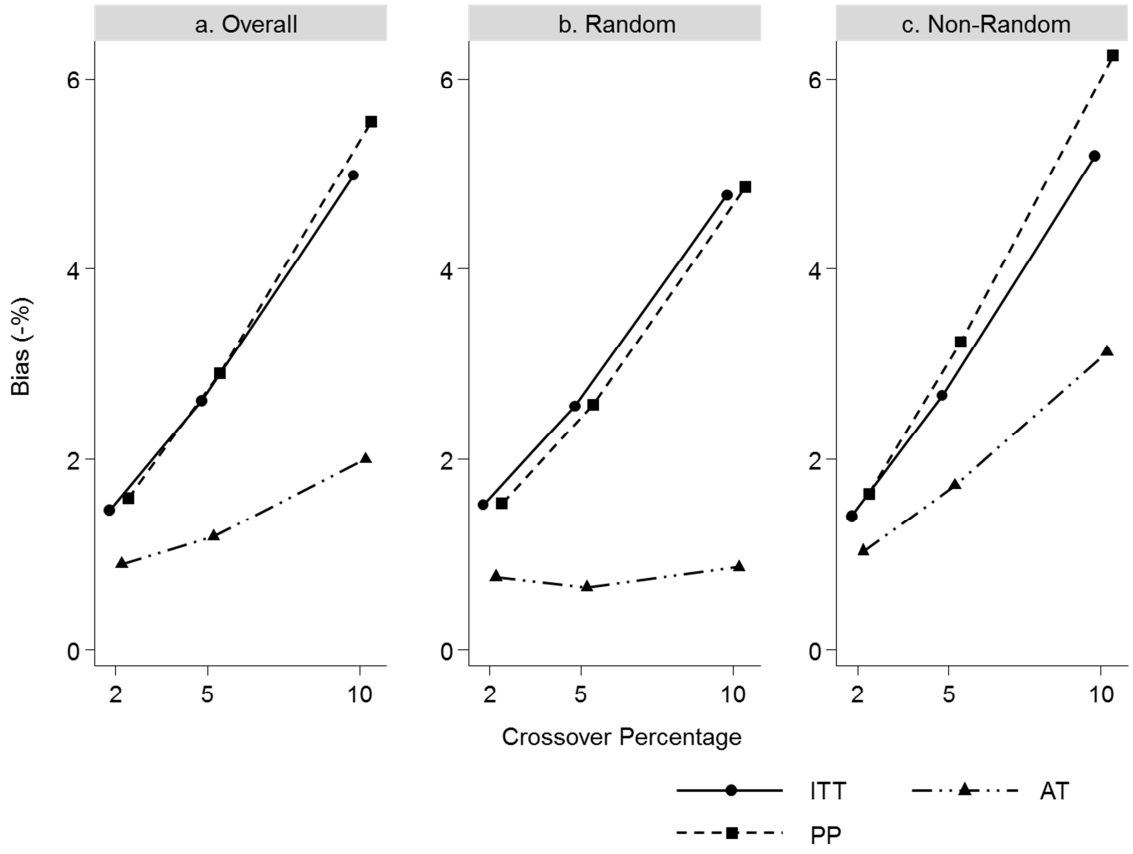
ITT = intention-to-treat; PP = per-protocol; AT = as-treated; ITT+PP = intention-to-treat and per-protocol combination

Figure 1. Type I error rates for the ITT, PP, AT and combined ITT+PP approaches by crossover type and percentage.



ITT = intention-to-treat; PP = per-protocol; AT = as-treated; ITT+PP = intention-to-treat and per-protocol combination

Figure 2. Bias for the ITT, PP and AT approaches by crossover type and percentage.



ITT = intention-to-treat; PP = per-protocol; AT = as-treated

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The non-inferiority randomized controlled trial (RCT) design has been used extensively in medical research to evaluate new therapies for the treatment of cancer and many other diseases. With the ever-increasing development of new technologies, the non-inferiority design has specifically played a prominent role in the evaluation of new radiotherapy techniques for the treatment of breast cancer in women diagnosed with early stage breast cancer [1-5].

This design, however, is not without controversy, with some authors even describing it as unethical [6]. Its main criticisms, though, stem from the many methodological issues that arise with its design and its analysis. During our experience with non-inferiority trials of radiotherapy techniques for the treatment of early stage breast cancer, we identified and focussed on a subset of these issues: 1) analyzing multiple time-to-event outcomes; 2) conducting an interim analysis of a binary outcome that is assessed repeatedly at pre-specified times; and 3) determining the optimal analysis population for dealing with subjects who cross over from the experimental to the standard arm. We have studied these important topics in a manuscript-based thesis, with a chapter dedicated to each of the issues. Although these issues are not limited to non-inferiority trials of radiotherapy techniques in breast cancer, many of the inputs for the simulations in this thesis have been extracted from such trials.

In Chapter 2, we studied four methods for analyzing multiple time-to-event outcomes in the context of a breast cancer non-inferiority trial where subjects were at risk for local recurrence, distant recurrence and death. These methods included the Cox proportional hazards (Cox-PH) model [7], the competing risk (CR) model [8], the marginal model [9] and the frailty model [10]. Although the primary outcome in this hypofractionation trial was local recurrence in the treated breast, the other events cannot be ignored, since they are related. For example, a subject is at higher risk of death once they have a distant recurrence. On the other hand, a subject cannot have a local or distant recurrence if they die. The Cox-PH model, which is the most commonly used method to analyze such data, ignores these relationships between events.

Our re-analysis of the previously reported hypofractionation trial, however, showed that the Cox-PH model was robust to the issue of multiple events in this specific trial. As well, the results of each of the Cox-PH, CR and event-specific marginal model were comparable, confirming that hypofractionation radiotherapy did not increase the risk of any of the events individually. Similarly, the results of both the overall marginal and the frailty model suggest that subjects treated with the hypofractionated radiotherapy had a comparable overall risk of any event compared with those treated by the standard radiotherapy technique.

However, our simulation results also showed that if the hazards of distant recurrence and death differ between treatment groups, the Cox-PH and CR models can yield conflicting results. The main reason for the contrasting results is in the methodology that each model

uses for handling competing events. The Cox-PH model censors competing events, whereas the CR model assumes that the subject is no longer at risk for an event once a competing event has occurred.

Some have suggested that designing the trial using a composite outcome such as disease-free survival (i.e. a combination of local recurrence, distant recurrence and disease-specific death) resolves the problem of competing events. As suggested by Friedlin and Korn [11], this may be the most meaningful outcome from a patient's perspective or from a general health perspective. However, irradiation of the breast after breast-conserving surgery is primarily aimed at preventing local recurrence and, therefore, radiation oncologists are generally interested in estimating the effect of treatment in preventing local recurrences without the background *noise* of other events that may occur but are not affected by the treatment. Moreover, since the combined event rates of distant recurrence and disease-specific death are much higher than local recurrence in this population, there exists the possibility that non-inferiority may be concluded using a disease-free survival outcome, but the experimental radiotherapy may still be inferior with respect to just local recurrence.

The overall marginal and frailty models, which showed similar results in our simulations, can also be used in a trial with multiple outcomes. The main argument against such an approach is that the models do not distinguish between event types and, therefore, assume that all events have the same biological etiology. In reality, the etiology of local and distant recurrences may be very different. Others have suggested using the marginal and

frailty models as a tool for assessing homogeneity of the treatment effects between the event types [12]. However, they also point out that the power of a trial to detect heterogeneity is extremely low.

A limitation of this project was the use of a latent failure times model to generate the joint distribution of failure times corresponding to the events. This approach generates data under the assumption that failure times exist for all events for each individual. This was not biologically plausible in the population studied. Subjects who died without experiencing a recurrence did not have a recurrence time and, consequently, the relationship structure between latent failure times could not be identified from observed data [13, 14]. Therefore, our assumptions for the correlations between local recurrence, distant recurrence and death used in our simulation were based on observed data from trials that do not directly fit the latent framework. However, we used only the latent model to generate data and none of its assumptions were used in the analysis. Further work is needed to assess these models under a framework where data is generated using other approaches that allow for additional correlation structures such as joint frailty survival models for recurrent events and a terminating event [15].

In Chapter 3, we examined the operating characteristics of three methods for estimating the interim event proportion for an interim analysis in RCTs with a binary outcome assessed repeatedly at pre-specified times and where the subject is considered to have experienced a failure at the first occurrence of the outcome. Our interest was motivated by trials where it is inefficient to wait until a considerable percentage of subjects have

completed their fixed follow-up duration in order to conduct an interim analysis, since there is a possibility that the trial will have completed accrual and all subjects have will have been treated prior to the interim analysis. We considered: 1) estimation of the event proportion based on subjects who have been followed for a pre-specified time (less than the full follow-up duration) or who experienced the outcome; 2) estimation of the event proportion based on all randomized subjects; and 3) the Kaplan-Meier approach to estimate the event proportion.

Our findings showed that in RCTs where waiting for half or more of the subjects to complete full follow-up prior to conducting an interim analysis is inefficient, performing an interim analysis when 50% of subjects have completed a portion of their fixed follow-up can be an effective strategy under certain scenarios where event distribution probabilities are equivalent between treatment groups. Moreover, under these scenarios, all three methods preserved type I and II errors and the Kaplan-Meier approach had the largest probability of stopping early when a treatment effect exists. Therefore, our results have shown that without inflating the type I error, the interim analysis can be performed earlier using the Kaplan-Meier approach and, if a statistically significant effect is found, the trial may be stopped and all future subjects would receive the experimental therapy. This approach not only protects future subjects in the trial from receiving an inferior treatment, but also avoids additional expenditure to complete the trial.

However, our findings also showed that none of the methods preserves type I error, and cannot be used when the experimental treatment delays the occurrence of the event. Some

may argue that the delay in the occurrence of the event is an effect of treatment and, therefore, cannot be considered under the null hypothesis. Furthermore, they may argue that a time-to-event analysis should be used in such a scenario. However, when clinicians are concerned with the occurrence of an event during a fixed follow-up period, regardless of when the event occurred, this scenario falls under the null hypothesis.

Our approach had limitations. We restricted our simulations to one interim analysis using either the O'Brien-Fleming [16] or Haybittle-Peto [17, 18] monitoring boundaries. Other boundaries such as Pocock's [19] were not evaluated because they are less conservative. In addition, our results cannot be generalized to trials that propose to have more than one interim analysis during the course of the trial. Furthermore, the alpha spending approach, which does not require either the number or the time of the interim analysis to be specified in advance, was also not evaluated as part of this study [20]. Further research is needed to address this issue.

The major challenge in proposing this type of interim analysis is in determining the event distribution probabilities for the experimental and standard treatments. Depending on the disease and the therapies being evaluated, the event distribution probabilities may vary and, therefore, the proposed approaches need to be assessed to fit the needs of the trial prior to being implemented. While the event distribution probabilities for the standard treatment group can be estimated from previous trials, estimation of the event distribution probabilities for a new therapy that has not been previously tested can be more challenging. As a consequence, during the design phase of a trial, incorrect specification

of the event distribution probabilities that does not mirror what subsequently occurs in the trial may provide error rates that do not match the error rates in the trial, creating a major flaw in the design of the trial. If, however, the event distribution probabilities can be estimated accurately, then the methods outlined in our proposed strategy can be used to undertake an interim analysis, provided they preserve the error rates when evaluated to fit the needs of the trial.

In Chapter 4, we studied analysis populations for non-inferiority trials of radiotherapy in women with early stage breast cancer in which subjects crossed over from the experimental arm to the standard arm prior to the initiation of therapy. Crossovers in drug trials can occur after a subject has initiated their assigned treatment (i.e. subjects can have both drugs during the trial). However, in trials comparing two radiotherapy techniques, crossovers generally occur prior to initiation of any treatment, and subjects are unable to cross over once they have had a dose of their assigned treatment. We evaluated the *intention-to-treat* (ITT), *per-protocol* (PP), *as-treated* (AT) and combined ITT+PP populations under scenarios of random and non-random crossovers in non-inferiority trials.

Our results showed that the AT population had the best performance in terms of type I error but it can be recommended only under the scenarios of random crossover, as type I error was generally not preserved when crossover was non-random. Furthermore, our results showed that both the ITT and PP populations can produce erroneous results under

scenarios of both random and non-random crossovers, with the PP population having the worst performance in the presence of non-random crossover.

This project had limitations. Firstly, determination of non-random crossover was based on a single binary covariate. In practice, crossovers can occur for multiple reasons and hazards of event may vary depending on the cause of crossover. Further research using a more complex model with multiple covariates and varying hazards is needed to investigate the effect of different causes of crossover. Secondly, the Loeys and Goetghebeur model which has the capacity to account for crossover irregularities was not evaluated [21]. This is primarily because this model is not readily available in standard statistical software.

Our research has added to the current literature on the optimal choice of analysis population for non-inferiority trials. We have shown that crossovers are a major problem especially when non-random, regardless of the analysis population. If crossovers are random, we could argue that the AT population is the optimal choice for analysis. However, determining whether crossovers are random or non-random is challenging. One could compare baseline characteristics of the crossover subjects with the non-crossover subjects to determine randomness. Even if none of these comparisons turns out to be statistically different, there is still no guarantee that the crossover was random since the crossover may be linked to an unmeasured variable. Therefore, it is extremely difficult to recommend a single analysis population for all non-inferiority trials.

Hence, the conduct of non-inferiority trials should be performed with extra rigour and to the highest standards, and attempts to prevent crossovers and other protocol deviations should be maximized. For radiotherapy trials, one strategy to avoid treatment planning complexity crossover would be in incorporate the double enrolment plan whereby a subject is first registered, and randomized only if they are able to receive both the experimental and standard therapy. Fleming [22] suggests that " *the preferred approach to enhancing the integrity and interpretability of the non-inferiority trial should be to establish performance standards for measures of quality of trial conduct when designing the trial, and then to provide careful oversight during the trial to ensure these standards are met*". Given this, he argues that that the ITT analysis which preserves the integrity of randomization should be the primary analysis with the AT and PP acting as supportive analyses [22].

Non-inferiority trials that evaluate new radiotherapy techniques will continue to play an important role in the treatment of early stage breast cancer. New radiotherapy technologies that reduce the treatment duration and are believed to have better toxicity profiles while being equally efficacious as current techniques are being developed continuously. It is critical that these treatments are rigorously evaluated through non-inferiority RCTs before implementing them into standard patient care. Hence, it is also vital that efficient approaches to the design and analysis of non-inferiority trials are developed. In this dissertation, we have studied and proposed recommendations with respect to the analysis of multiple events, early interim analysis and the optimal analysis population. Further research is needed to develop efficient designs of non-inferiority trials

that incorporate the relationships or correlations between multiple events rather than adjusting for these relationships at the analysis stage. Furthermore, research is needed on interim analysis approaches that are robust to varying event distribution probabilities between treatment groups. This may be accomplished by differentially weighting subjects that have been followed for a long period compared with those that have been followed for a shorter duration. Lastly, methodologies such as the double enrolment should be studied and implemented to reduce treatment crossover.

In summary, this PhD dissertation identified and investigated issues related to non-inferiority trials of radiotherapy for the prevention of local recurrence in women with early stage breast cancer that were encountered during my experience working at the Ontario Clinical Oncology Group. The three papers make contributions by exploring each issue using statistical simulations. Although the bases for the simulations were data from non-inferiority radiotherapy breast cancer trials, the issues investigated in this thesis are applicable to other treatment modalities and diseases. Therefore, this dissertation adds to the more general literature on non-inferiority RCT methodology.

References

1. Vicini FA, Chen P, Wallace M, Mitchell C, Hasan Y, Grills I, Kestin L, Schell S, Goldstein NS, Kunzman J, Gilbert S, Martinez A: Interim cosmetic results and toxicity using 3D conformal external beam radiotherapy to deliver accelerated partial breast irradiation in patients with early-stage breast cancer treated with breast-conserving therapy. *Int J Radiat Oncol Biol Phys* 2007, 69:1124-1130.

2. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P, Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial A of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet Oncol* 2008, 9:331-341.

3. START Trialists' Group, Bentzen SM, Agrawal RK, Aird EG, Barrett JM, Barrett-Lee PJ, Bentzen SM, Bliss JM, Brown J, Dewar JA, Dobbs HJ, Haviland JS, Hoskin PJ, Hopwood P, Lawton PA, Magee BJ, Mills J, Morgan DA, Owen JR, Simmons S, Sumo G, Sydenham MA, Venables K, Yarnold JR: The UK Standardisation of Breast Radiotherapy (START) Trial B of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial. *Lancet* 2008, 371:1098-1107.

4. Whelan TJ, Pignol JP, Levine MN, Julian JA, MacKenzie R, Parpia S, Shelley W, Grimard L, Bowen J, Lukka H, Perera F, Fyles A, Schneider K, Gulavita S, Freeman C: Long-term results of hypofractionated radiation therapy for breast cancer. *N Engl J Med* 2010, 362:513-520.

5. Olivotto IA, Whelan TJ, Parpia S, Kim DH, Berrang T, Truong PT, Kong I, Cochrane B, Nichol A, Roy I, Germain I, Akra M, Reed M, Fyles A, Trotter T, Perera F, Beckham W, Levine MN, Julian JA: Interim cosmetic and toxicity results from RAPID: a

randomized trial of accelerated partial breast irradiation using three-dimensional conformal external beam radiation therapy. *J Clin Oncol* 2013, 31:4038-4045.

6. Garattini S, Bertele' V: Non-inferiority trials are unethical because they disregard patients' interests. *Lancet* 2007, 370:1875-1877.

7. Cox DR: Regression models and life-tables. *J Royal Stat Soc Series B (Methodol)* 1972, 43:187-220.

8. Fine J, Gray RJ: A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Ass* 1999, 94:496-509.

9. Wei LJ, Glidden DV: An overview of statistical methods for multiple failure time data in clinical trials. *Stat Med* 1997, 16(8):833-9

10. Wienke A: *Frailty Models in Survival Analysis.* Boca Raton, FL: Chapman & Hall/CRC; 2010

11. Freidlin B, Korn EL: Testing treatment effects in the presence of competing risks. *Stat Med* 2005, 24:1703-1712.

12. Pogue JM: Testing for Treatment Heterogeneity between the Individual Outcomes within a Composite Outcome. *McMaster University* 2012

13. Prentice RL, Kalbfleisch JD, Peterson AV,Jr, Flournoy N, Farewell VT, Breslow NE: The analysis of failure times in the presence of competing risks. *Biometrics* 1978, 34:541-554.

14. Andersen PK, Keiding N: Multi-state models for event history analysis. *Stat Meth Med Res* 2002, 11:91-115.

15. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P: Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics* 2007, 8:708-721.

16. O'Brien P, Fleming T: A multiple testing procedure for clinical trials. *Biometrics* 1979, 35:549-556.

17. Haybittle JL: Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971, 44:793-797.

18. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976, 34:585-612.

19. Pocock S: Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977, 64:191-199.

20. DeMets DL, Lan KK: Interim analysis: the alpha spending function approach. *Stat Med* 1994, 13(13-14):1341-52.

21. Loeys T, Goetghebeur E: A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics* 2003, 59:100-105.

22. Fleming TR, Odem-Davis K, Rothmann MD, Li Shen Y: Some essential considerations in the design and conduct of non-inferiority trials. *Clin Trials* 2011, 8:432-439.